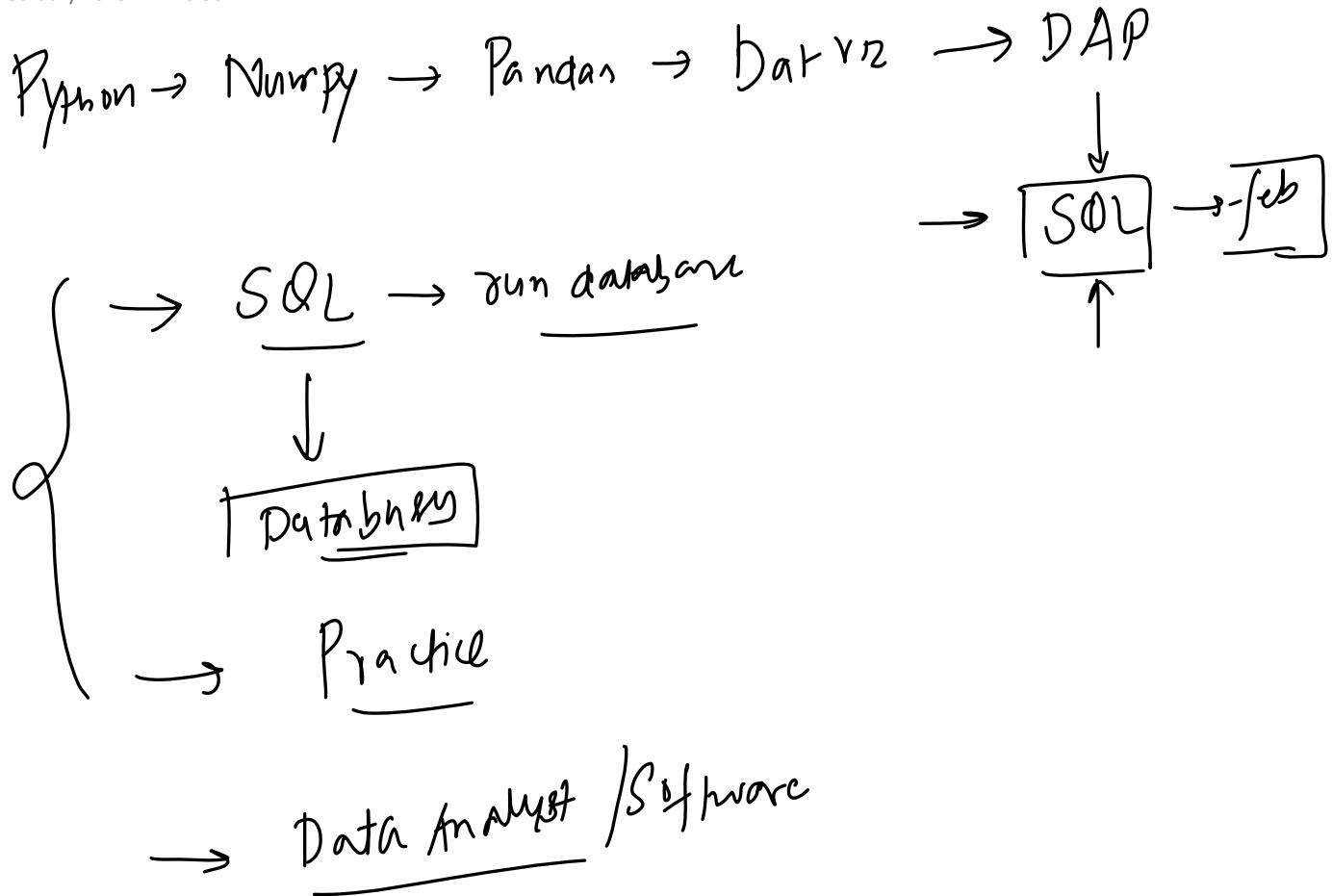


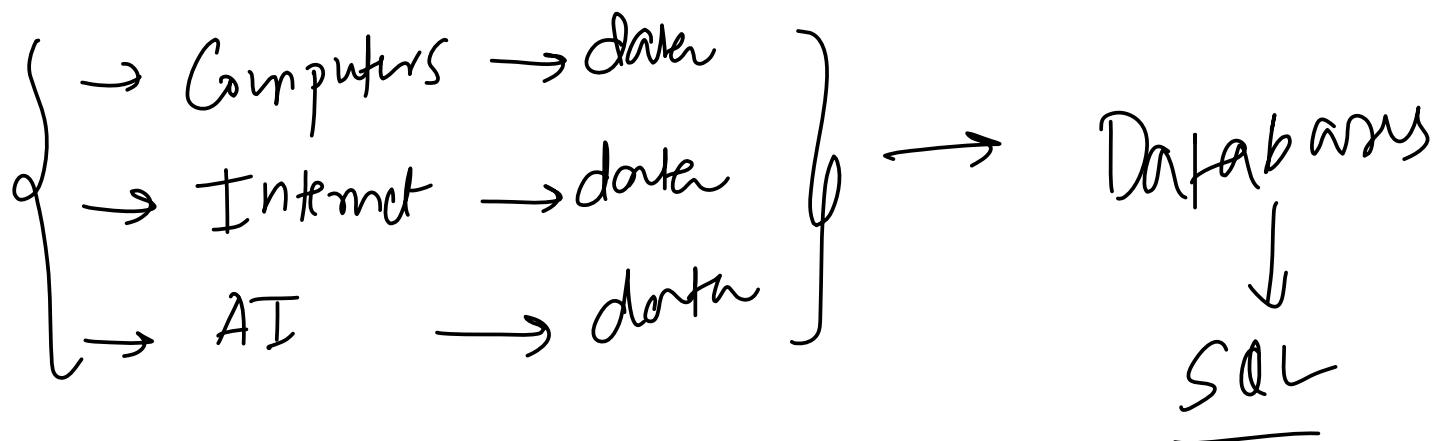
1. Before starting

06 February 2023 16:36



2. Importance of Data

06 February 2023 16:36



3. What are Databases?

06 February 2023 16:37

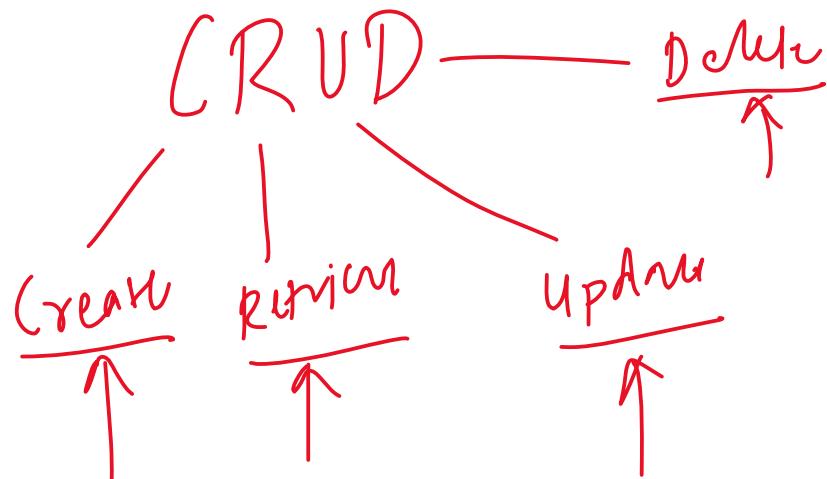
A Database is a shared collection of logically related data and description of these data, designed to meet the information needs of an organization

Data Storage: A database is used to store large amounts of structured data, making it easily accessible, searchable, and retrievable.

Data Analysis: A database can be used to perform complex data analysis, generate reports, and provide insights into the data.

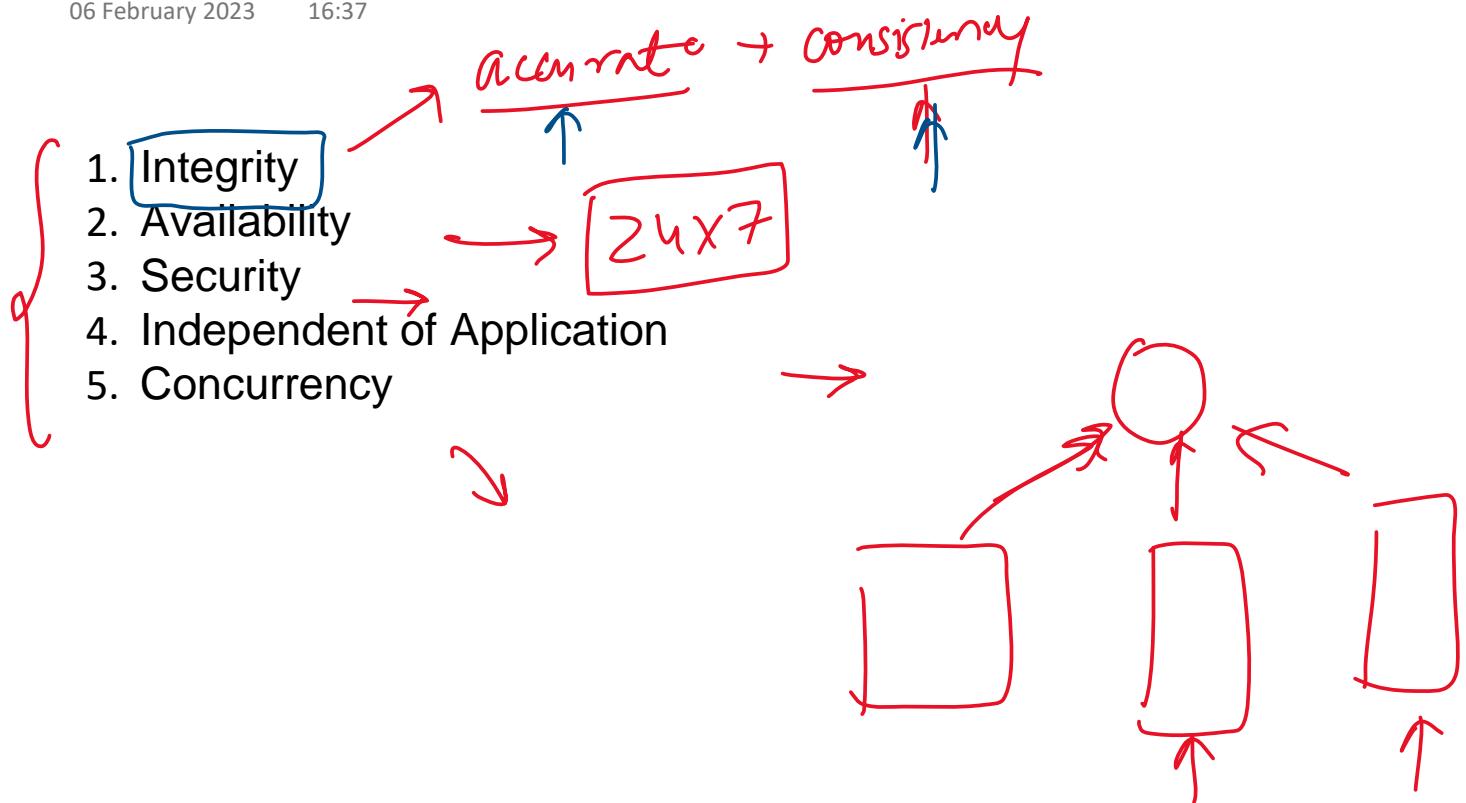
Record Keeping: A database is often used to keep track of important records, such as financial transactions, customer information, and inventory levels.

Web Applications: Databases are an essential component of many web applications, providing dynamic content and user management.



4. Properties of an Ideal Database

06 February 2023 16:37

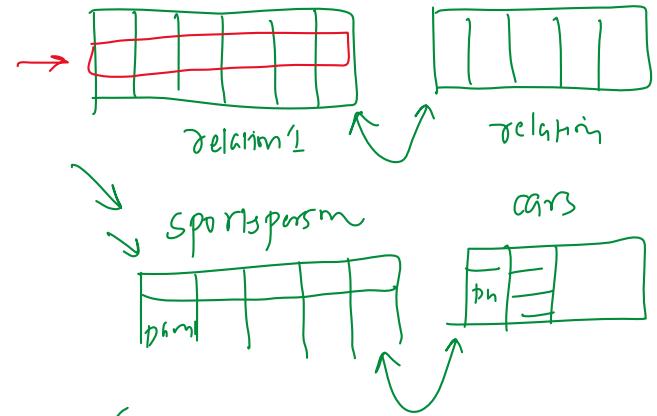


5. Types of Databases

06 February 2023 16:42

1. Relational Databases - (RDB)

Also known as SQL databases, these databases use a relational model to organize data into tables with rows and columns.



2. NoSQL Databases -

These databases are designed to handle large amounts of unstructured or semi-structured data, such as documents, images, or videos. (MongoDB)

3. Column Databases -

These databases store data in columns rather than rows, making them well-suited for data warehousing and analytical applications. (Amazon Redshift, Google BigQuery)

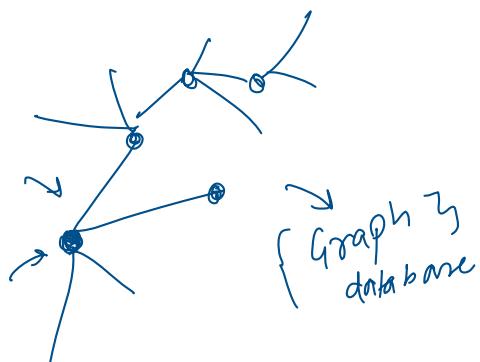
4. Graph Databases -

These databases are used to store and query graph-structured data, such as social network connections or recommendation systems. (Neo4j, Amazon Neptune)

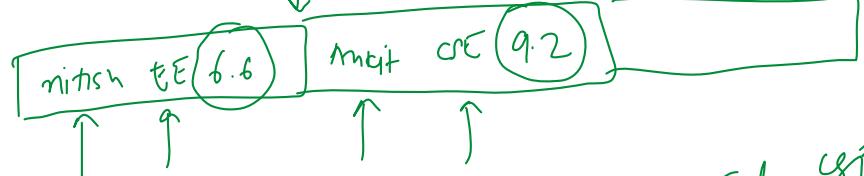
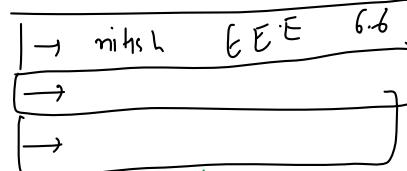
5. Key-value databases -

These databases store data as a collection of keys and values, making them well-suited for caching and simple data storage needs (Redis and Amazon DynamoDB)

Which one should you use?



1000 students



std	name	branch	CGPA
1	nitish	EEE	6.6
2	Ankit	CSE	9.2
3
4

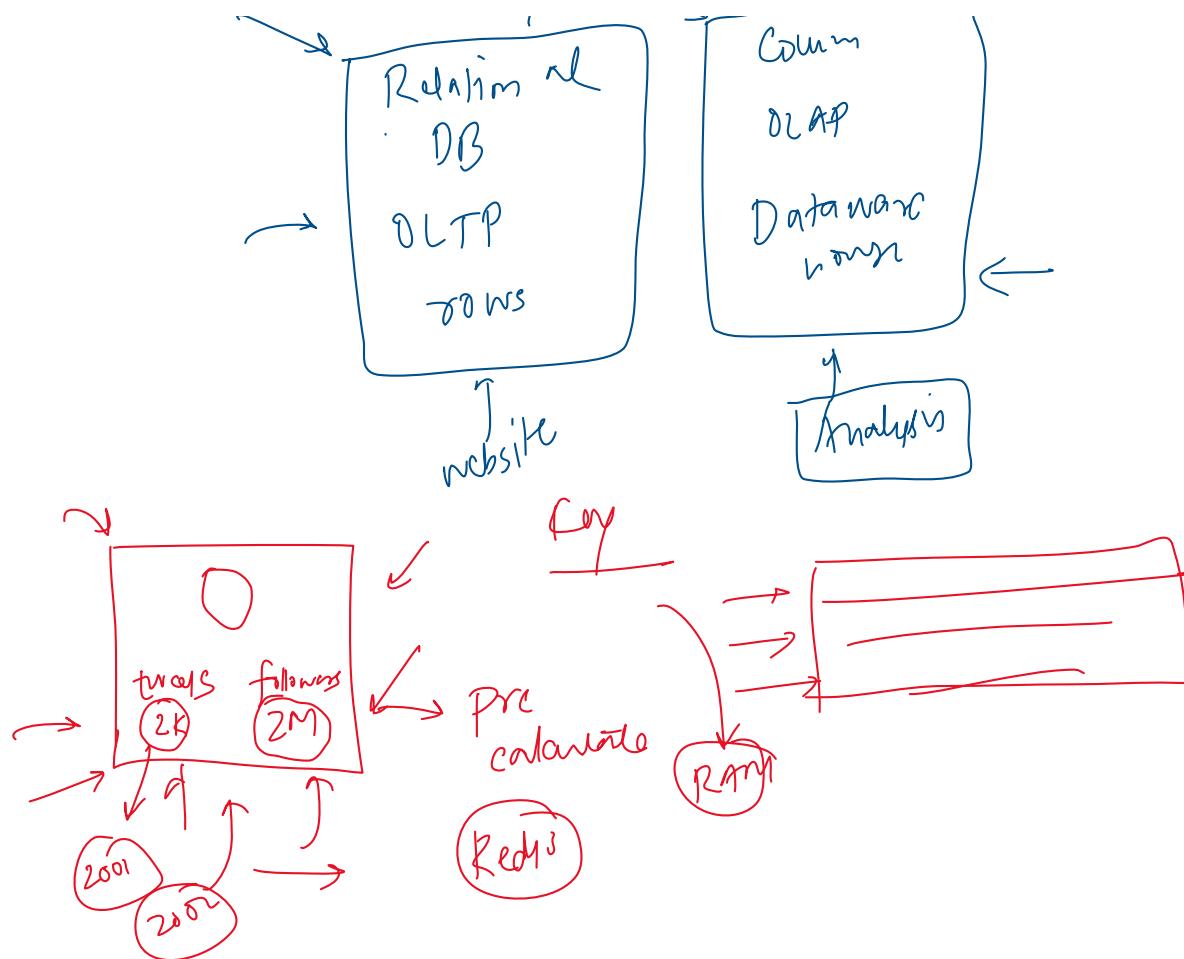
nitish and now ... EEE CSE ...

std	branch	CGPA
1	EEE	6.6
2	CSE	9.2
3
4

6.6 9.2 ...

Relational

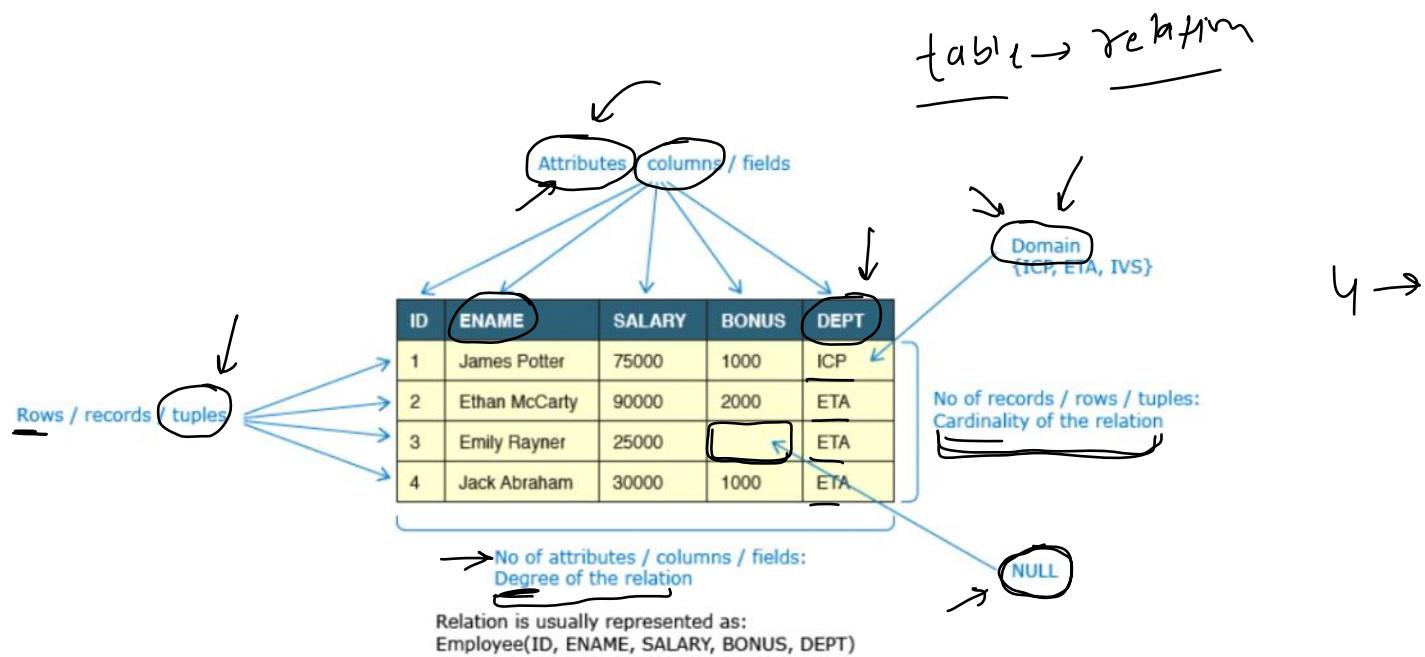
Column



6. Relational Databases

06 February 2023 16:42

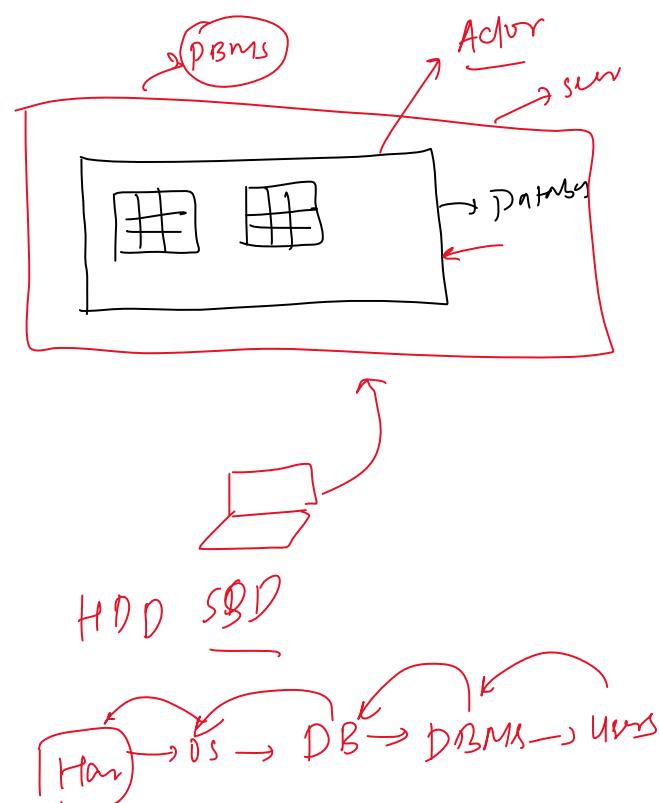
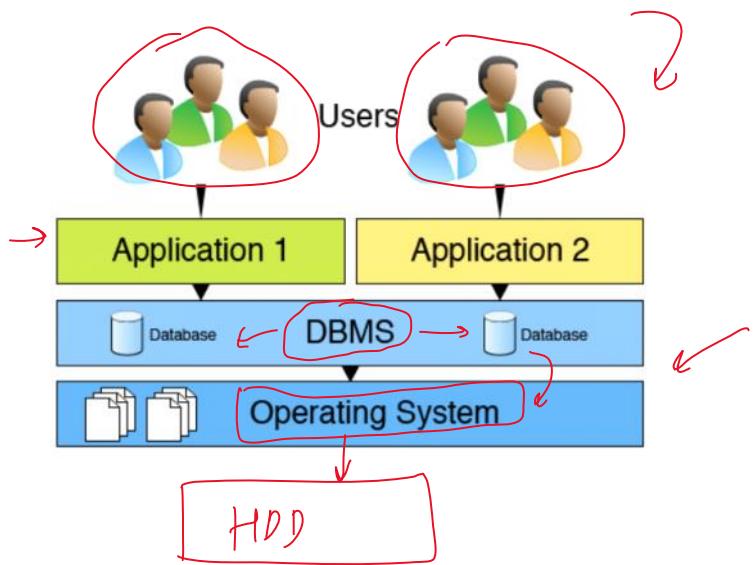
Also known as SQL databases, these databases use a relational model to organize data into tables with rows and columns.



7. What is a DBMS

06 February 2023 16:41

A database management system (DBMS) is a software system that provides the interfaces and tools needed to store, organize, and manage data in a database. A DBMS acts as an intermediary between the database and the applications or users that access the data stored in the database.



8. Core Functionalities of a DBMS

06 February 2023 16:41

Functions of DBMS

CRUD

Data Management - Store, retrieve and modify data

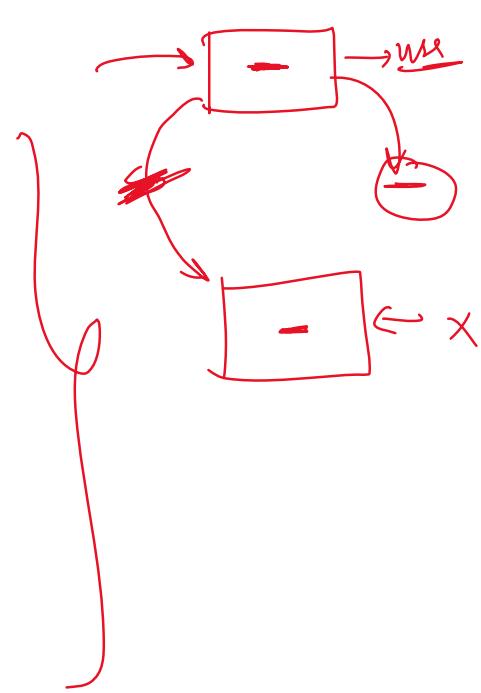
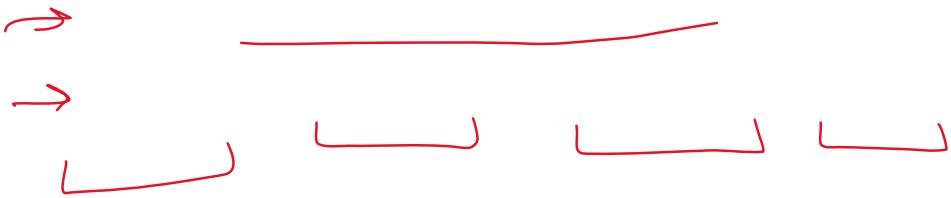
Integrity - Maintain accuracy of data

Concurrency - Simultaneous data access for multiple users

→ **Transaction** - Modification to database must either be successful or must not happen at all

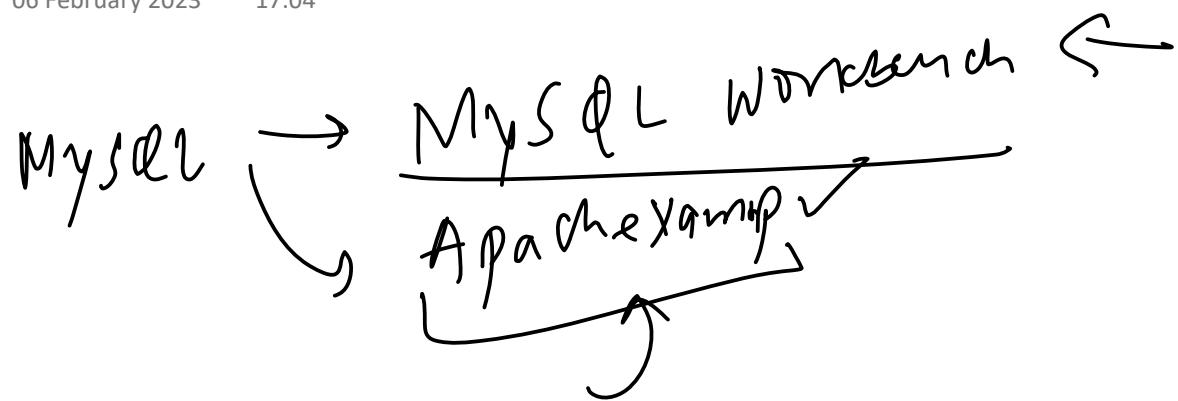
→ **Security** - Access to authorized users only

Utilities - Data import/export, user management, backup, logging



9. Practical

06 February 2023 17:04



10. Database Keys

06 February 2023 17:07

A key in a database is an attribute or a set of attributes that uniquely identifies a tuple (row) in a table. Keys play a crucial role in ensuring the integrity and reliability of a database by enforcing unique constraints on the data and establishing relationships between tables.

1. **Super Key** → amta
A Super key is a combination of columns that uniquely identifies any row within a relational database management system (RDBMS) table

2. **Candidate key** → Unmeadw
A candidate key is a minimal Super key, meaning it has no redundant attributes. In other words, it's the smallest set of attributes that can be used to uniquely identify a tuple (row) in the table
→ vidya

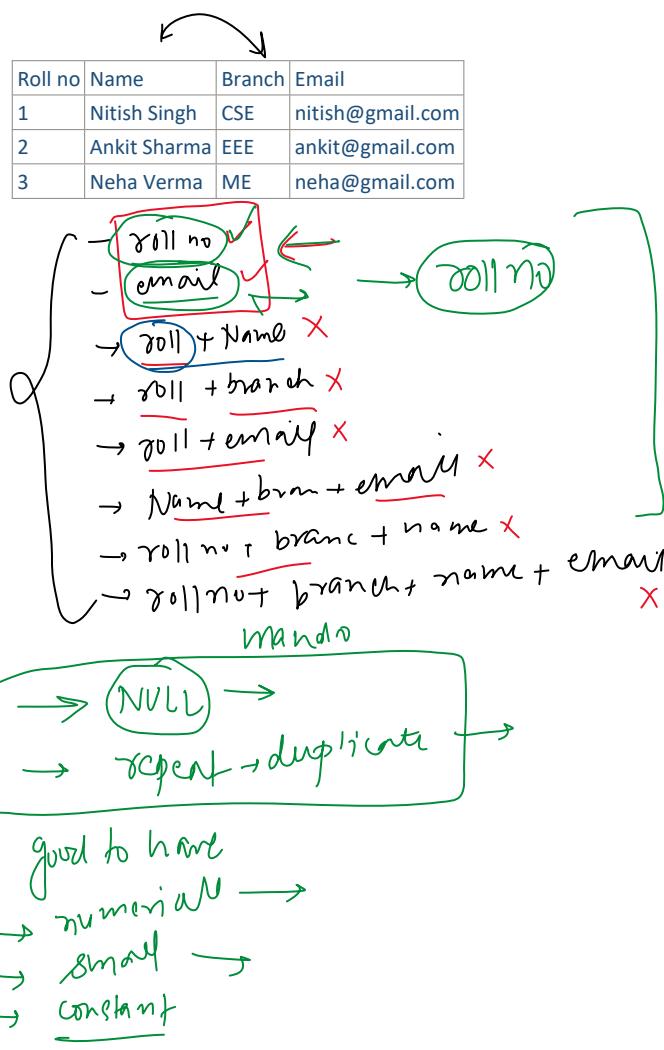
3. **Primary Key** → vidya
A primary key is a unique identifier for each tuple in a table. There can only be one primary key in a table, and it cannot contain null values.

4. **Alternate Key** →
An alternate key is a candidate key that is not used as the primary key.

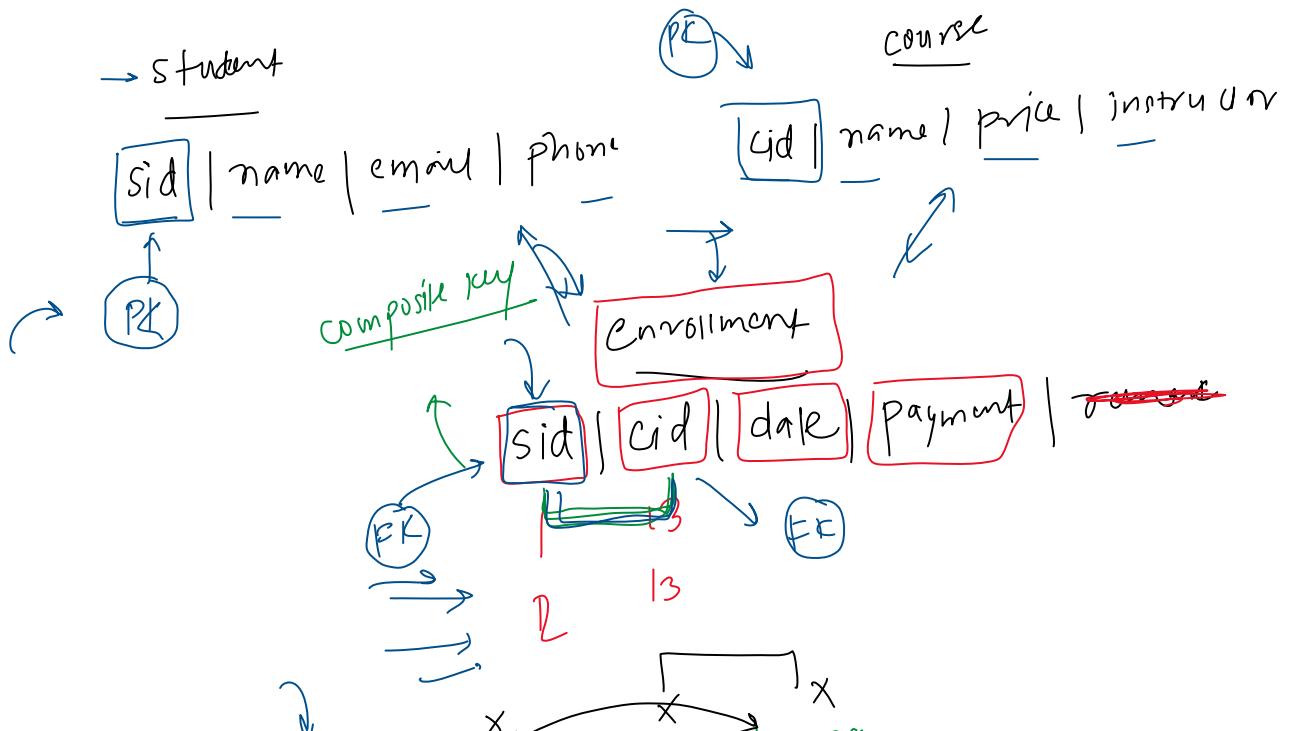
5. **Composite Key** →
A composite key is a primary key that is made up of two or more attributes. Composite keys are used when a single attribute is not sufficient to uniquely identify a tuple in a table.

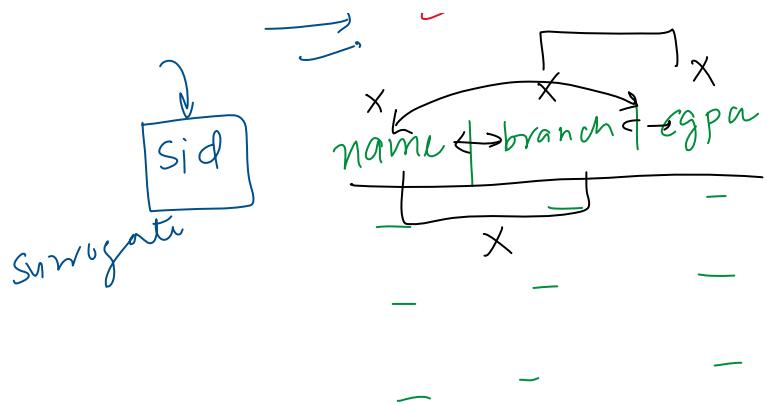
6. **Surrogate Key** →

7. **Foreign Key** →
A foreign key is a primary key from one table that is used to establish a relationship with another table.



$$CK - PK = AF$$



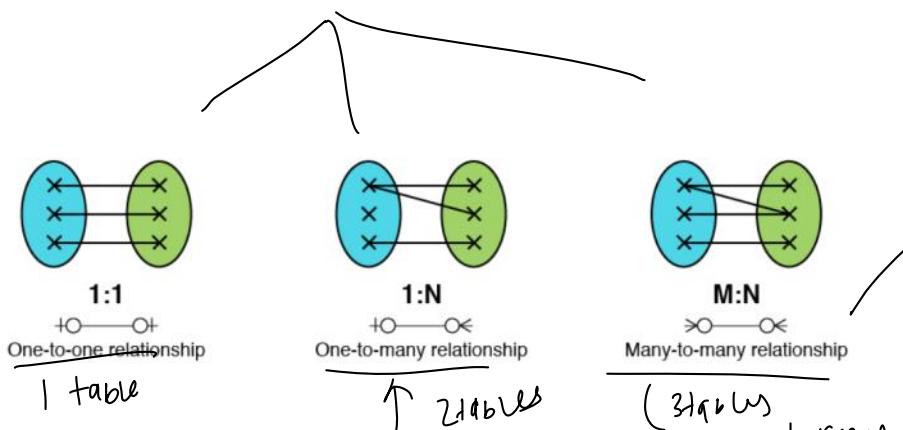


11. Cardinality of Relationships

06 February 2023 16:43

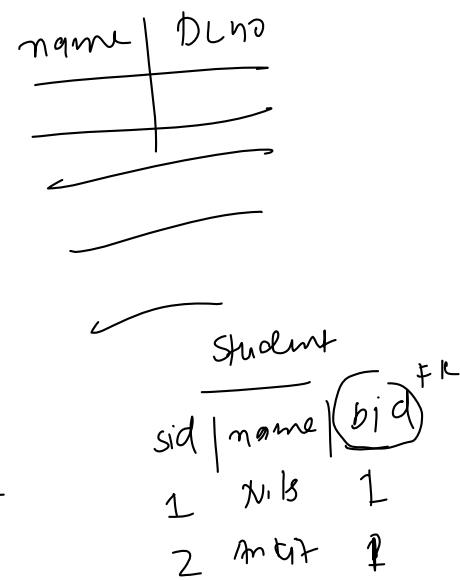
Cardinality in database relationships refers to the number of occurrences of an entity in a relationship with another entity. Cardinality defines the number of instances of one entity that can be associated with a single instance of the related entity.

Entity
↓
fact



Examples

1. Person → Driving License Number
 2. Student → college branch
 3. Restaurants → orders
 4. Restaurants → menu
 5. Students → courses
- (Handwritten notes: circled 'bid' under example 1, circled 'branch' under example 2, circled 'cid' under example 5.)*



sid | name

cid | course | DNA

sid | cid | date

12. Drawbacks of Databases

06 February 2023 16:39

Complexity: Setting up and maintaining a database can be complex and time-consuming, especially for large and complex systems.

Cost: The cost of setting up and maintaining a database, including hardware, software, and personnel, can be high.

Scalability: As the amount of data stored in a database grows, it can become more difficult to manage, leading to performance and scalability issues.

Data Integrity: Ensuring the accuracy and consistency of data stored in a database can be a challenge, especially when multiple users are updating the data simultaneously.

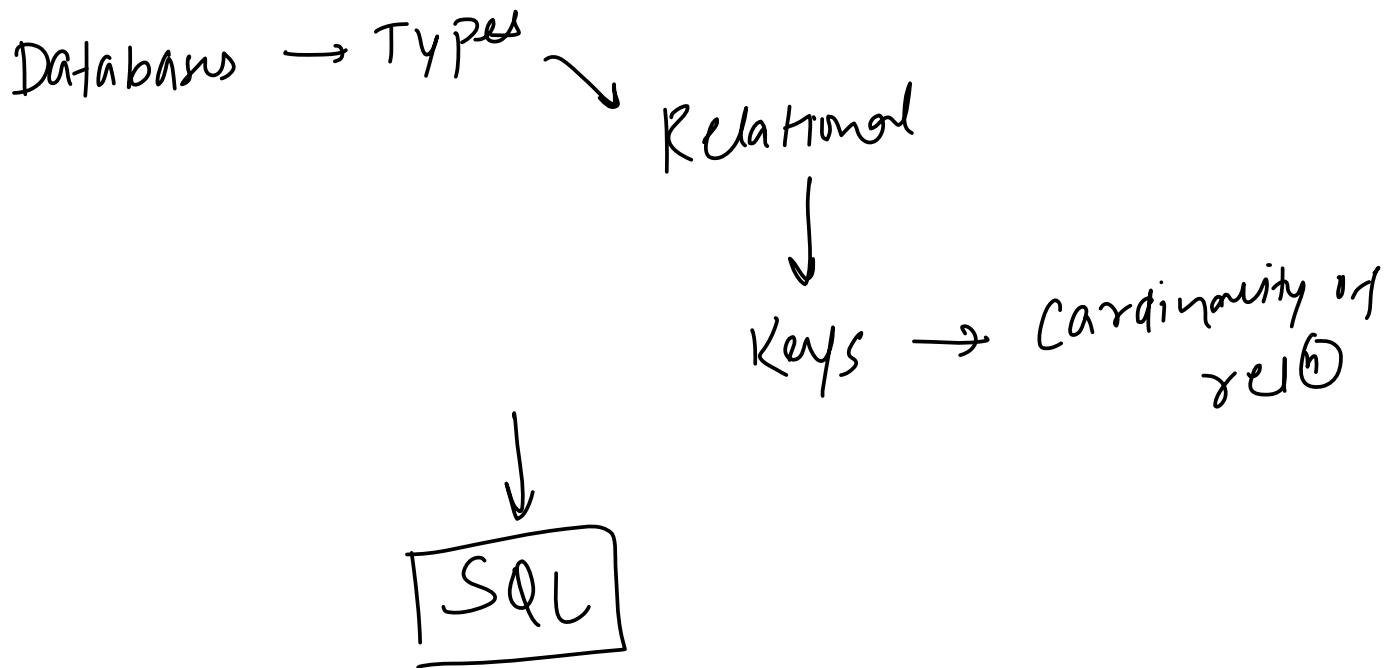
Security: Securing a database from unauthorized access and protecting sensitive information can be difficult, especially with the increasing threat of cyber attacks.

Data Migration: Moving data from one database to another or upgrading to a new database can be a complex and time-consuming process.

Flexibility: The structure of a database is often rigid and inflexible, making it difficult to adapt to changing requirements or to accommodate new types of data.

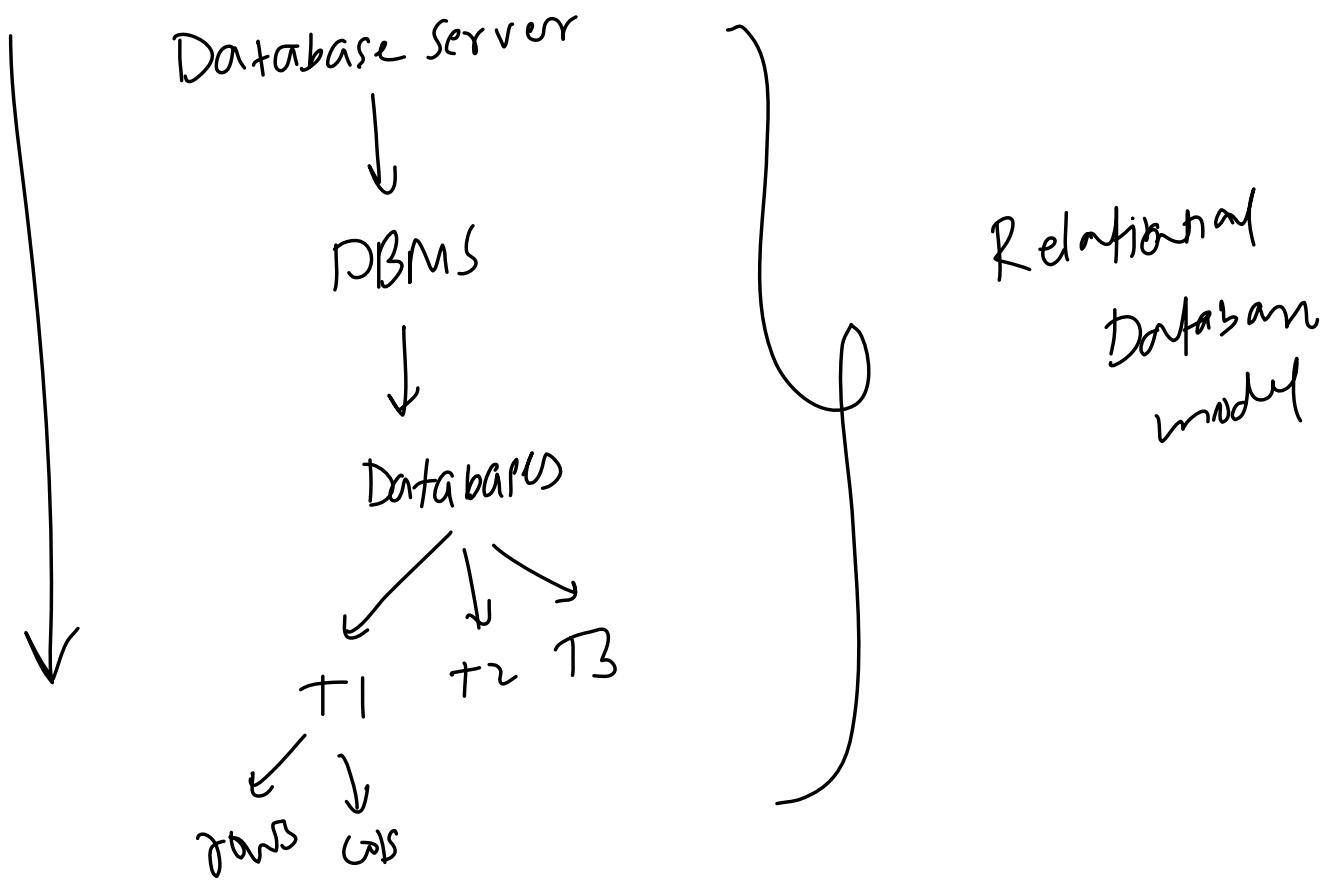
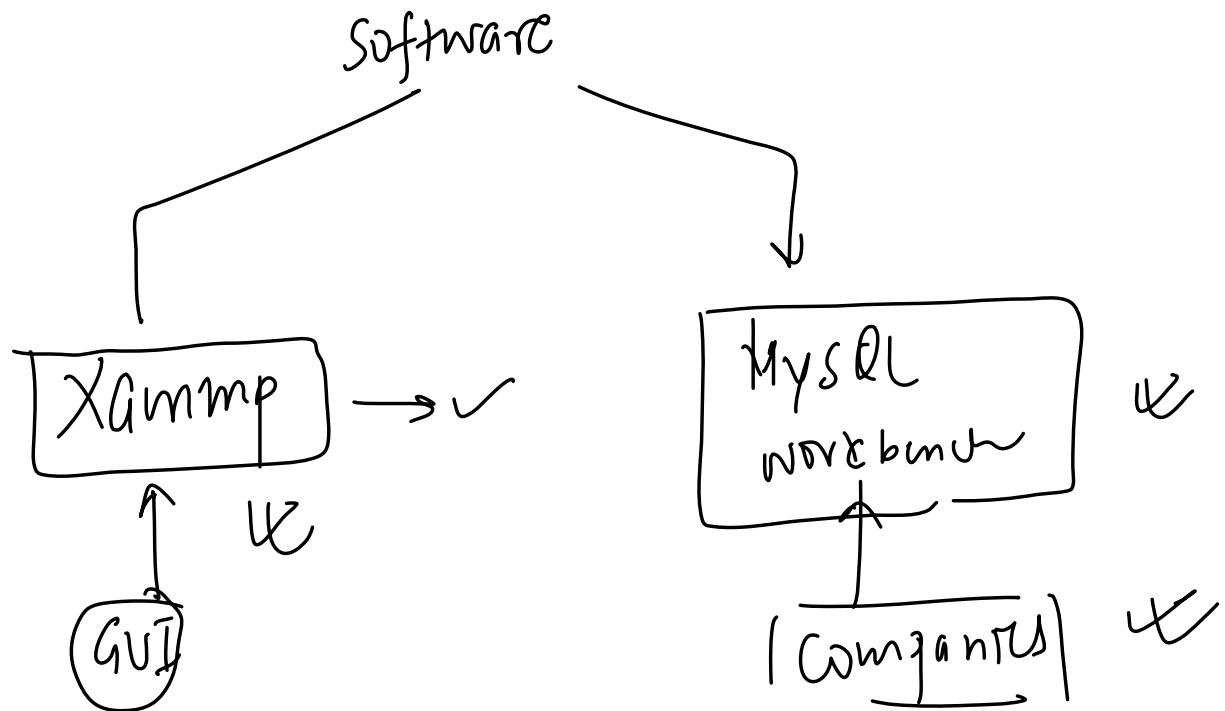
Till now?

08 February 2023 14:56



Installation

08 February 2023 16:05



What is SQL?

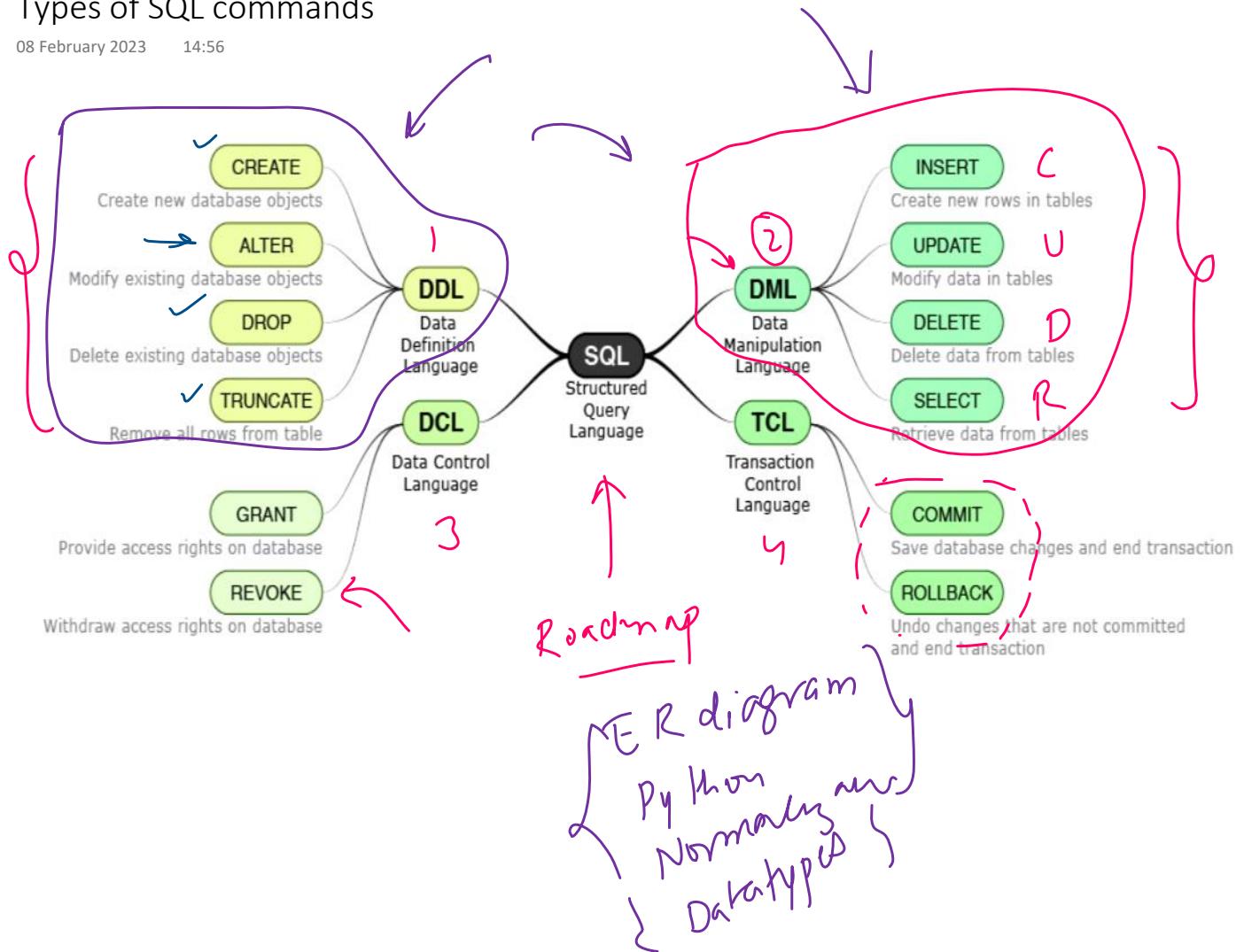
08 February 2023 14:56

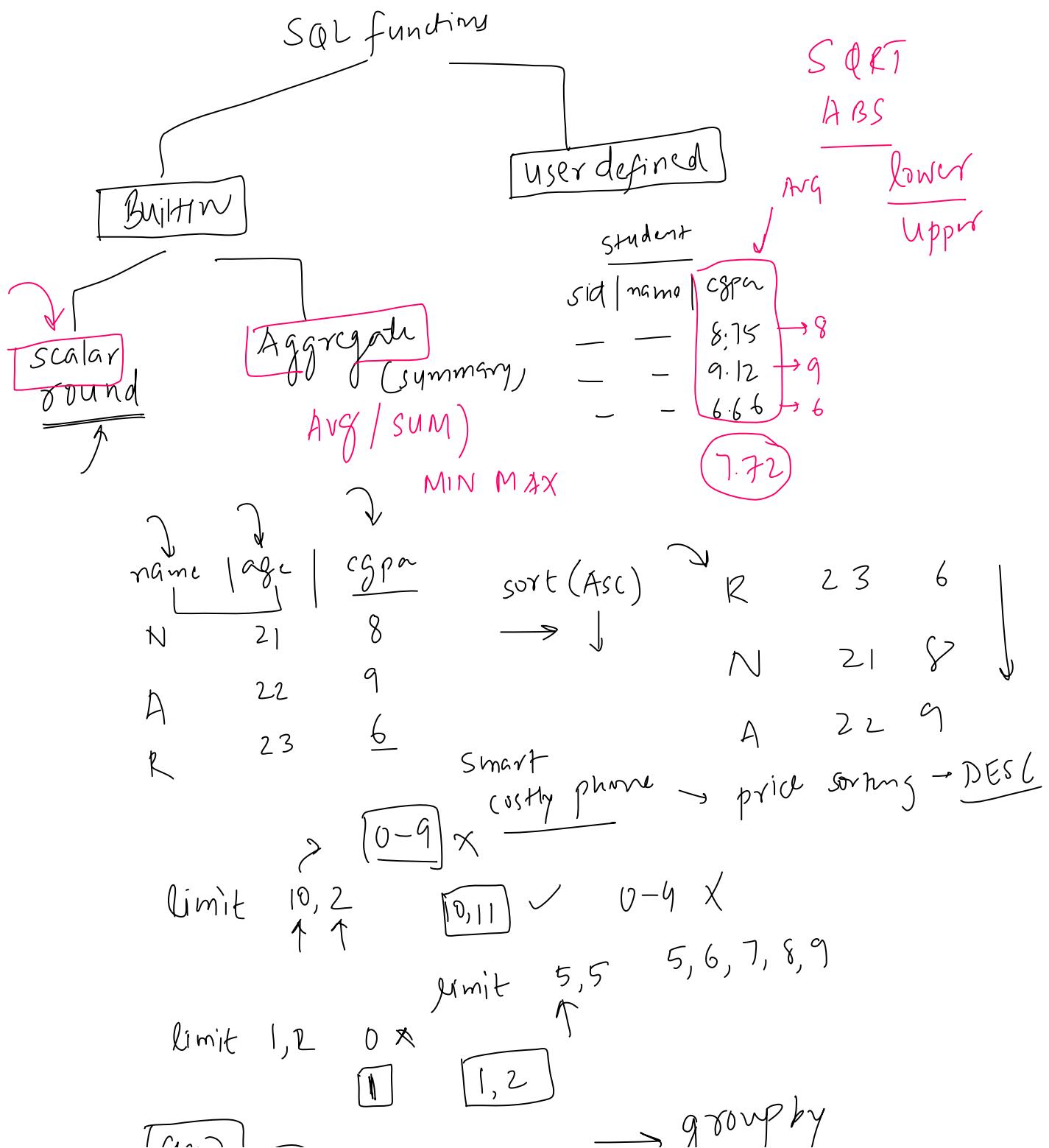
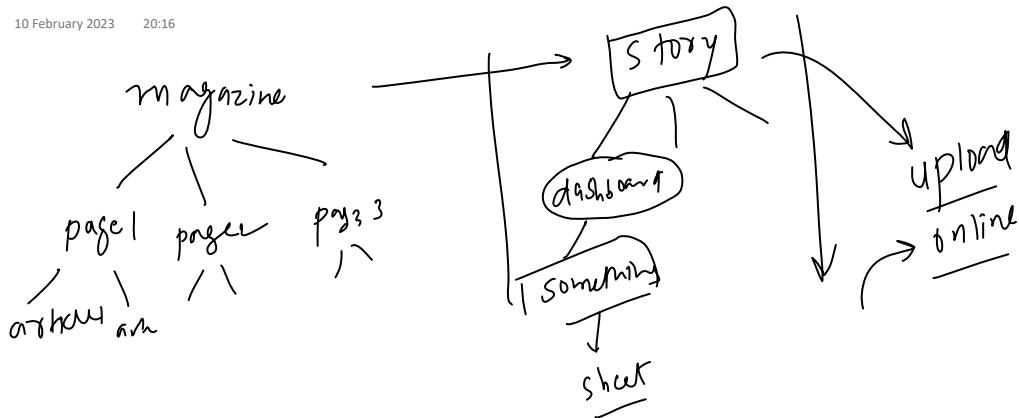
CRUD

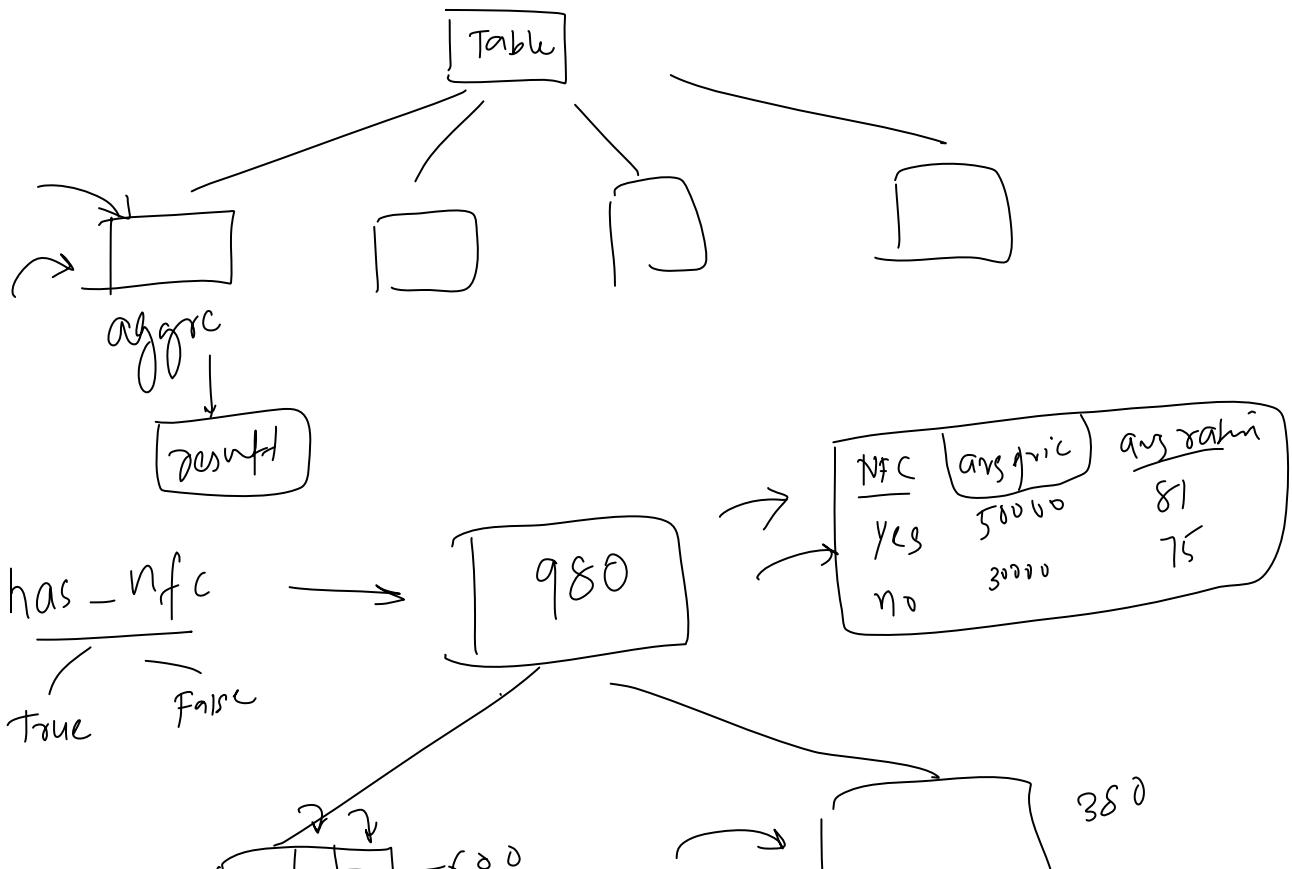
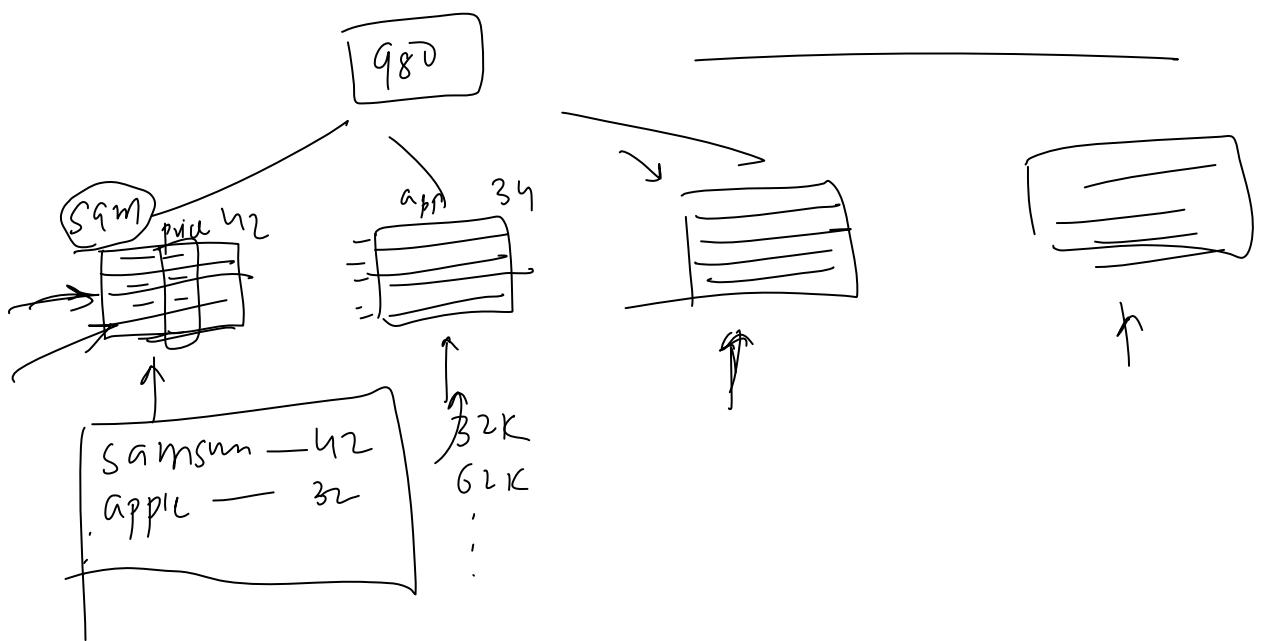
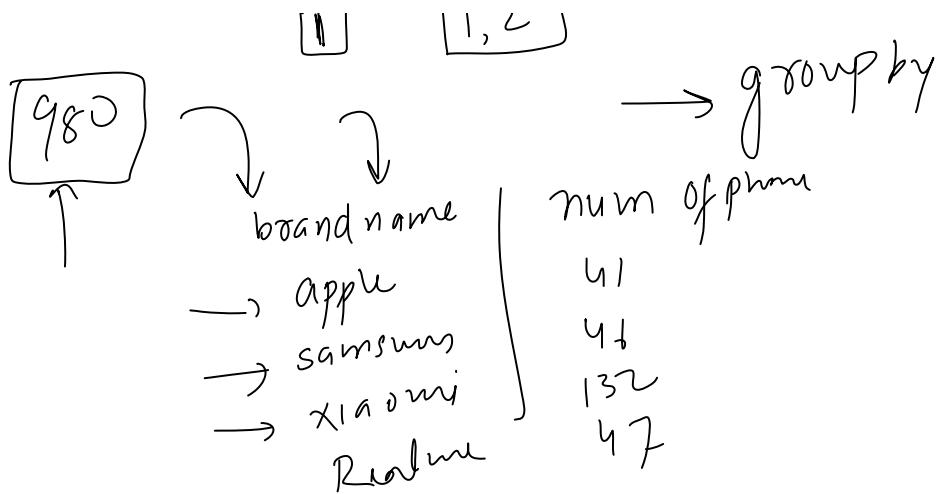
SQL (Structured Query Language) is a programming language used for managing and manipulating data in relational databases. It allows you to insert, update, retrieve, and delete data in a database. It is widely used for data management in many applications, websites, and businesses. In simple terms, SQL is used to communicate with and control databases.

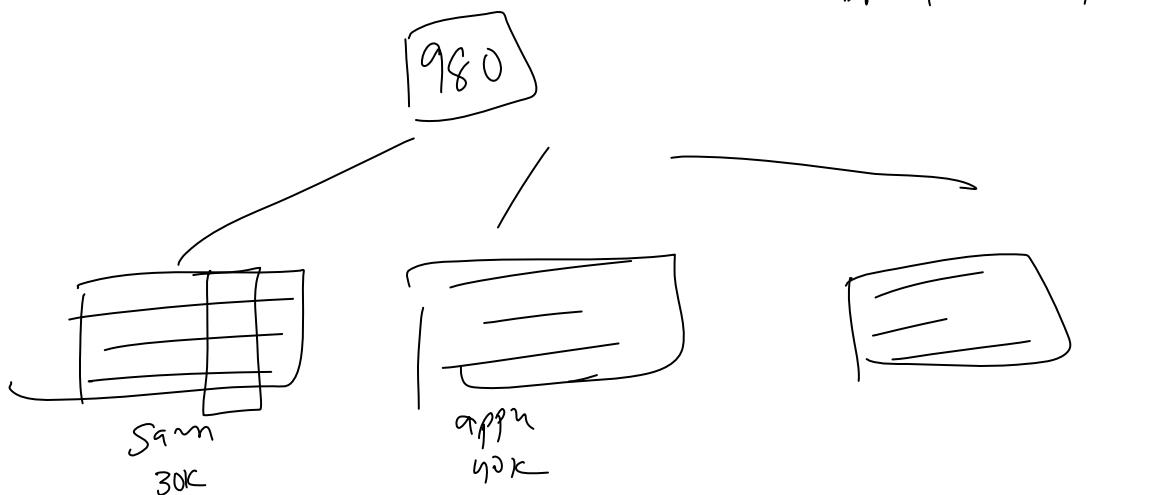
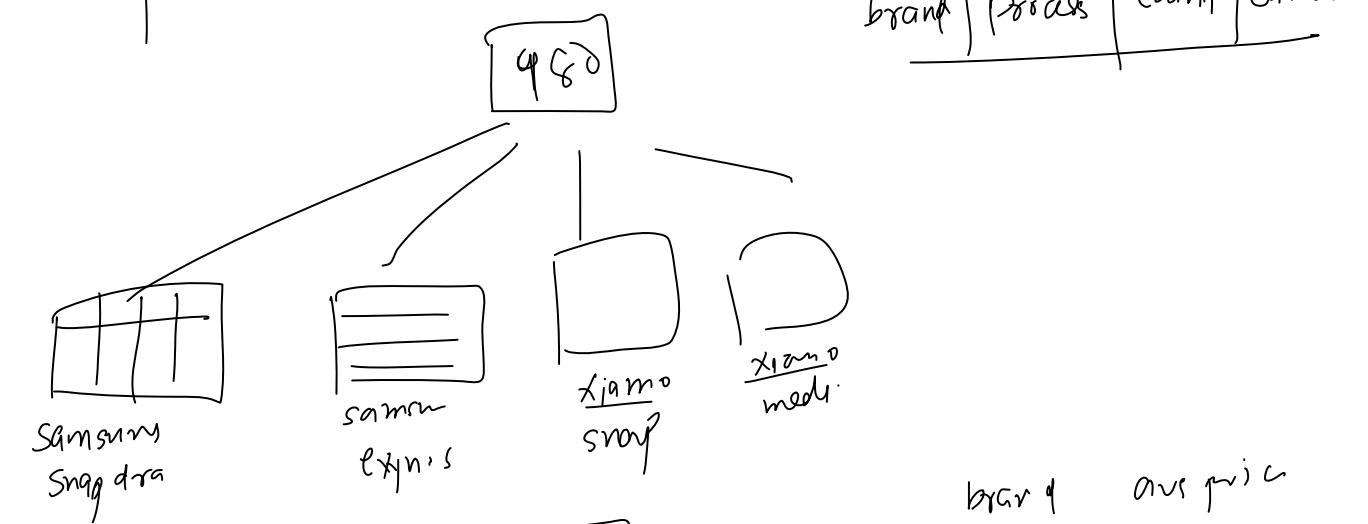
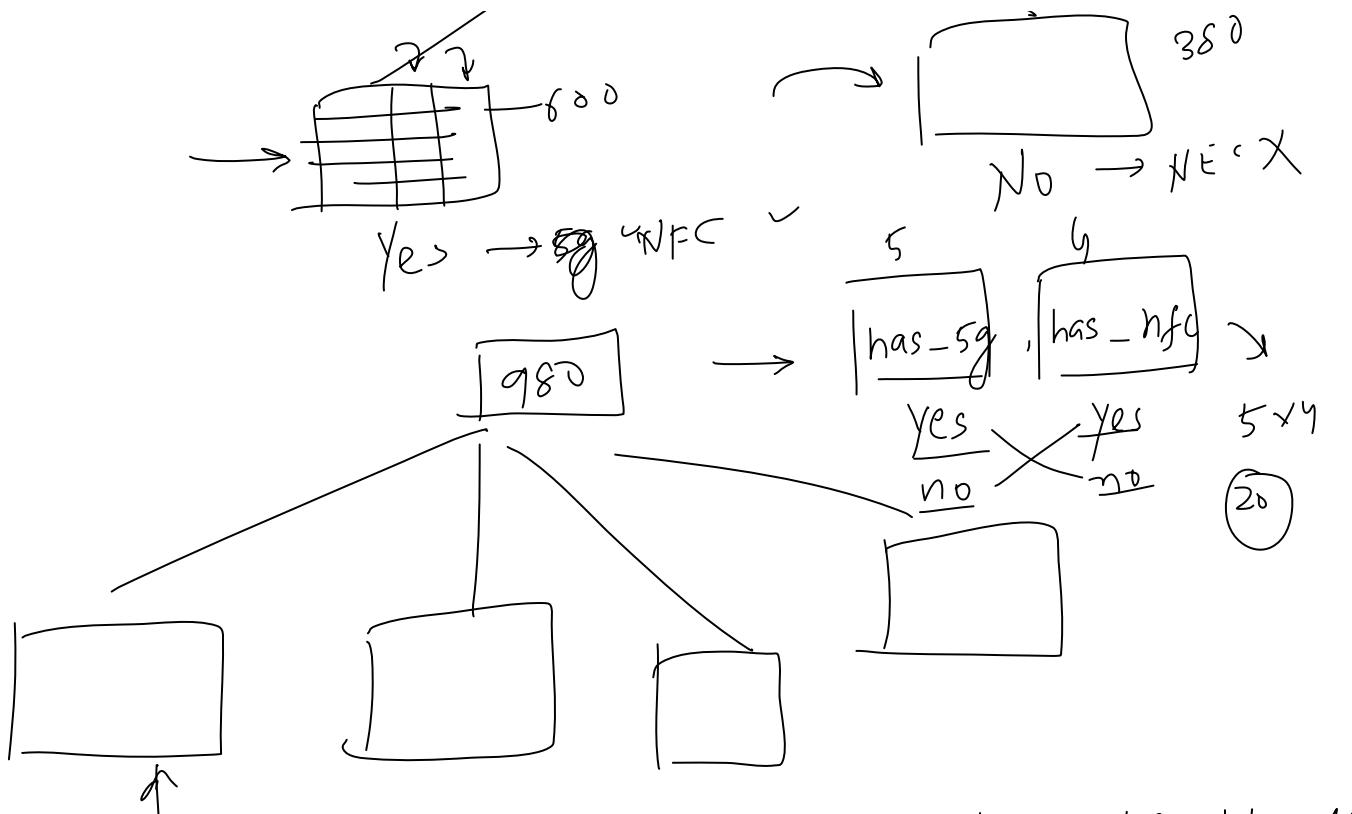
Types of SQL commands

08 February 2023 14:56







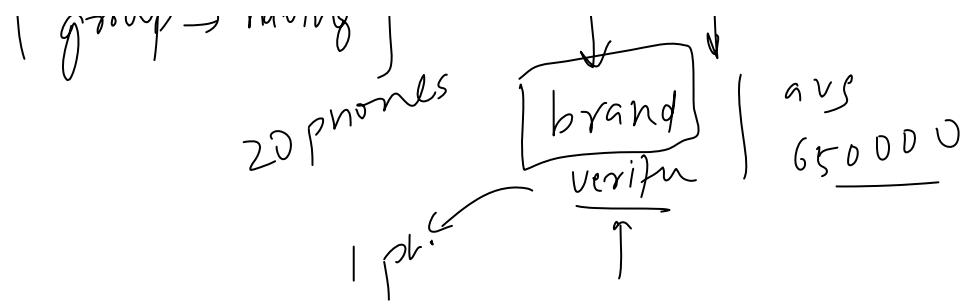


Select → Where
group → having

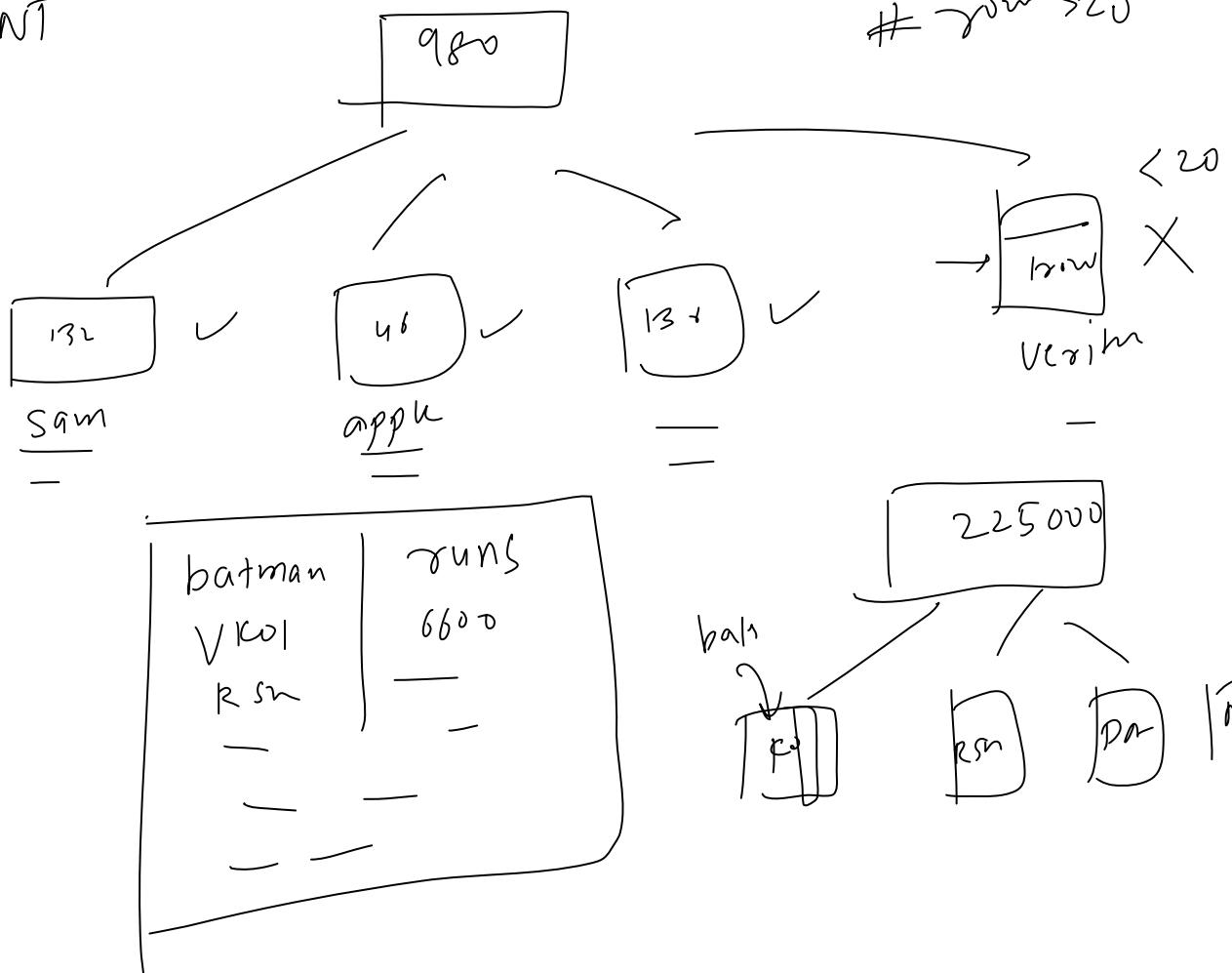
...ands

11 - and

1 avg



COUNT



DDL commands for Databases

08 February 2023 14:57

1. CREATE
2. DROP



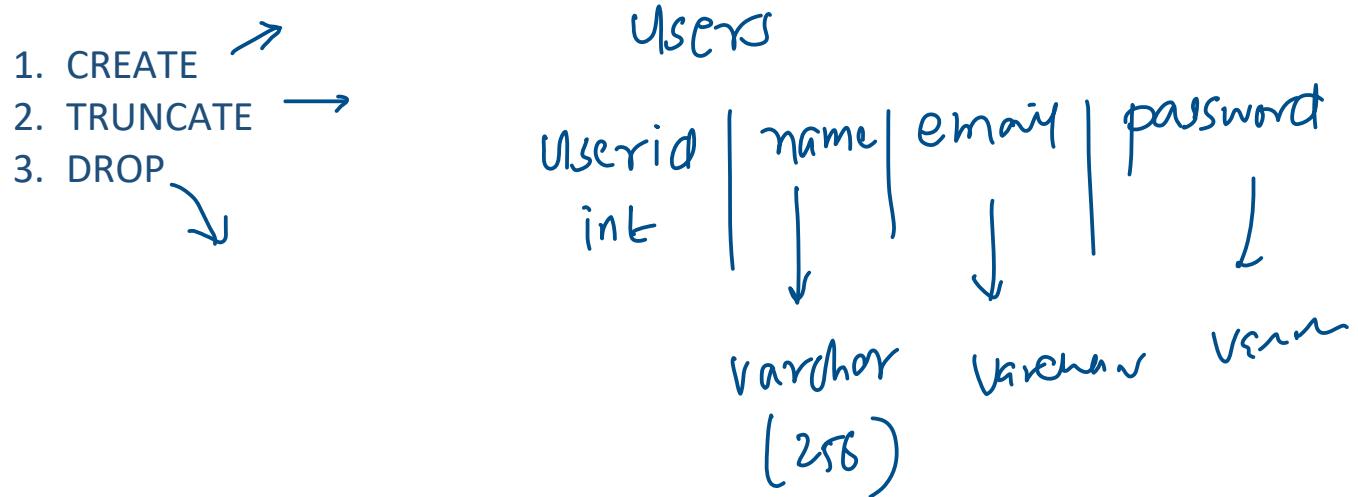
CREATE DATABASE camphix



DROP DATABASE camphix

DDL commands for Tables

08 February 2023 15:00



Data Integrity ✎

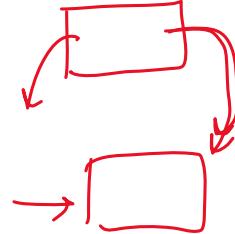
08 February 2023 15:00

Data integrity in databases refers to the accuracy, completeness, and consistency of the data stored in a database. It is a measure of the reliability and trustworthiness of the data and ensures that the data in a database is protected from errors, corruption, or unauthorized changes.

There are various methods used to ensure data integrity, including:

↳ **Constraints:** → DDL

Constraints in databases are rules or conditions that must be met for data to be inserted, updated, or deleted in a database table. They are used to enforce the integrity of the data stored in a database and to prevent data from becoming inconsistent or corrupted.



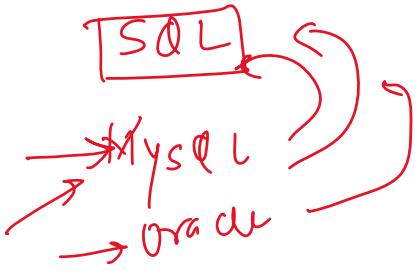
→ **Transactions:** a sequence of database operations that are treated as a single unit of work.

↳ **Normalization:** a design technique that minimizes data redundancy and ensures data consistency by organizing data into separate tables.

Constraints in MySQL →

08 February 2023 15:01

Constraints in databases are rules or conditions that must be met for data to be inserted, updated, or deleted in a database table. They are used to enforce the integrity of the data stored in a database and to prevent data from becoming inconsistent or corrupted.



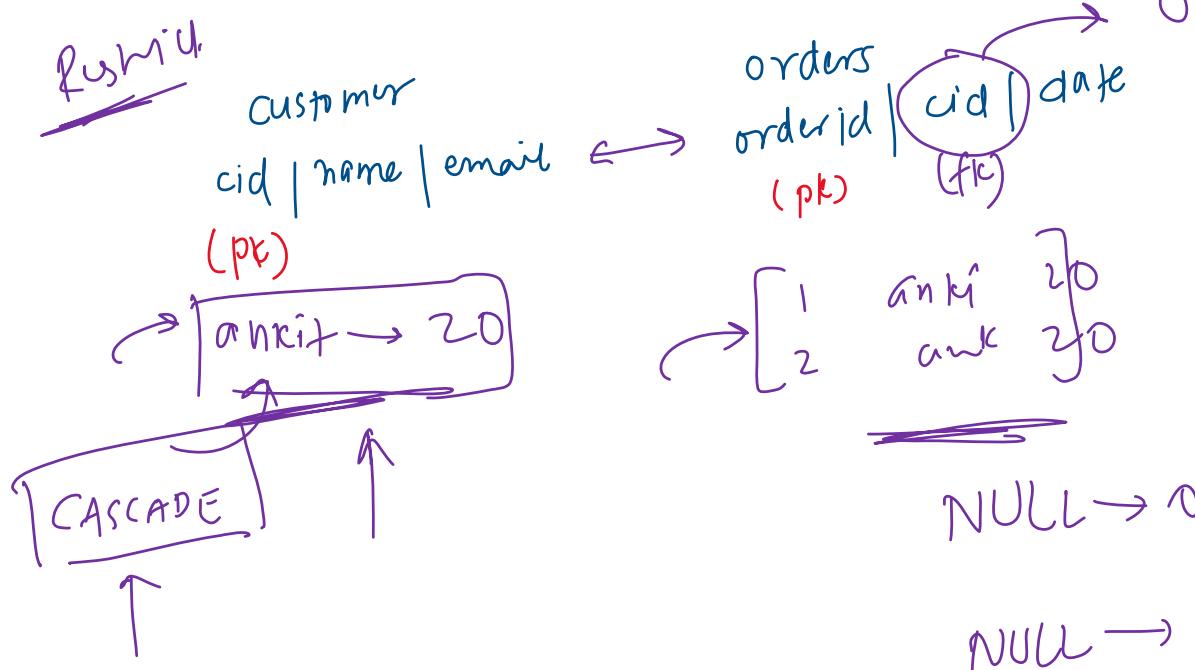
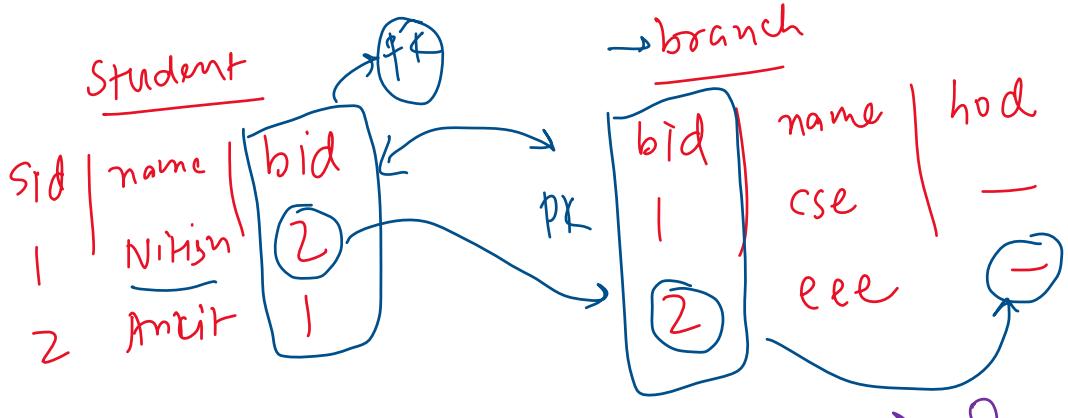
1. NOT NULL →
 2. UNIQUE(combo) →
→ Another way of creating constraint
 3. PRIMARY KEY
 4. AUTO INCREMENT
 5. CHECK
 6. DEFAULT
 7. FOREIGN KEY
- Referential Actions
1. RESTRICT
 2. CASCADE
 3. SET NULL
 4. SET DEFAULT

MySQL users table diagram:

User Id	Name	Email	Age	Reg Date
70	-	-	-	-
71	-	-	-	-
72	-	-	-	-

Annotations:

- NULL:** Points to the first row (User Id 70).
- PK:** Points to the primary key constraint on the "User Id" column.
- UNQ:** Points to the unique constraint on the "Email" column.
- PNR (NOT NULL):** Points to the NOT NULL constraint on the "Age" column.
- age > 13:** Points to the CHECK constraint on the "Age" column.
- reg date:** Points to the "Reg Date" column.



ALTER TABLE command

08 February 2023 15:02

The "**ALTER TABLE**" statement in SQL is used to modify the structure of an existing table. Some of the things that can be done using the ALTER TABLE statement include

1. Add columns
 2. Delete columns
 3. Modify columns
- 
- -
 -
- 

Editing and Deleting Constraints

08 February 2023 15:02

1. Add ↗
2. Delete ✓
3. ~~Edit~~ nahi hota

What are SQL joins

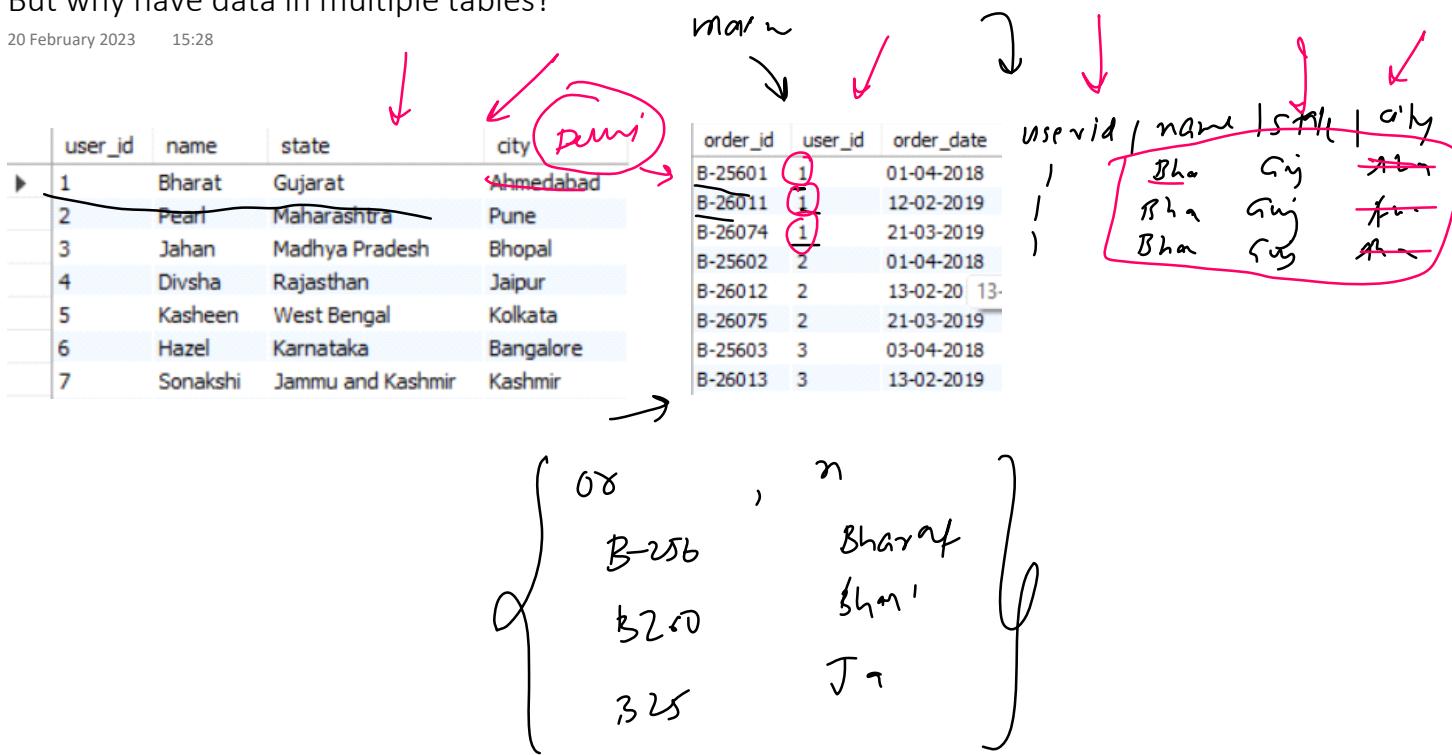
20 February 2023 15:28

In SQL (Structured Query Language), a join is a way to combine data from two or more database tables based on a related column between them.

Joins are used when we want to query information that is distributed across multiple tables in a database, and the information we need is not contained in a single table. By joining tables together, we can create a virtual table that contains all of the information we need for our query.

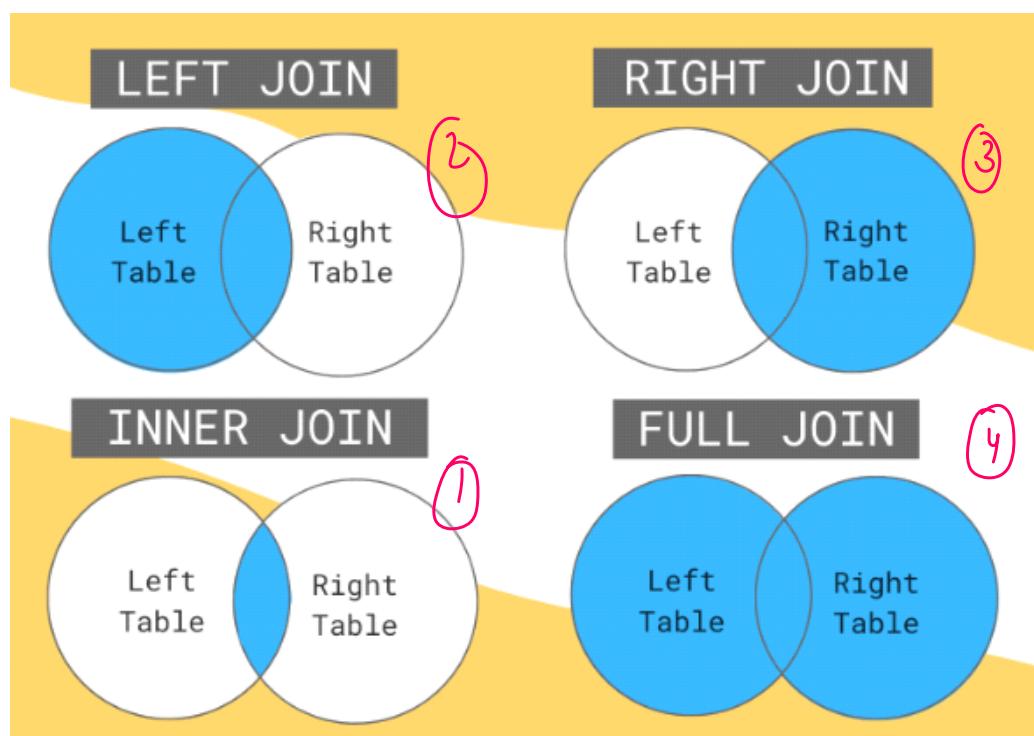
But why have data in multiple tables?

20 February 2023 15:28



Types of Joins

20 February 2023 15:30



Cross join
Self join

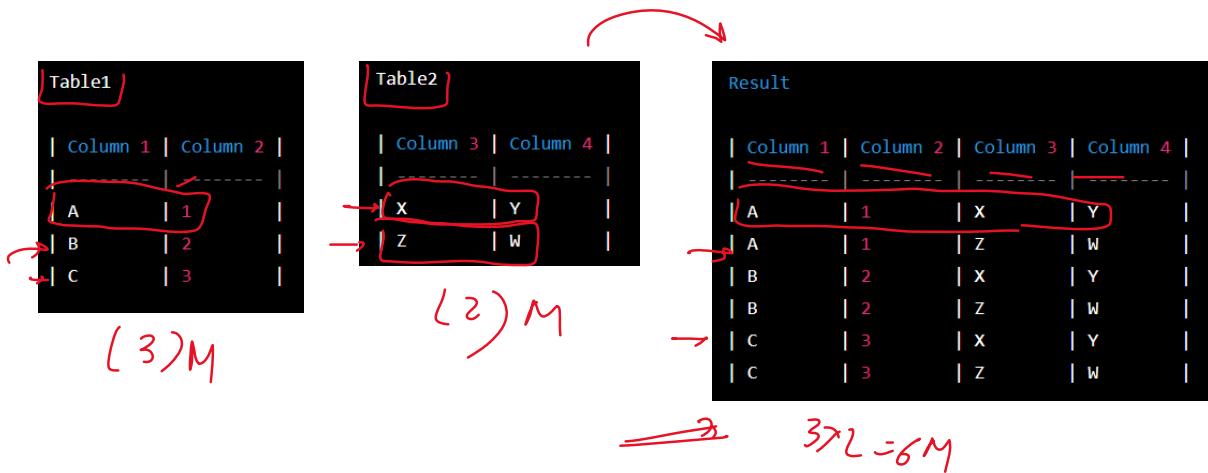
Cross Joins -> Cartesian Products

20 February 2023 15:29

In SQL, a cross join (also known as a Cartesian product) is a type of join that returns the Cartesian product of the two tables being joined. In other words, it returns all possible combinations of rows from the two tables.

Cross joins are not commonly used in practice, but they can be useful in certain scenarios, such as generating test data or exploring all possible combinations of items in a product catalogue. However, it's important to be cautious when using cross joins with large tables, as they can generate a very large result set, which can be resource-intensive and slow to process.

$$A = \{1, 2\}$$
$$B = \{3, 4\}$$
$$(1,3) (1,4) (2,3) (2,4)$$

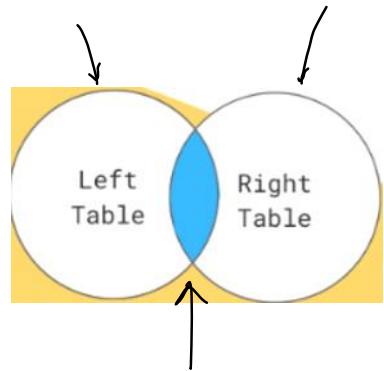


Inner Joins

20 February 2023 15:30

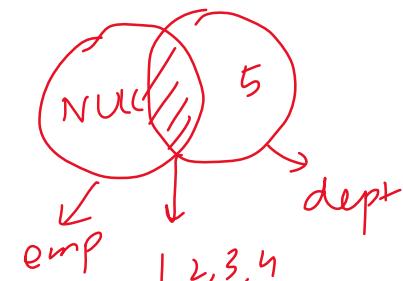
In SQL, an inner join is a type of join operation that combines data from two or more tables based on a specified condition. The inner join returns only the rows from both tables that satisfy the specified condition, i.e., the matching rows.

When you perform an inner join on two tables, the result set will only contain rows where there is a match between the joining columns in both tables. If there is no match, then the row will not be included in the result set.



Employee ID	Name	Department ID	Salary
1	John Smith	1	100000
2	Jane Doe	2	50000
3	Bob Johnson	3	75000
4	Lisa Wong	1	90000
5	Mike Lee	2	120000
6	Tim Davis	4	60000
7	Sarah Chen	NULL	80000

Department ID	Department Name
1	Engineering
2	Sales
3	Finance
4	Marketing
5	Operations



Employee ID	Name	Department ID	Salary	Department ID	Department
1	John Smith	1	100000	1	Engineering
4	<u>Lisa Wong</u>	1	90000	1	Engineering
2	<u>Jane Doe</u>	2	50000	2	Sales
5	<u>Mike Lee</u>	2	120000	2	Sales
3	<u>Bob Johnson</u>	3	75000	3	Finance

Left Join

20 February 2023 15:30

A left join, also known as a left outer join, is a type of SQL join operation that returns all the rows from the left table (also known as the "first" table) and matching rows from the right table (also known as the "second" table). If there are no matching rows in the right table, the result will contain NULL values in the columns that come from the right table.

In other words, a left join combines the rows from both tables based on a common column, but it also includes all the rows from the left table, even if there are no matches in the right table. This is useful when you want to include all the records from the first table, but only some records from the second table.

The diagram shows two tables, `emp` and `dep2`. The `emp` table has 8 rows with Employee IDs 1 through 8. The `dep2` table has 3 rows with Department IDs 1, 2, and 3. A red circle highlights the first row of the `emp` table (Employee ID 1). Red arrows point from the `emp` table to the `dep2` table, indicating the mapping of Employee ID 1 to Department ID 1. The resulting joined table is shown below.

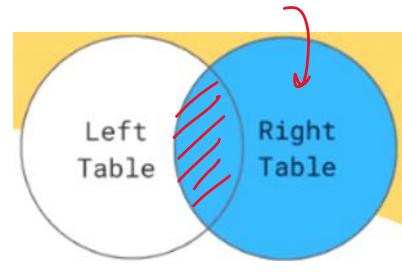
Employee ID	Name	Department ID	Salary	Department ID	Department Name
1	John Smith	1	100000	1	Engineering
2	Jane Doe	2	50000	2	Sales
3	Bob Johnson	3	75000	3	Finance
4	Lisa Wong	1	90000		
5	Mike Lee	2	120000		
6	Tim Davis	4	60000		
7	Sarah Brown	5	80000		
8	Mark Wilson	2	95000		

Employee ID	Name	Department ID	Salary	Department ID	Department
1	John Smith	1	100000	1	Engineering
2	Jane Doe	2	50000	2	Sales
3	Bob Johnson	3	75000	3	Finance
4	Lisa Wong	1	90000	1	Engineering
5	Mike Lee	2	120000	2	Sales
6	Tim Davis	4	60000	NULL	NULL
7	Sarah Brown	5	80000	NULL	NULL
8	Mark Wilson	2	95000	2	Sales

Right Join

20 February 2023 15:31

A right join, also known as a right outer join, is a type of join operation in SQL that returns all the rows from the right table and matching rows from the left table. If there are no matches in the left table, the result will still contain all the rows from the right table, with NULL values for the columns from the left table.



emp

Employee ID	Name	Department ID	Salary
1	John Smith	1 ✓	100000
2	Jane Doe	2 ✓	50000
3	Bob Johnson	3 ✓	75000
4	Lisa Wong	1 -	90000
5	Mike Lee	2 -	120000
7	Sarah Brown	NULL	80000
8	Mark Wilson	2 ✓	95000

dept

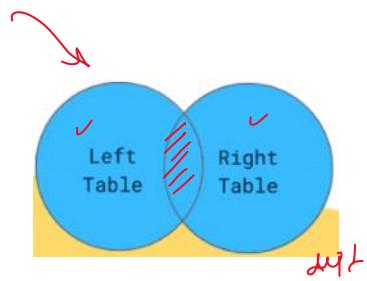
Department ID	Department Name
1	Engineering
2	Sales
3	Finance
4	Marketing
5	HR

Employee ID	Name	Department ID	Salary	Department Name
1	John Smith	1	100000	Engineering
4	Lisa Wong	1	90000	Engineering
2	Jane Doe	2	50000	Sales
5	Mike Lee	2	120000	Sales
8	Mark Wilson	2	95000	Sales
3	Bob Johnson	3	75000	Finance
NULL	NULL	4	NULL	Marketing
NULL	NULL	5	NULL	HR

Full Outer Join

20 February 2023 15:31

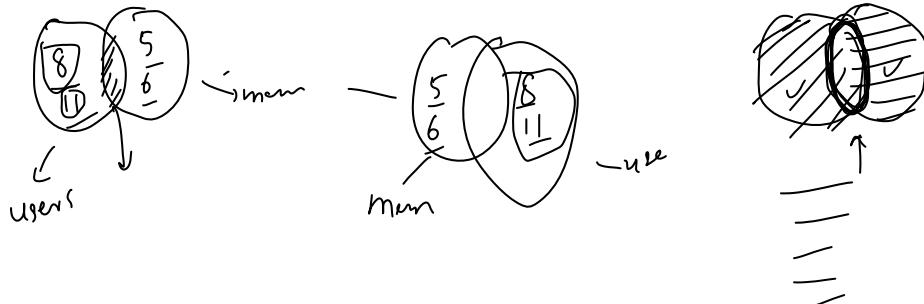
A full outer join, sometimes called a full join, is a type of join operation in SQL that returns all matching rows from both the left and right tables, as well as any non-matching rows from either table. In other words, a full outer join returns all the rows from both tables and matches rows with common values in the specified columns, and fills in NULL values for columns where there is no match.



emp		
emp_id	emp_name	dept_id
1	Alice	1
2	Bob	1
3	Charlie	2
4	Dave	null
5	Eve	3

dept	
dept_id	dept_name
1	Sales
2	Marketing
3	Finance
4	IT
5	HR

emp_id	emp_name	dept_id	dept_id	dept_name
1	Alice	1 ✓	1	Sales
2	Bob	1 ↘	1	Sales
3	Charlie	2 ↘	2	Marketing
4	Dave	null	null	null
5	Eve	3 ↘	3	Finance
null ↗	null ↗	null	4 ↗	IT
null ↗	null ↗	null	5 ↗	HR



SQL Set Operations

20 February 2023 15:35

1. **UNION:** The UNION operator is used to combine the results of two or more SELECT statements into a single result set. The UNION operator removes duplicate rows between the various SELECT statements.
 2. **UNION ALL:** The UNION ALL operator is similar to the UNION operator, but it does not remove duplicate rows from the result set.
 3. **INTERSECT:** The INTERSECT operator returns only the rows that appear in both result sets of two SELECT statements.
 4. **EXCEPT:** The EXCEPT or MINUS operator returns only the distinct rows that appear in the first result set but not in the second result set of two SELECT statements.

person

id	name
1	Alice
2	Bob
3	Charlie

except

person

<u>id</u>	<u>name</u>
3	Charlie ✓
4	David ✓
5	Emily ✓

person

4 h | Dr

id	name
1	Alice ✓
2	Bob ✓
3	Charlie ✓
4	David ✓
5	Emily ✓

unison all

id	name
1	Alice ✓
2	Bob ✓
3	Charlie ✓
3	Charlie ✓
4	David ✓
5	Emily ✓

'interse 4-

id	name
3	Charlie <i>←</i>

id	name
1	Alice <i>✓</i>
2	Bob <i>✓</i>

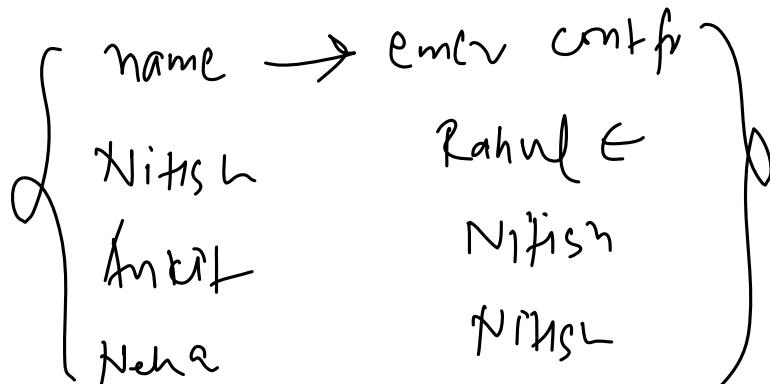
Self Joins

20 February 2023 15:34

A self join is a type of join in which a table is joined with itself. This means that the table is treated as two separate tables, with each row in the table being compared to every other row in the same table.

Self joins are used when you want to compare the values of two different rows within the same table. For example, you might use a self join to compare the salaries of two employees who work in the same department, or to find all pairs of customers who have the same billing address.

t1				t2			
user_id	name	age	emergency_contact	user_id	name	age	emergency_contact
1 Nitish	34	11		1 Nitish	34	11	
2 Ankit	32	1		2 Ankit	32	1	
3 Neha	23		1	3 Neha	23		1
4 Radhika	34	3		4 Radhika	34		3
8 Abhinav	31	11		8 Abhinav	31		11
11 Rahul	29	8		11 Rahul	29		8



Joining on more than one cols

20 February 2023 16:58

student

student_id	first_name	last_name	class_id	enrollment_year
1	John	Smith	1	2021
2	Jane	Doe	2	2020
3	Bob	Johnson	1	2021
4	Sally	Brown	3	2022
5	Tom	Williams	2	2022
6	Alice	Davis	4	2020

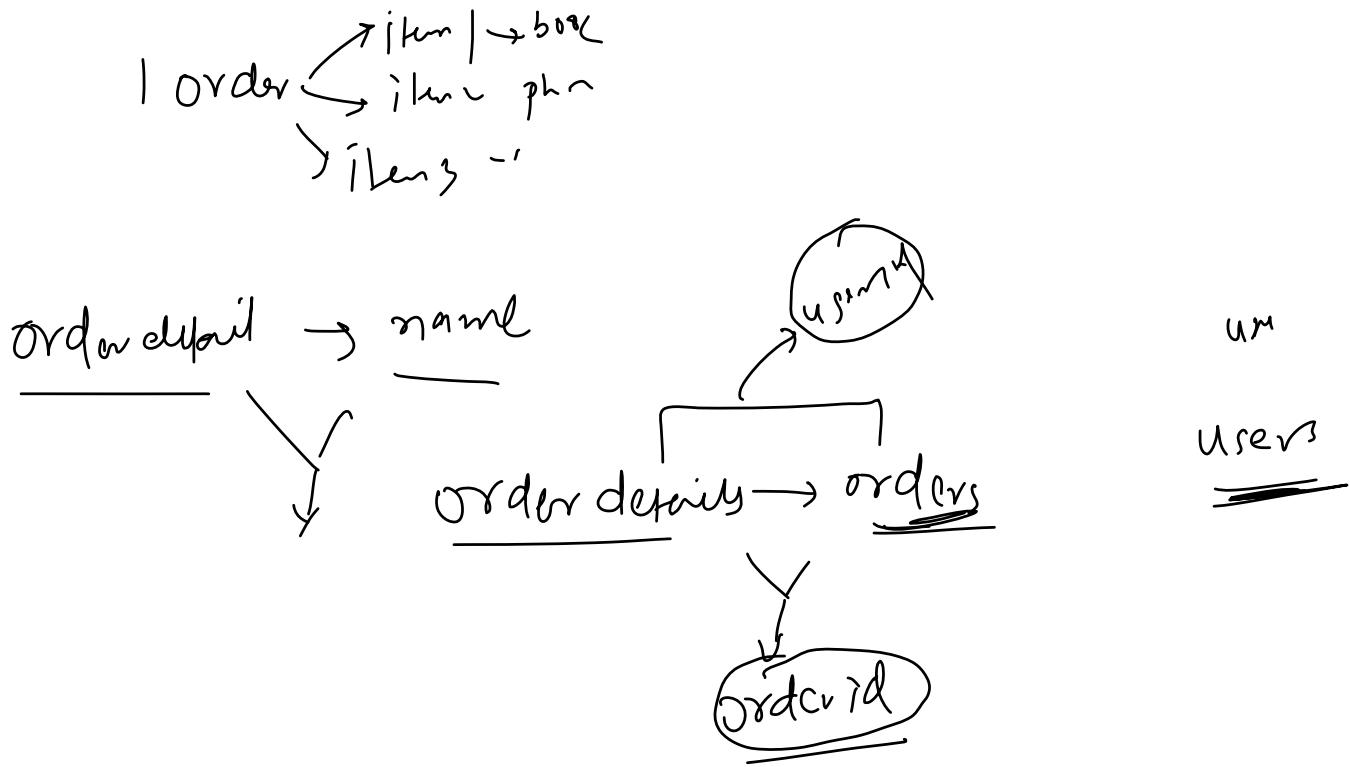
Class - brn

class_id	class_name	teacher	class_year
1	Math 101	Mr. Smith	2021
2	English 1	Ms. Johnson	2021
3	Science 1	Dr. Lee	2022
4	History 1	Ms. Williams	2022

Joining more than 2 tables

20 February 2023 15:33

1. Find order name and corresponding category name



Filtering Columns

20 February 2023 15:31

1. Find order_id, name and city by joining users and orders.
2. Find order_id, product category by joining order_details and category

Filtering Rows

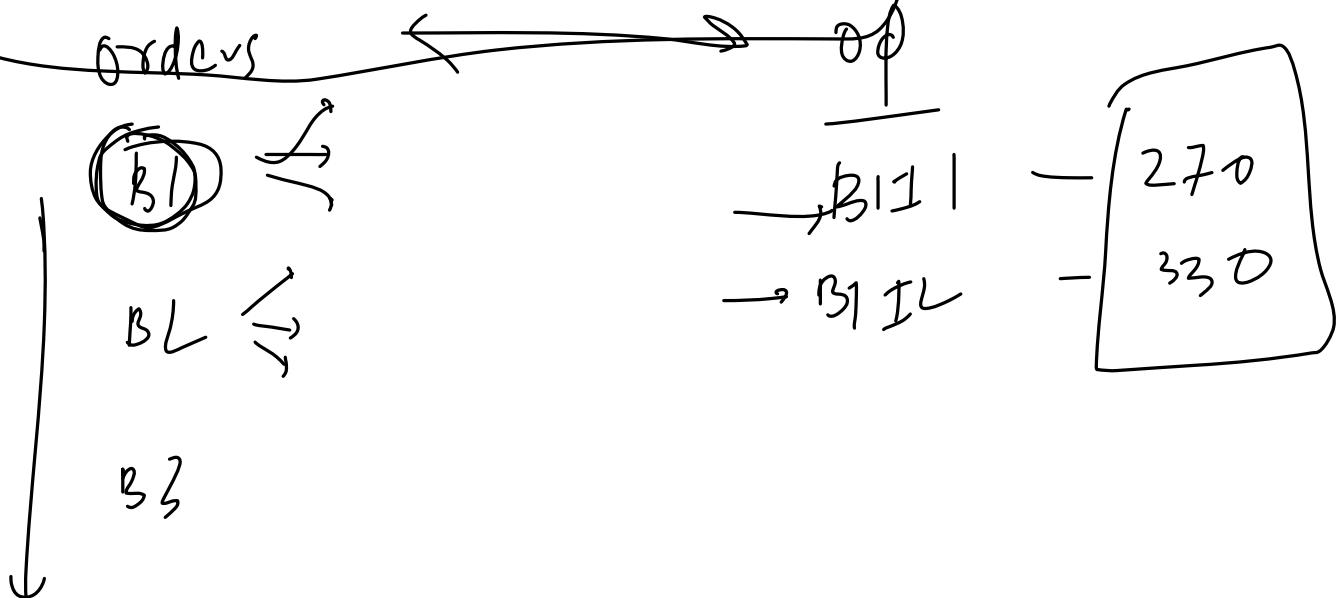
20 February 2023 15:32

1. Find all the orders placed in pune
2. Find all orders under Chairs category

Practice Questions

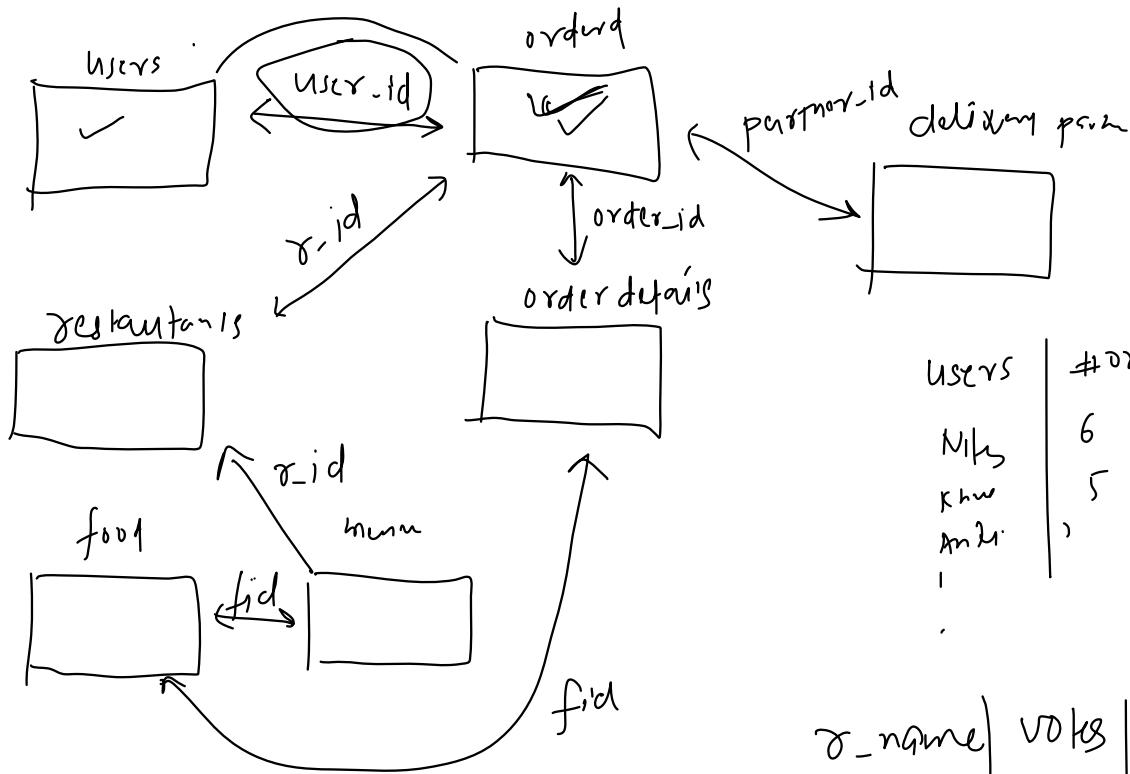
20 February 2023 15:35

1. Find all profitable orders
2. Find the customer who has placed max number of orders
3. Which is the most profitable category
4. Which is the most profitable state
5. Find all categories with profit higher than 5000



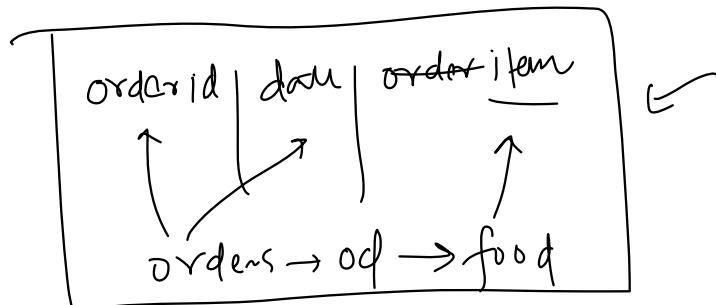
Zomato Case Study

21 February 2023 19:46



Users	#order
Niks	6
Khw	5
Anil	1

r_name	votes	avg_review
Domir	500	3.9
Br.c	600	4.1



15 may - 15 june

User	f_name	num
Niks	Piza	3
Khw	Br	4

Name	salary

What is a Subquery

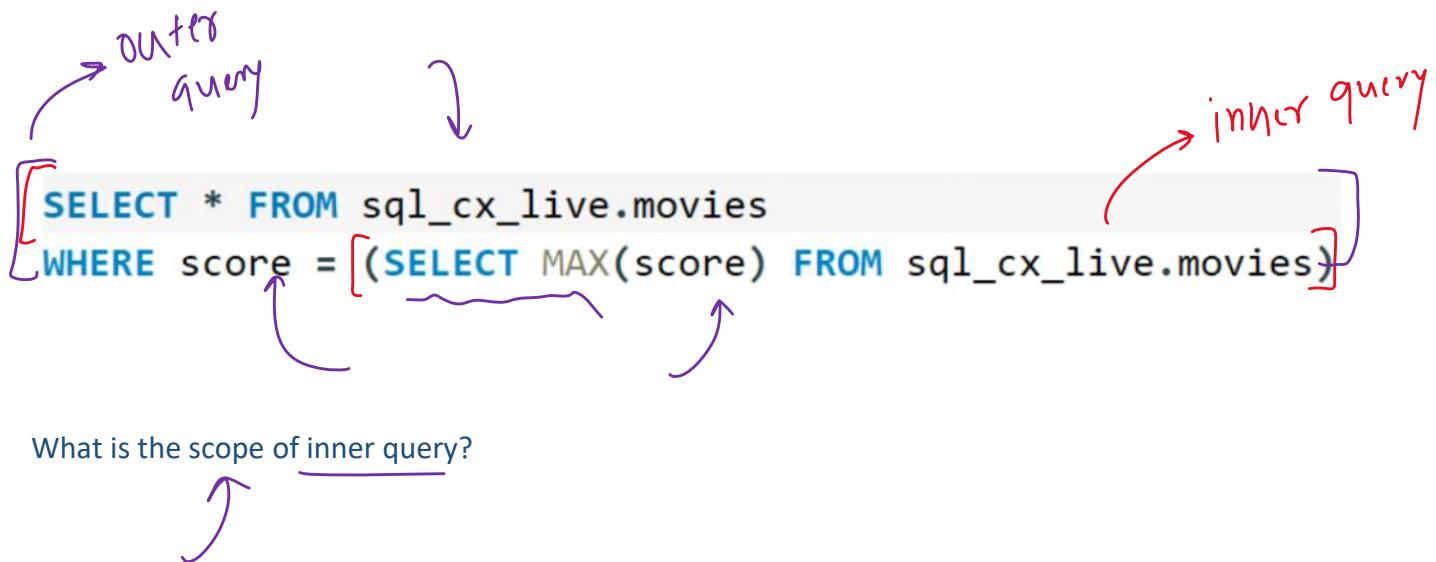
22 February 2023 14:25

In SQL, a subquery is a query within another query. It is a SELECT statement that is nested inside another SELECT, INSERT, UPDATE, or DELETE statement. The subquery is executed first, and its result is then used as a parameter or condition for the outer query.

Note - The topic is slightly difficult and needs a lot of practice

Example - Find the movie with highest rating

name	rating	genre	year	released	score	votes	director	writer
The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King
The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole
Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett
Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams
Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray
Friday the 13th	R	Horror	1980	May 9, 1980 (United States)	6.4	123000.0	Sean S. Cunningham	Victor Miller
The Blues Brothers	R	Action	1980	June 20, 1980 (United States)	7.9	188000.0	John Landis	Dan Aykroyd
Raging Bull	R	Biography	1980	December 19, 1980 (United States)	8.2	330000.0	Martin Scorsese	Jake LaMotta
Superman II	PG	Action	1980	June 19, 1981 (United States)	6.8	101000.0	Richard Lester	Jerry Siegel
The Long Riders	R	Biography	1980	May 16, 1980 (United States)	7.0	10000.0	Walter Hill	Bill Bryden
Any Which Way You Can	PG	Action	1980	December 17, 1980 (United States)	6.1	18000.0	Buddy Van Horn	Stanford Sherman



What is the scope of inner query?

Types of Subqueries

22 February 2023 14:26

Based on:

1. The result it returns ✓
2. Based on working

Returned Data

Scalar ✓
Subquery
(9.3)
horror

Row
Subquery

genre
horror
romance
action

Table
Subquery

genre	avg rating
horror	6.5
romance	7
action	6.8

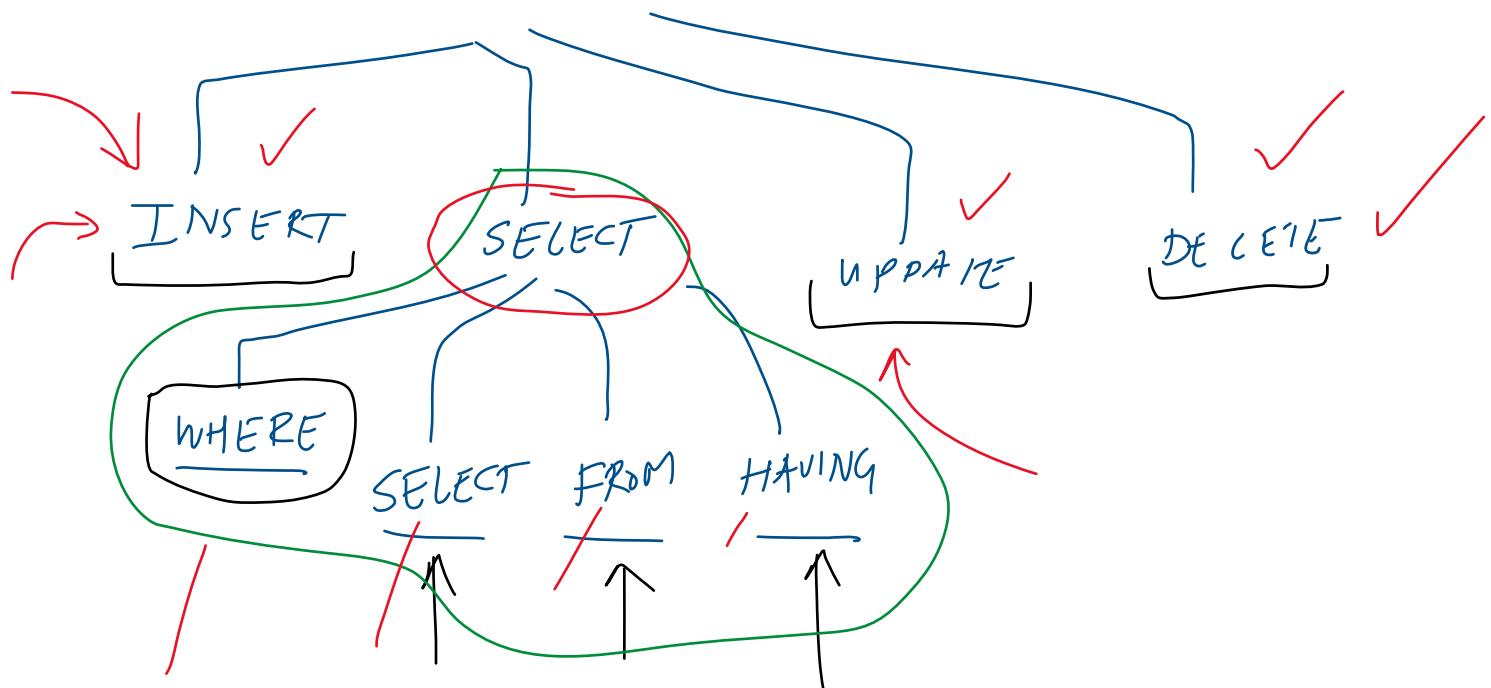
Working

Independent

Correlated

Where can subqueries be used?

22 February 2023 14:27



Independent Subquery - Scalar Subquery

22 February 2023 14:28

1. Find the movie with highest profit(vs order by)
2. Find how many movies have a rating > the avg of all the movie ratings(Find the count of above average movies) ✓
3. Find the highest rated movie of 2000
4. Find the highest rated movie among all movies whose number of votes are > the dataset avg votes

$$O(n) + O(n) = O(2n) \times$$

$$O(n \log n) + O(1)$$

$$\frac{O(n \log n)}{\text{indexing}} \times$$

Independent Subquery - Row Subquery(One Col Multi Rows)

22 February 2023 14:29

NO 1

IN

1. Find all users who never ordered
2. Find all the movies made by top 3 directors(in terms of total gross income)
3. Find all movies of all those actors whose filmography's avg rating > 8.5(take 25000 votes as cutoff)

Independent Subquery - Table Subquery(Multi Col Multi Row)

22 February 2023 14:29

1. Find the most profitable movie of each year
2. Find the highest rated movie of each genre votes cutoff of 25000
3. Find the highest grossing movies of top 5 actor/director combo in terms of total gross income



Correlated Subquery

22 February 2023 14:29

1. Find all the movies that have a rating higher than the average rating of movies in the same genre.[Animation]
2. Find the favorite food of each customer.

Usage with SELECT

22 February 2023 14:29

1. Get the percentage of votes for each movie compared to the total number of votes.

2. Display all movie names ,genre, score and avg(score) of genre

-> Why this is inefficient?

20 + w -

name	percentage of votes
------	---------------------

Usage with FROM

22 February 2023 14:30

1. Display average rating of all the restaurants

Usage with HAVING

22 February 2023 14:30

1. Find genres having avg score > avg score of all the movies

Subquery In INSERT

22 February 2023 14:30

Populate a already created loyal_customers table with records of only those customers who have ordered food more than 3 times.

Subquery in UPDATE

22 February 2023 14:31

Populate the money col of loyal_customer table using the orders table. Provide a 10% app money to all customers based on their order value.

Subquery in DELETE

22 February 2023 14:31

Delete all the customers record who have never ordered.

Database Server Vs Database Client

25 February 2023 17:35

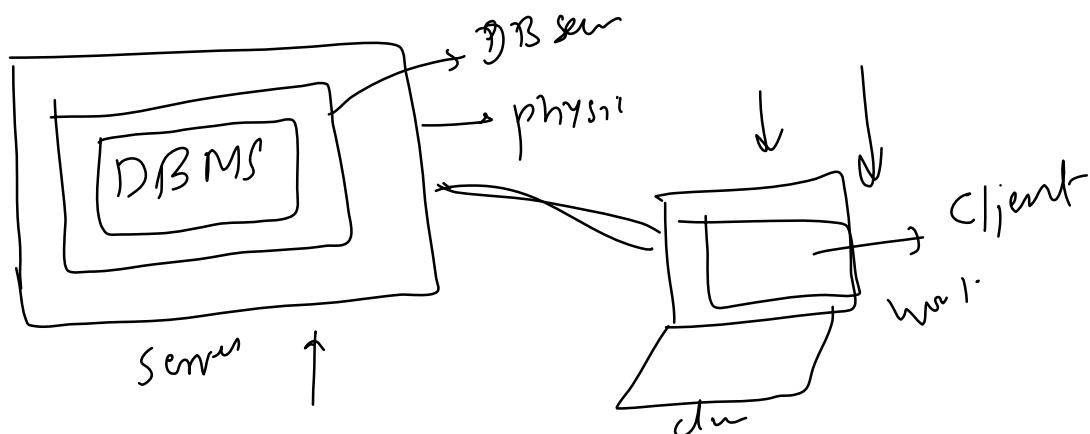
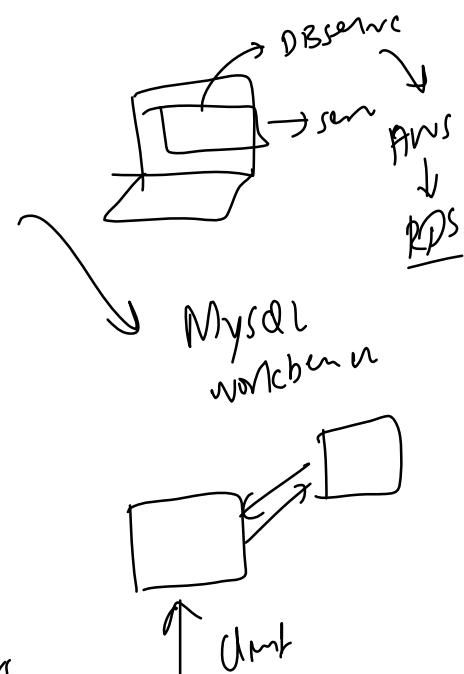
1) **Database Server:** A database server is a software application that provides access to a database over a network. It manages and processes requests from multiple clients, and performs tasks such as managing user authentication, concurrency control, and transaction management. A database server is typically installed on a dedicated computer or server, and is responsible for storing, managing, and processing data.

2) **Database Client:** A database client is a software application that connects to a database server and sends requests for data or database operations. It is the front-end interface that allows users to interact with the database, and provides tools such as query editors, data visualization, and reporting.

3. **Database:** A database is a collection of related data that is organized and stored in a structured format. It is designed to efficiently store, retrieve, and manage large amounts of data. A database can include multiple tables, indexes, and other objects that are used to manage and manipulate data.

4. **DBMS (Database Management System):** A DBMS is a software application that provides tools for creating, managing, and manipulating databases. It includes a variety of functions and features, such as data storage, data retrieval, data backup and recovery, user management, and security. The DBMS is responsible for managing the underlying database, and provides a mechanism for users to interact with the data.

Mysql, Oracle, SQL, PostgreSQL

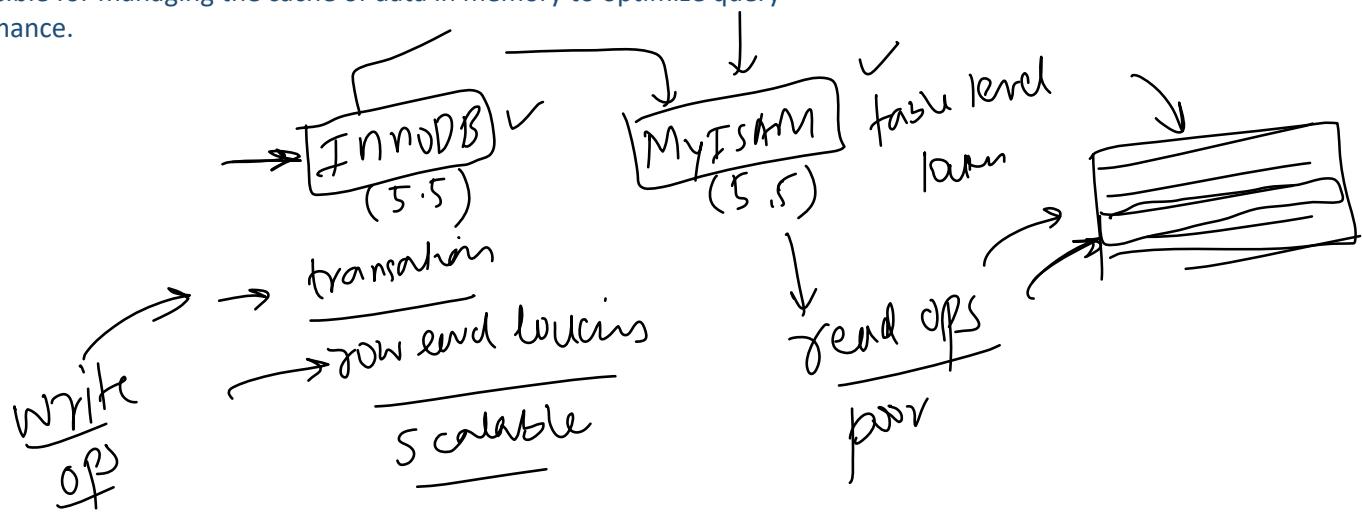


What are Database Engines

25 February 2023 17:35

A database engine, also known as a database management system (DBMS) engine, is the software component of a DBMS that is responsible for managing the storage, organization, and retrieval of data in a database. It is the core component of a DBMS that provides the necessary tools and services for creating, modifying, and querying the database.

A database engine typically includes several key components, such as a query optimizer, transaction manager, storage manager, and buffer manager. The query optimizer is responsible for optimizing SQL queries to retrieve data from the database efficiently, while the transaction manager ensures that multiple transactions are executed correctly and consistently. The storage manager handles the physical storage of the data on disk or in memory, and the buffer manager is responsible for managing the cache of data in memory to optimize query performance.



Components of DBMS

25 February 2023 19:19

1. Database Engine
2. Security and Access Control - used to manage user permissions and access rights to the database.
3. Backup and Recovery - used to create backups of the database and recover the data in case of failures.
4. Data Dictionary - used to store metadata about the database schema and data
5. User Interface - used to provide a graphical interface to interact with the database.

What is Collation

25 February 2023 17:35

Collation refers to the rules and algorithms used to compare and sort characters in a database. It determines how character strings are compared and sorted, including the order of the characters, the treatment of case sensitivity, and the handling of accent marks or other special characters.

Collation is important in database management because it affects the way queries are executed and results are returned. If the collation of a database is not set correctly, queries may return incorrect results or the database may not sort data properly.

1. Binary: Compares strings byte by byte. It is case-sensitive and accent-sensitive.
2. Case-insensitive: Compares strings without regard to case, but is accent-sensitive.
3. Accent-insensitive["café" and "cafe"]: Compares strings without regard to accents, but is case-sensitive.
4. Case- and accent-insensitive: Compares strings without regard to case or accents.
5. Unicode: Supports Unicode character sets, and is available in multiple variants, such as utf8mb4_unicode_ci, utf8mb4_unicode_520_ci, utf8mb4_unicode_520_ci_ai, etc.

Diff between COUNT(*) and COUNT(col) ←

25 February 2023 17:35

Select COUNT(*) from Table
↓ rows #

Dealing with NULL values

25 February 2023 17:36

1. How it deals with NON NULL values
2. Order By
3. Group By - NULL values are treated as a separate group and are not included in any group that contains non-NULL values.
4. Aggregate

When performing aggregate operations in MySQL, NULL values are treated differently depending on whether or not the GROUP BY clause is used.

Without GROUP BY:

- If the aggregate function is SUM, AVG, MAX, MIN, or COUNT, NULL values are ignored and not included in the calculation.
- If the aggregate function is GROUP_CONCAT or CONCAT, NULL values are included in the result, but a NULL value is returned if all the values being concatenated are NULL.

With GROUP BY:

- If the aggregate function is COUNT, NULL values are not included in the count for each group. However, if you use COUNT(*) instead of COUNT(column), then NULL values are included in the count.
- If the aggregate function is SUM, AVG, MAX, or MIN, NULL values are ignored and not included in the calculation for each group. If a group contains only NULL values, then the result for that group will be NULL.
- If the aggregate function is GROUP_CONCAT or CONCAT, NULL values are included in the result for each group, but a NULL value is returned if all the values being concatenated in a group are NULL.

- How to find null values?
- How to replace null values?

DELETE Vs TRUNCATE

25 February 2023 17:36

- DELETE is a Data Manipulation Language (DML) statement, whereas TRUNCATE is a Data Definition Language (DDL) statement. This means that TRUNCATE requires the ALTER TABLE privilege, whereas DELETE requires the DELETE privilege on the table.
- DELETE can be rolled back using a transaction log, which means that you can undo the changes made by DELETE if necessary. TRUNCATE, on the other hand, cannot be rolled back because it does not generate a transaction log.
- DELETE is slower than TRUNCATE because it generates transaction log entries for each deleted row. If you need to delete a large number of rows, TRUNCATE may be a better option for performance reasons.
- If you use foreign key constraints in your database, DELETE can cause integrity issues if you delete rows that are referenced by other tables. In this case, you should use TRUNCATE or disable the foreign key constraints before using DELETE.

Non-equi joins

25 February 2023 17:37

In a non equi join, the join condition is based on operators other than equality. Specifically, the join condition can use operators such as greater than, less than, or not equal to, among others. Non equi joins are useful when you need to join tables on columns with similar but not identical data, or when you need to join tables based on a range of values rather than an exact match.

Natural Joins

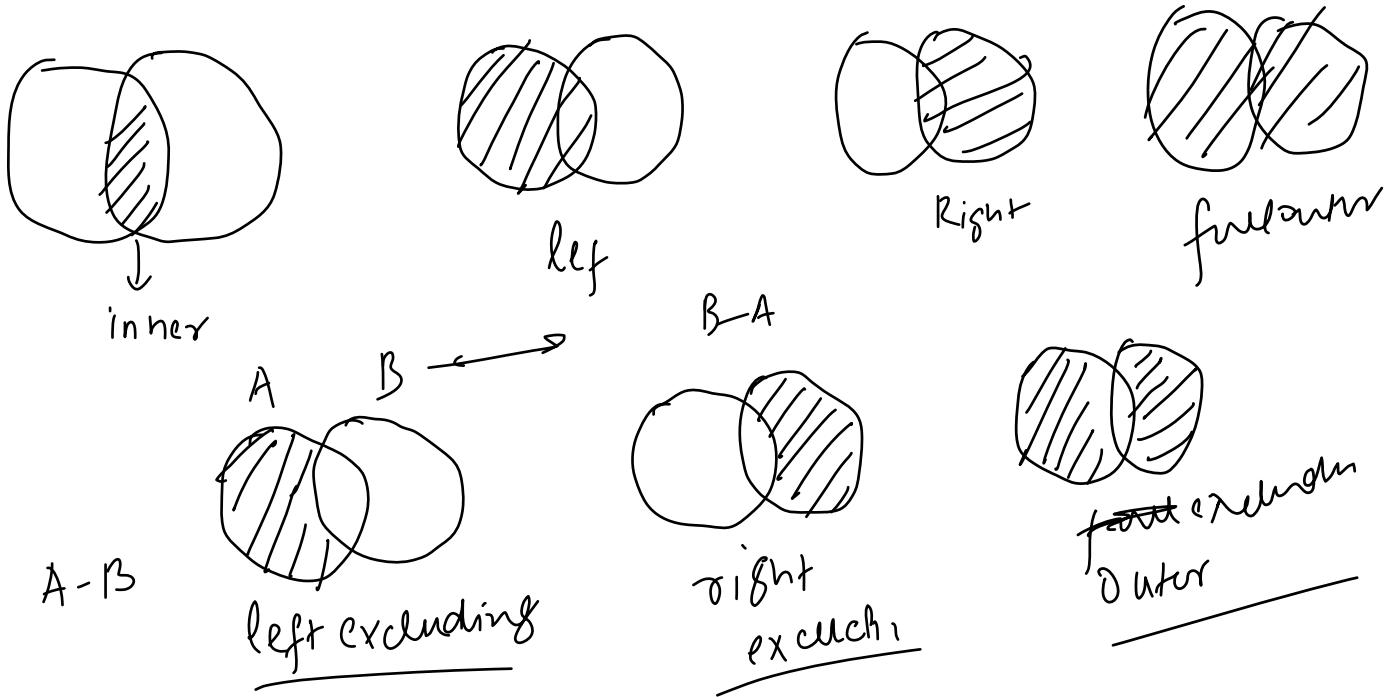
25 February 2023 17:37

A natural join is a type of join in SQL where two tables are joined based on the columns with the same name and data type. In other words, it is a join where the join condition is implicitly based on the column names that exist in both tables, and it eliminates the duplicate columns from the result set.

Anti Joins

25 February 2023 17:37

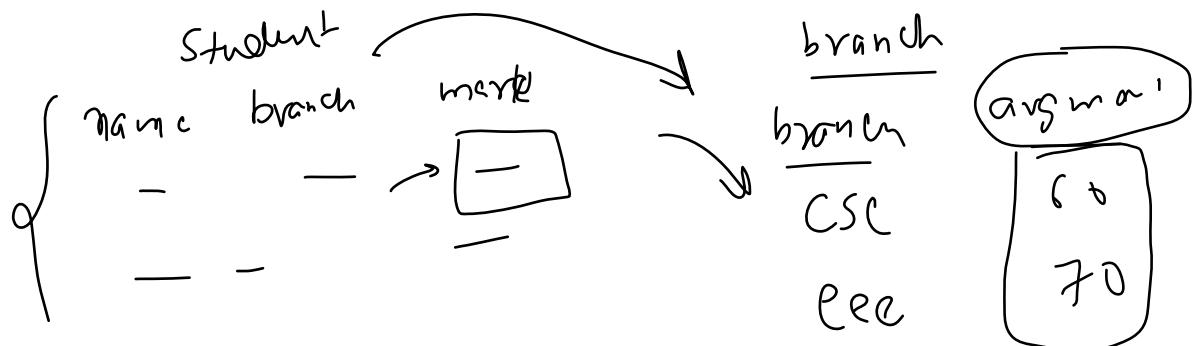
An excluding join, also known as an anti-join, is a type of join operation in SQL that returns only the rows from one table that do not have any matching rows in another table. In other words, it returns the rows that are not included in the result set of an inner join between the two tables.



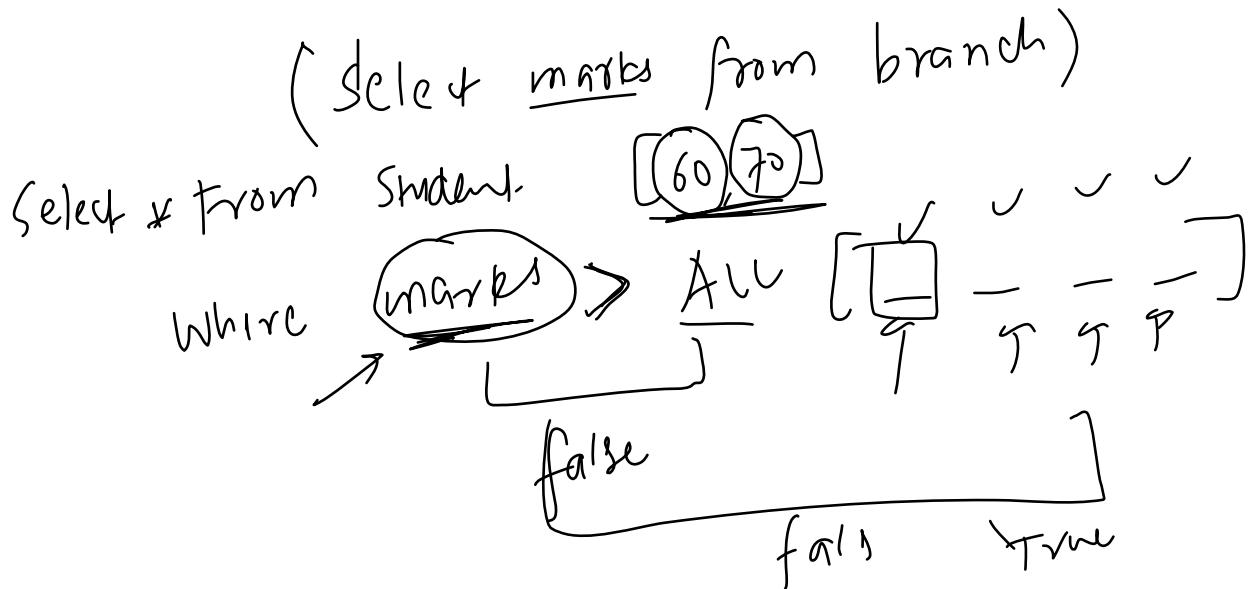
All and Any Operators

25 February 2023 17:37

All Any



ANY



Select * from students



Removing Duplicate Rows

25 February 2023 17:42

1. Find duplicate values
2. Delete duplicate

Metadata Queries

25 February 2023 17:42

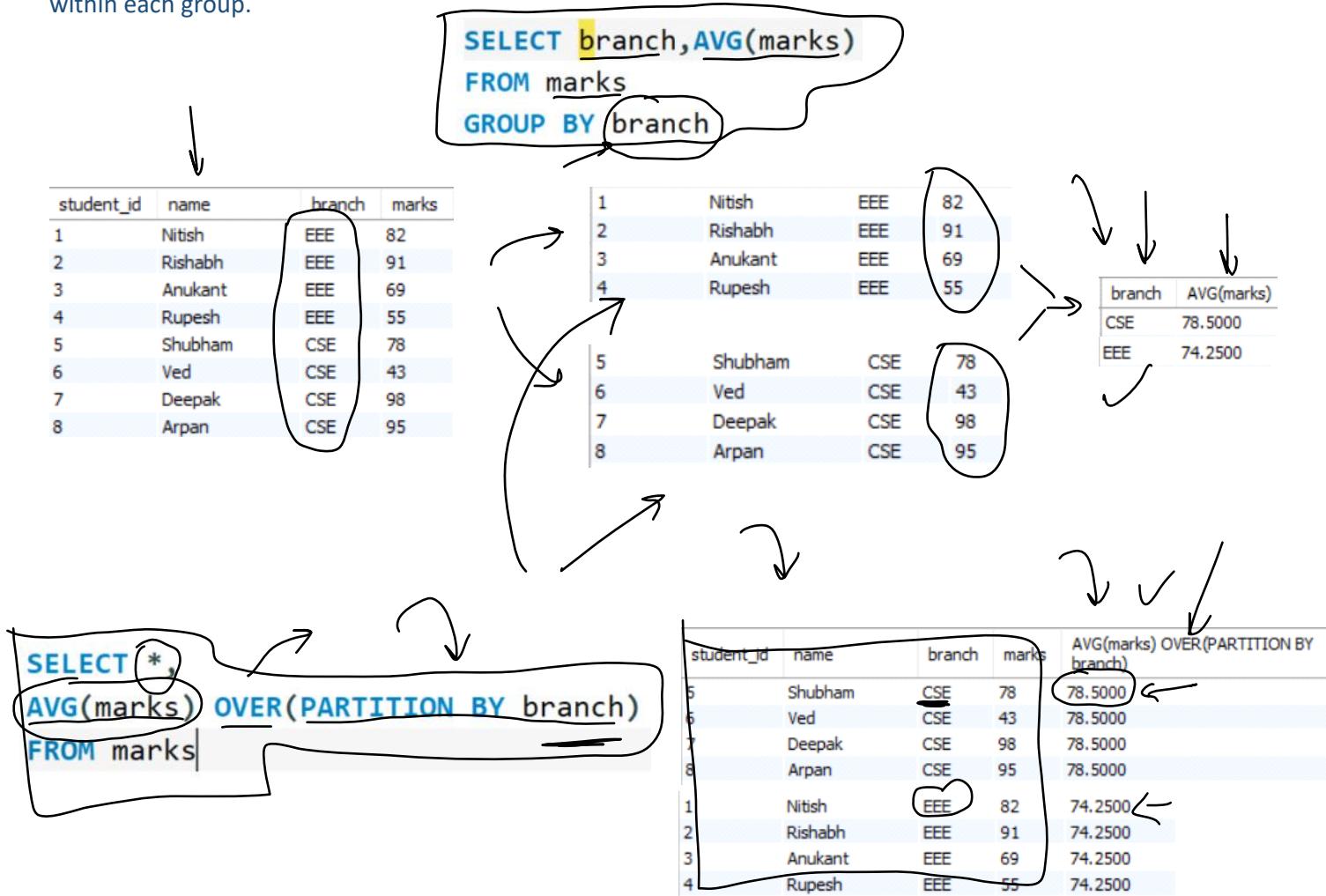
- How to see all the tables of a database
- How to print col names
- How to see all the constraints of a table (Homework)
- copy table definition -> How to create empty tables with the same structure as another table?

What are Window Functions?

27 February 2023 14:18

Window functions in SQL are a type of analytical function that perform calculations across a set of rows that are related to the current row, called a "window". A window function calculates a value for each row in the result set based on a subset of the rows that are defined by a window specification.

The window specification is defined using the OVER() clause in SQL, which specifies the partitioning and ordering of the rows that the window function will operate on. The partitioning divides the rows into groups based on a specific column or expression, while the ordering defines the order in which the rows are processed within each group.



Aggregate Function with OVER()

27 February 2023 16:41

Find all the students who have marks higher than the avg marks of their respective branch

RANK/DENSE_RANK/ROW_NUMBER

27 February 2023 16:56

1. Find top 2 most paying customers of each month
2. Create roll no from branch and marks

mark	ranc	dense_ran
95	—	—
95	—	—
89	— 3	— 2

row-number

FIRST_VALUE/LAST VALUE/NTH_VALUE

27 February 2023 16:56

1. Find the branch toppers
2. FRAME Clause
3. Find the last guy of each branch
4. Alternate way of writing Window functions
5. Find the 2nd last guy of each branch, 5th topper of each branch

Frames

27 February 2023 19:08

A frame in a window function is a subset of rows within the partition that determines the scope of the window function calculation. The frame is defined using a combination of two clauses in the window function: **ROWS** and **BETWEEN**.

The **ROWS** clause specifies how many rows should be included in the frame relative to the current row. For example, **ROWS 3 PRECEDING** means that the frame includes the current row and the three rows that precede it in the partition.

The **BETWEEN** clause specifies the boundaries of the frame.

Examples

- **ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW**: means that the frame includes all rows from the beginning of the partition up to and including the current row.
- **ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING**: the frame includes the current row and the row immediately before and after it.
- **ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING**: the frame includes all rows in the partition.
- **ROWS BETWEEN 3 PRECEDING AND 2 FOLLOWING**: the frame includes the current row and the three rows before it and the two rows after it.

The diagram illustrates the concept of a window function frame using a table of student marks and a vertical stack of marks.

Table:

	name	branch	marks
1	Nitish	EEE	82
2	Rishabh	EEE	91
3	Anukant	EEE	69
4	Rupesh	EEE	55
5	Shubham	CSE	78
6	Ved	CSE	43
7	Deepak	CSE	98
8	Arpan	CSE	95

Vertical Stack:

A vertical stack of marks is shown on the right, representing the rows of the table. The marks are grouped into four categories: 1st (R), N, A, and 2nd (Ru). The marks are: 91, 82, 69, 55, 78, 43, 98, and 95. Arrows point from the table rows to their corresponding marks in the stack.

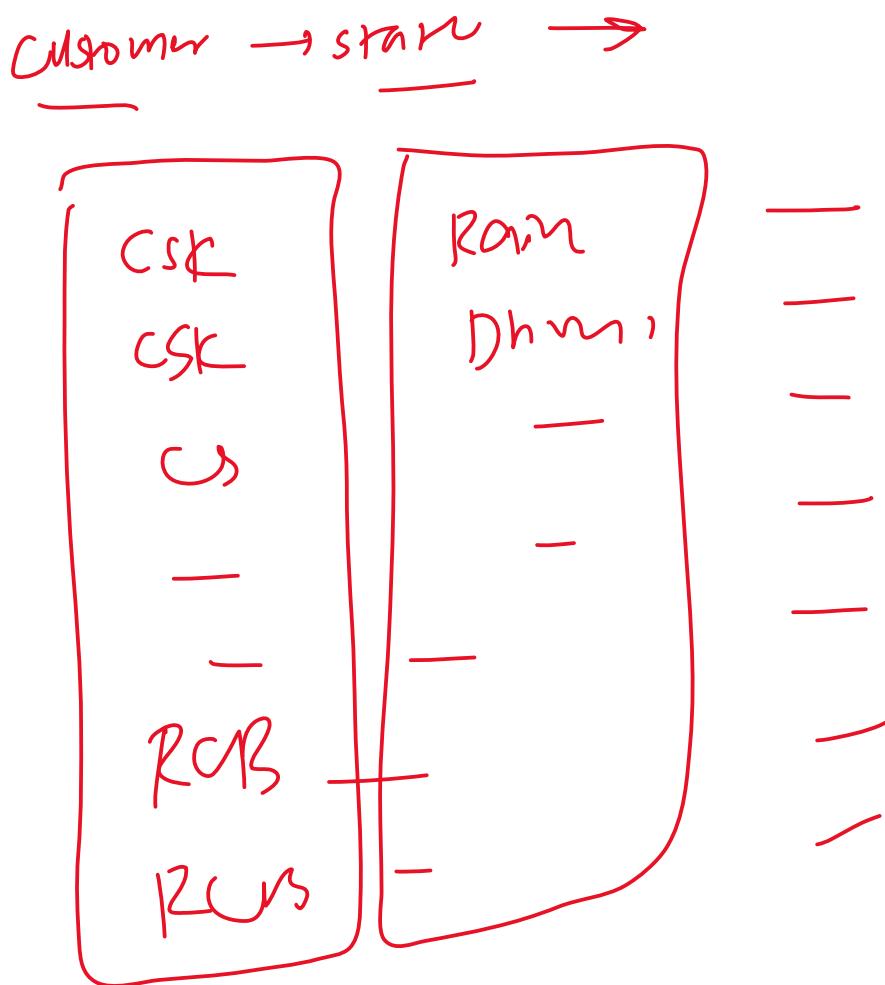
LEAD & LAG

27 February 2023 17:12

Find the MoM revenue growth of Zomato

Ranking

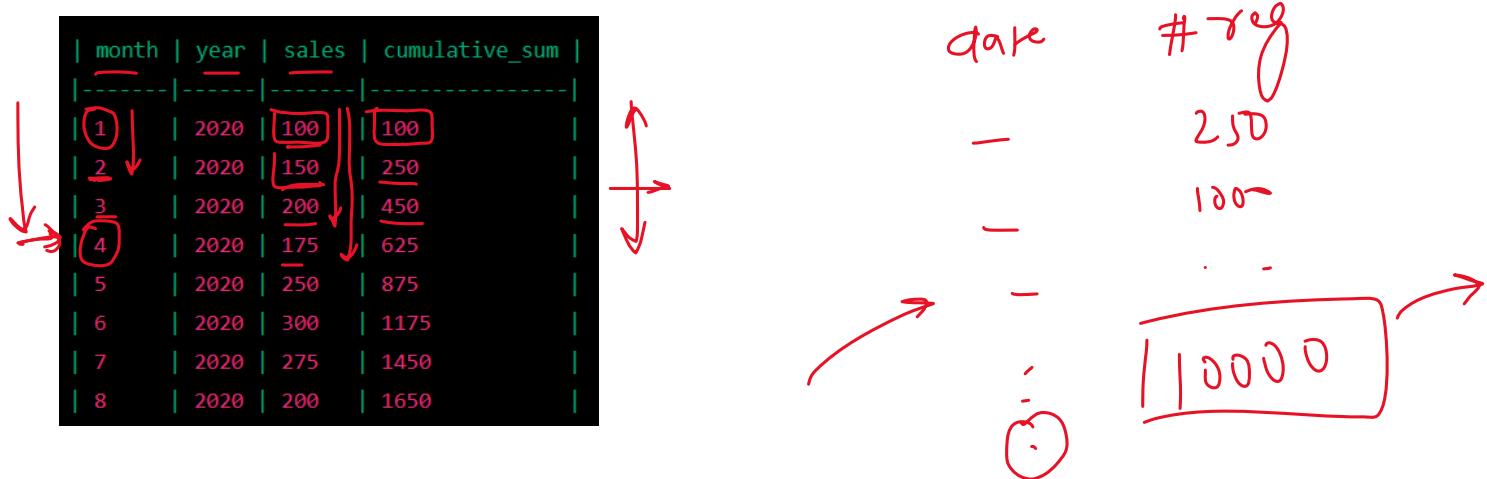
01 March 2023 13:57



Cumulative Sum

01 March 2023 13:58

Cumulative sum is another type of calculation that can be performed using window functions. A cumulative sum calculates the sum of a set of values up to a given point in time, and includes all previous values in the calculation.



→ VKohli

50th, 100th, 200th

↓ ↓ ↓

curr val curr curr

Cumulative Average

01 March 2023 13:58

Cumulative average is another type of average that can be calculated using window functions. A cumulative average calculates the average of a set of values up to a given point in time, and includes all previous values in the calculation.

Student_ID	Test_Number	Score	Cumulative_Avg
1	1	85	85.00
1	2	90	87.50
1	3	80	85.00
2	1	70	70.00
2	2	75	72.50
2	3	80	75.00
3	1	90	90.00
3	2	95	92.50
3	3	90	91.67

V Kohli →

Running Average

01 March 2023 13:58

Running average (also known as moving average) is a statistical technique that calculates the average value of a dataset over a moving window of consecutive data points.

The window size determines the number of data points used to calculate the average, and as the window moves forward in time, the average is recalculated using the new data points and dropping the oldest one. This means that the running average is continuously updated and reflects the most recent trends in the data.

For example, a running average of a batsman's runs scored over a window of 10 matches will calculate the average runs scored in the last 10 matches, then move the window one match forward and recalculate the average for the new set of 10 matches, and so on.

Running averages are often used in finance, economics, and engineering to smooth out noisy or volatile data series, and to identify trends or patterns that may be obscured by random fluctuations in the data.

window = 5 match

current form

match_id	runs_scored	running_avg	cumulative_avg
1	52	52.0	52.0
2	41	46.5	46.5
3	17	36.7	36.7
4	68	44.5	44.5
5	36	42.8	42.8
6	91	49.2	50.0
7	22	44.0	45.1
8	55	44.9	45.6
9	81	51.2	48.9
10	13	41.6	45.6
11	29	41.5	45.3
12	44	42.3	45.2
13	36	41.4	44.8
14	72	47.9	45.8
15	87	56.0	48.7

Percent of total

01 March 2023 13:59

Percent of total refers to the percentage or proportion of a specific value in relation to the total value. It is a commonly used metric to represent the relative importance or contribution of a particular value within a larger group or population.

category	total_sales	percent_of_total
Category A	500	50%
Category B	300	30%
Category C	200	20%

$\mu_m = 36.1.$

$\sigma_m = 10.1.$

Swiggy

domino's

Food

Percent Change

01 March 2023 13:59

Percent change is a way of expressing the difference between two values as a percentage of the original value. It is often used to measure how much a value has increased or decreased over a given period of time, or to compare two different values.



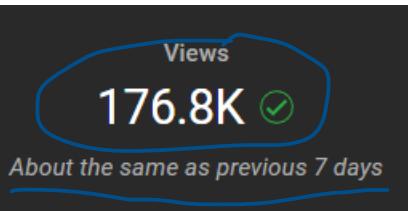
$$\text{percent change} = ((\text{new value} - \text{old value}) / \text{old value}) \times 100$$



Dec



A Jan



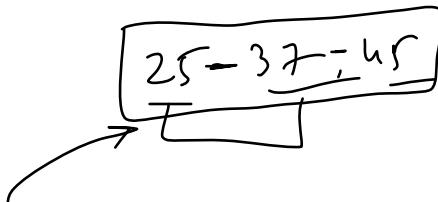
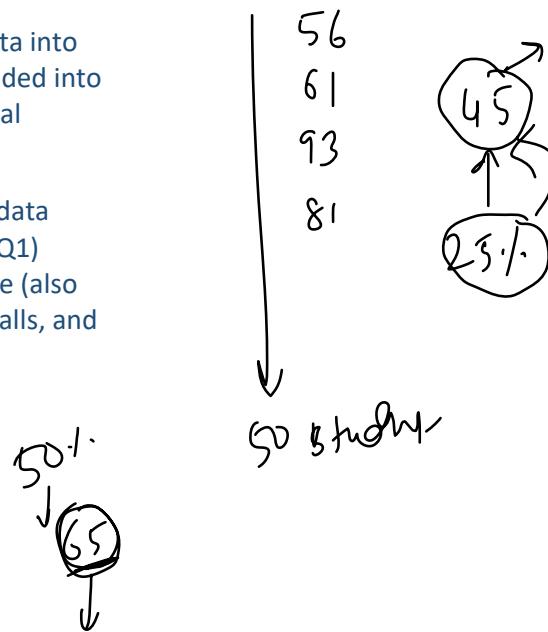
Percentiles & Quantiles

01 March 2023 13:59

A **Quantile** is a measure of the distribution of a dataset that divides the data into any number of equally sized intervals. For example, a dataset could be divided into **deciles** (ten equal parts), **quartiles** (four equal parts), **percentiles** (100 equal parts), or any other number of intervals.

Each quantile represents a value below which a certain percentage of the data falls. For example, the 25th percentile (also known as the first quartile, or Q1) represents the value below which 25% of the data falls. The 50th percentile (also known as the median) represents the value below which 50% of the data falls, and so on.

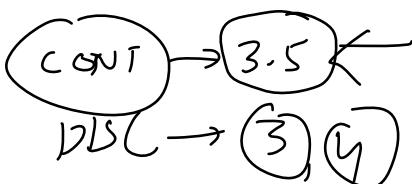
- Q1. Find the median marks of all the students ↪
Q2. Find branch wise median of student marks.



PERCENTILE_CONT calculates the continuous percentile value, which returns the interpolated value between adjacent data points. In other words, it estimates the percentile value by assuming that the values between data points are distributed uniformly. This function returns a value that may not be present in the original dataset.

PERCENTILE_DISC, on the other hand, calculates the discrete percentile value, which returns the value of the nearest data point. This function returns a value that is present in the original dataset.

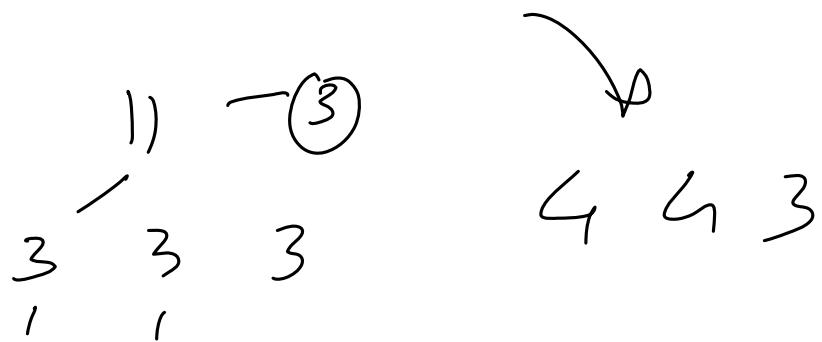
For example if we have 1,2,3,4,4,5



Segmentation

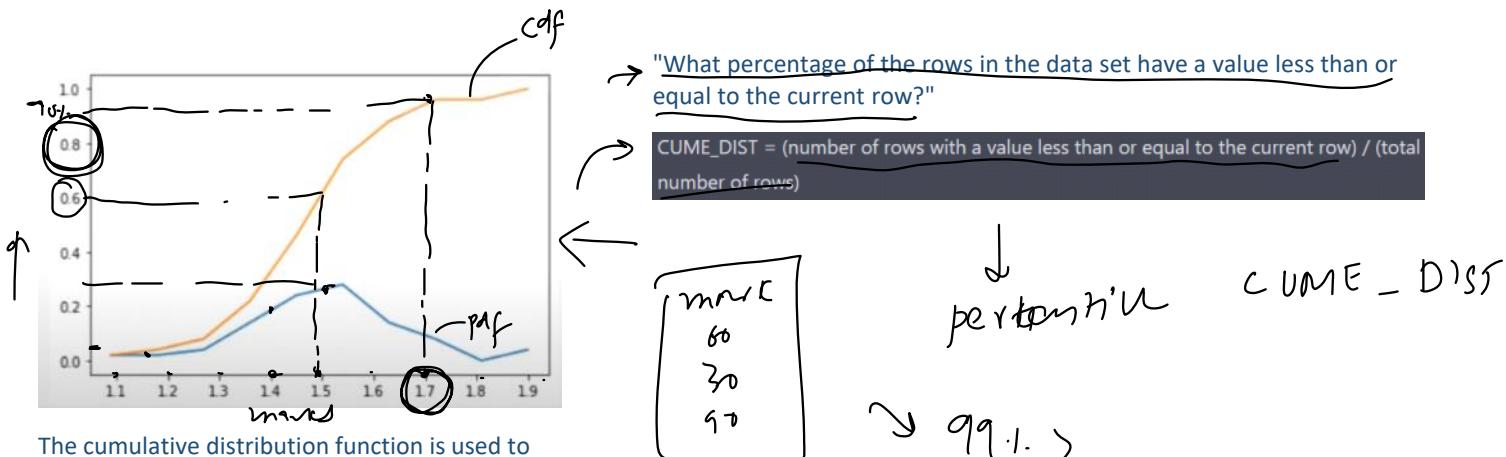
02 March 2023 09:30

Segmentation using NTILE is a technique in SQL for dividing a dataset into equal-sized groups based on some criteria or conditions, and then performing calculations or analysis on each group separately using window functions.



Cumulative Distribution

02 March 2023 09:03



The cumulative distribution function is used to describe the probability distribution of random variables. It can be used to describe the probability for a discrete, continuous or mixed variable. It is obtained by summing up the probability density function and getting the cumulative probability for a random variable

Partition By multiple columns

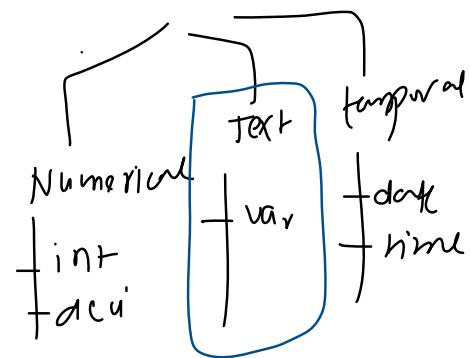
01 March 2023 14:00

String Data Types

03 March 2023 14:44

— — — (1) (2)

1. **CHAR:** This data type is used to store fixed-length strings. The length of the string is specified when the table is created, and the field will always use that amount of space, regardless of whether the string stored in it is shorter or longer. For example, if you define a CHAR(10) field and store the string "hello" in it, MySQL will pad the string with spaces so that it takes up 10 characters. CHAR fields are useful when you have a field that always contains the same length of data, such as a state abbreviation or a phone number.
2. **VARCHAR:** This data type is used to store variable-length strings. The length of the string can be up to a specified maximum, but the field will only use as much space as it needs to store the actual data. For example, if you define a VARCHAR(10) field and store the string "hello" in it, MySQL will only use 5 characters to store the data. VARCHAR fields are useful when you have a field that can contain varying amounts of data, such as a user's name or address.
3. **TEXT:** This data type is used to store larger amounts of variable-length string data than VARCHAR. It can store up to 65,535 characters. TEXT fields are useful when you need to store large amounts of text data, such as blog posts or comments.
4. **MEDIUMTEXT:** This data type is used to store even larger amounts of text data than TEXT. It can store up to 16,777,215 characters. MEDIUMTEXT fields are useful when you need to store very large amounts of text data, such as long-form articles or legal documents.
5. **LONGTEXT:** This data type is used to store the largest amounts of text data. It can store up to 4,294,967,295 characters. LONGTEXT fields are useful when you need to store extremely large amounts of text data, such as entire books or large collections of data.



char(10)

hello - - -

Text

Wildcards

03 March 2023 14:47

The LIKE operator in MySQL is used to match a string value against a pattern using wildcard characters. It is commonly used in SELECT, WHERE, and JOIN clauses to filter or join rows based on a pattern match.

The LIKE operator uses two wildcard characters: the **percent sign (%)** and the **underscore (_)**. The percent sign represents zero, one, or more characters, while the underscore represents a single character.

like

nit

wildcard

%

_

String Functions

03 March 2023 14:51

- **upper/lower**

- **concat & concat_ws**

- **substr -> last 5 chars**

- **replace**

- **reverse -> palindrome**

- **char_length vs length** -> where both are not same -

The main difference between CHAR_LENGTH and LENGTH is that CHAR_LENGTH returns the length of a string in characters, while LENGTH returns the length of a string in bytes. This difference is important when dealing with multi-byte character sets, such as UTF-8, where a single character may be represented by multiple bytes. Example - café

- **insert(str, pos, len newstr)**

- str: The original string to insert into.
- pos: The position at which to insert the new substring. The first position is 1.
- len: The number of characters to replace.
- newstr: The new substring to insert.

- **left and right**

- **repeat**

- **trim[ltrim and rtrim]**

- **substring_index(Split)** - www.campusx.in

- **strcmp** -

The STRCMP() function returns an integer that indicates the relationship between the two strings:

- If str1 is less than str2, the function returns a negative integer.
- If str1 is greater than str2, the function returns a positive integer.
- If str1 is equal to str2, the function returns 0.

- **locate("world", "hello world")**

- **lpad and rpad ('hello', 10, '*')**

Data Cleaning

03 March 2023 17:03

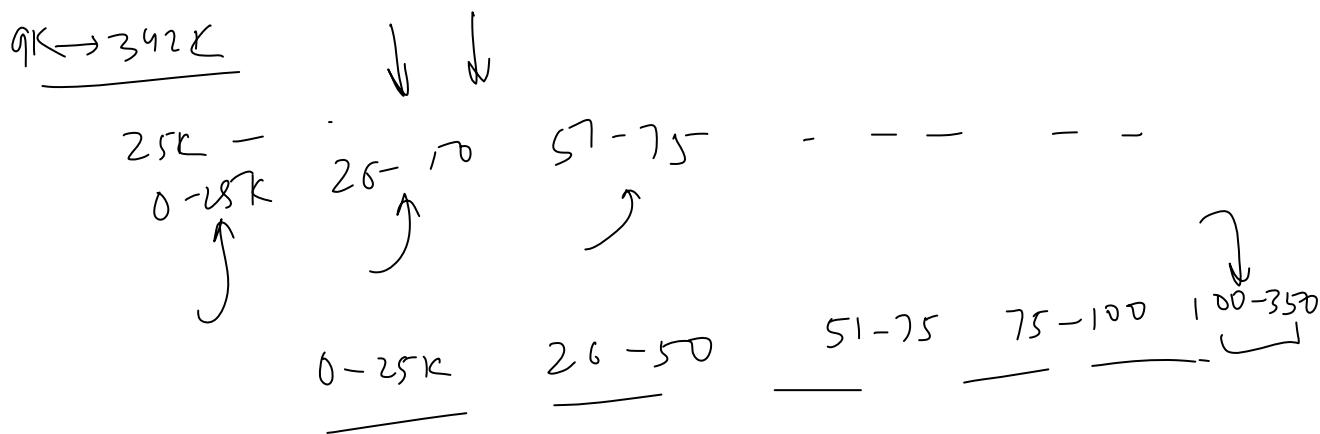
1. Create backup
2. Check number of rows
3. Check memory consumption for reference
4. Drop non important cols
5. Drop null values
6. Drop duplicates
7. Clean RAM -> change col data type
8. Clean weight -> change col type
9. ROUND price col and change to integer
10. Change the OpSys col
11. Gpu
12. Cpu

Extra

03 March 2023 17:14

```
SELECT * FROM laptops
WHERE Company IS NULL AND TypeName IS NULL AND Inches IS NULL
AND ScreenResolution IS NULL AND Cpu IS NULL AND Ram IS NULL
AND Memory IS NULL AND Gpu IS NULL AND OpSys IS NULL AND
WEIGHT IS NULL AND Price IS NULL
```

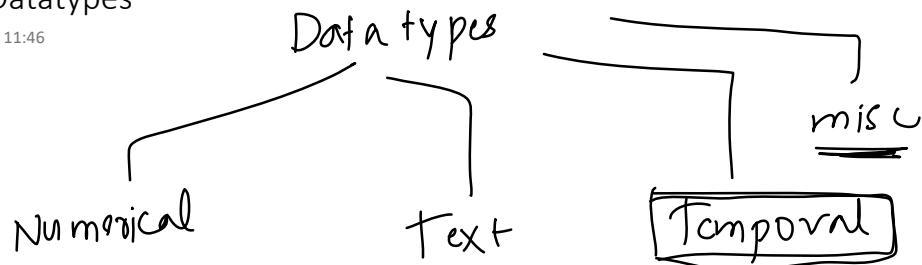
Memory	Type	primary-storage		secondary storage
		HDD	SSD	
	HDD	1024	0	1024
	Hybrid	128	0	1024
	SSD	256	0	0



App'c	0	
Dell	1	
brand	Touch	
Apple	0	
Dell	(37)	(43)
HP	(110)	(28)
	—	—

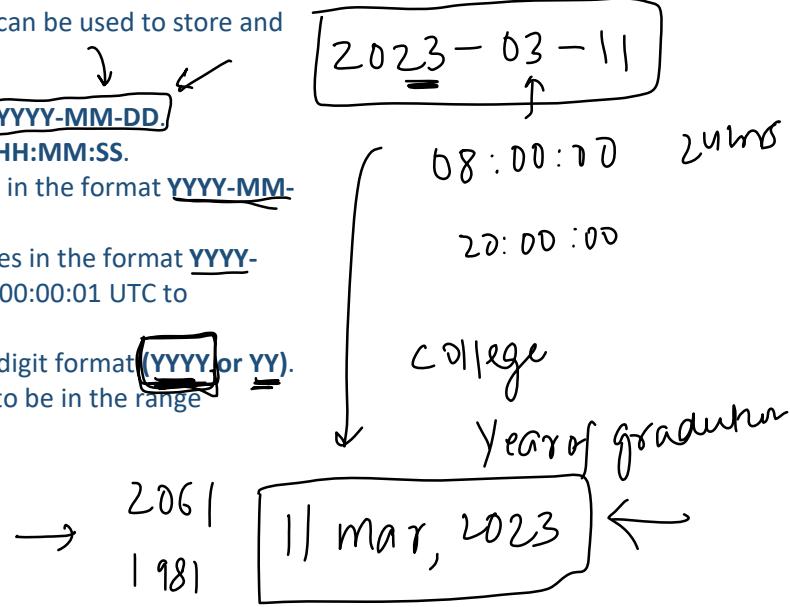
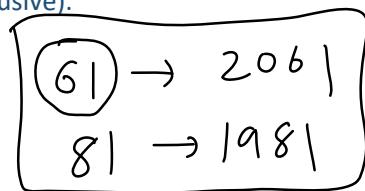
Temporal Datatypes

24 February 2023 11:46



In MySQL, there are several temporal data types that can be used to store and manipulate time and date values. These include:

1. **DATE** - used for storing date values in the format YYYY-MM-DD.
2. **TIME** - used for storing time values in the format HH:MM:SS.
3. **DATETIME** - used for storing date and time values in the format YYYY-MM-DD HH:MM:SS.
4. **TIMESTAMP** - used for storing date and time values in the format YYYY-MM-DD HH:MM:SS. It has a range of 1970-01-01 00:00:01 UTC to 2038-01-19 03:14:07 UTC.
5. **YEAR** - used for storing year values in 2-digit or 4-digit format (YYYY or YY). If the year is specified with 2 digits, it is assumed to be in the range 1970-2069 (inclusive).



Creating and Populating Temporal Tables

24 February 2023 11:47

1. Uber -> user_id, cab_id, start_time, end_time

DATETIME Functions

24 February 2023 11:50

1. CURR_DATE()
 2. CURR_TIME()
 3. NOW()
-

Extraction Function

1. DATE() and TIME()
2. YEAR()
3. DAY() or DAYOFMONTH()
4. DAYOFWEEK()
5. DAYOFYEAR()
6. MONTH() and MONTHNAME()
7. QUARTER()
8. WEEK() or WEEKOFYEAR()
9. HOUR() -> MINUTE() -> SECOND()
10. LAST_DAY()

Datetime Formatting

24 February 2023 11:48

`DATE_FORMAT()`

`TIME_FORMAT()`

Type conversion

24 February 2023 11:48

1. Implicit Type Conversion
2. Explicit Type Conversion -> STR_TO_DATE()

23:30:00 → 11:30 pm

11:45:56 → 11:45 am

DATETIME Arithmetic

24 February 2023 11:50

1. DATEDIFF()
2. TIMEDIFF()
3. DATE_ADD() and DATE_SUB() INTERVAL
4. ADDTIME() and SUBTIME()

TIMESTAMP VS DATETIME

11 March 2023 11:10

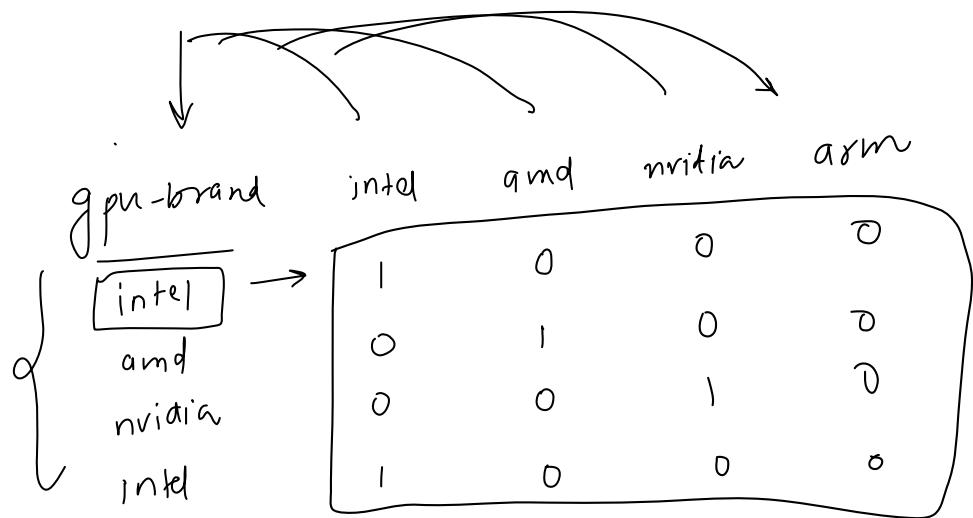
In MySQL, both DATETIME and TIMESTAMP are used to store date and time values, but they differ in their range, storage format, and behaviour.

Here are the main differences between DATETIME and TIMESTAMP:

1. **Range:** DATETIME supports a range of '1000-01-01 00:00:00' to '9999-12-31 23:59:59', while TIMESTAMP supports a range of '1970-01-01 00:00:01' UTC to '2038-01-19 03:14:07' UTC.
2. **Storage format:** DATETIME uses 8 bytes to store the date and time values, while TIMESTAMP uses 4 bytes.
3. **Behaviour on insertion/update:** DATETIME values are stored as-is, without any conversion, while TIMESTAMP values are converted from the current time zone to UTC when inserted, and converted back to the current time zone when retrieved.
4. **Precision:** DATETIME can store up to microseconds (6 digits after the decimal point), while TIMESTAMP can only store up to seconds.
5. **Auto-update:** TIMESTAMP columns can be set to update automatically whenever the row is inserted or updated, using the ON UPDATE CURRENT_TIMESTAMP clause.

In general, you should use DATETIME when you need to store date and time values outside the range of TIMESTAMP, or when you need to store values with greater precision than TIMESTAMP. You should use TIMESTAMP when you need to store values that can be automatically updated, or when you want to take advantage of its smaller storage format.

1. head - tail - sample
2. for numerical cols
 - 8 number summary [count, min, max, mean, std, q1, q2, q3]
 - missing values
 - outliers
 - horizontal/vertical histograms
3. for categorical cols
 - value counts
 - pie chart
 - missing value
4. numerical - numerical
 - side by side 8 number analysis--
 - scatterplot
 - correlation
5. categorical-categorical
 - contingency table
 - stacked bar chart
6. numerical-categorical
 - compare distribution across categories
8. missing value treatment
9. feature engineering
 - ppi
 - screen_size_bracket
10. one hot encoding



What is Statistics

09 March 2023 14:56

Statistics is a branch of mathematics that involves collecting, analysing, interpreting, and presenting data. It provides tools and methods to understand and make sense of large amounts of data and to draw conclusions and make decisions based on the data.

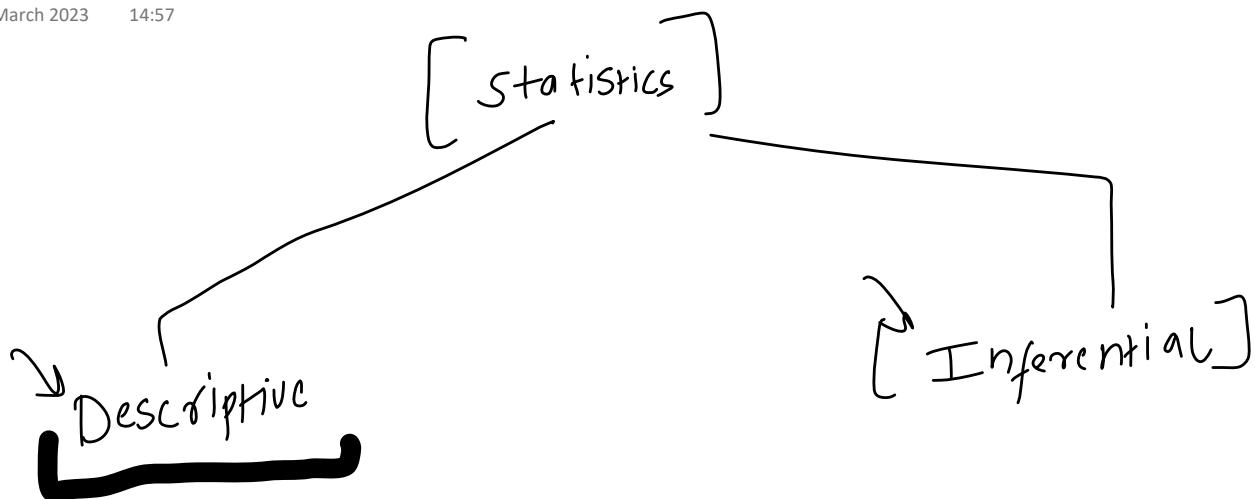
In practice, statistics is used in a wide range of fields, such as business, economics, social sciences, medicine, and engineering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions, and make forecasts and predictions.

Examples:

1. Business - Data Analysis(Identifying customer behavior) and Demand Forecasting
2. Medical - Identify efficacy of new medicines(Clinical trials), Identifying risk factor for diseases(Epidemiology)
3. Government & Politics - Conducting surveys, Polling
4. Environmental Science - Climate research

Types of Statistics

09 March 2023 14:57



Descriptive statistics deals with the collection, organization, analysis, interpretation, and presentation of data. It focuses on summarizing and describing the main features of a set of data, without making inferences or predictions about the larger population.

Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables

Population Vs Sample

09 March 2023 14:57

Population refers to the entire group of individuals or objects that we are interested in studying. It is the complete set of observations that we want to make inferences about. For example, the population might be all the students in a particular school or all the cars in a particular city.

A **sample**, on the other hand, is a subset of the population. It is a smaller group of individuals or objects that we select from the population to study. Samples are used to estimate characteristics of the population, such as the mean or the proportion with a certain attribute. For example, we might randomly select 100 students.

Examples

1. All cricket fans vs fans who were present in the stadium
2. All students vs who visit college for lectures

Things to be careful about when creating samples

- 
1. Sample Size ←
 2. Random ←
 3. Representative ←

Parameter Vs Statistics

A parameter is a characteristic of a population, while a statistic is a characteristic of a sample. Parameters are generally unknown and are estimated using statistics. The goal of statistical inference is to use the information obtained from the sample to make inferences about the population parameters.

Inferential Statistics

09 March 2023 14:57

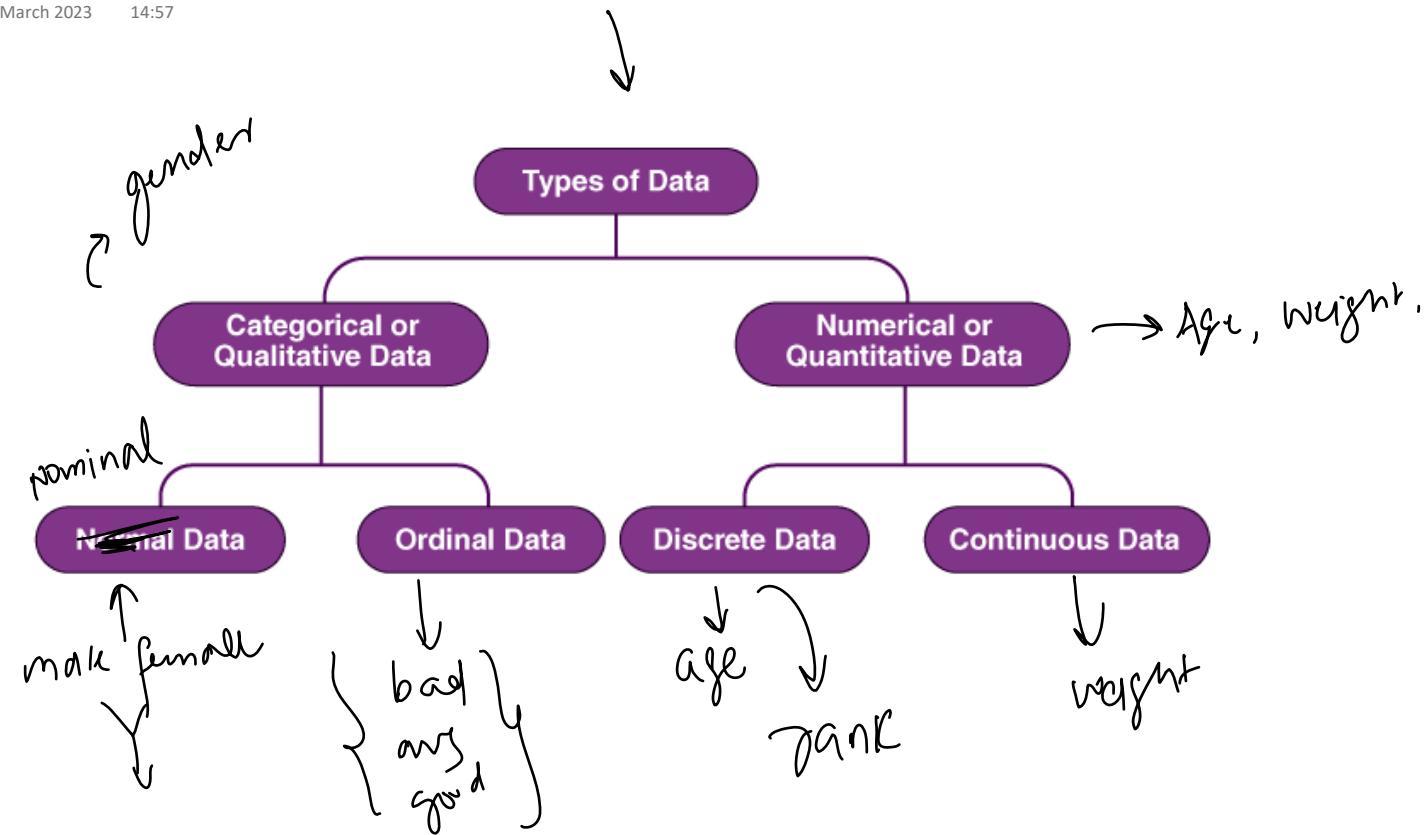
Inferential statistics is a branch of statistics that deals with making inferences or predictions about a larger population based on a sample of data. It involves using statistical techniques to test hypotheses and draw conclusions from data. Some of the topics that come under inferential statistics are:

1. **Hypothesis testing:** This involves testing a hypothesis about a population parameter based on a sample of data. For example, testing whether the mean height of a population is different from a given value.
2. **Confidence intervals:** This involves estimating the range of values that a population parameter could take based on a sample of data. For example, estimating the population mean height within a given confidence level.
3. **Analysis of variance (ANOVA):** This involves comparing means across multiple groups to determine if there are any significant differences. For example, comparing the mean height of individuals from different regions.
4. **Regression analysis:** This involves modelling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a product based on advertising expenditure.
5. **Chi-square tests:** This involves testing the independence or association between two categorical variables. For example, testing whether gender and occupation are independent variables.
6. **Sampling techniques:** This involves ensuring that the sample of data is representative of the population. For example, using random sampling to select individuals from a population.
7. **Bayesian statistics:** This is an alternative approach to statistical inference that involves updating beliefs about the probability of an event based on new evidence. For example, updating the probability of a disease given a positive test result.

Why ML is closely associated with statistics?

Types of Data

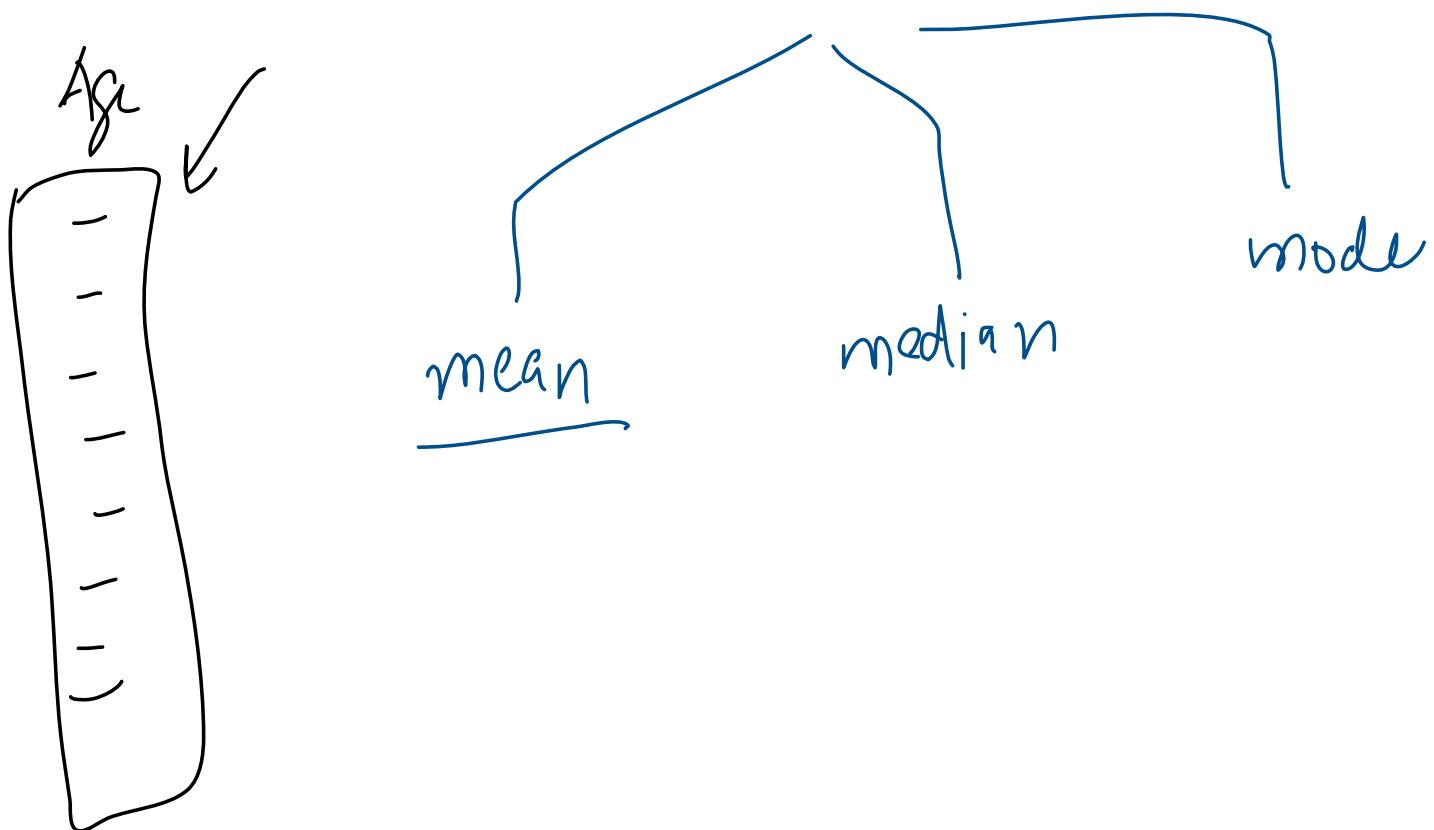
09 March 2023 14:57



Measure of Central Tendency

09 March 2023 14:58

A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.



1. Mean

09 March 2023 16:35

Mean: The mean is the sum of all values in the dataset divided by the number of values.

$$\frac{3+1+2}{5} = \frac{6}{5} = 1.2$$

2
5

$$\frac{\sum_{i=1}^N x_i}{N}$$

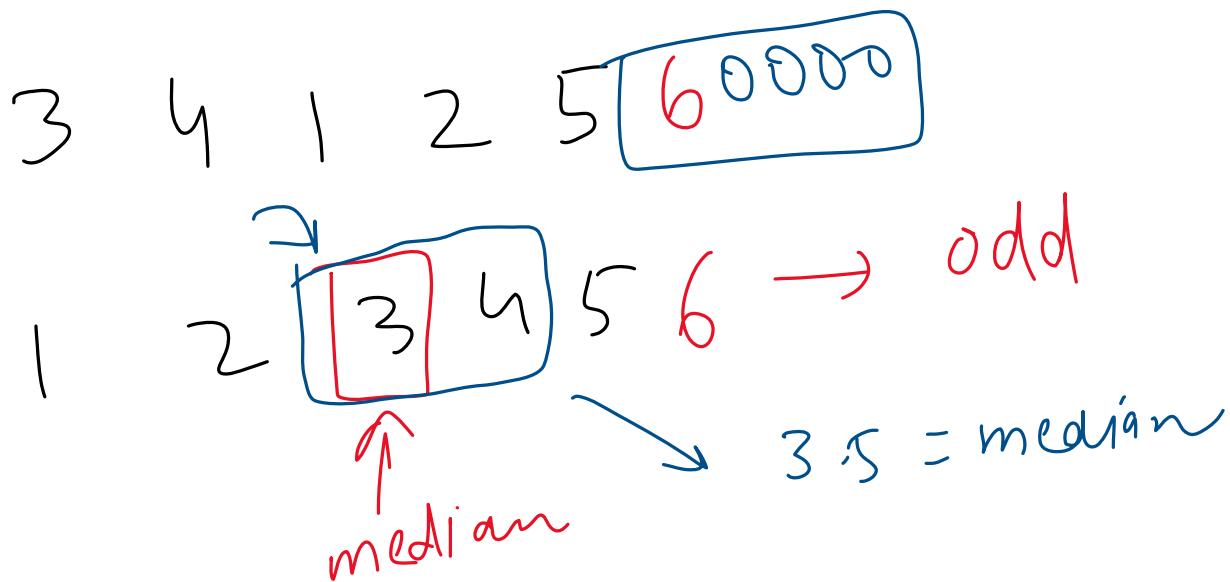
$$\frac{\sum_{i=1}^n x_i}{n}$$

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$
<u>$N = \text{number of items in the population}$</u>	<u>$n = \text{number of items in the sample}$</u>

2. Median

09 March 2023 16:36

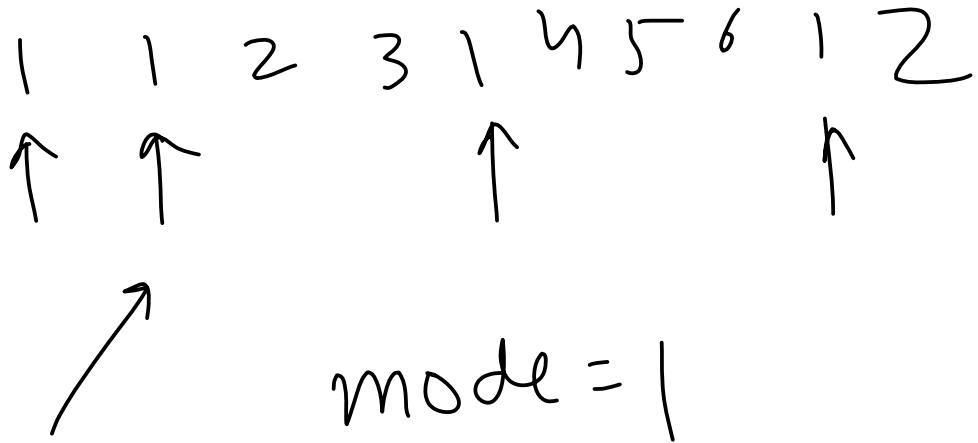
Median: The median is the middle value in the dataset when the data is arranged in order.



3. Mode

09 March 2023 16:36

Mode: The mode is the value that appears most frequently in the dataset.



4. Weighted Mean

09 March 2023 16:39

Weighted Mean: The weighted mean is the sum of the products of each value and its weight, divided by the sum of the weights. It is used to calculate a mean when the values in the dataset have different importance or frequency.

$$\boxed{1} \quad \boxed{2} \quad \boxed{3} = \boxed{\frac{1+2+3}{3}}$$

0 {

$$\begin{aligned} &\rightarrow \boxed{LR} \quad 0.2 \rightarrow 10L \\ &\rightarrow \boxed{RF} \quad 0.3 \rightarrow 15L \\ &\rightarrow \boxed{X_{fb50}} \quad 0.5 \rightarrow 12L \end{aligned}$$
$$\begin{aligned} &0.2 \times 10L + 0.3 \times 15L + 0.5 \times 12L \\ &0.2 + 0.3 + 0.5 \end{aligned}$$

5. Trimmed Mean

10 March 2023 09:37

A trimmed mean is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called the trimming percentage.

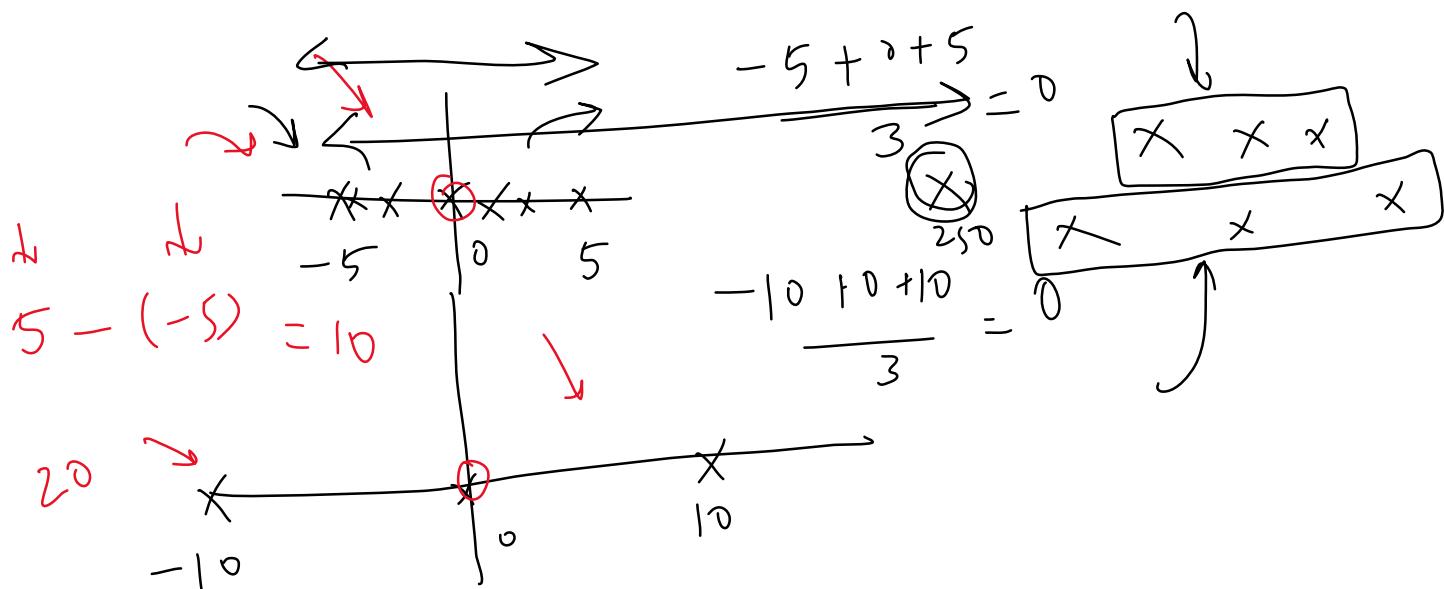
20.1.



Measure of Dispersion

09 March 2023 14:58

A measure of dispersion is a statistical measure that describes the spread or variability of a dataset. It provides information about how the data is distributed around the central tendency (mean, median or mode) of the dataset.



1. Range

09 March 2023 16:36

Range: The range is the difference between the maximum and minimum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.

2. Variance

09 March 2023 16:36

Variance: The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

(3)

$$\bar{x} = 3 \quad \sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

add -5, 0, 5

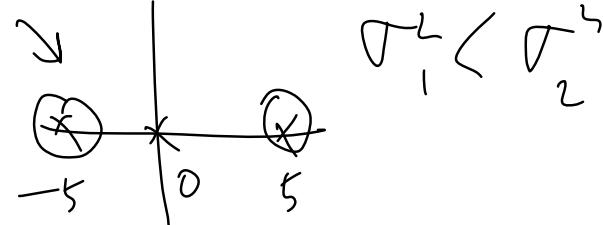
	X-mean	(X-mean) ²
3 .	3-3	0
2 .	2-3	1
1 .	1-3	4
5 .	5-3	4
4 .	4-3	1

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}{n}$$

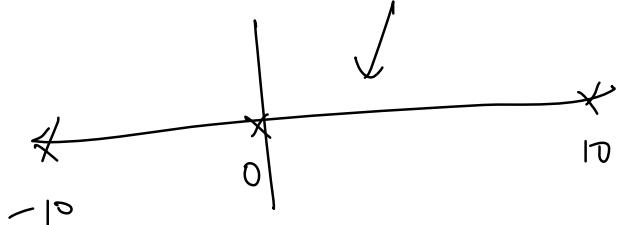
$$= \frac{(-5)^2 + (0)^2 + (5)^2}{3} = \frac{50}{3}$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad \text{Population Variance}$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \quad \text{Sample Variance}$$



Mean Absolute Deviation



$$MAD = \frac{\sum |x_i - \bar{x}|}{n}$$

inference



3. Standard Deviation

09 March 2023 16:37

$$\sqrt{2} \quad 1.41$$

Standard Deviation: The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

SD unit → same → data

$$\sigma^2 = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Population Variance

$$s^2 = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Sample Variance

SD

SD

15

17

13

14

→ 15

$$\frac{(15-15)^2 + (17-15)^2 + (13-15)^2 + (14-15)^2}{4}$$

4. Coefficient of Variation

09 March 2023 16:37

Salary

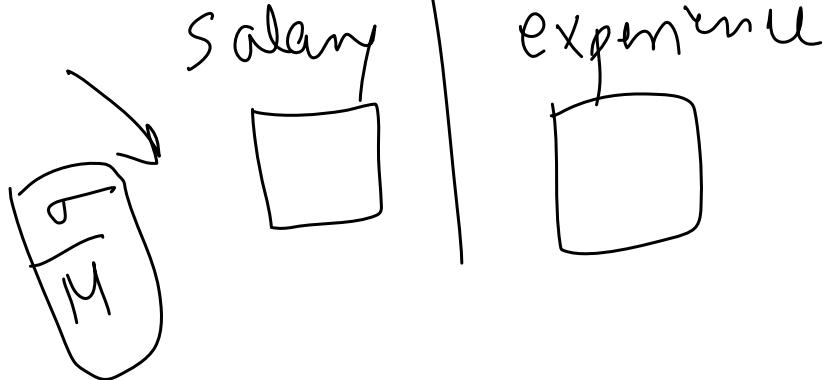
Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in fields such as biology, chemistry, and engineering.

The coefficient of variation (CV) is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is a dimensionless quantity that is expressed as a percentage.

The formula for calculating the coefficient of variation is:

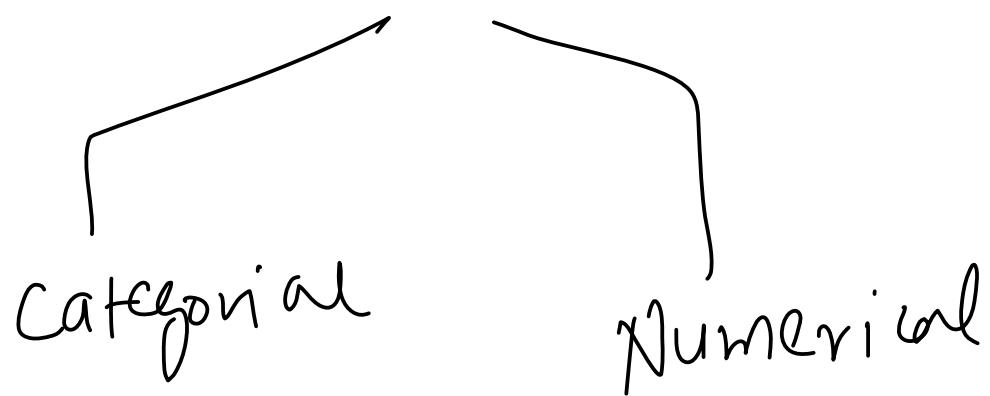
$$CV = \underbrace{(\text{standard deviation} / \text{mean}) \times 100\%}$$

CV σ / μ



Graphs for Univariate Analysis

09 March 2023 14:58



1. Categorical - Frequency Distribution Table & Cumulative Frequency

09 March 2023 16:50

A **frequency distribution table** is a table that summarizes the number of times (or frequency) that each value occurs in a dataset.

Let's say we have a survey of 200 people and we ask them about their favourite type of vacation, which could be one of six categories: Beach, City, Adventure, Nature, Cruise, or Other

Type of Vacation	Frequency
Beach	60
City	40
Adventure	30
Nature	35
Cruise	20
Other	15

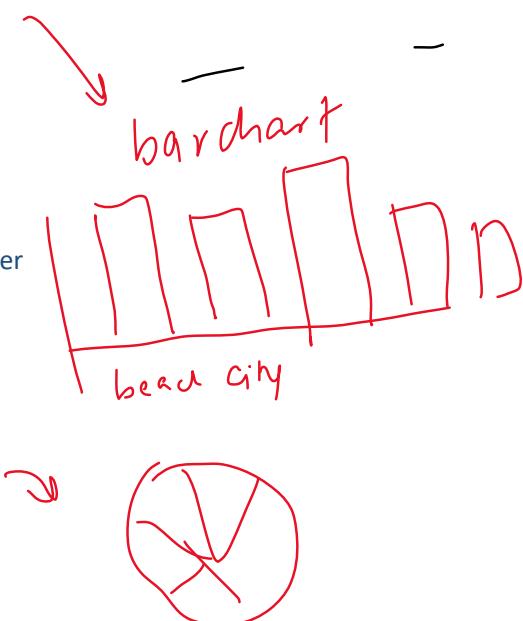
Relative frequency is the proportion or percentage of a category in a dataset or sample. It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample.

Type of Vacation	Frequency	Relative Frequency
Beach	60	0.3
City	40	0.2
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
Other	15	0.075

Cumulative frequency is the running total of frequencies of a variable or category in a dataset or sample. It is calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample.

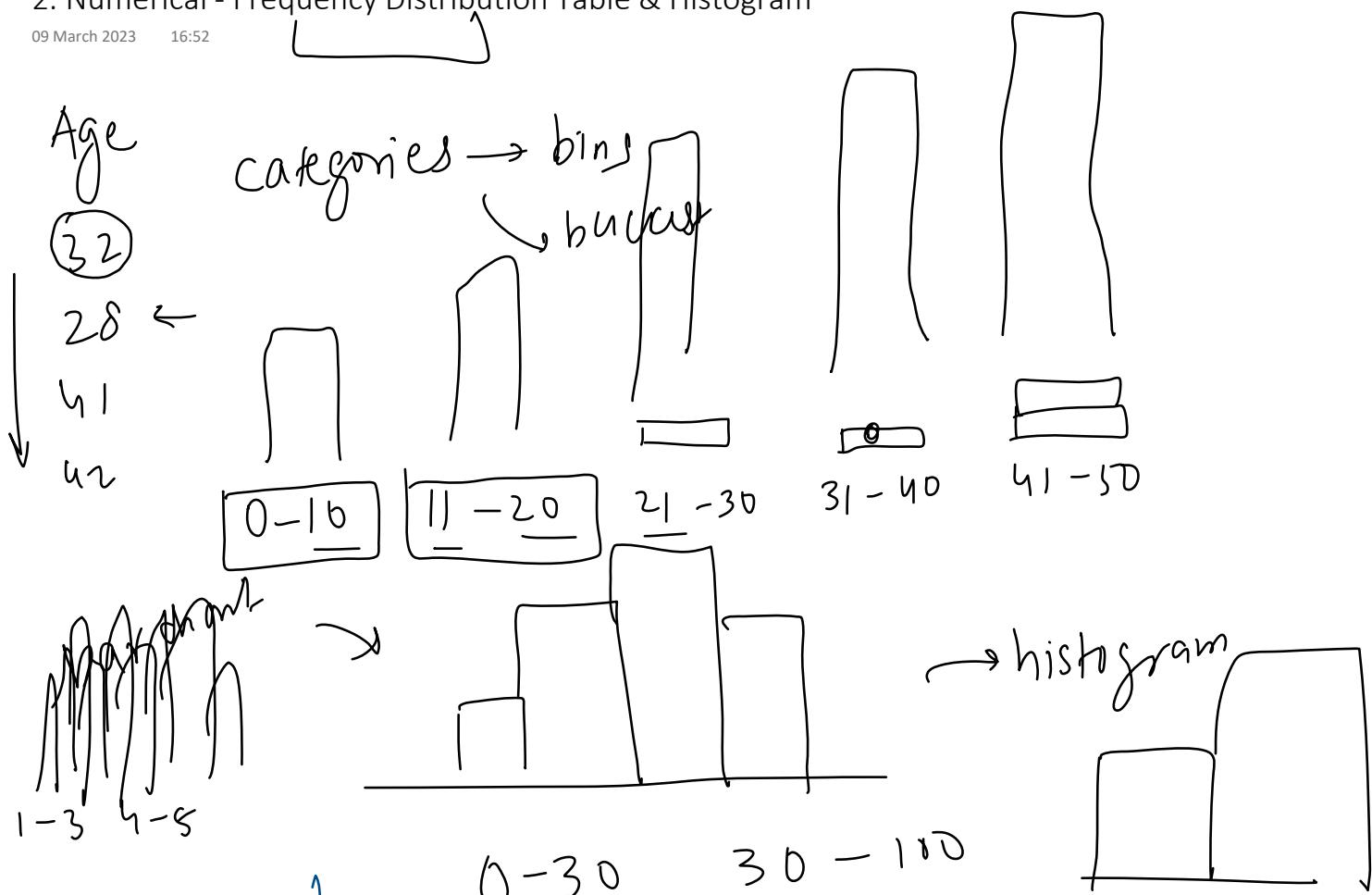
Type of Vacation	Frequency	Relative Frequency	Cumulative Frequency
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200

name prefer
Nish Beach

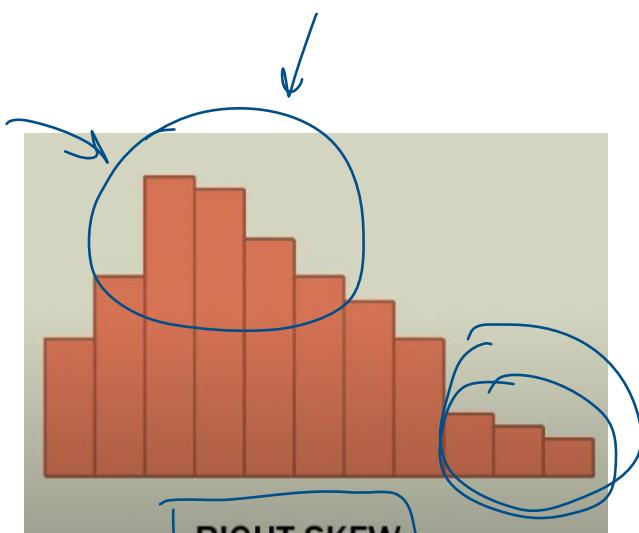
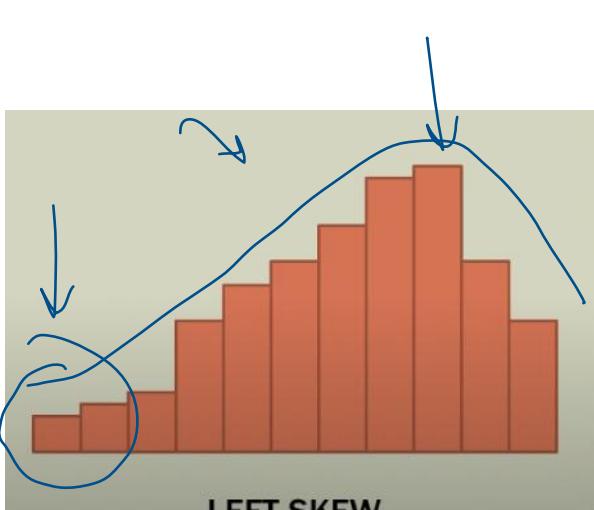
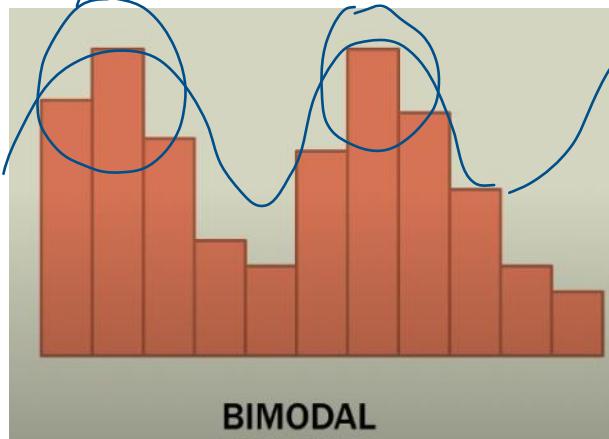
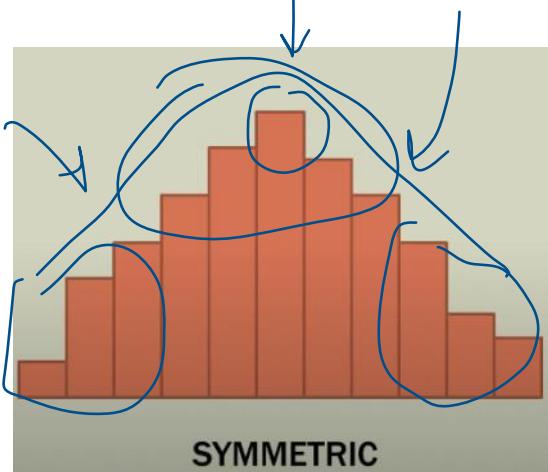


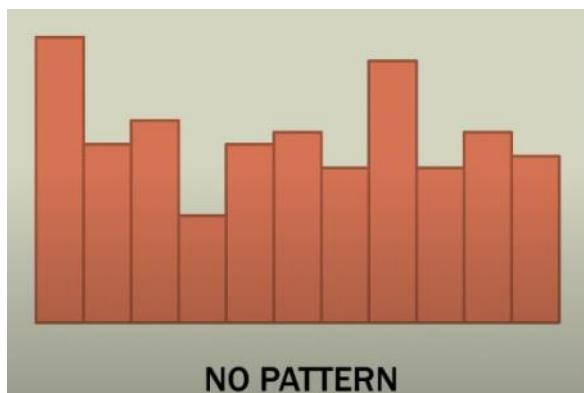
2. Numerical - Frequency Distribution Table & Histogram

09 March 2023 16:52



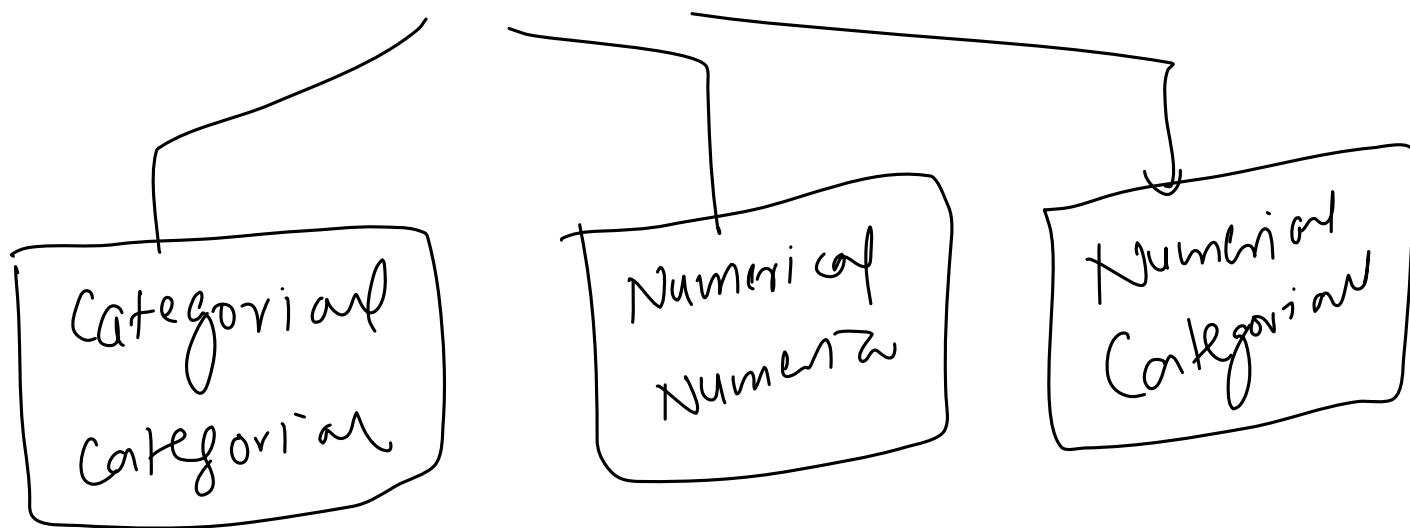
Shapes of Histogram





Graphs for Bivariate Analysis

09 March 2023 14:59



1. Categorical - Categorical

09 March 2023 16:58

Contingency Table/Crosstab

A contingency table, also known as a cross-tabulation or crosstab, is a type of table used in statistics to summarize the relationship between two categorical variables. A contingency table displays the frequencies or relative frequencies of the observed values of the two variables, organized into rows and columns.

<u>Survived</u>	<u>P class</u>
0	1
1	2
3	

P class



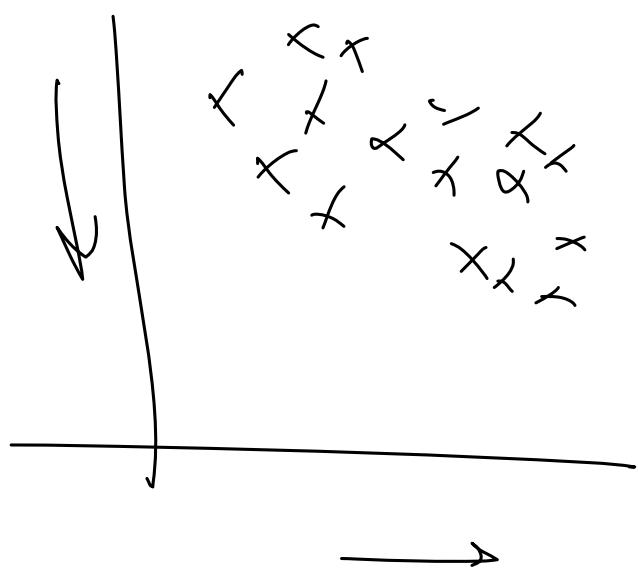
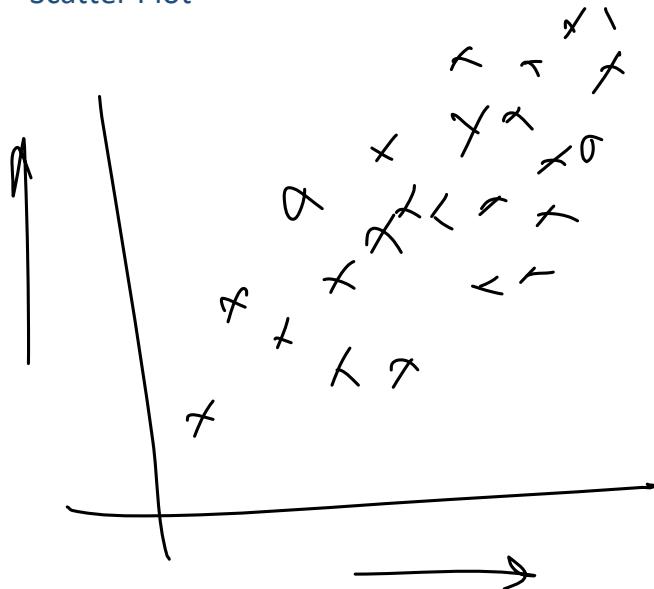
<u>Survived</u>	<u>P class</u>	1	2	3
0	0	42	31	63
1	1	71	118	13
3				

$$2 \times 3 = 6$$

2. Numerical - Numerical

09 March 2023 16:58

Scatter Plot



3. Categorical - Numerical

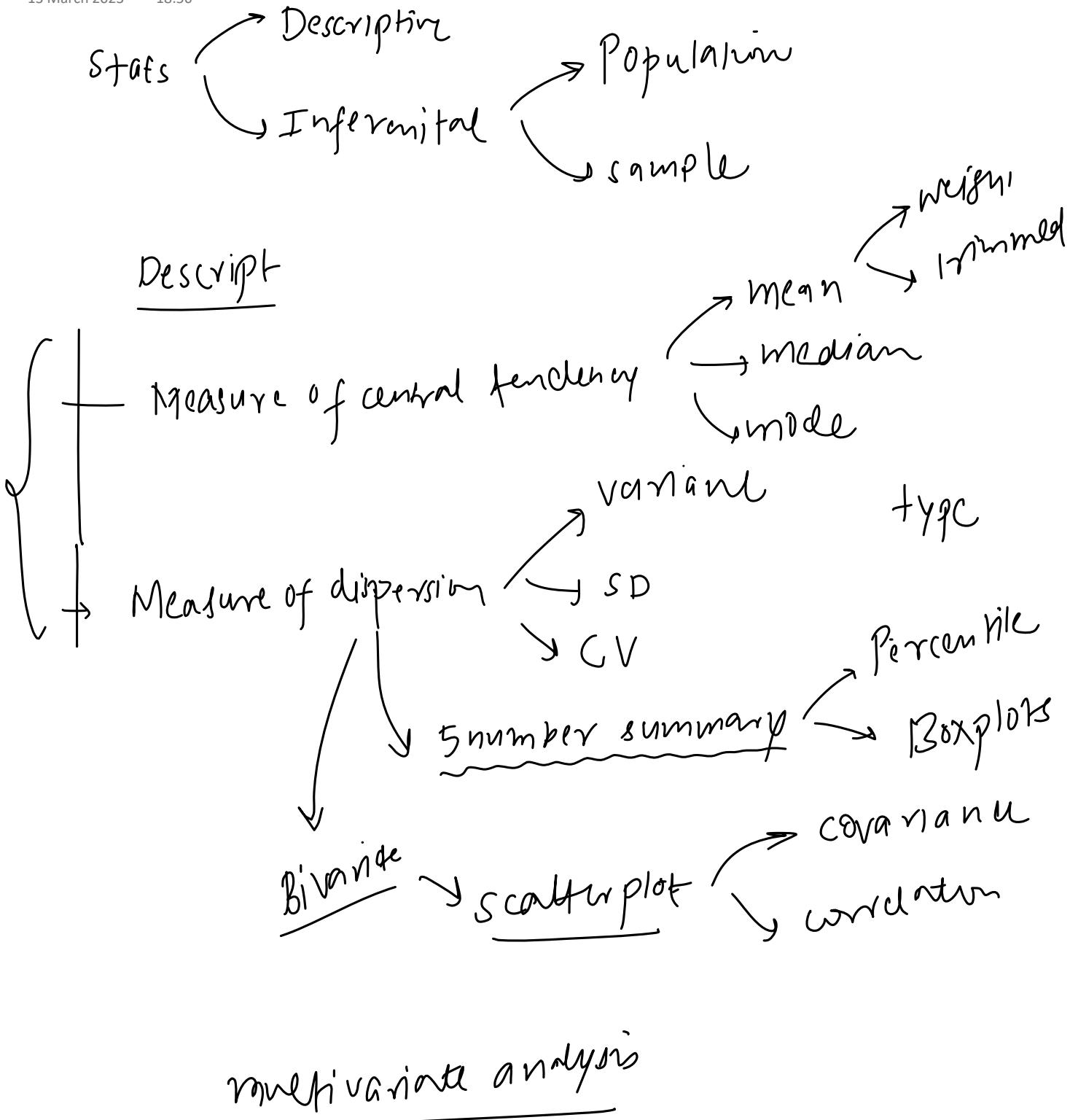
09 March 2023 16:58

Contingency

	0-10	<u>11-20</u>	<u>21-30</u>
male	32	41	110
female	15	18	120

Recap

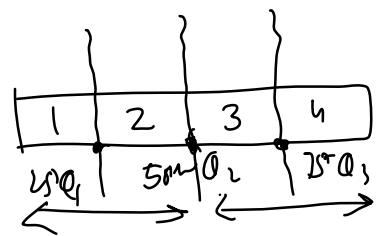
13 March 2023 18:56



Quantiles and Percentiles

13 March 2023 06:57

Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.



Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles used in statistical analysis, including:

- Quartiles: Divide the data into four equal parts, Q₁ (25th percentile), Q₂ (50th percentile or median), and Q₃ (75th percentile).
- Deciles: Divide the data into ten equal parts, D₁ (10th percentile), D₂ (20th percentile), ..., D₉ (90th percentile).
- Percentiles: Divide the data into 100 equal parts, P₁ (1st percentile), P₂ (2nd percentile), ..., P₉₉ (99th percentile).
- Quintiles: Divides the data into 5 equal parts

Things to remember while calculating these measures:

- Data should be sorted from low to high
- You are basically finding the location of an observation
- They are not actual values in the data
- All other tiles can be easily derived from Percentiles

Percentile

A percentile is a statistical measure that represents the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

Formula to calculate the percentile value:

$$PL = \frac{p}{100} (N+1)$$

where:

- PL = the desired percentile value location
- N = the total number of observations in the dataset
- p = the percentile rank (expressed as a percentage)

Example:

Find the 75th percentile score from the below data

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Step 1 - Sort the data (Asc)

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

$$PL = \frac{75}{100} (10+1) = \frac{3}{4} \times 11 = \frac{33}{4} = 8.25$$

96 — 98

075

91 — 93

100

$$\begin{array}{r} 96 \xrightarrow{\quad} 98 \\ 8 \xrightarrow{\quad} 9 \\ \uparrow \end{array}$$

(0.75)

$$\begin{array}{r} 91 \xrightarrow{\quad} 93 \\ 5 \xrightarrow{\quad} 6 \end{array}$$

→ $96 + 0.25(98 - 96) = 96 + 0.25 \times 2 = 96.5$

75th percentile = 96.5

$$P_L = \frac{50}{100} (10+1) = \frac{1}{2} \times 11 = 5.5$$

$$\begin{array}{l} 91 + 0.5(93 - 91) \\ 91 + 0.5 \times 2 = 92 \end{array}$$

↗ Percentile of a value

$$\text{Percentile rank} = \frac{x + 0.5y}{n} \quad \left(\frac{n}{N} \right)$$

X = number of values below the given valueY = number of values equal to the given valueN = total number of values in the dataset

78, 82, 84, 88, 91, 93, 94, 96, 98, 99
 ↓ ↓
 1 2 3

$$\frac{9}{16} \quad 0.9$$

$$= \frac{3 + 0.5 \times 1}{10} = \frac{3.5}{10}$$

$$\frac{0.35}{10} \rightarrow 0.35 \rightarrow 35\%$$

$$\frac{9 + 0.5 \times 1}{10} = \frac{9.5}{10} = 0.95 =$$

5 number summary

13 March 2023 06:57

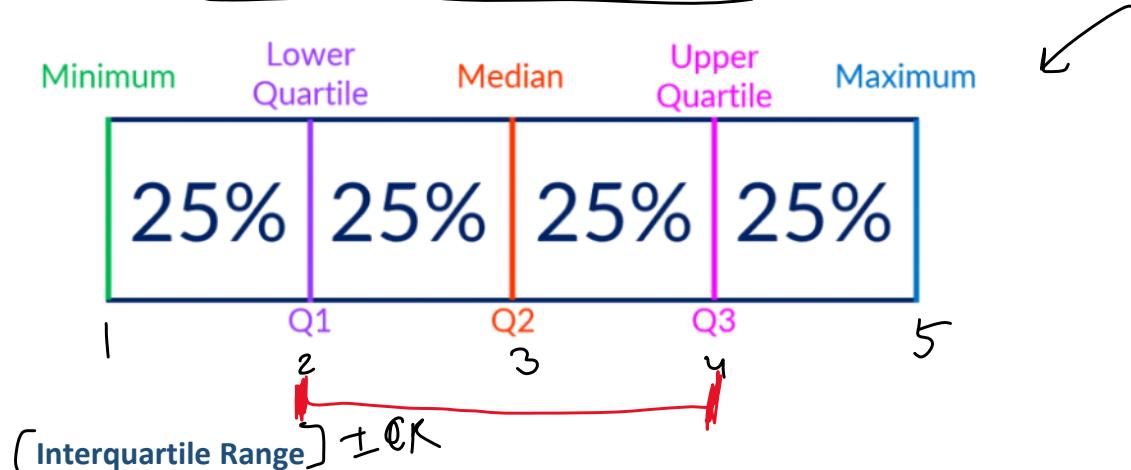
The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

(descriptive)

1. **Minimum value:** The smallest value in the dataset.
2. **First quartile (Q1):** The value that separates the lowest 25% of the data from the rest of the dataset.
3. **Median (Q2):** The value that separates the lowest 50% from the highest 50% of the data.
4. **Third quartile (Q3):** The value that separates the lowest 75% of the data from the highest 25% of the data.
5. **Maximum value:** The largest value in the dataset.

The five-number summary is often represented visually using a **box plot**, which displays the range of the dataset, the median, and the quartiles.

The five-number summary is a useful way to quickly summarize the central tendency, variability, and distribution of a dataset.



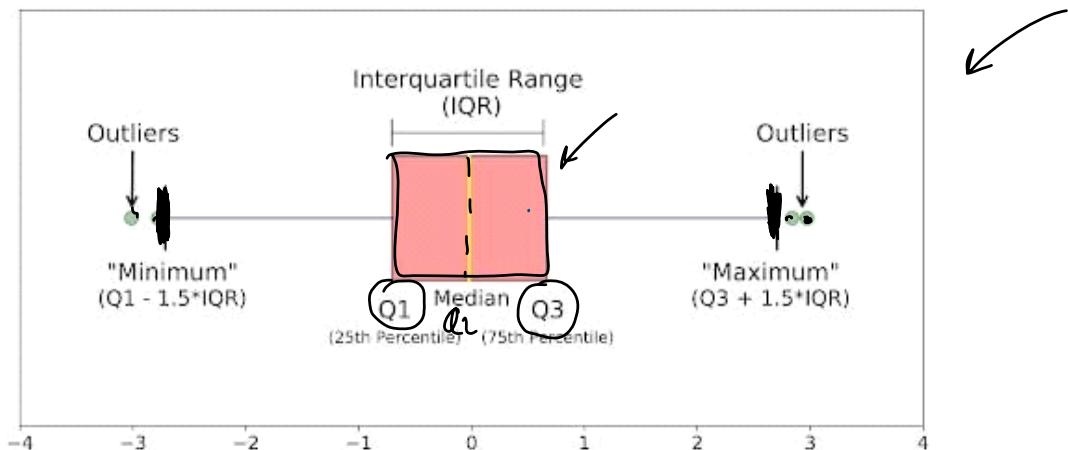
The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.

Boxplots

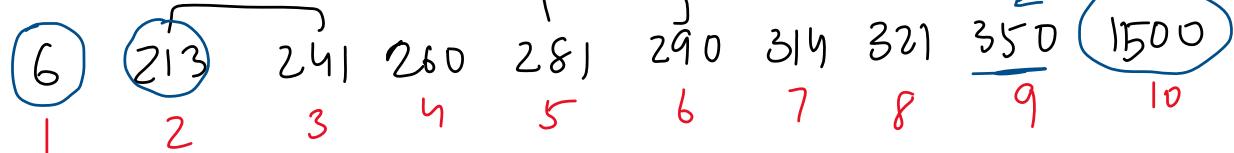
13 March 2023 06:57

1. What is a boxplot

A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that shows the distribution of the data. The box plot displays a summary of the data, including the minimum and maximum values, the first quartile (Q1), the median (Q2), and the third quartile (Q3).



2. How to create a boxplot with example



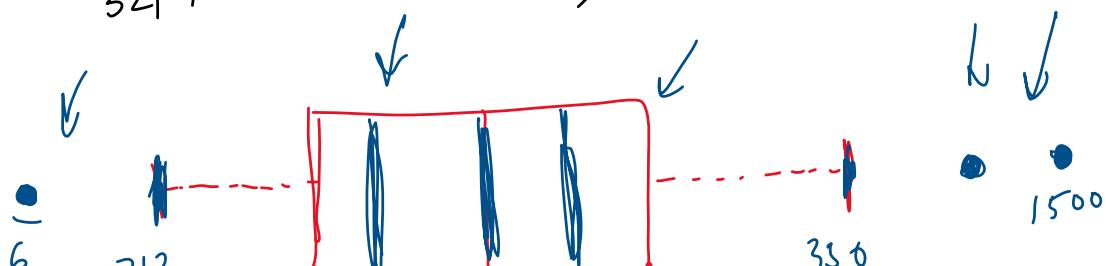
$$Q_2 = \frac{50}{100} (11) = 5.5 = \boxed{285.5}$$

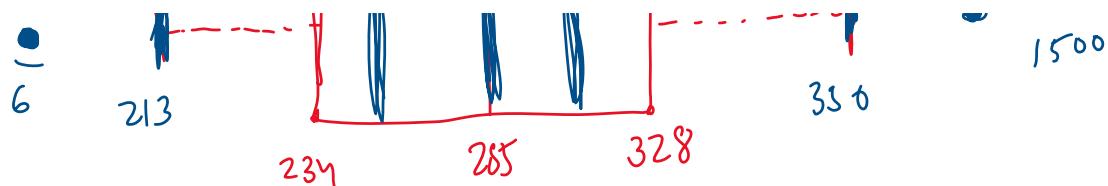
$$Q_1 = \frac{25}{100} \times 11 = \frac{11}{4} = 2.75$$

$$213 + 0.75(241 - 213) = \boxed{234}$$

$$Q_3 = \frac{75}{100} \times 11 = \frac{33}{4} = 8.25$$

$$321 + 0.25(350 - 321) = \boxed{328.25}$$





min and max

$$IQR = 94$$

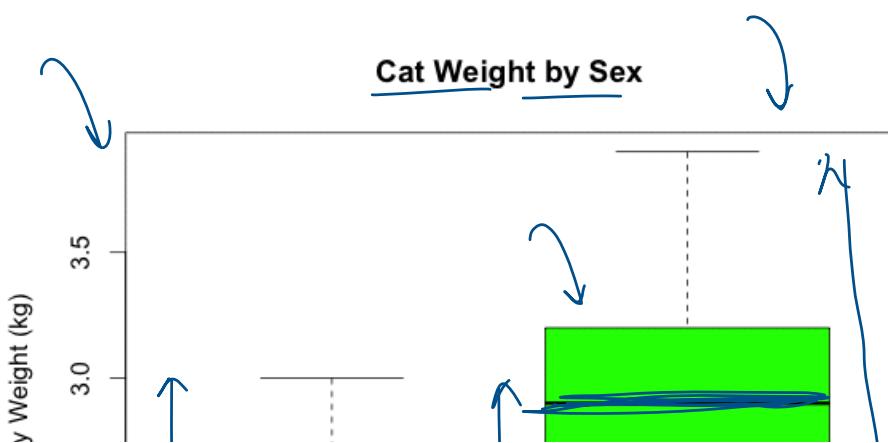
$$\text{min} = Q_1 - 1.5(IQR) = 93$$

$$\text{max} = Q_3 + 1.5(IQR) = 469$$

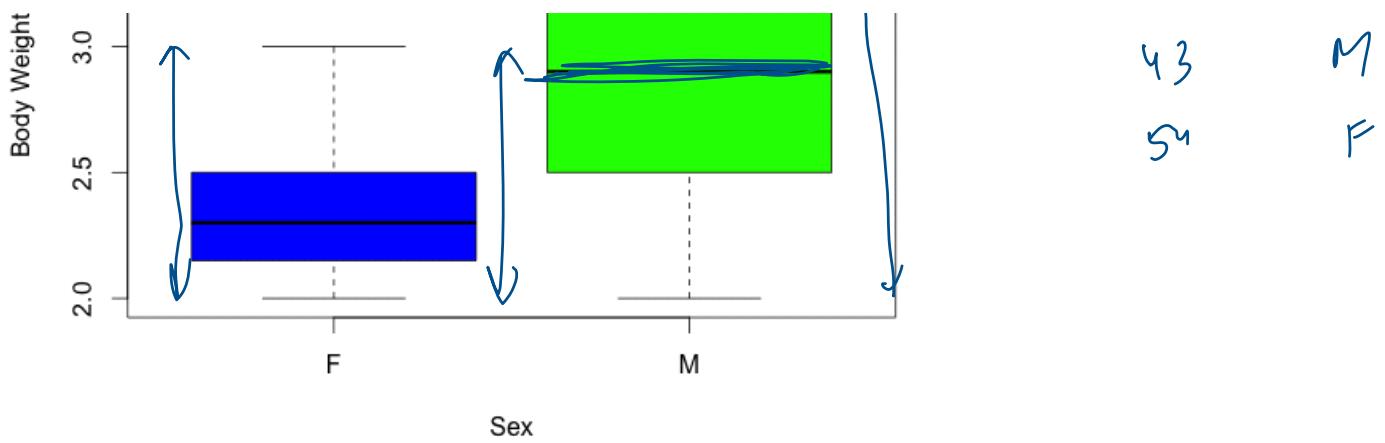
1. Benefits of a Boxplot

- Easy way to see the distribution of data
- Tells about skewness of data
- Can identify outliers
- Compare 2 categories of data

2. Side by side boxplot

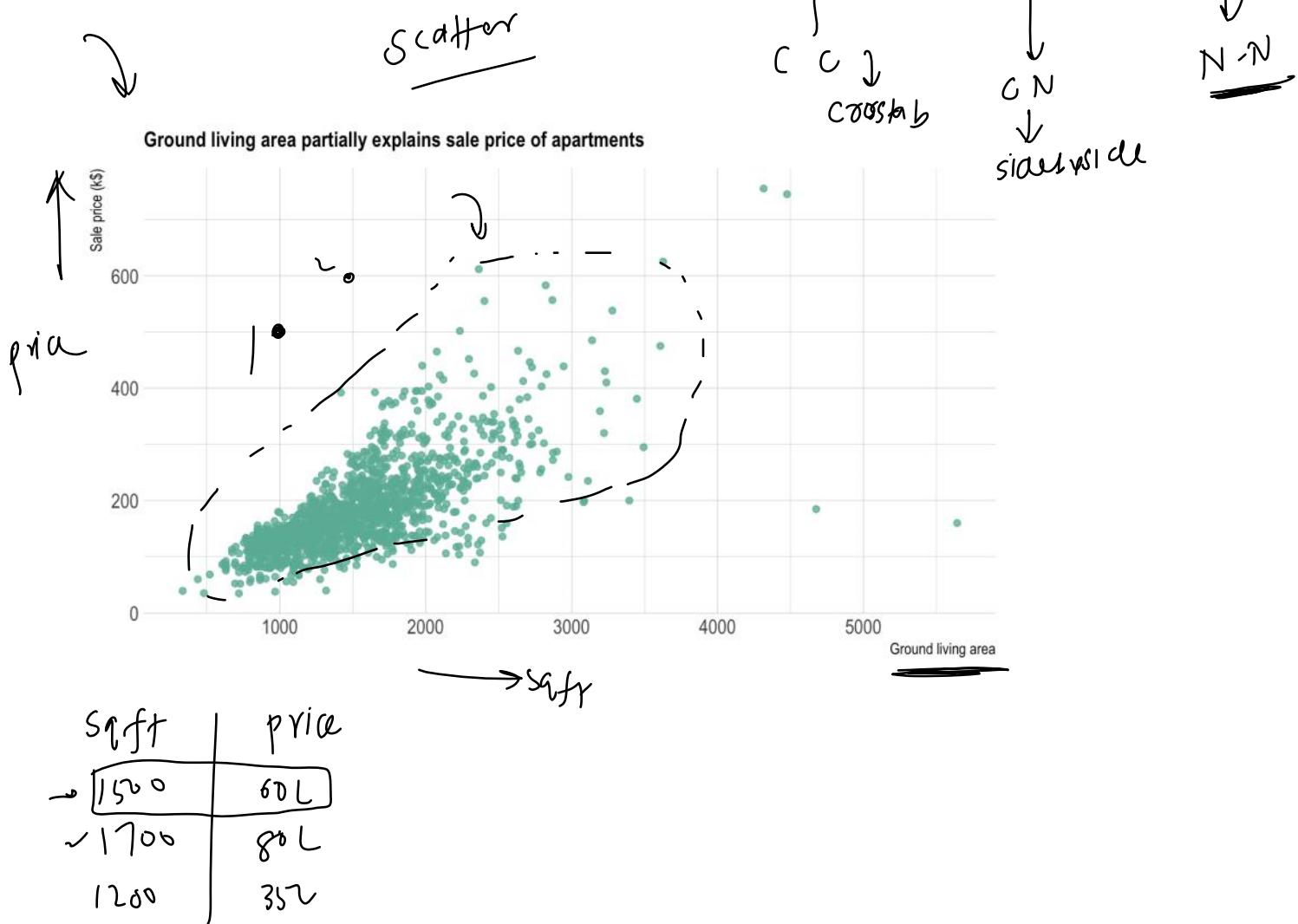


weight
Age | gender
21 M
39 F
43 M



Scatterplots

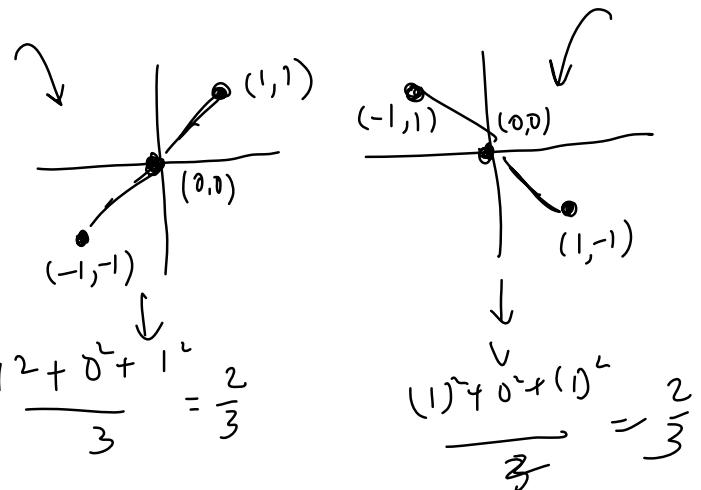
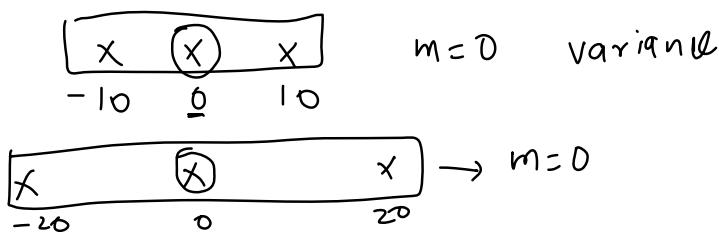
13 March 2023 06:58



Covariance

13 March 2023 06:57

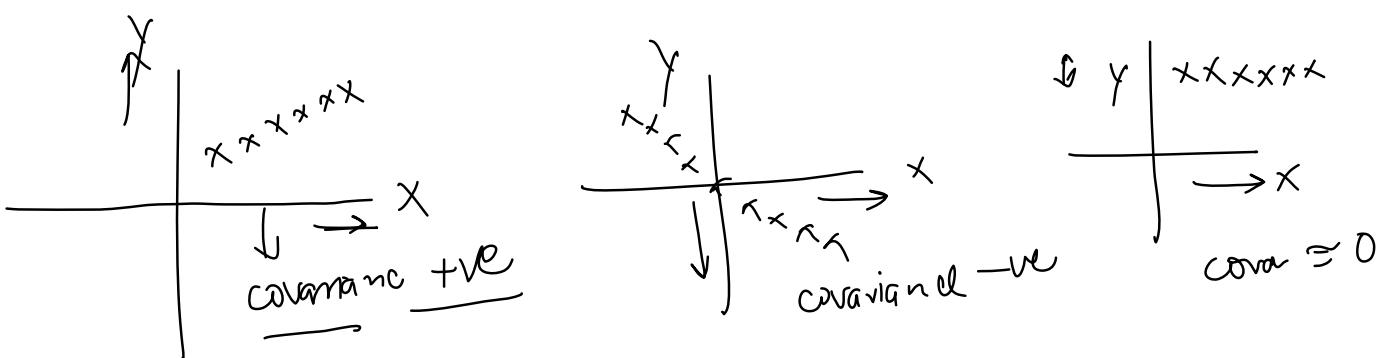
- What problem does Covariance solve?



- What is covariance and how is it interpreted?

Covariance is a statistical measure that describes the degree to which two variables are linearly related. It measures how much two variables change together, such that when one variable increases, does the other variable also increase, or does it decrease?

If the covariance between two variables is positive, it means that the variables tend to move together in the same direction. If the covariance is negative, it means that the variables tend to move in opposite directions. A covariance of zero indicates that the variables are not linearly related.



- How is it calculated?

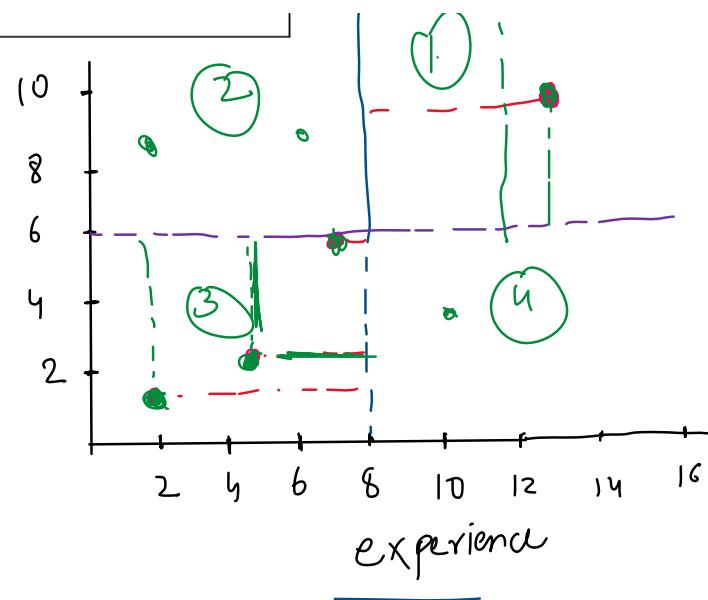
Covariance Formula	
Population	Sample
$\sigma_{xy} = \frac{\sum(X - \mu_x)(Y - \mu_y)}{N}$	$s_{xy} = \frac{\sum(X - \bar{x})(Y - \bar{y})}{n - 1}$
X, Y – The Value of X and Y in the Population μ_x, μ_y – The population Mean of X and Y N – Total Number of Observations	X, Y – The Value of X and Y in the Sample Data \bar{x}, \bar{y} – The Sample Mean of X and Y n – Total Number of Observations



Exp(x)	Salary(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	-1
12	12	9	6	54
13	10	5	4	20

$$\bar{x} = 8 \quad \bar{y} = 6$$

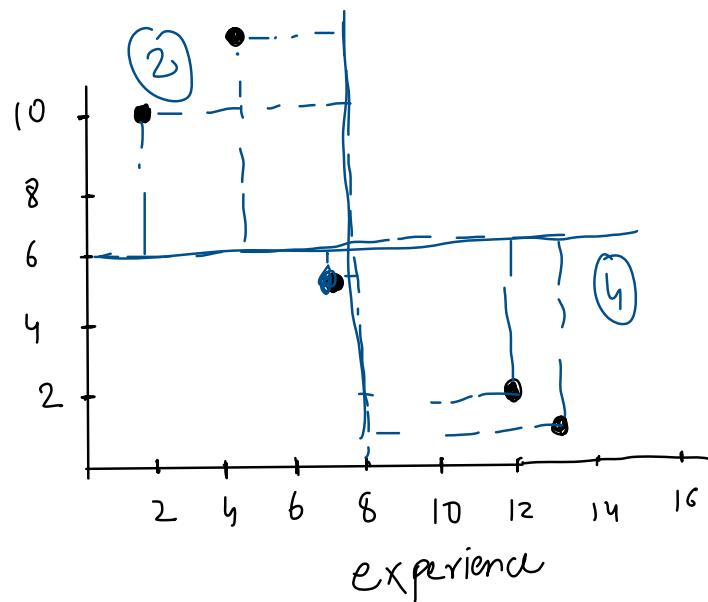
$$\text{COV} = \frac{\sum_{i=1}^{n-1} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



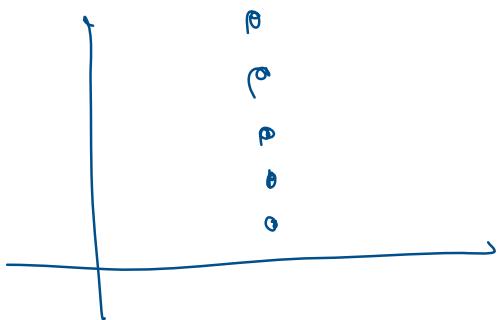
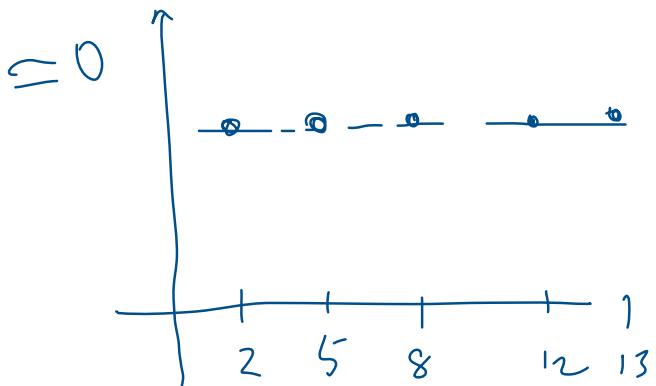
Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10	-6	4	-24
5	12	-3	6	-18
8	5	0	1	0
12	2	9	-4	-36
13	1	5	-5	-25

$$\bar{x} = 8 \quad \bar{y} = 6$$

$$\text{COV} = \frac{-83}{4}$$

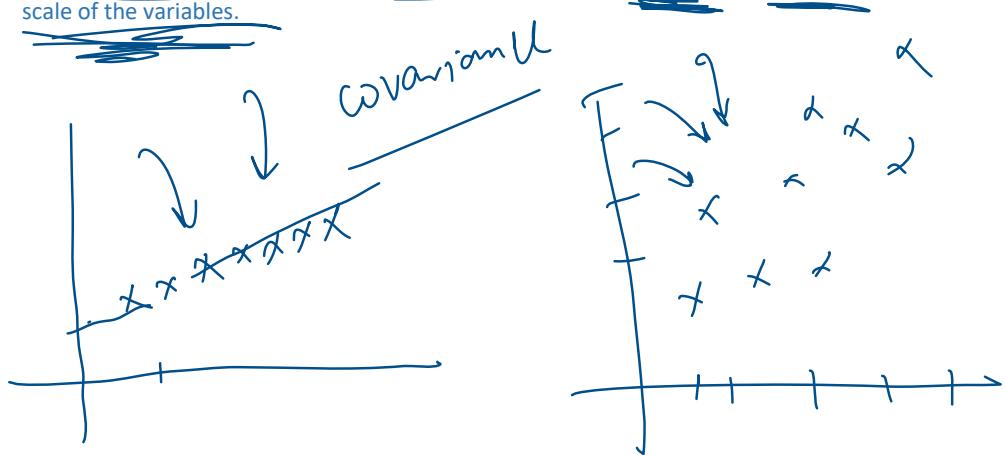


Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10			
5	10			
8	10			
12	10			
13	10			



- Disadvantages of using Covariance

One limitation of covariance is that it does not tell us about the strength of the relationship between two variables, since the magnitude of covariance is affected by the scale of the variables.



- Covariance of a variable with itself

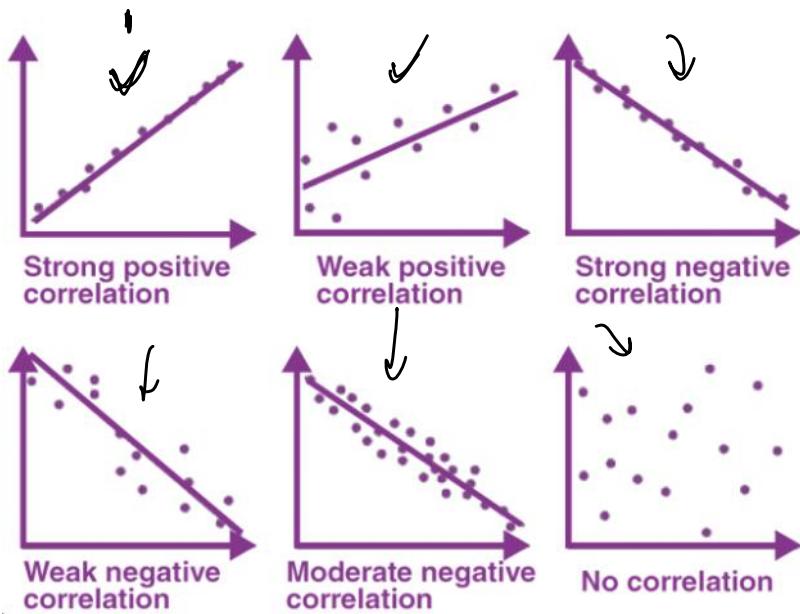
$$\sum (x - \bar{x})(y - \bar{y})$$

$$\begin{aligned}
 & \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\
 &= \sum_{i=1}^n (x_i - \bar{x})(\bar{y} - \bar{x}) \\
 &= \boxed{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}}
 \end{aligned}$$

Correlation

13 March 2023 06:58

1. What problem does Correlation solve?

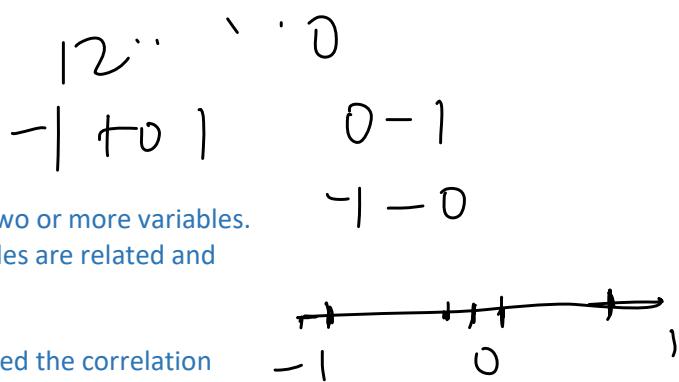


Can we quantify this weak and strong relationship?

2. What is correlation?

Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together.

Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of -1 indicates a perfect negative correlation, a correlation coefficient of 0 indicates no correlation, and a correlation coefficient of 1 indicates a perfect positive correlation.



$$\boxed{\text{Correlation} = \frac{\text{Cov}(x, y)}{\sigma_x * \sigma_y}}$$

Correlation and Causation

13 March 2023 18:31

The phrase "**correlation does not imply causation**" means that just because two variables are associated with each other, it does not necessarily mean that one causes the other. In other words, a correlation between two variables does not necessarily imply that one variable is the reason for the other variable's behaviour.

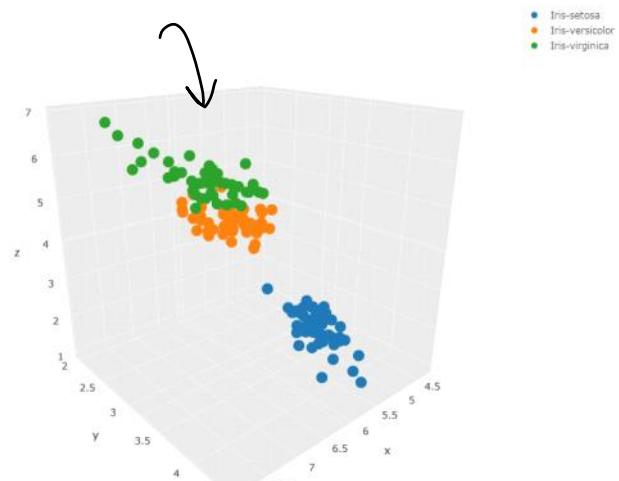
Suppose there is a positive correlation between the number of firefighters present at a fire and the amount of damage caused by the fire. One might be tempted to conclude that the presence of firefighters causes more damage. However, this correlation could be explained by a third variable - the severity of the fire. More severe fires might require more firefighters to be present, and also cause more damage.

Thus, while correlations can provide valuable insights into how different variables are related, they cannot be used to establish causality. Establishing causality often requires additional evidence such as experiments, randomized controlled trials, or well-designed observational studies.

Visualizing Multiple Variables

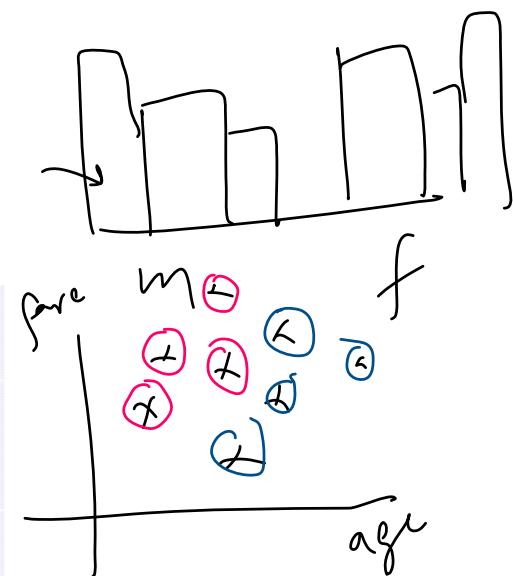
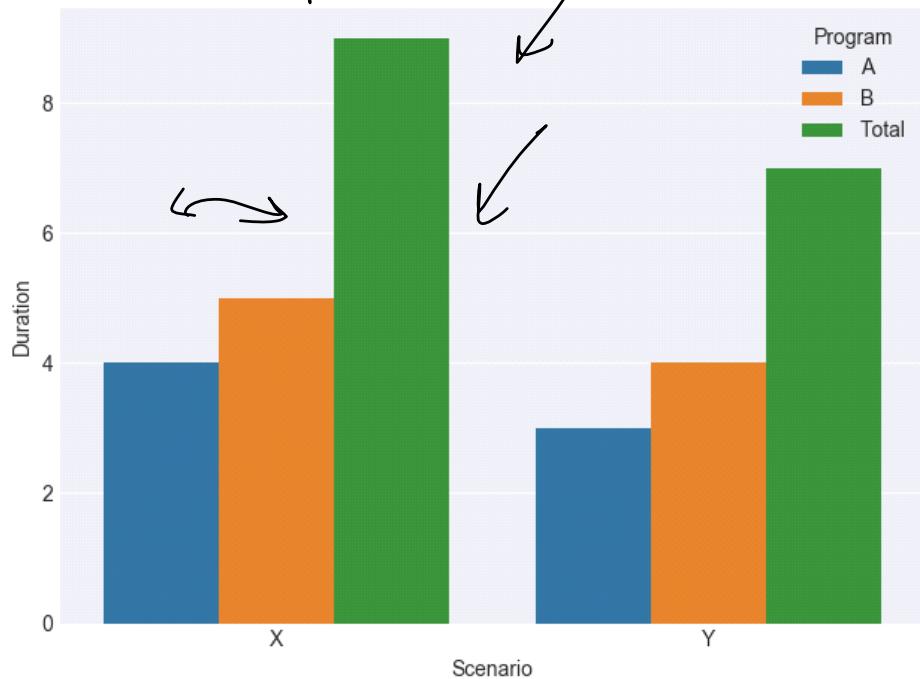
13 March 2023 06:58

1. 3D Scatter Plots

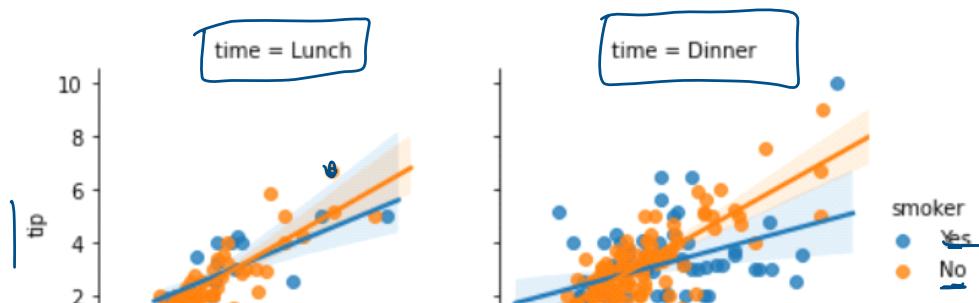


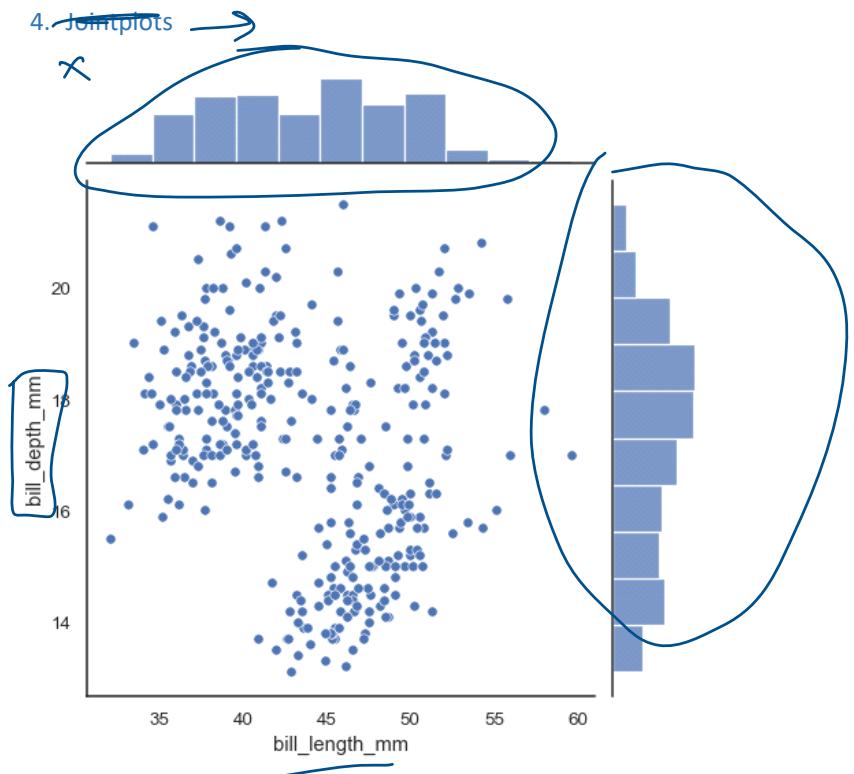
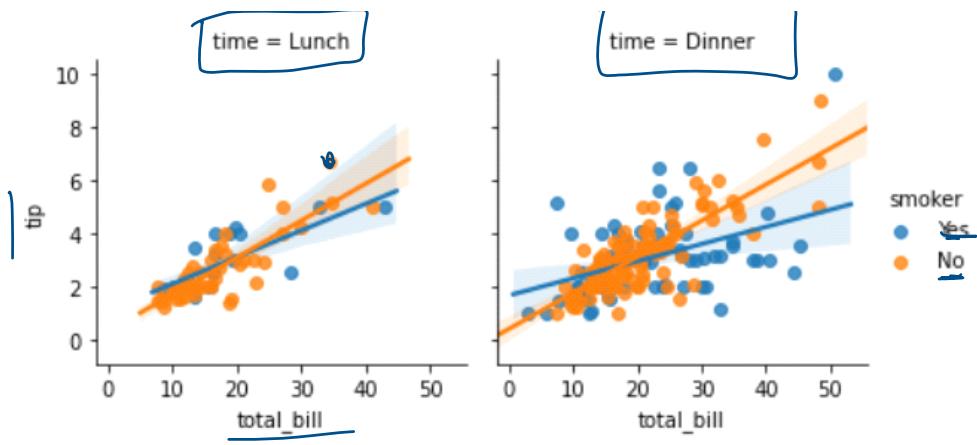
wt | ht | age)

2. Hue Parameter

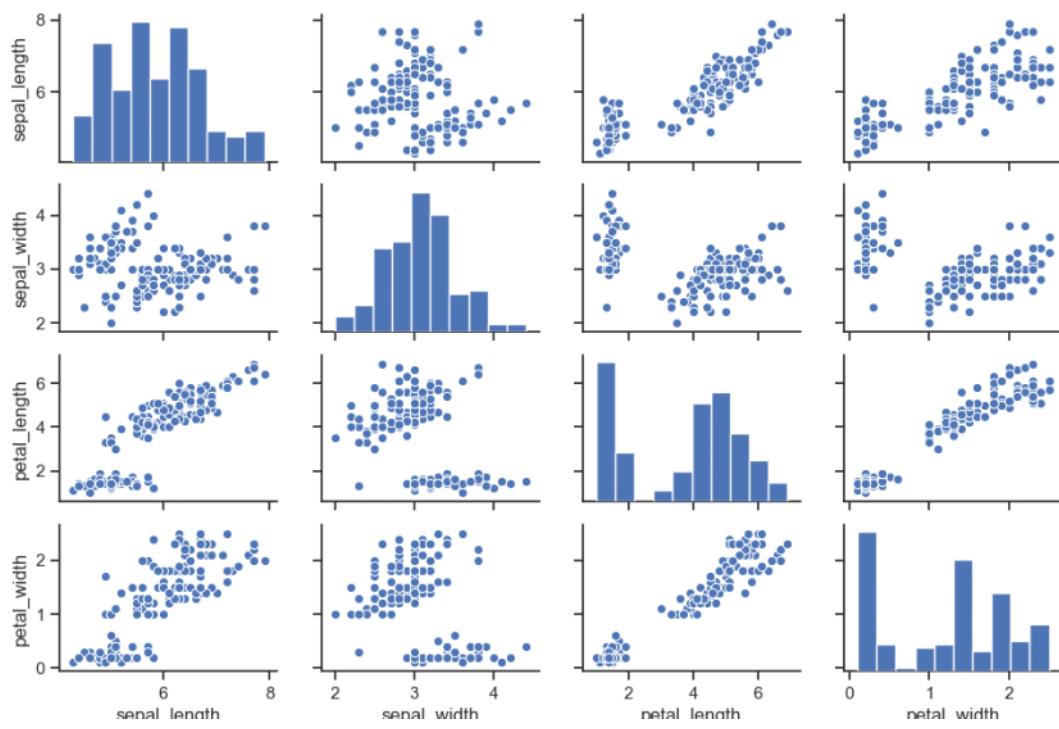


3. Facetgrids

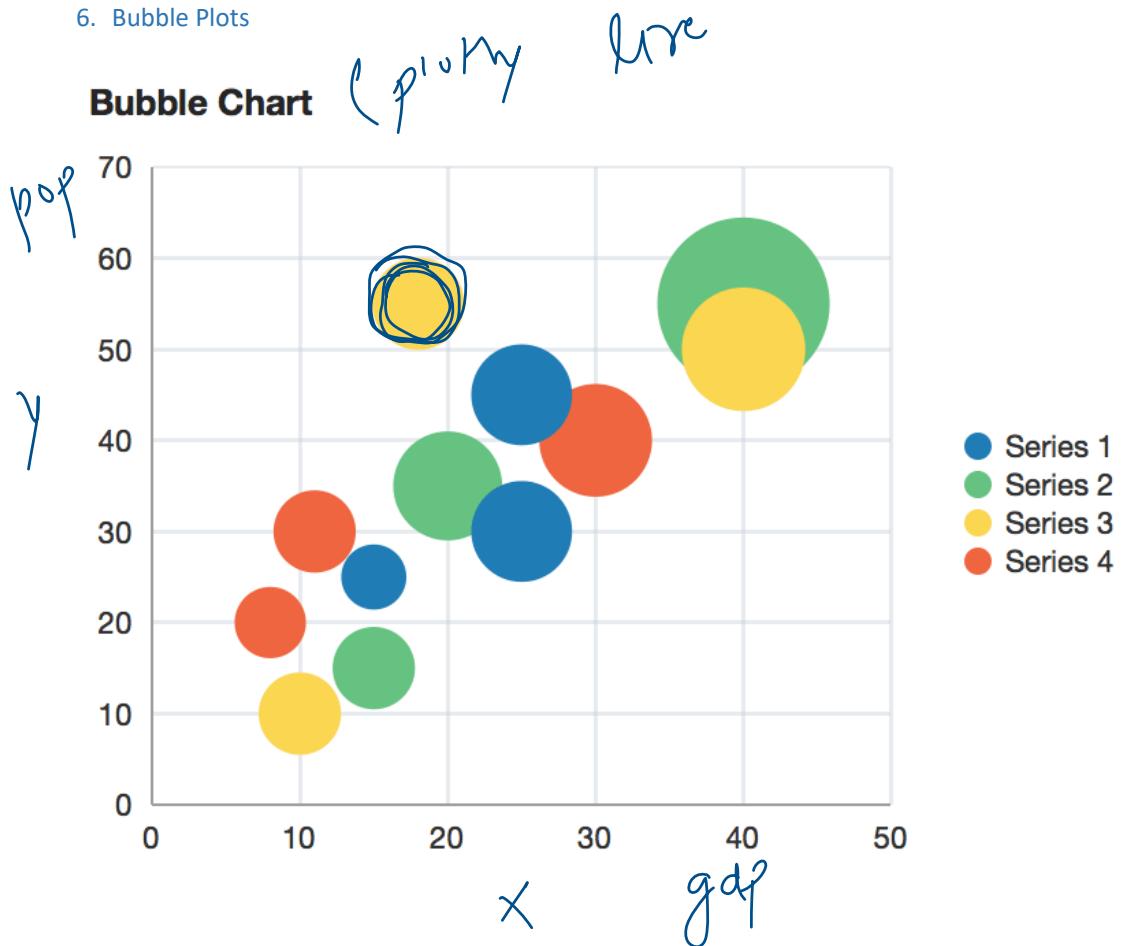




5. Pairplots



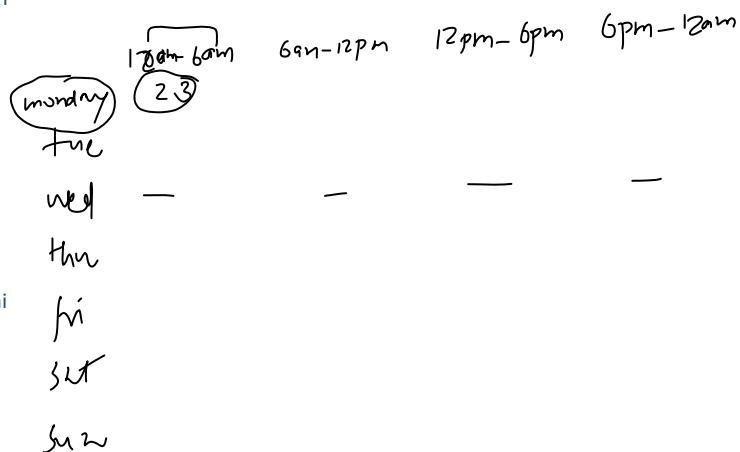
6. Bubble Plots



Flights Case Study

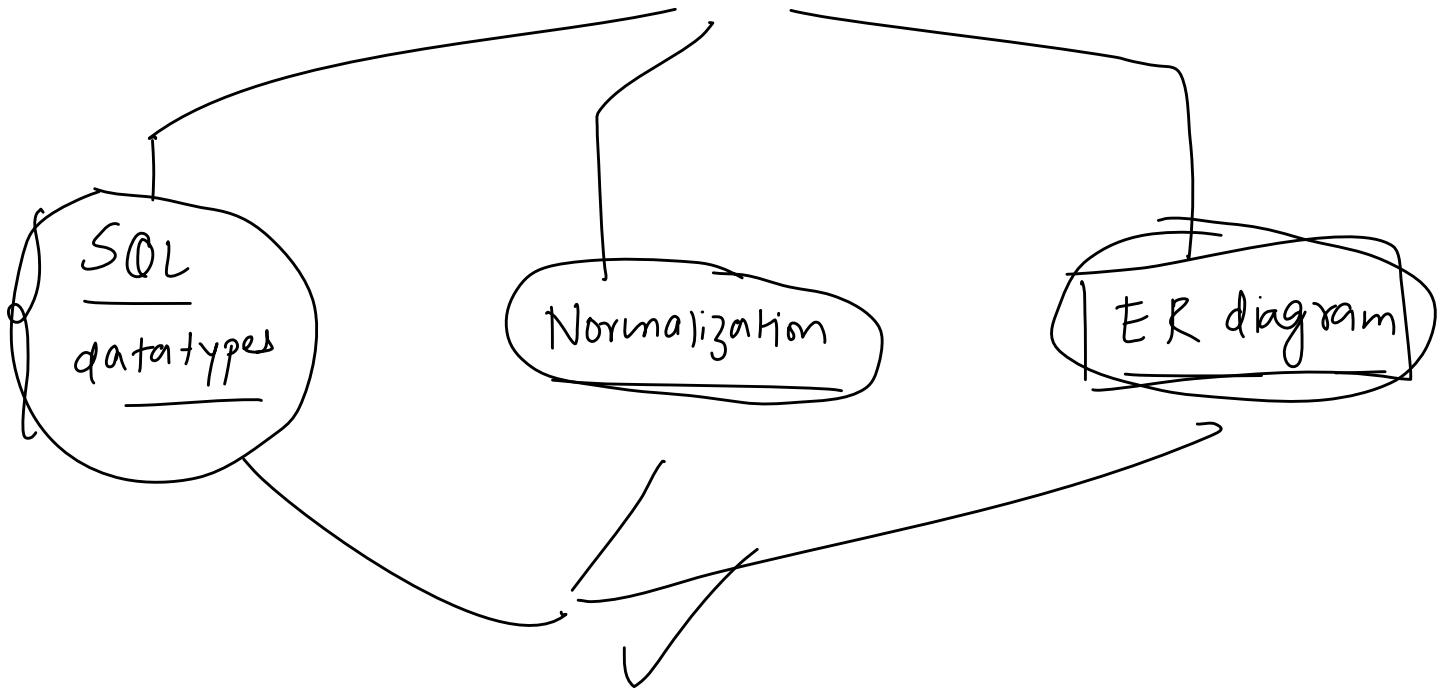
17 March 2023 14:52

1. Find the month with most number of flights
2. Which week day has most costly flights
3. Find number of indigo flights every month
4. Find list of all flights that depart between 10AM and 2PM from Delhi to Bangalore
5. Find the number of flights departing on weekends from Bangalore
6. Calculate the arrival time for all flights by adding the duration to the departure time.
7. Calculate the arrival date for all the flights
8. Find the number of flights which travel on multiple dates.
9. Calculate the average duration of flights between all city pairs. The answer should be in $xh\,ym$ format
10. Find all flights which departed before midnight but arrived at their destination after midnight having only 0 stops.
11. Find quarter wise number of flights for each airline
12. Find the longest flight distance(between cities in terms of time) in India
13. Average time duration for flights that have 1 stop vs more than 1 stops
14. Find all Air India flights in a given date range originating from Delhi
15. Find the longest flight of each airline
16. Find all the pair of cities having average time duration > 3 hours
17. Make a weekday vs time grid showing frequency of flights from Bangalore and Delhi
18. Make a weekday vs time grid showing avg flight price from Bangalore and Delhi



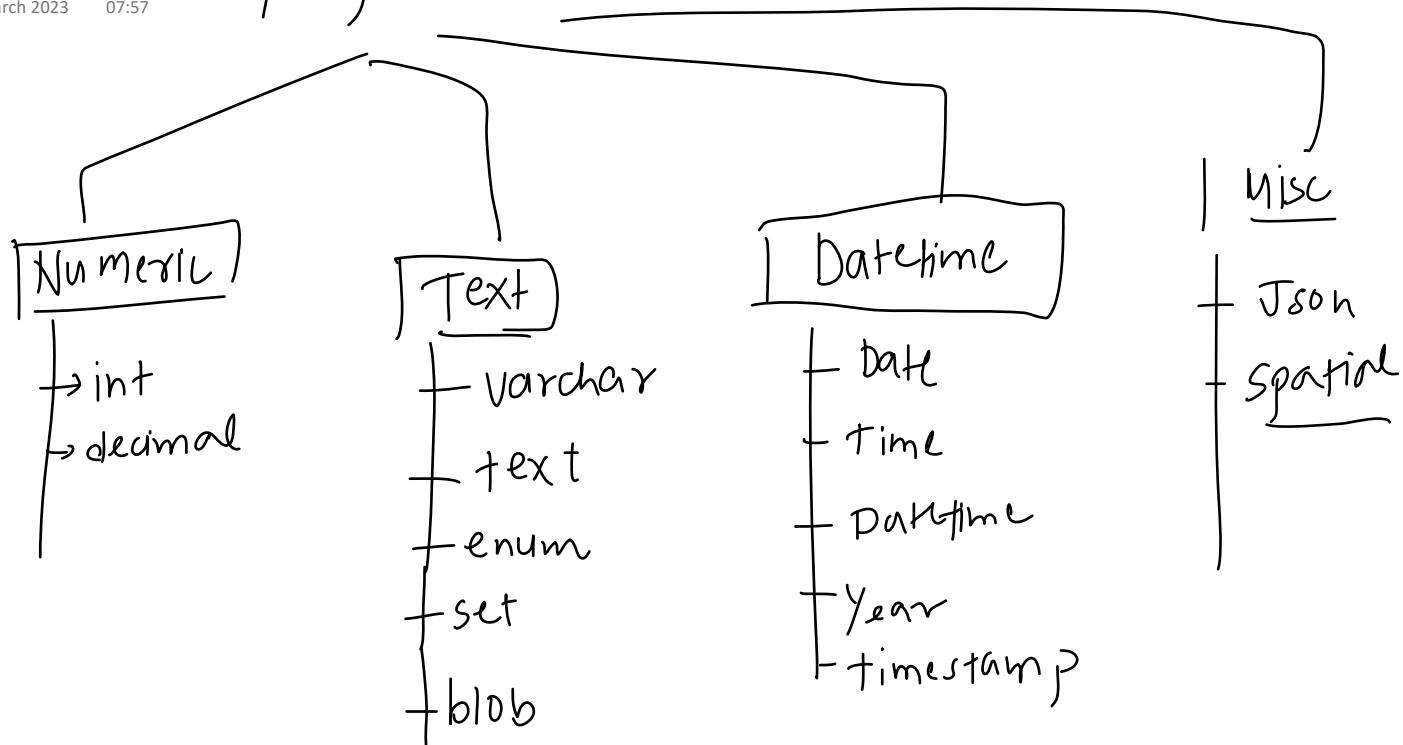
Plan of Attack

18 March 2023 18:25



SQL Datatypes (MySQL)

18 March 2023 07:57

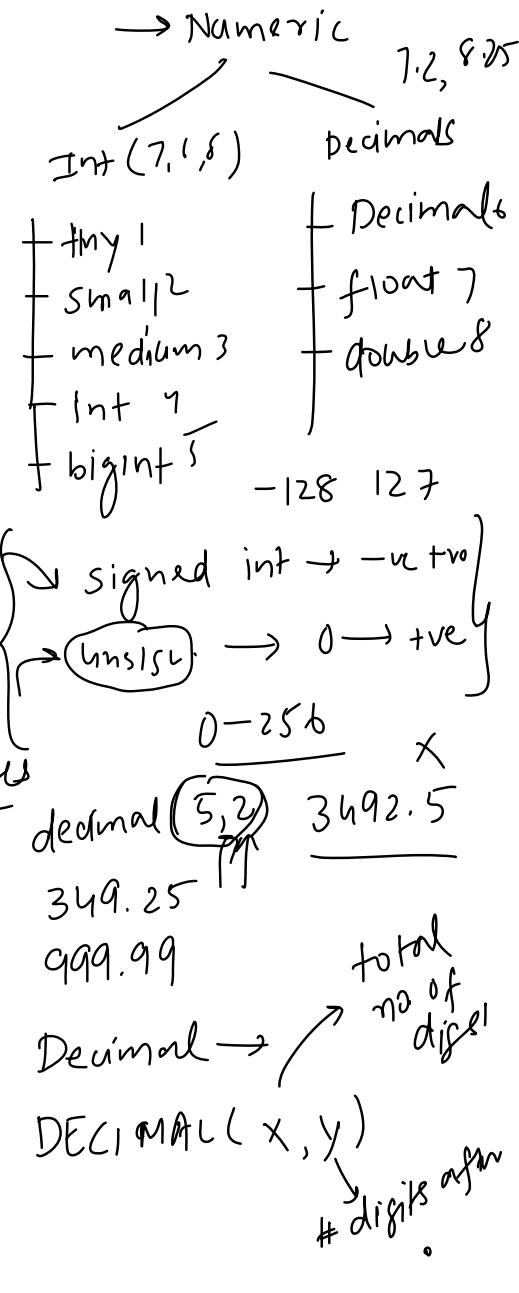


Numeric Types

18 March 2023 08:09

1. **INT:** The INT data type is used to store integers with a maximum value of 2147483647 and a minimum value of -2147483648. Examples of data that can be stored in INT include employee IDs, order numbers, and product IDs.
2. **TINYINT:** The TINYINT data type is used to store integers with a maximum value of 127 and a minimum value of -128. Examples of data that can be stored in TINYINT include Boolean values, such as 0 for false and 1 for true.
3. **SMALLINT:** The SMALLINT data type is used to store integers with a maximum value of 32767 and a minimum value of -32768. Examples of data that can be stored in SMALLINT include quantities of items, such as the number of products sold in a transaction.
4. **MEDIUMINT:** The MEDIUMINT data type is used to store integers with a maximum value of 8388607 and a minimum value of -8388608. Examples of data that can be stored in MEDIUMINT include the number of visitors to a website or the number of followers on a social media platform.
5. **BIGINT:** The BIGINT data type is used to store integers with a maximum value of 9223372036854775807 and a minimum value of -9223372036854775808. Examples of data that can be stored in BIGINT include the total revenue generated by a company or the number of views on a YouTube video.
6. **FLOAT:** The FLOAT data type is used to store single-precision floating-point numbers, which are numbers with a decimal point. Examples of data that can be stored in FLOAT include the price of a product or the temperature of a room.
7. **DOUBLE:** The DOUBLE data type is used to store double-precision floating-point numbers, which are numbers with a decimal point that can store more digits than FLOAT. Examples of data that can be stored in DOUBLE include very large or very small numbers, such as the distance between planets in the solar system or the size of an atom.
8. **DECIMAL:** The DECIMAL data type is used to store exact decimal values with a fixed number of digits before and after the decimal point. Examples of data that can be stored in DECIMAL include financial values, such as the cost of an item or the total balance in a bank account.

Type	Storage (Bytes)	Minimum Value Signed	Minimum Value Unsigned	Maximum Value Signed	Maximum Value Unsigned
TINYINT	1	-128	0	127	255
SMALLINT	2	-32768	0	32767	65535
MEDIUMINT	3	-8388608	0	8388607	16777215
INT	4	-2147483648	0	2147483647	4294967295
BIGINT	8	-2^{63}	0	$2^{63}-1$	$2^{64}-1$



String Data Type

18 March 2023 08:15

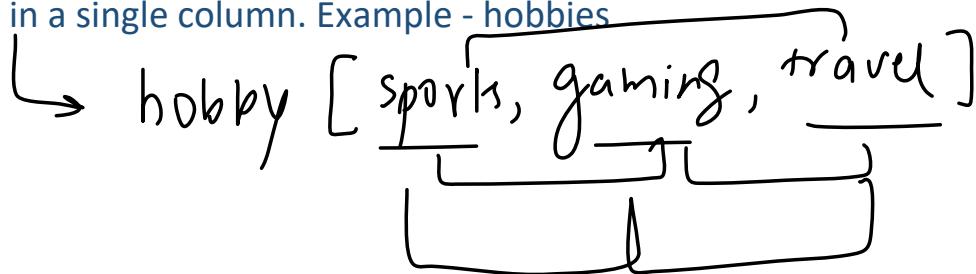
1. **CHAR:** This data type is used to store fixed-length strings. The length of the string is specified when the table is created, and the field will always use that amount of space, regardless of whether the string stored in it is shorter or longer. For example, if you define a CHAR(10) field and store the string "hello" in it, MySQL will pad the string with spaces so that it takes up 10 characters. CHAR fields are useful when you have a field that always contains the same length of data, such as a state abbreviation or a phone number.
2. **VARCHAR:** This data type is used to store variable-length strings. The length of the string can be up to a specified maximum, but the field will only use as much space as it needs to store the actual data. For example, if you define a VARCHAR(10) field and store the string "hello" in it, MySQL will only use 5 characters to store the data. VARCHAR fields are useful when you have a field that can contain varying amounts of data, such as a user's name or address.
3. **TEXT:** This data type is used to store larger amounts of variable-length string data than VARCHAR. It can store up to 65,535 characters. TEXT fields are useful when you need to store large amounts of text data, such as blog posts or comments.
4. **MEDIUMTEXT:** This data type is used to store even larger amounts of text data than TEXT. It can store up to 16,777,215 characters. MEDIUMTEXT fields are useful when you need to store very large amounts of text data, such as long-form articles or legal documents.
5. **LONGTEXT:** This data type is used to store the largest amounts of text data. It can store up to 4,294,967,295 characters. LONGTEXT fields are useful when you need to store extremely large amounts of text data, such as entire books or large collections of data.

ENUM and SET

18 March 2023 08:16

male, female

1. **ENUM:** The ENUM data type is used to store a set of predefined values. You can specify a list of possible values for an ENUM column, and the column can only store one of these values. The ENUM data type can be used to ensure that only valid values are stored in a column, and it can also save storage space compared to storing string values. Example - gender
2. **SET:** The SET data type is similar to ENUM, but it can store multiple values. You can specify a list of possible values for a SET column, and the column can store any combination of these values. The SET data type can be used to store sets of values, such as tags or categories, in a single column. Example - hobbies



BLOB

18 March 2023 08:16

Text

The BLOB (Binary Large Object) data type in MySQL is used to store large binary data, such as images, audio, video, or other multimedia content.

In MySQL, there are four types of BLOB data types that can be used to store binary data with different maximum sizes:

- TINYBLOB: Maximum length of 255 bytes. TINYBLOB is the smallest BLOB data type in MySQL. It can be used to store small binary data, such as icons, small images, or serialized objects.
- BLOB: Maximum length of 65,535 bytes (64 KB). BLOB is a medium-sized BLOB data type that can be used to store larger binary data, such as images, audio, video, or other multimedia files.
- MEDIUMBLOB: Maximum length of 16,777,215 bytes (16 MB). MEDIUMBLOB is a larger BLOB data type that can be used to store even larger binary data, such as high-resolution images or longer audio or video files.
- LONGBLOB: Maximum length of 4,294,967,295 bytes (4 GB). LONGBLOB is the largest BLOB data type in MySQL, and it can be used to store very large binary data, such as very high-resolution images, long audio or video files, or even entire documents.

LOAD_FILE(PATH)

Pros of storing files in BLOB columns:

- BLOB columns allow you to store binary data directly in the database, without needing to store the file externally.
- Storing files in the database can simplify backup and restore procedures, as all the data is in one place.
- Access to BLOB data can be controlled through database user permissions.

Cons of storing files in BLOB columns:

- Storing large files in the database can slow down database performance and increase storage requirements.
- If you need to access the file outside of the database (e.g. to share it with another application or user), you'll need to extract it from the database.
- Some file types may not be well-suited for storage in BLOB columns, depending on their size, structure, and how they are accessed.

Datetime

18 March 2023 08:17

In MySQL, there are several temporal data types that can be used to store and manipulate time and date values. These include:

1. **DATE** - used for storing date values in the format **YYYY-MM-DD**.
2. **TIME** - used for storing time values in the format **HH:MM:SS**.
3. **DATETIME** - used for storing date and time values in the format **YYYY-MM-DD HH:MM:SS**.
4. **TIMESTAMP** - used for storing date and time values in the format **YYYY-MM-DD HH:MM:SS**. It has a range of 1970-01-01 00:00:01 UTC to 2038-01-19 03:14:07 UTC.
5. **YEAR** - used for storing year values in 2-digit or 4-digit format (**YYYY** or **YY**). If the year is specified with 2 digits, it is assumed to be in the range 1970-2069 (inclusive).

Spatial Datatypes

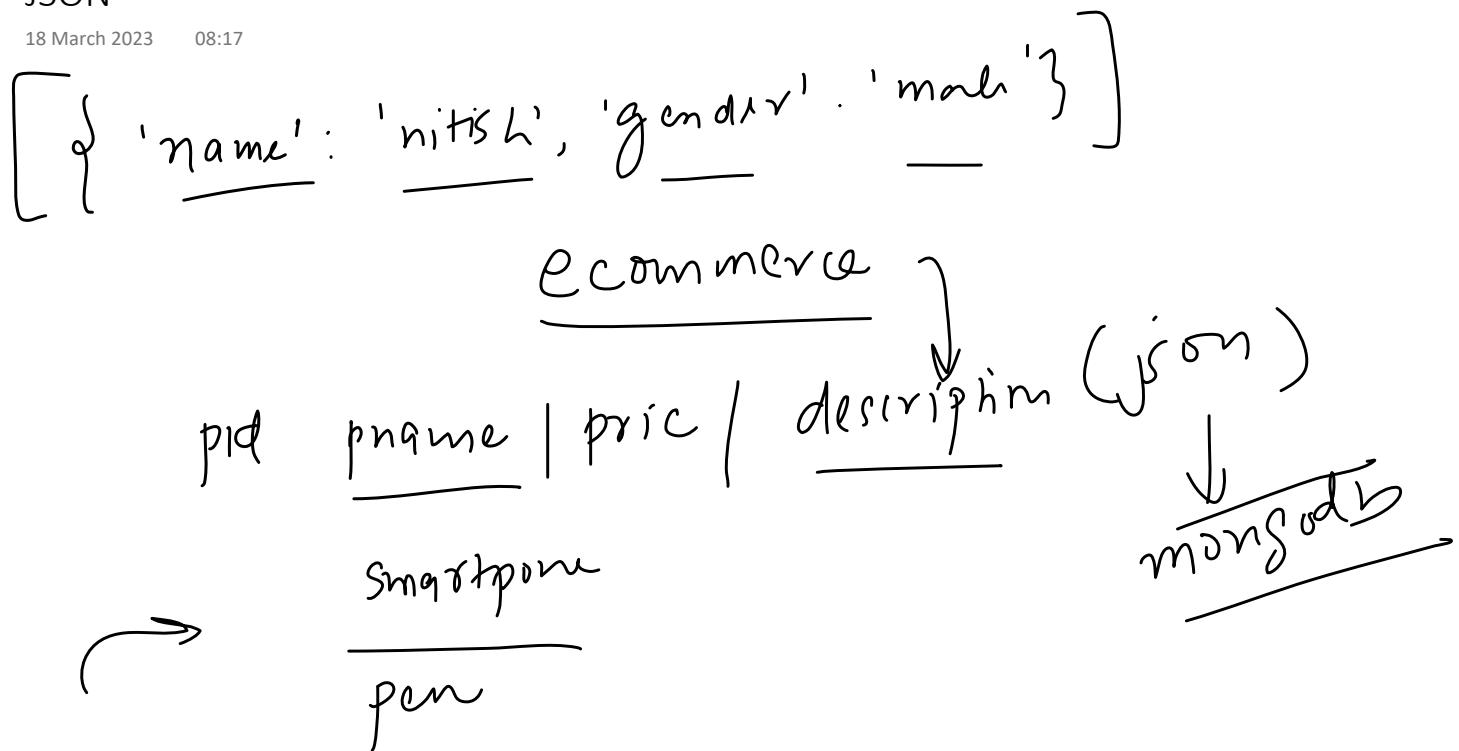
18 March 2023 08:17

GEOMETRY - The GEOMETRY data type is a generic spatial data type that can store any type of geometric data, including points, lines, and polygons.

`ST_ASTEXT()`, `ST_X()`,`ST_Y()`

JSON

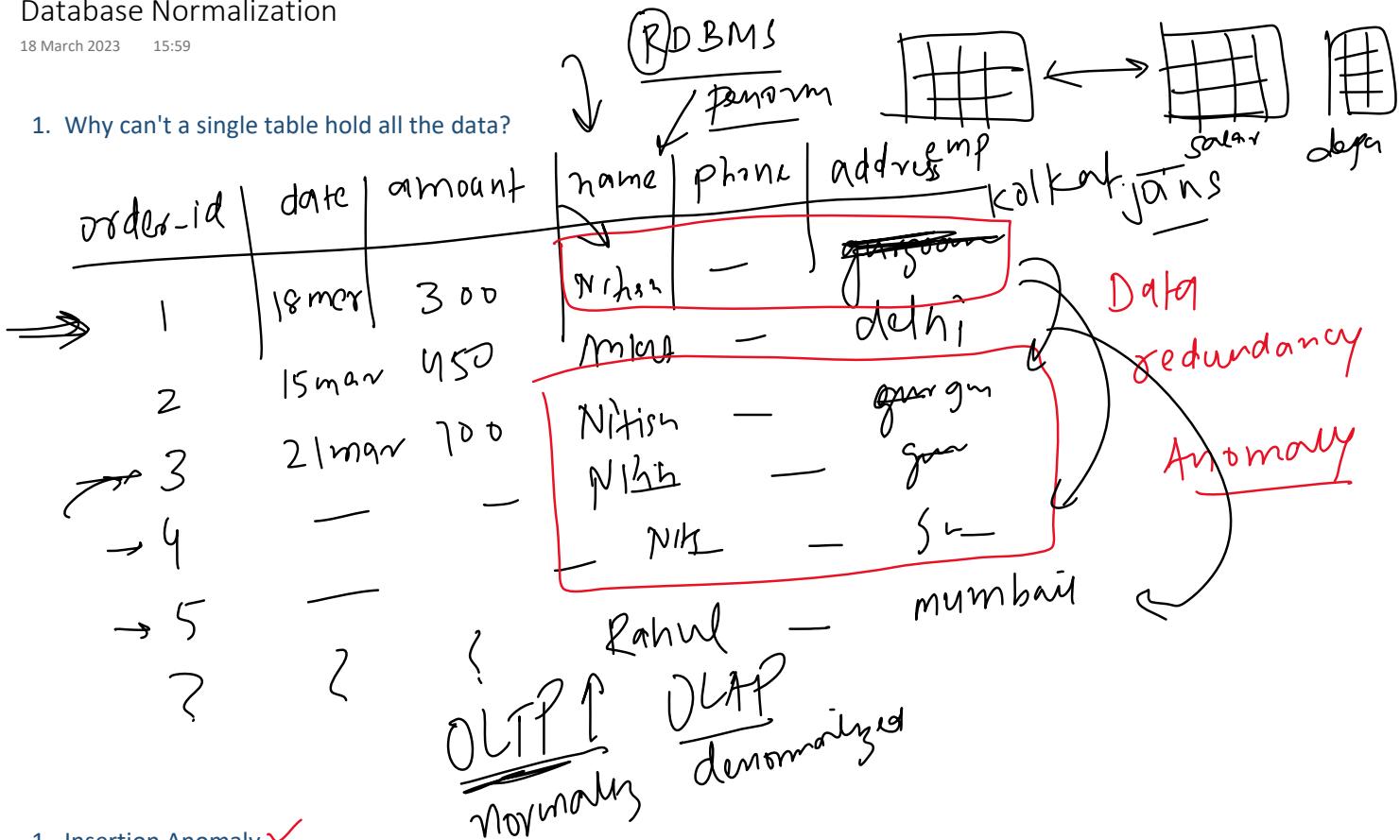
18 March 2023 08:17



Database Normalization

18 March 2023 15:59

- Why can't a single table hold all the data?



- Insertion Anomaly ✓
- Deletion Anomaly ✓
- Update Anomaly ✓
- What is the solution? -> Normalization

Database normalization is a process used to organize data in a database to reduce data redundancy and dependency. The goal of normalization is to ensure that each piece of data is stored in one place, in a structured way, to minimize the risk of inconsistencies and improve the overall efficiency and usability of the database.

There are several levels of database normalization, each with its own set of rules and guidelines. The most commonly used levels of normalization are:

- a. **First Normal Form (1NF)**: This level requires that all data in a table is stored in a way that each column contains only atomic (indivisible) values, and there are no repeating groups or arrays.
- b. **Second Normal Form (2NF)**: This level requires that each non-key attribute in a table is dependent on the entire primary key, not just a part of it.
- c. **Third Normal Form (3NF)**: This level requires that each non-key attribute in a table is dependent only on the primary key and not on any other non-key attributes.

There are higher levels of normalization, such as Fourth Normal Form (4NF) and Fifth Normal Form (5NF), but they are less commonly used in practice.

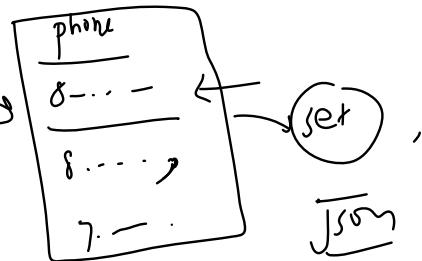
BCNF

1st Normal Form

18 March 2023 15:59

A table is in 1 NF if:

- a. There are only Single Valued Attributes or each col should contain atomic values.
- b. Attribute Domain does not change -> data type should not change.
- c. There is a unique name for every Attribute/Column.
- d. The order in which data is stored does not matter.



Employee ID	First Name	Last Name	Address	Skills
1	John	Smith	123 Main St, Anytown	Programming, Database Management
2	Jane	Doe	456 Elm St, Othertown	Programming, Project Management
3	Bob	Johnson	789 Oak St, Thirdtown	Database Management, Networking

Employee ID	First Name	Last Name	House Address	City	Skill
1	John	Smith	123 Main St	Anytown	Programming
1	John	Smith	123 Main St	Anytown	Database Management
2	Jane	Doe	456 Elm St	Othertown	Programming
2	Jane	Doe	456 Elm St	Othertown	Project Management
3	Bob	Johnson	789 Oak St	Thirdtown	Database Management
3	Bob	Johnson	789 Oak St	Thirdtown	Networking

Employee ID	First Name	Last Name	House Address	City
1	John	Smith	123 Main St	Anytown
2	Jane	Doe	456 Elm St	Othertown
3	Bob	Johnson	789 Oak St	Thirdtown

Skill ID	Skill
1	Programming
2	Database Management
3	Project Management
4	Networking

Employee ID	Skill ID
1	1
1	2
2	1
2	3
3	2

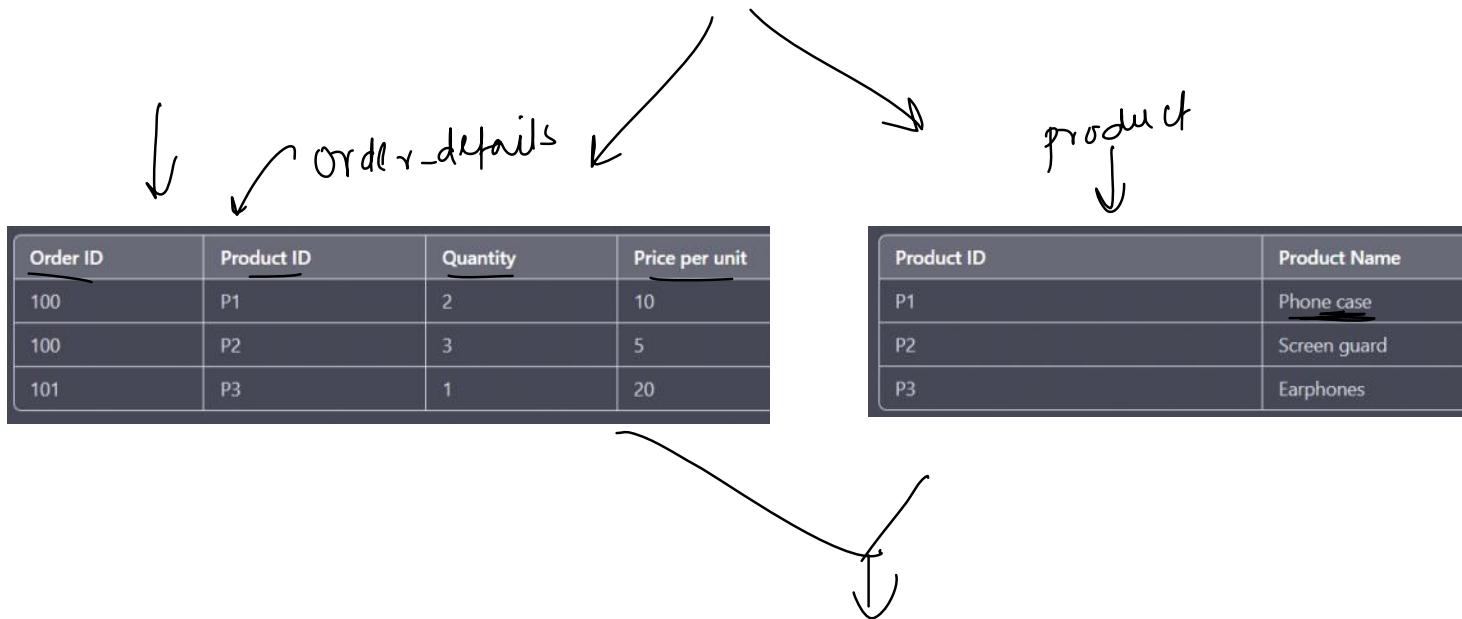
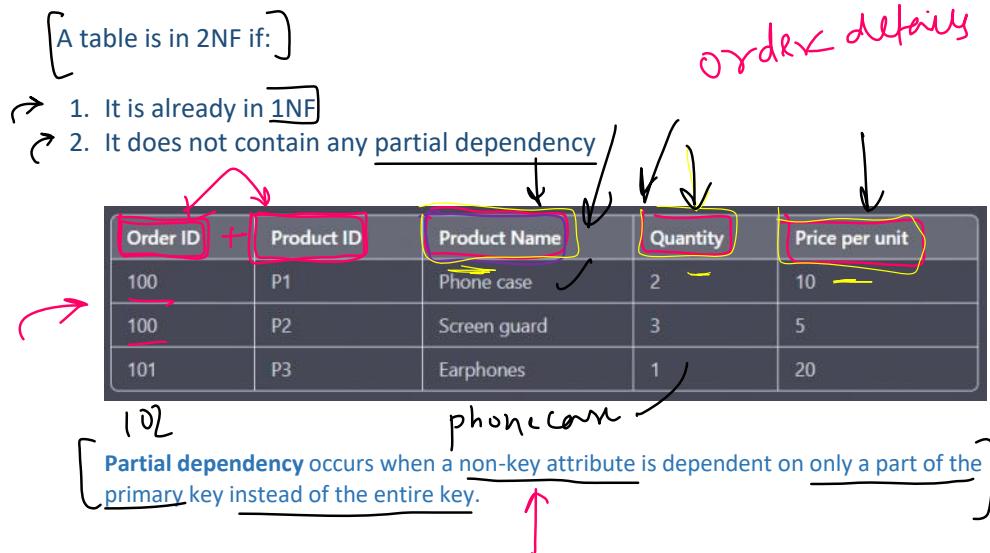
3	Project Management
4	Networking



2	3
3	2
3	4

2nd Normal Form

18 March 2023 16:01



3rd Normal Form

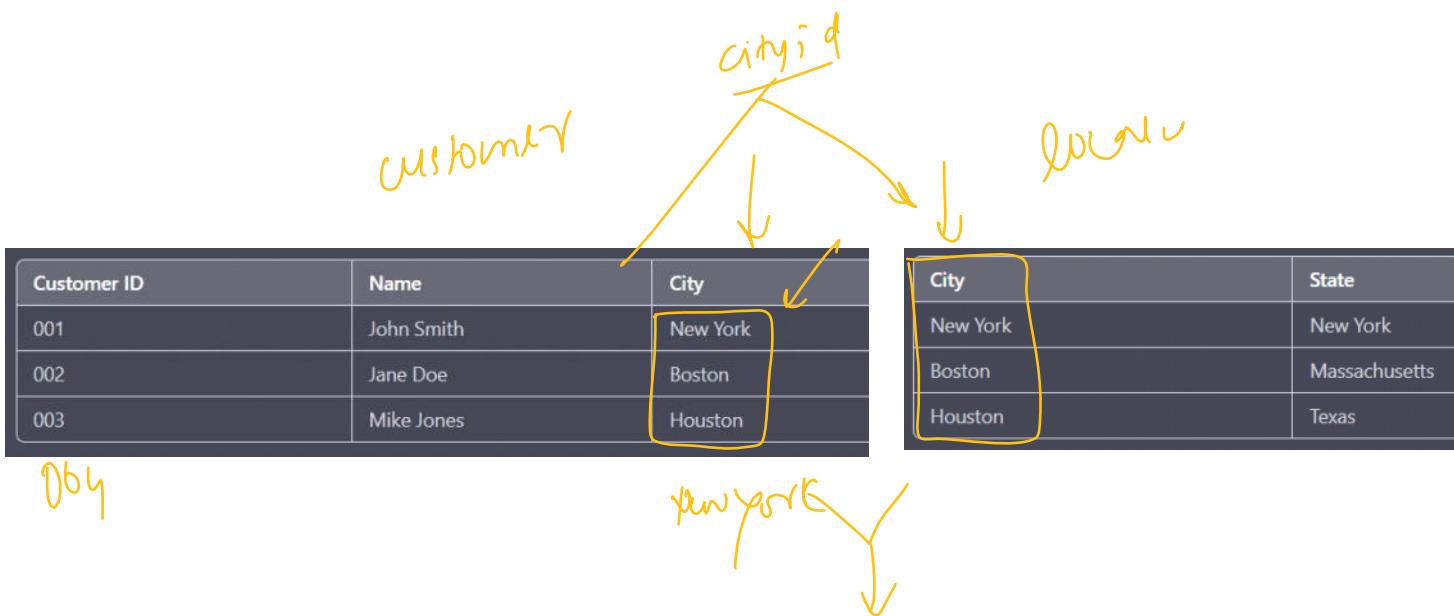
18 March 2023 16:01

A table is in 3NF if:

- 1. If it is already in 2NF
- 2. There is no transitive dependency.

A transitive dependency exists when a non-key attribute depends on another non-key attribute, which is not a part of the primary key.

Customer ID	Name	City	State
001	John Smith	New York	New York
002	Jane Doe	Boston	Massachusetts
003	Mike Jones	Houston	Texas



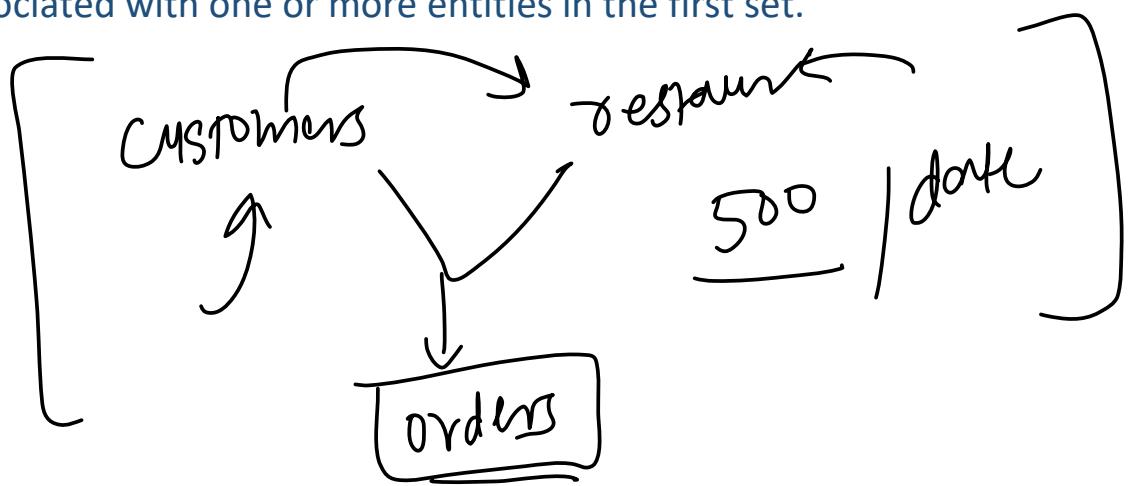
ER Diagram

18 March 2023 17:47

ER diagram stands for Entity-Relationship diagram. It is a graphical representation of entities and their relationships to each other. ER diagrams are used in database design to visualize the entities, attributes, and relationships involved in a system.

There are three basic types of relationships in an ER diagram:

- a. One-to-One (1:1): Each entity in one set is associated with only one entity in the other set, and vice versa.
- b. One-to-Many (1:N): Each entity in one set is associated with one or more entities in the other set, but each entity in the other set is associated with only one entity in the first set.
- c. Many-to-Many (N:M): Each entity in one set is associated with one or more entities in the other set, and each entity in the other set is associated with one or more entities in the first set.



Random Variables

15 March 2023 11:43

- What are Algebraic Variables?

In Algebra a variable, like x , is an unknown value

$$x + 5 = 10 \Rightarrow x = 5$$

Dice $\begin{matrix} 1 & -4 \\ 2 & -5 \\ 3 & -6 \end{matrix}$

- What are Random Variables in Stats and Probability?

A Random Variable is a set of possible values from a random experiment.

coin toss H

$$\underline{x} = \{1, 0\} \quad \underline{y} = \{1, 2, 3, 4, 5, 6\}$$

$H=1 \quad T=0$ randomly \hookrightarrow sample space

$$\boxed{X \ Y \ Z} \quad x, y, z$$

- { • Types of Random Variables? }

Discrete
RV

$$\{H, T\}$$

$$\{1, 2, 3, 4, 5, 6\}$$

$\uparrow \uparrow \uparrow$

Continuous
RV

$$X = \{0, 10\}$$

Probability Distributions

15 March 2023 11:53

1. What are Probability Distributions?

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values.

coin toss	1 (H)	0 (T)
probab	$\frac{1}{2}$	$\frac{1}{2}$

dice

2 dice \rightarrow

2 3 4 5 6 7 8 9

1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

1	1	2	3	4	5	6	7
2	3	4	5	6	7	8	
3	4	5	6	7	8	9	
4	5	6	7	8	9	10	
5	6	7	8	9	10	11	
6	7	8	9	10	11	12	

Problem with Distribution?

2 \rightarrow	$\frac{1}{36}$
3 \rightarrow	$\frac{2}{36}$
4 \rightarrow	$\frac{3}{36}$
5 \rightarrow	$\frac{4}{36}$
6 \rightarrow	$\frac{5}{36}$

7 \rightarrow $\frac{6}{36}$

8 \rightarrow $\frac{5}{36}$

9 \rightarrow $\frac{4}{36}$

10 \rightarrow $\frac{3}{36}$

11 \rightarrow $\frac{2}{36}$

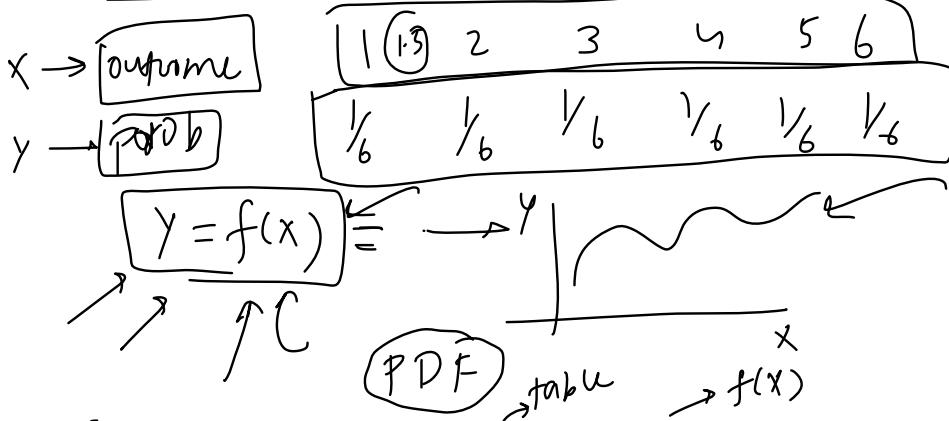
12 \rightarrow $\frac{1}{36}$

In many scenarios, the number of outcomes can be much larger and hence a table would be tedious to write down. Worse still, the number of possible outcomes could be infinite, in which case, good luck writing a table for that.

Example - Height of people, Rolling 10 dice together

→ Solution - Function?

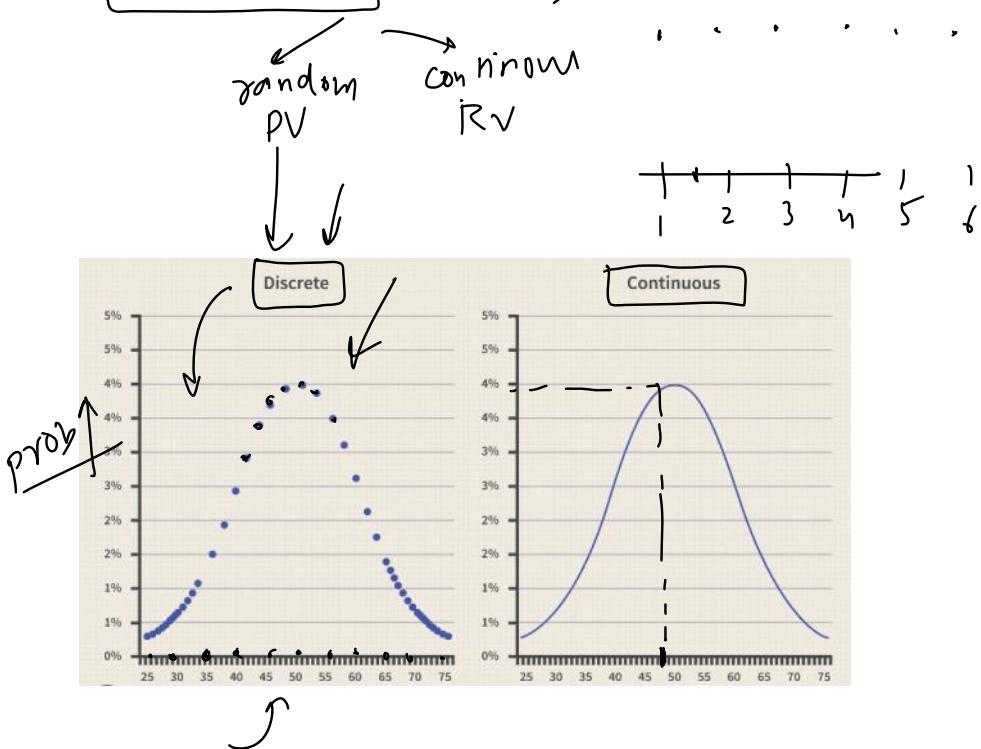
→ What if we use a mathematical function to model the relationship between outcome and probability?



∫ Note - A lot of time Probability Distribution and Probability Distribution Functions are ↴

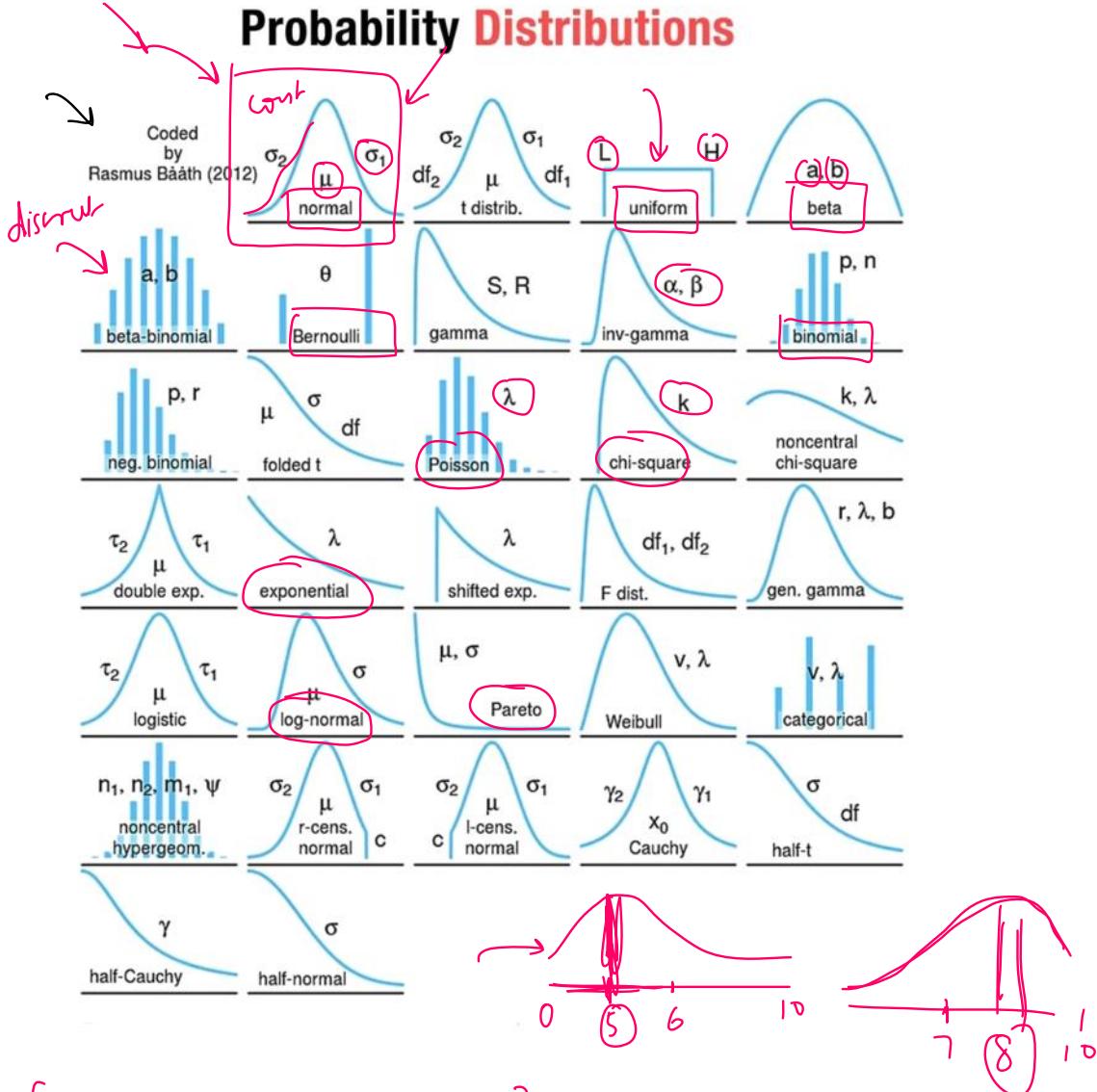
(PDF) \rightarrow tabular $\rightarrow f(x)$
 Note - A lot of time Probability Distribution and Probability Distribution Functions are used interchangeably.

1. Types of Probability Distributions (PDF)



Famous Probability Distributions

Probability Distributions



Why are Probability Distributions important?

- Gives an idea about the shape/distribution of the data.
- And if our data follows a famous distribution then we automatically know a lot about the data.

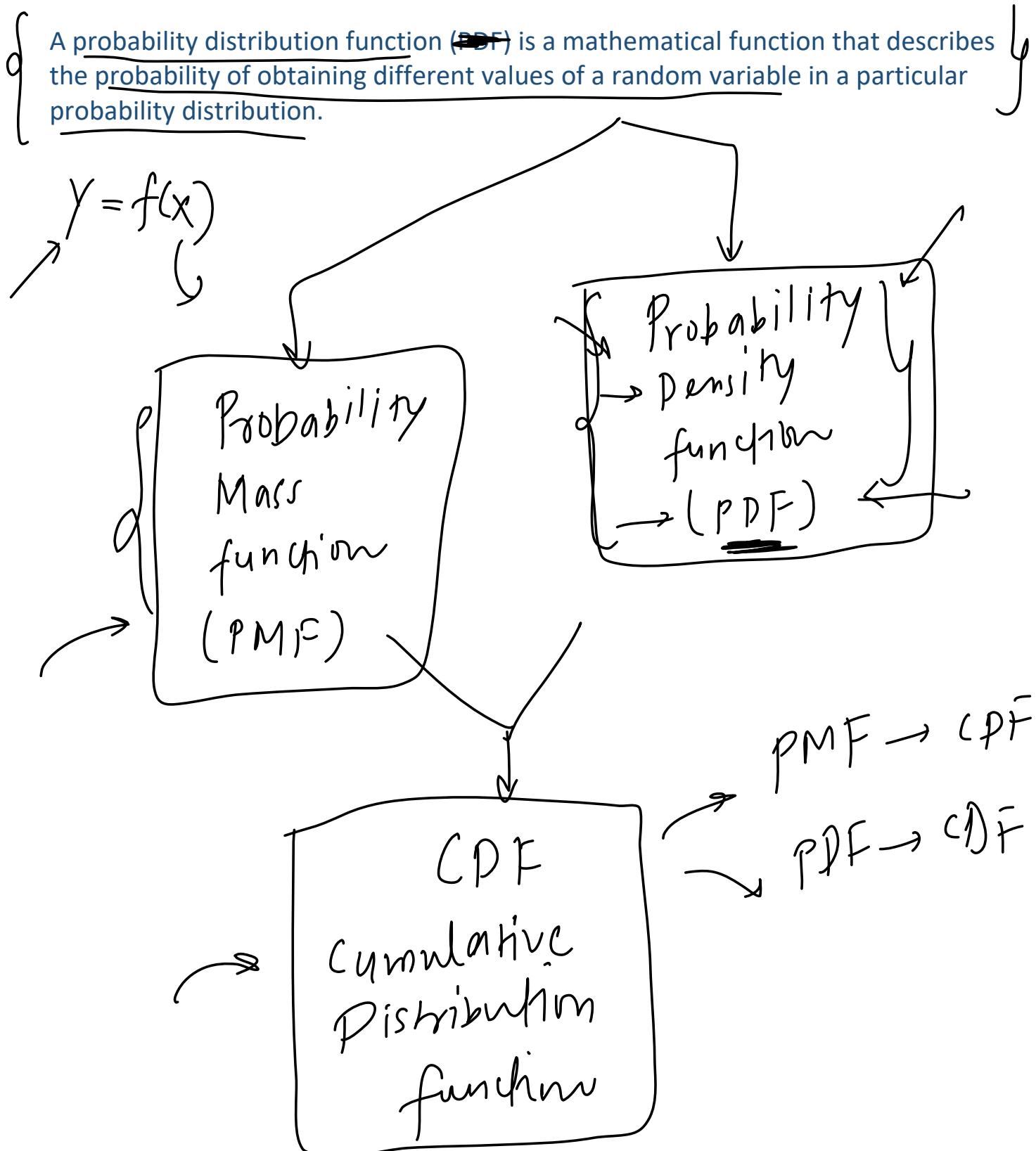
A note on Parameters (PDF)

Parameters in probability distributions are numerical values that determine the shape, location, and scale of the distribution.

Different probability distributions have different sets of parameters that determine their shape and characteristics, and understanding these parameters is essential in statistical analysis and inference.

[Probability Distribution Functions] → PDF

15 March 2023 20:08



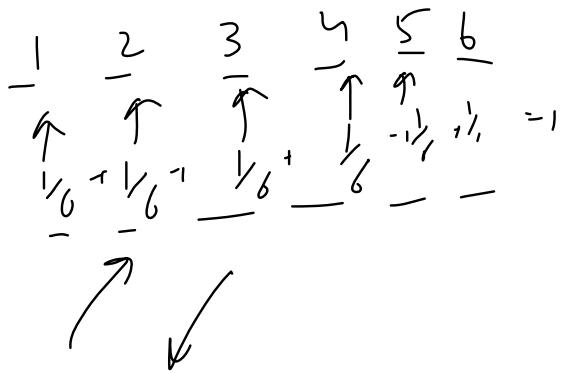
Probability Mass Function (PMF)

15 March 2023 15:25

PMF stands for Probability Mass Function. It is a mathematical function that describes the probability distribution of a **discrete random variable**.

The PMF of a discrete random variable assigns a probability to each possible value of the random variable. The probabilities assigned by the PMF must satisfy two conditions:

- a. The probability assigned to each value must be non-negative (i.e., greater than or equal to zero).
- b. The sum of the probabilities assigned to all possible values must equal 1.



$$y = f(x) \rightarrow y = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

pmf → $y = \begin{cases} \frac{1}{36} & x \in \{2, 12\} \\ \frac{2}{36} & x \in \{3, 11\} \\ 0 & \text{otherwise} \end{cases}$

Examples

https://en.wikipedia.org/wiki/Bernoulli_distribution

https://en.wikipedia.org/wiki/Binomial_distribution

Cumulative Distribution Function(CDF) of PMF

15 March 2023 20:09

The cumulative distribution function (CDF) $F(x)$ describes the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x

$$F(x) = P(X \leq x)$$

\downarrow \downarrow

PMF $f(x)$

$$f(x \leq 4) = f(x=4) + f(x=3) + f(x=2) + f(x=1)$$

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

Examples:

https://en.wikipedia.org/wiki/Bernoulli_distribution

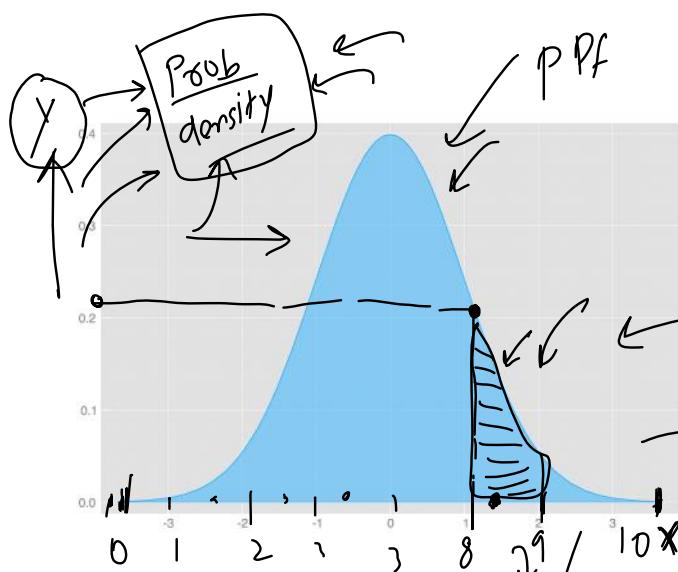
https://en.wikipedia.org/wiki/Binomial_distribution

	PMF	CDF
1	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{2}{6}$
3	$\frac{1}{6}$	$\frac{3}{6}$
4	$\frac{1}{6}$	$\frac{4}{6}$
5	$\frac{1}{6}$	$\frac{5}{6}$
6	$\frac{1}{6}$	$\frac{6}{6}$

Probability Density Function (PDF)]

15 March 2023 15:25

PDF stands for Probability Density Function. It is a mathematical function that describes the probability distribution of a **continuous random variable**.



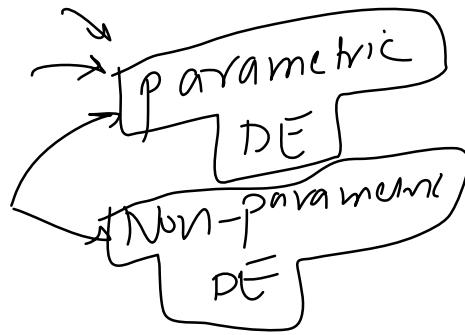
$$\begin{aligned}
 & \text{Sample PDF} \\
 & \text{cgp} = \boxed{7.912} \\
 & 7.9123 \leftarrow \text{FG} \\
 \hline
 & \overbrace{\text{p}(0 \leq x \leq 1)}^{\text{area}} = 1 \\
 & \xrightarrow{\text{marks}} 8 \rightarrow 8.1 \\
 & \quad 8 \rightarrow 8.01
 \end{aligned}$$

1. Why Probability Density and why not Probability
 2. What does the area of this graph represents?
 3. How to calculate Probability then?
 4. Examples of PDF
 - a. https://en.wikipedia.org/wiki/Normal_distribution
 - b. https://en.wikipedia.org/wiki/Log-normal_distribution
 - c. https://en.wikipedia.org/wiki/Poisson_distribution
 5. How is graph calculated?

Density Estimation

16 March 2023 06:54

Density estimation is a statistical technique used to estimate the probability density function (PDF) of a random variable based on a set of observations or data. In simpler terms, it involves estimating the underlying distribution of a set of data points.



→ Density estimation can be used for a variety of purposes, such as hypothesis testing, data analysis, and data visualization. It is particularly useful in areas such as machine learning, where it is often used to estimate the probability distribution of input data or to model the likelihood of certain events or outcomes.

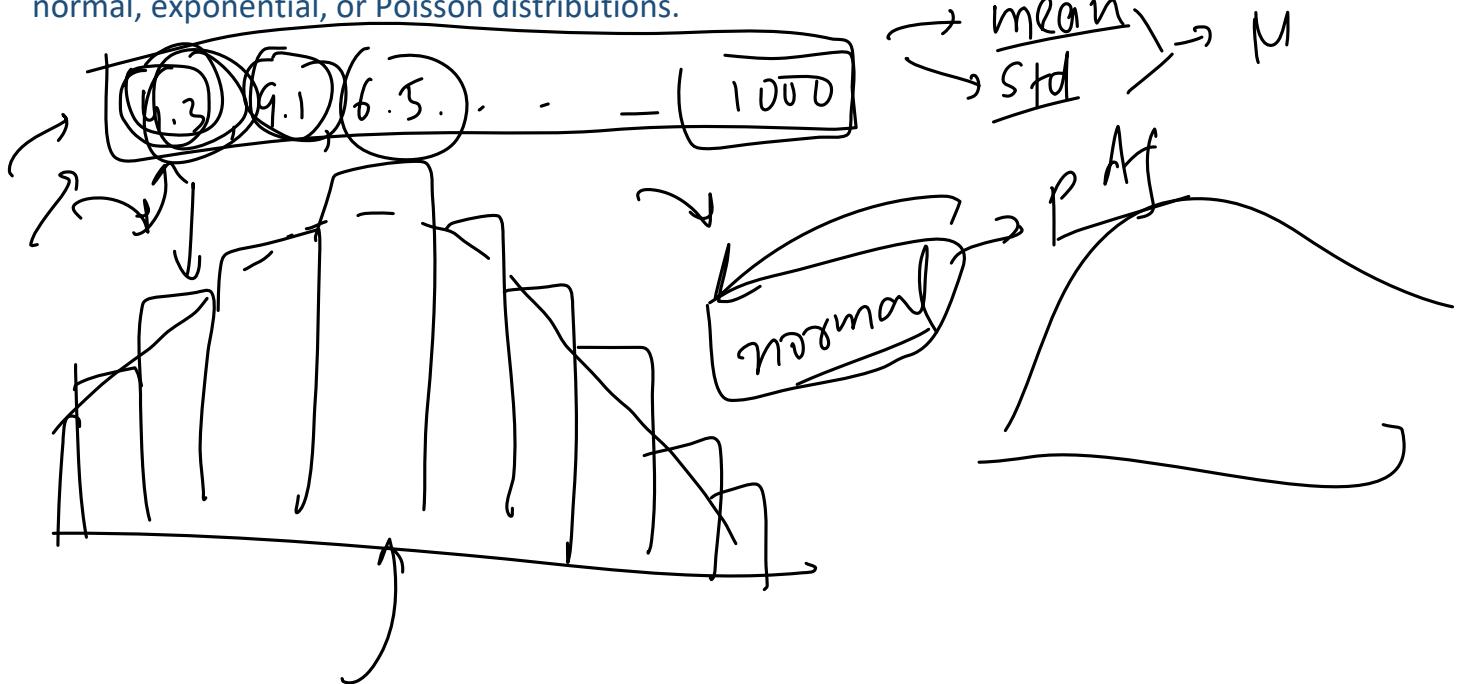
There are various methods for density estimation, including **parametric** and **non-parametric approaches**. Parametric methods assume that the data follows a specific probability distribution (such as a normal distribution), while non-parametric methods do not make any assumptions about the distribution and instead estimate it directly from the data.

Commonly used techniques for density estimation include kernel density estimation (KDE), histogram estimation, and Gaussian mixture models (GMMs). The choice of method depends on the specific characteristics of the data and the intended use of the density estimate.

Parametric Density Estimation

16 March 2023 06:54

Parametric density estimation is a method of estimating the probability density function (PDF) of a random variable by assuming that the underlying distribution belongs to a specific parametric family of probability distributions, such as the normal, exponential, or Poisson distributions.



Non-Parametric Density Estimation (KDE)

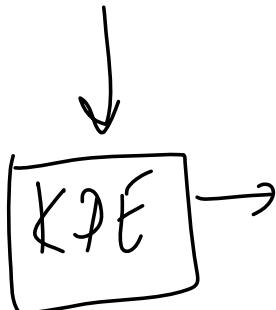
16 March 2023 06:55

But sometimes the distribution is not clear or it's not one of the famous distributions.

- Non-parametric density estimation is a statistical technique used to estimate the probability density function of a random variable without making any assumptions about the underlying distribution. It is also referred to as non-parametric density estimation because it does not require the use of a predefined probability distribution function, as opposed to parametric methods such as the Gaussian distribution.

The non-parametric density estimation technique involves constructing an estimate of the probability density function using the available data. This is typically done by creating a kernel density estimate

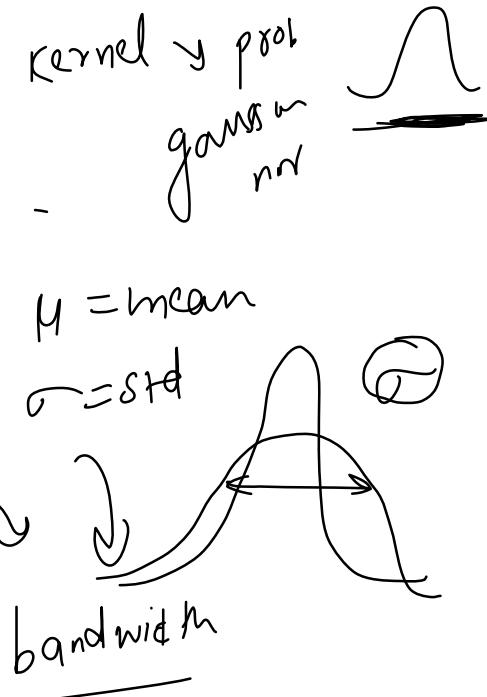
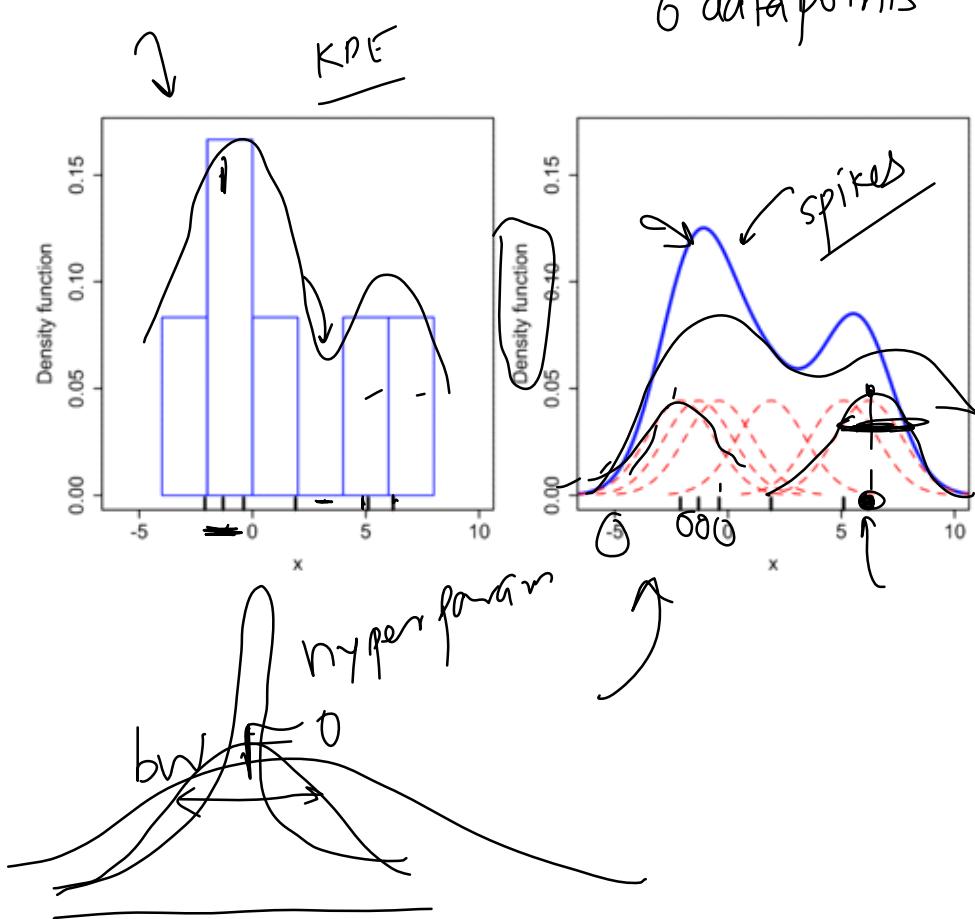
{ Non-parametric density estimation has several advantages over parametric density estimation. One of the main advantages is that it does not require the assumption of a specific distribution, which allows for more flexible and accurate estimation in situations where the underlying distribution is unknown or complex. However, non-parametric density estimation can be computationally intensive and may require more data to achieve accurate estimates compared to parametric methods.



Kernel Density Estimate(KDE)

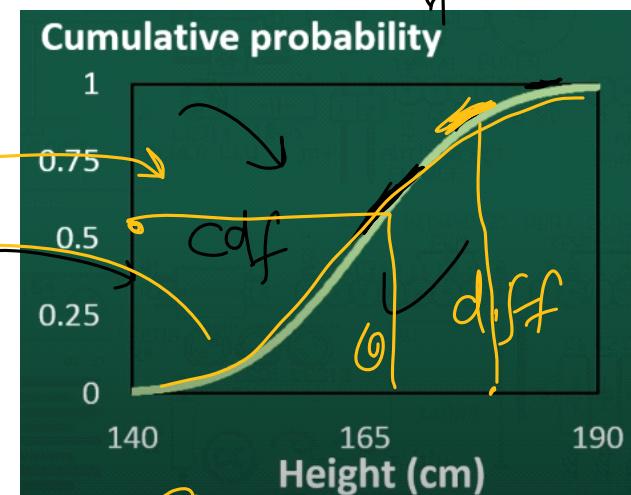
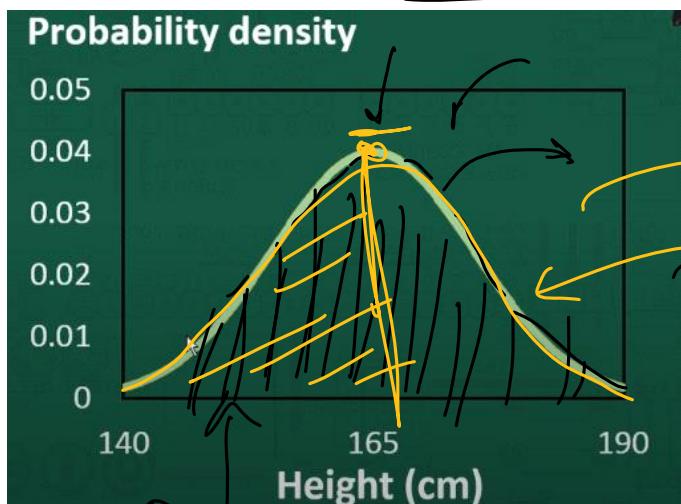
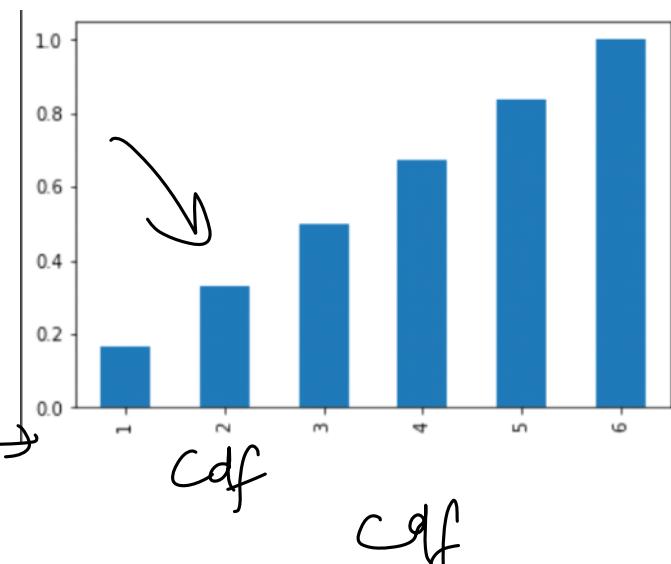
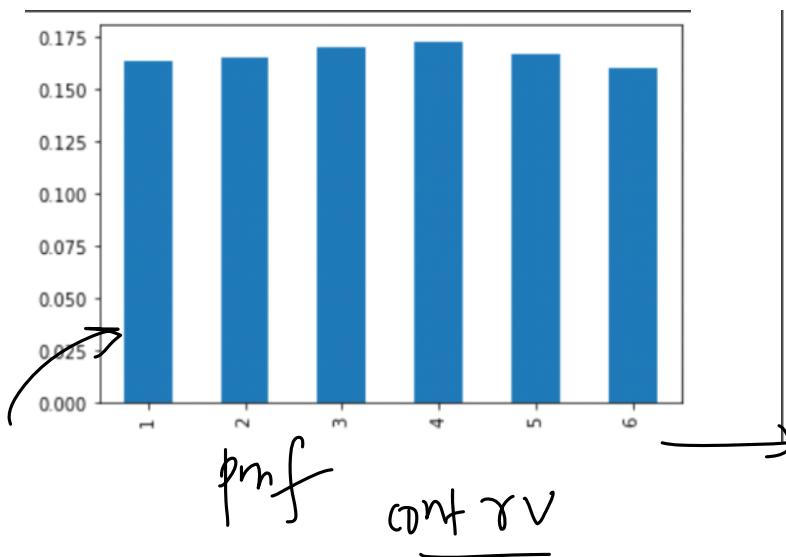
16 March 2023 16:08

The KDE technique involves using a kernel function to smooth out the data and create a continuous estimate of the underlying density function.



Cumulative Distribution Function(CDF) of PDF

15 March 2023 15:25



integrate w.r.t.

How to use PDF and CDF in Data Analysis

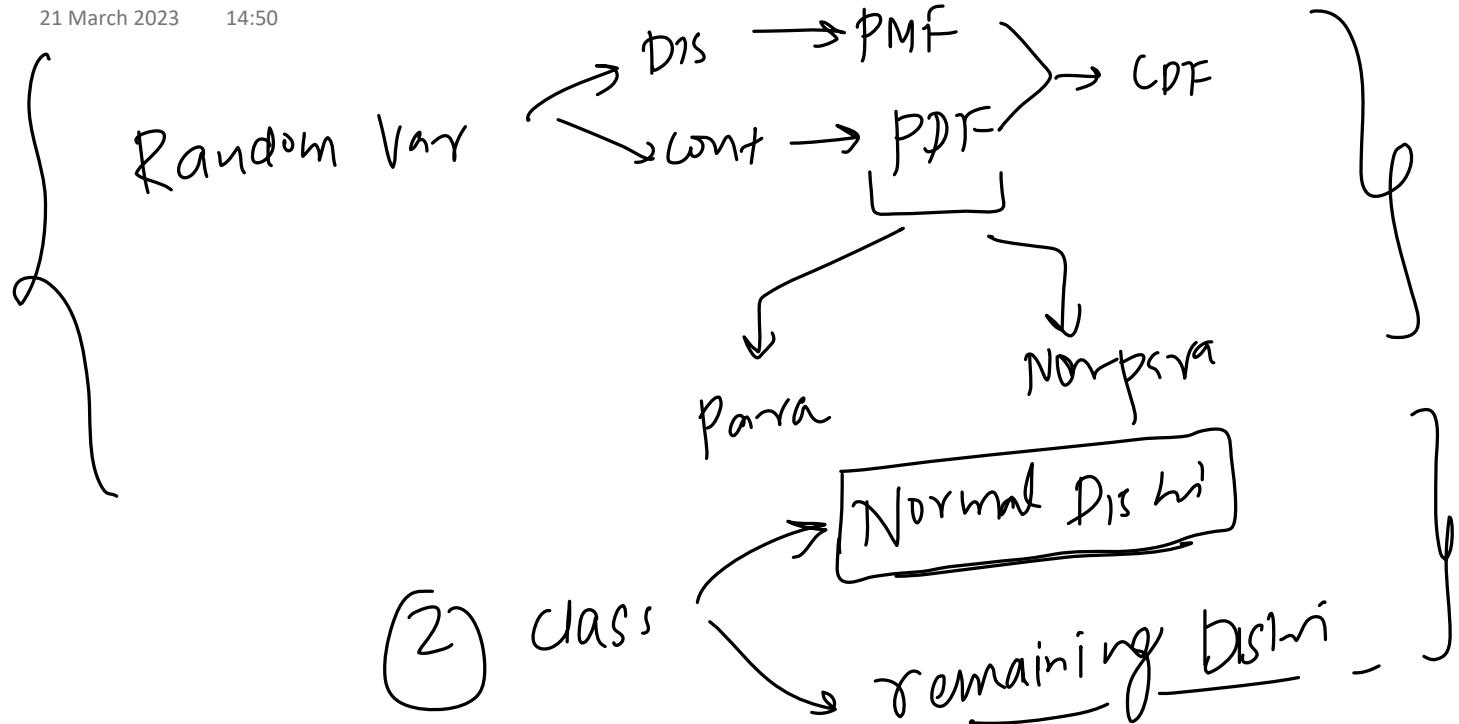
15 March 2023 20:10

2D Probability Density Plots

16 March 2023 06:50

Recap

21 March 2023 14:50



How to use PDF in Data Science

20 March 2023 18:11

PDF →

CDF

PDF

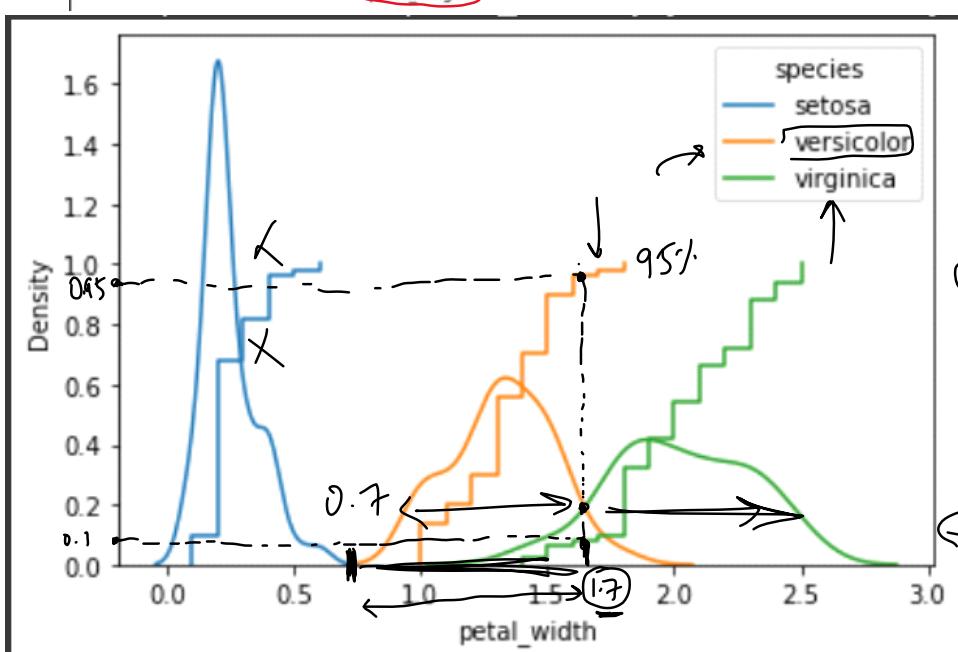
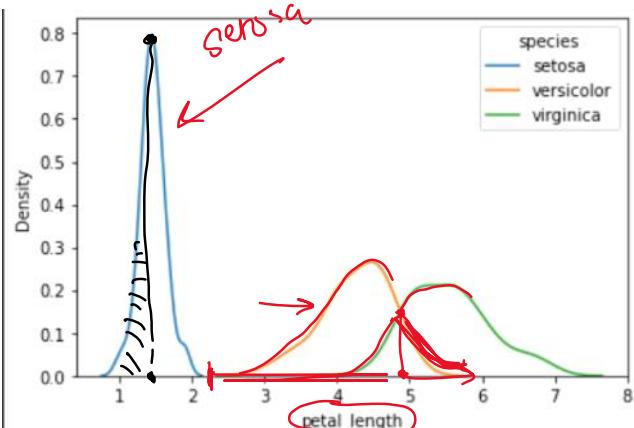
$P(X=x)$

$2.3 < pl < 5 \rightarrow$ versicolor

$P(X \leq x)$

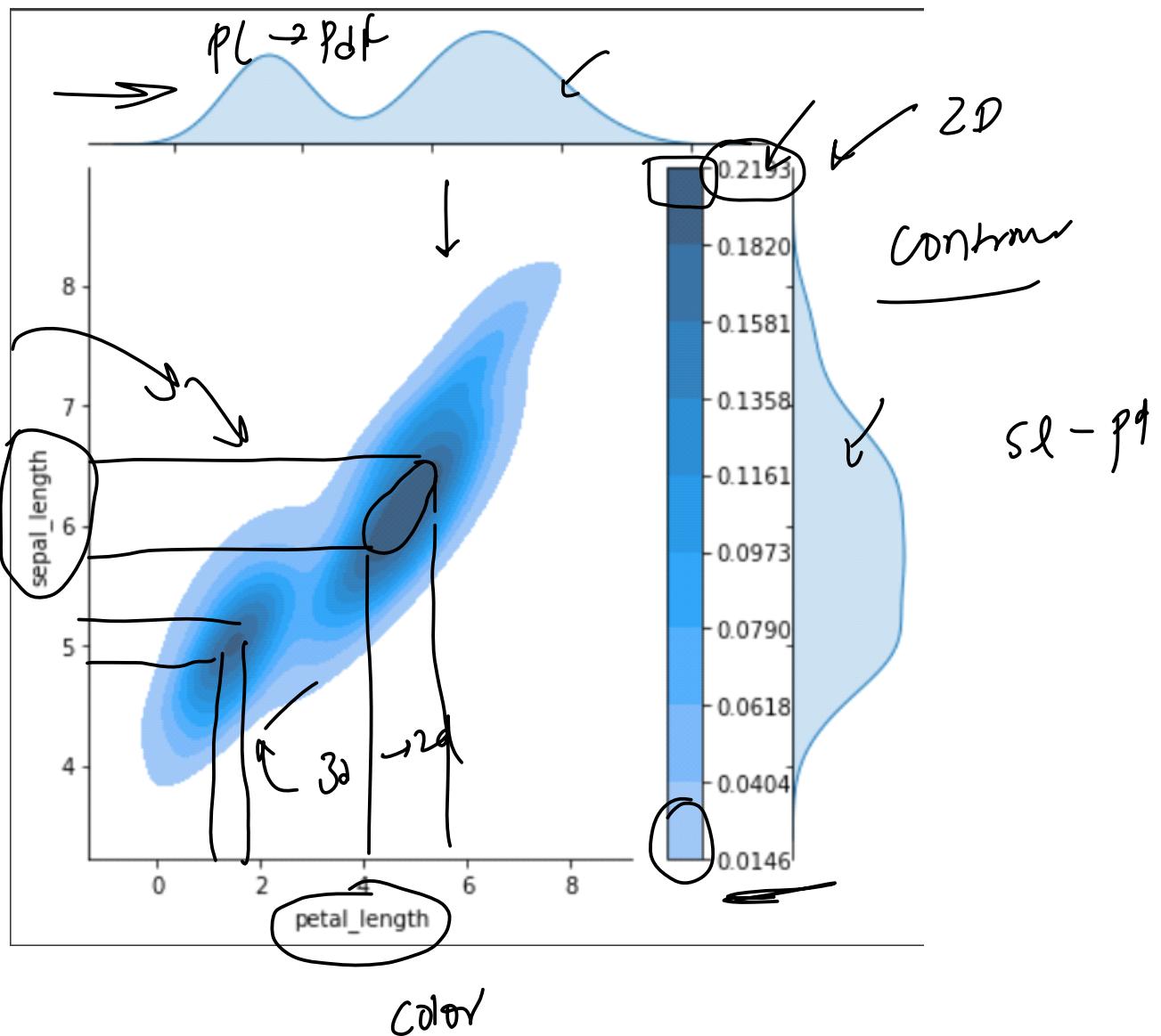
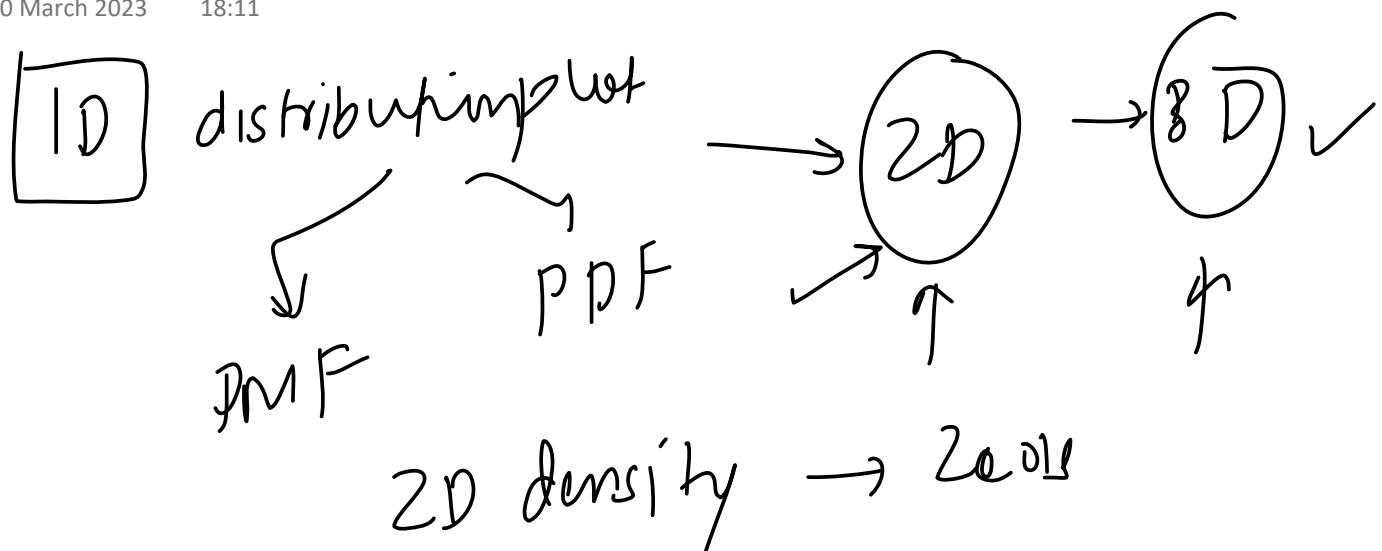
$pl > 5 \rightarrow$ virginica

$pl < 2.3$
 > 2.3



2D Density Plots

20 March 2023 18:11



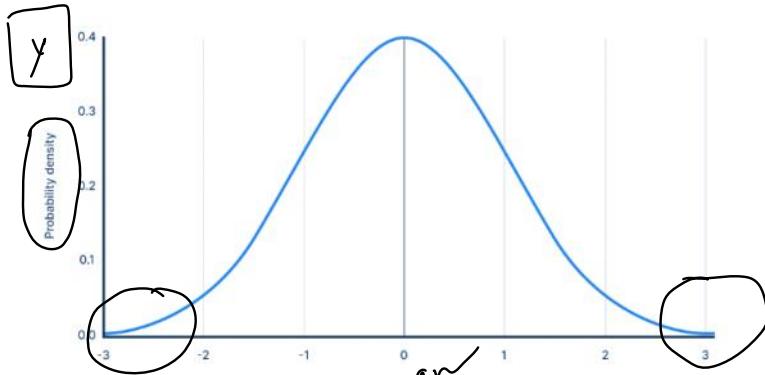
Normal Distribution

20 March 2023 18:06

1. What is normal distribution?

Normal distribution, also known as Gaussian distribution, is a probability distribution that is commonly used in statistical analysis. It is a continuous probability distribution that is symmetrical around the mean, with a bell-shaped curve.

→ pdf



- > Tail
- > Asymptotic in nature
- > Lots of points near the mean and very few far away

X

The normal distribution is characterized by two parameters: the mean (μ) and the standard deviation (σ). The mean represents the centre of the distribution, while the standard deviation represents the spread of the distribution.

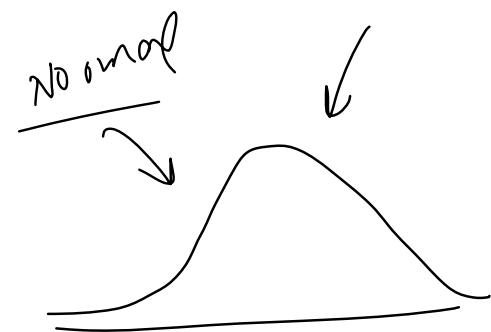
Denoted as:

$$X \sim N(\mu, \sigma)$$

$\mu \rightarrow \text{mean}$
 $\sigma \rightarrow \text{std}$

Why is it so important?

Commonality in Nature: Many natural phenomena follow a normal distribution, such as the heights of people, the weights of objects, the IQ scores of a population, and many more. Thus, the normal distribution provides a convenient way to model and analyse such data.



PDF Equation of Normal Distribution

$$y = f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

μ, σ

data → Normal → μ, σ

$\sigma \sqrt{2\pi}$

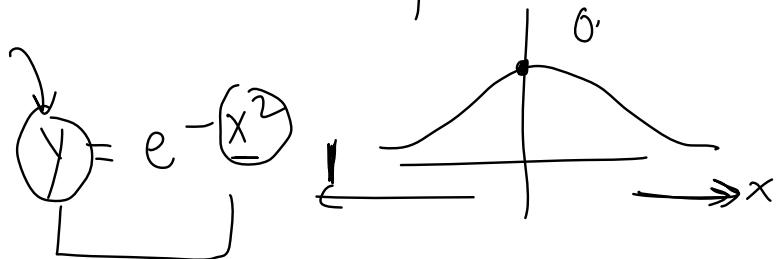
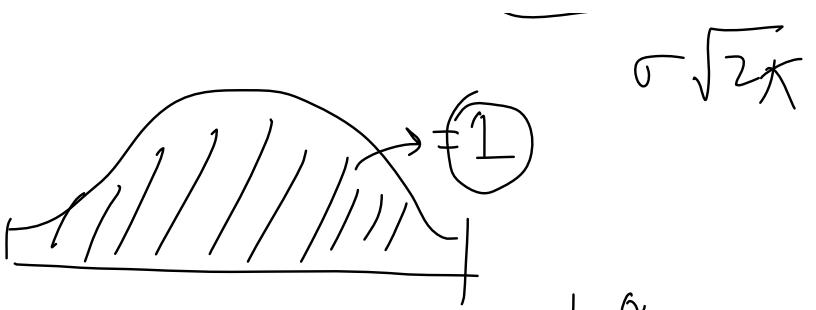
Parameters in Normal Distribution

<https://samp-suman-normal-dist-visualize-app-lkntug.streamlit.app/>

Equation in detail:

$$y = e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

$\sigma \sqrt{2\pi}$



$$y = \frac{1}{e^{x^2}}$$

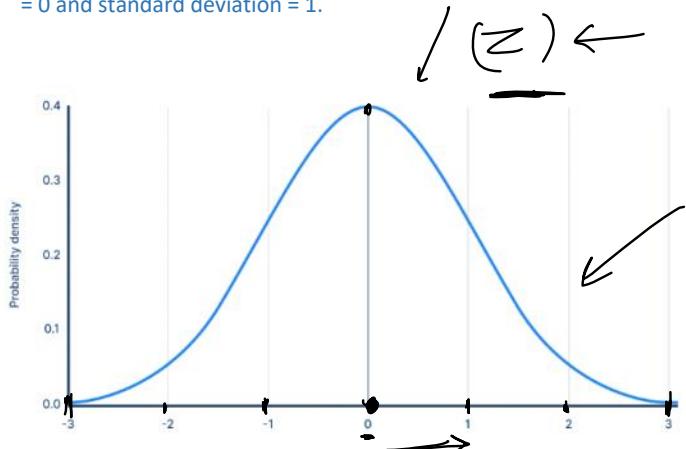
Standard Normal Variate (Z) \rightarrow Standard Normal distribution

20 March 2023 18:08

$X \sim N(\mu, \sigma)$ $\mu = 0$
 $\sigma = 1$

- What is Standard Normal Variate

A Standard Normal Variate (Z) is a standardized form of the normal distribution with mean = 0 and standard deviation = 1.



$$Z \sim N(0, 1)$$

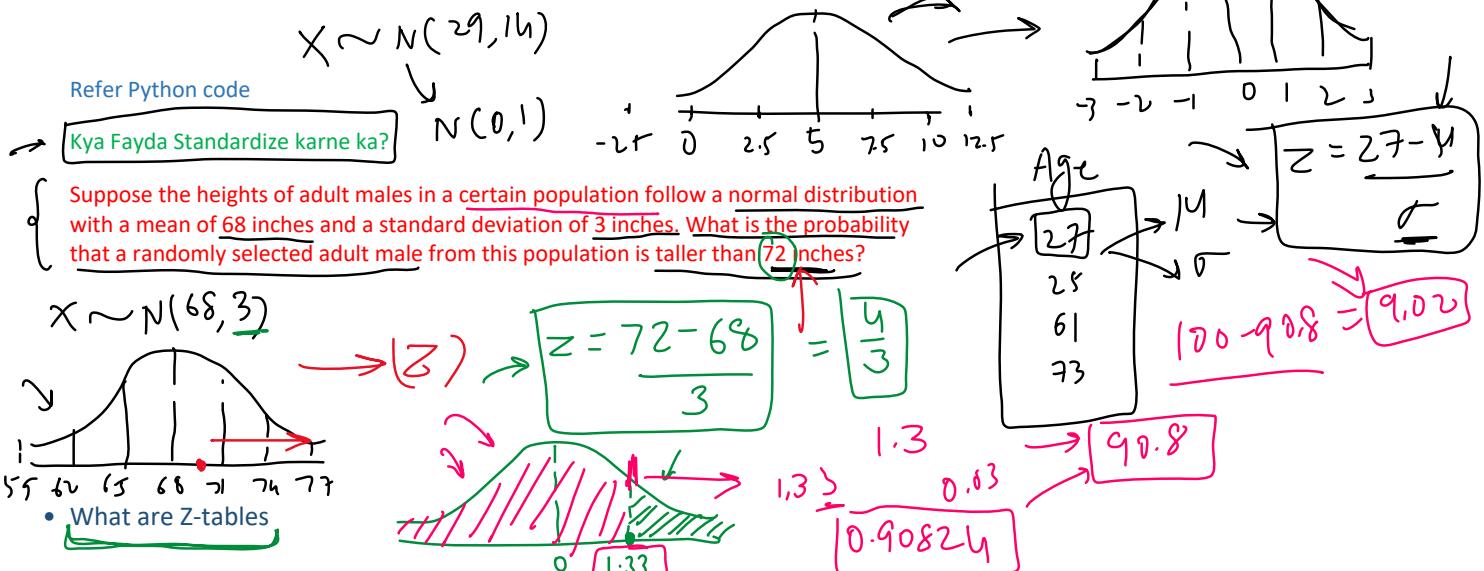
{ Standardizing a normal distribution allows us to compare different distributions with each other, and to calculate probabilities using standardized tables or software.

Equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

$$X \sim N(5, 2.5)$$

- How to transform a normal distribution to Standard Normal Variate



A z-table tells you the area underneath a normal distribution curve, to the left of the z-score

<https://www.ztable.net/>

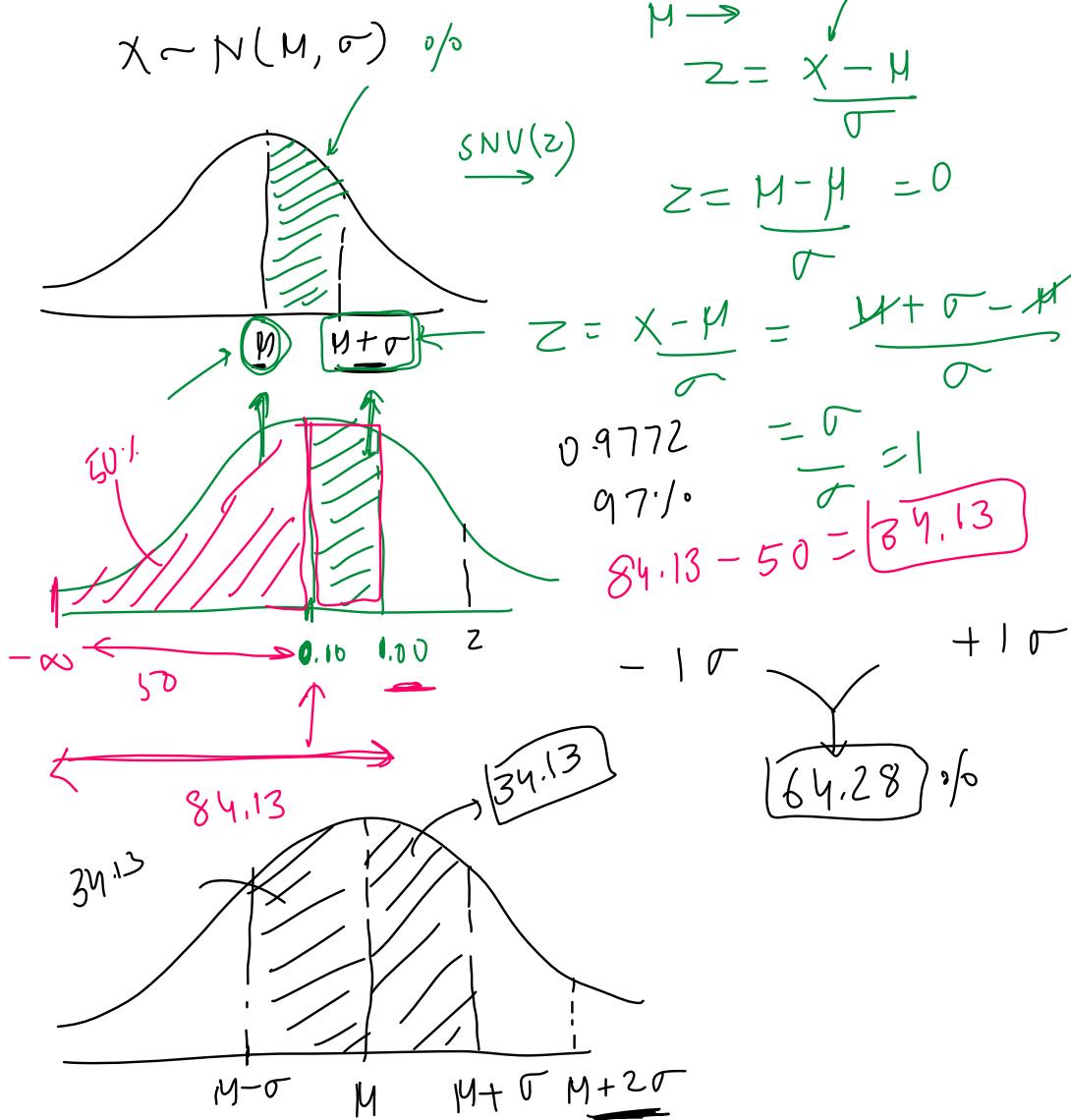
For a Normal Distribution $X \sim (\mu, \sigma)$ what percent of population lie between mean and 1 standard deviation, 2 std and 3 std?

$$X \sim N(\mu, \sigma) \%$$

$$\mu \rightarrow \checkmark$$

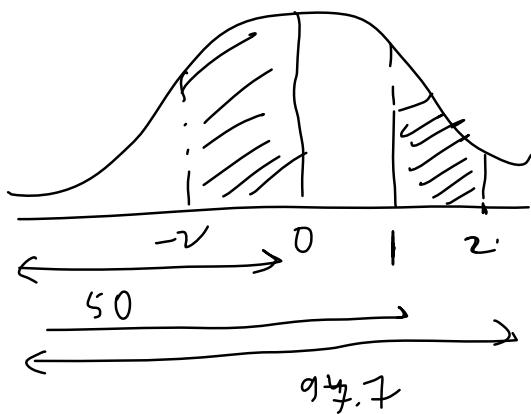
$$Z = X - \mu$$

Standard deviation, Z score and S.D.



$$Z = \frac{\mu + 2\sigma - \mu}{\sigma} = 2$$

f.g.

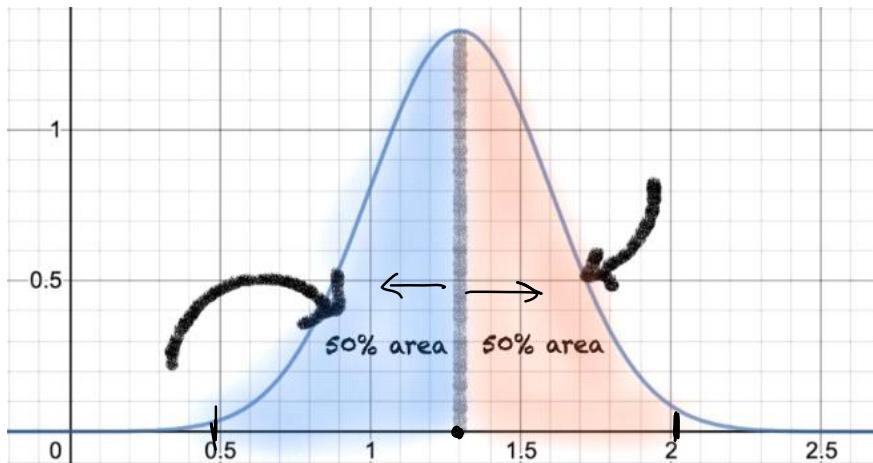


Properties of Normal Distribution

20 March 2023 18:06

1. Symmetry

The normal distribution is symmetric about its mean, which means that the probability of observing a value above the mean is the same as the probability of observing a value below the mean. The bell-shaped curve of the normal distribution reflects this symmetry.

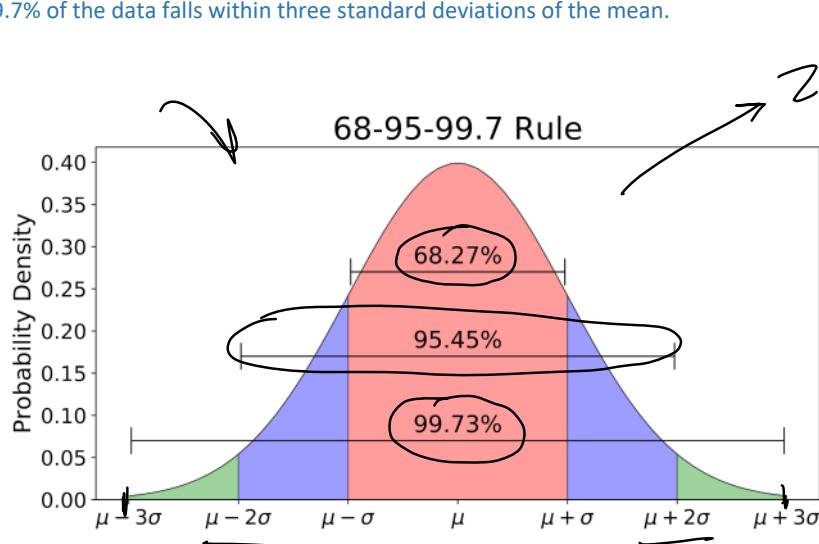


2. Measures of Central Tendencies are equal → mean → median → mode

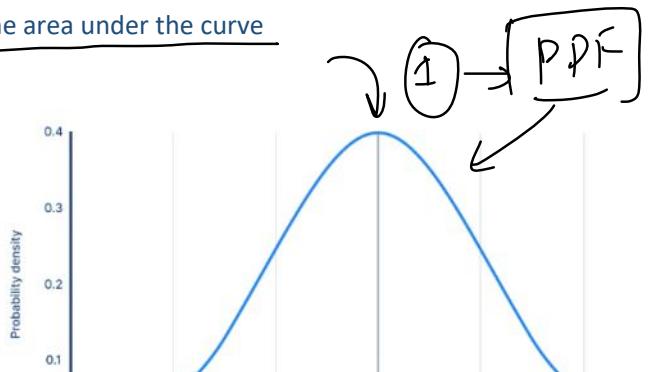
3. Empirical Rule

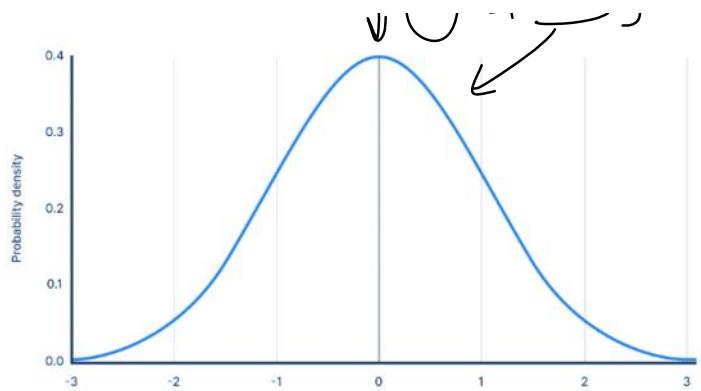
The normal distribution has a well-known empirical rule, also called the 68-95-99.7 rule, which states that approximately 68% of the data falls within one standard deviation of the mean, about 95% of the data falls within two standard deviations of the mean, and about 99.7% of the data falls within three standard deviations of the mean.

Standard deviation
~ a vertical line



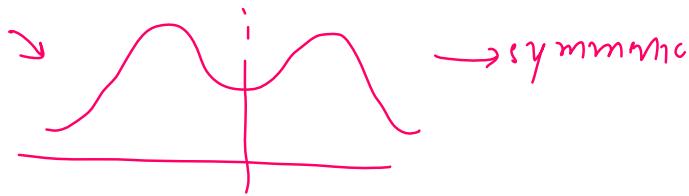
4. The area under the curve





Skewness

20 March 2023 18:07



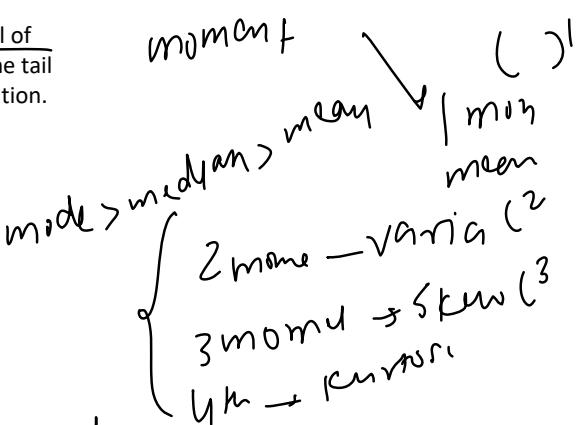
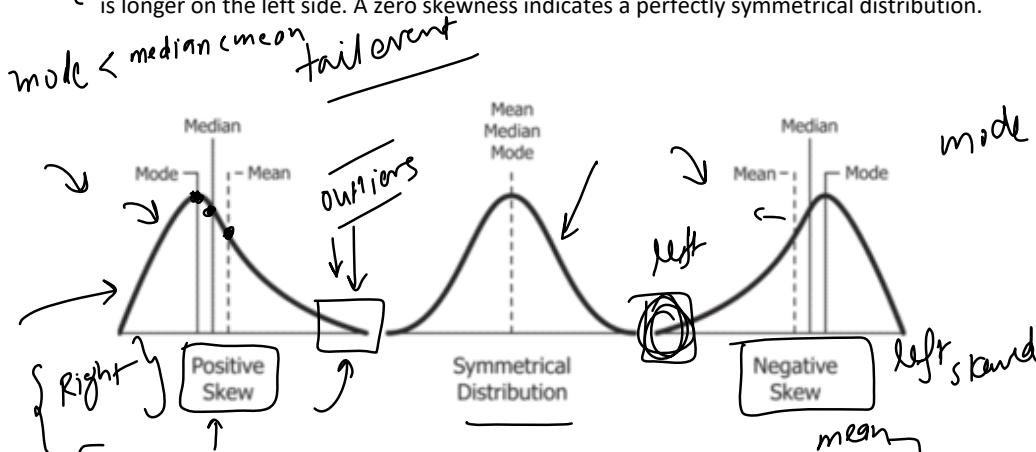
- What is skewness?

A normal distribution is a bell-shaped, symmetrical distribution with a specific mathematical formula that describes how the data is spread out. Skewness indicates that the data is not symmetrical, which means it is not normally distributed.

Skewness is a measure of the asymmetry of a probability distribution. It is a statistical measure that describes the degree to which a dataset deviates from the normal distribution.

In a symmetrical distribution, the mean, median, and mode are all equal. In contrast, in a skewed distribution, the mean, median, and mode are not equal, and the distribution tends to have a longer tail on one side than the other.

Skewness can be positive, negative, or zero. A positive skewness means that the tail of the distribution is longer on the right side, while a negative skewness means that the tail is longer on the left side. A zero skewness indicates a perfectly symmetrical distribution.

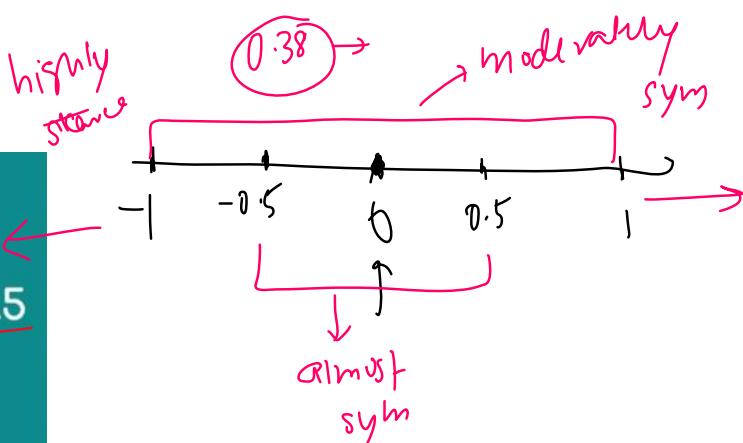
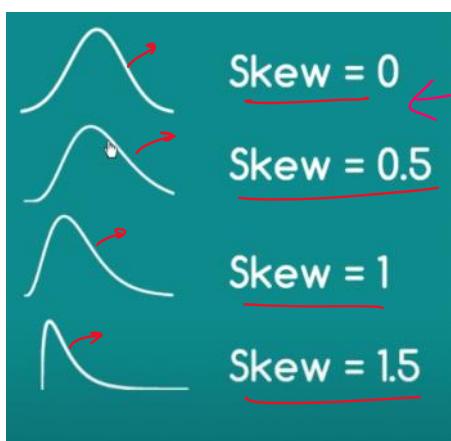


The greater the skew the greater the distance between mode, median and mean.

- How skewness is calculated?

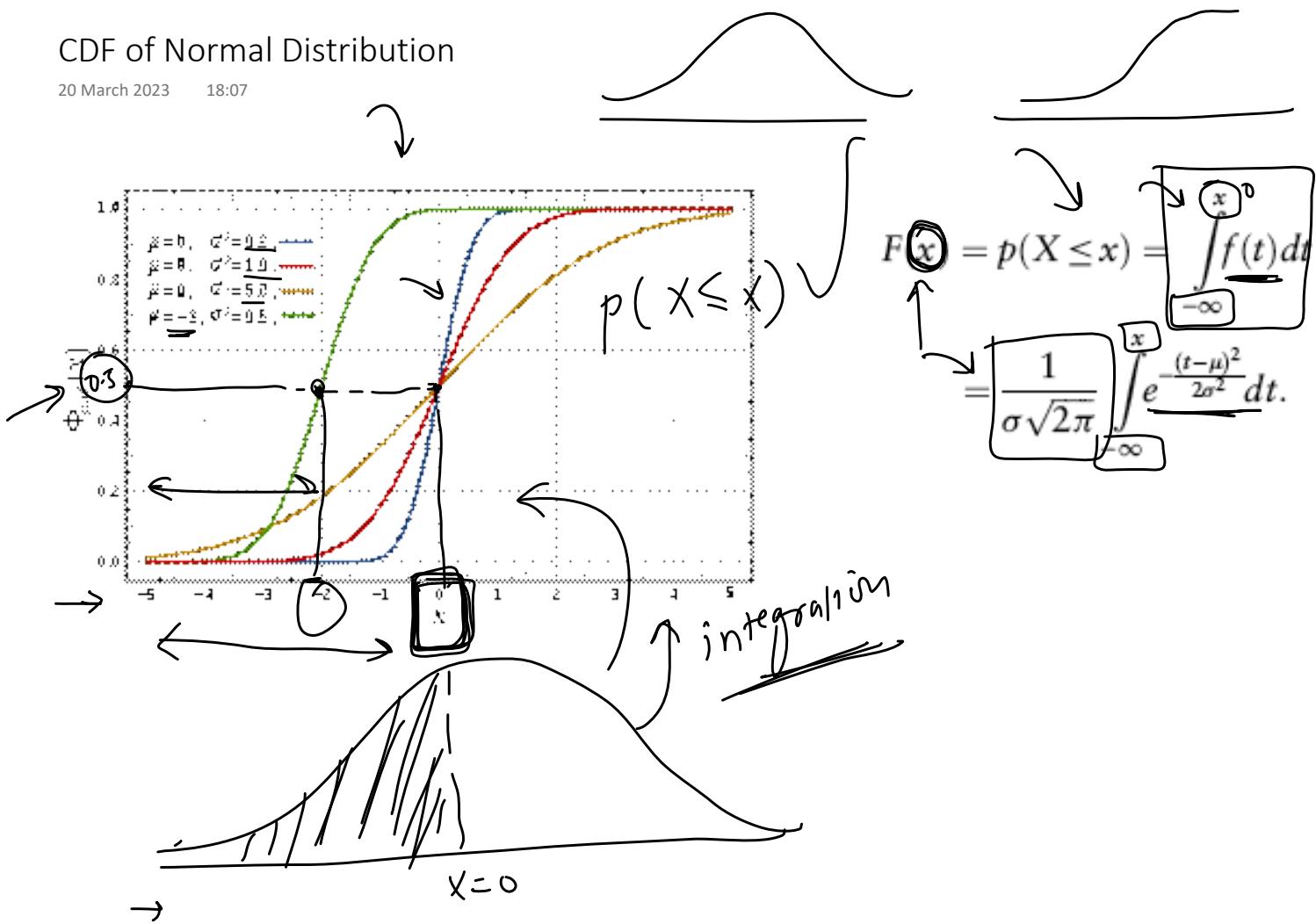
$$\rightarrow \frac{n}{(n-1)(n-2)} \sum \left(\frac{(x - \bar{x})}{s} \right)^3$$

- Python Example
- Interpretation



CDF of Normal Distribution

20 March 2023 18:07



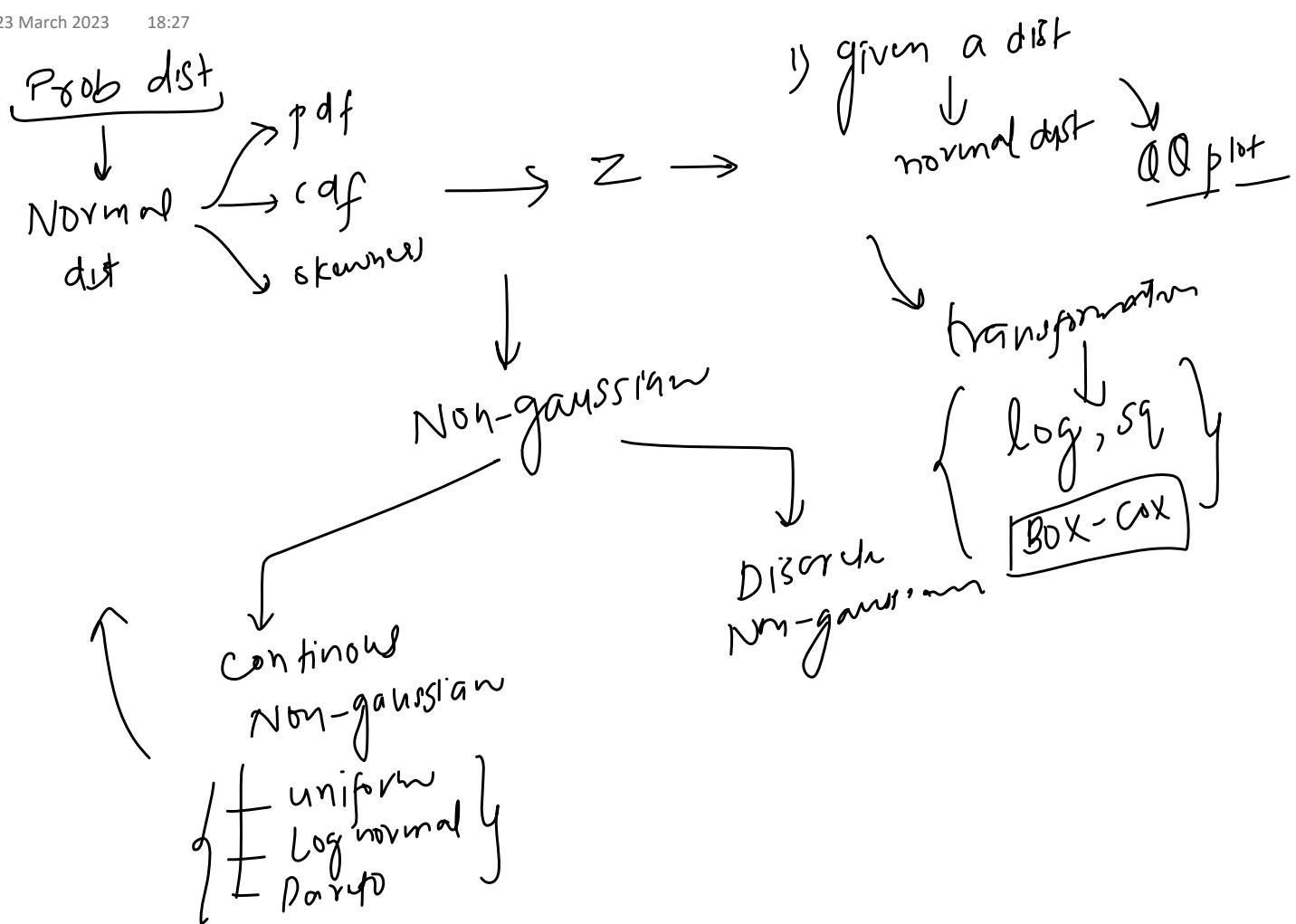
Use in Data Science

20 March 2023 18:08

- Outlier detection
- Assumptions on data for ML algorithms -> Linear Regression and GMM
- Hypothesis Testing
- Central Limit Theorem

Recap

23 March 2023 18:27

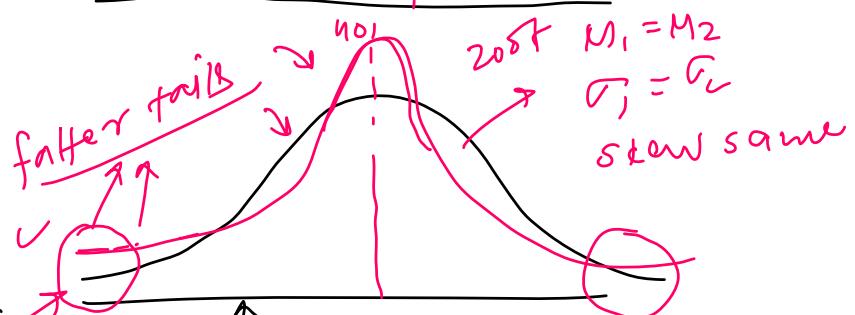
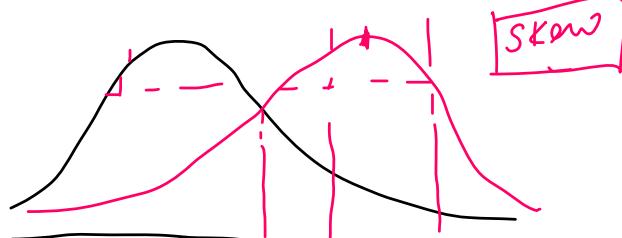
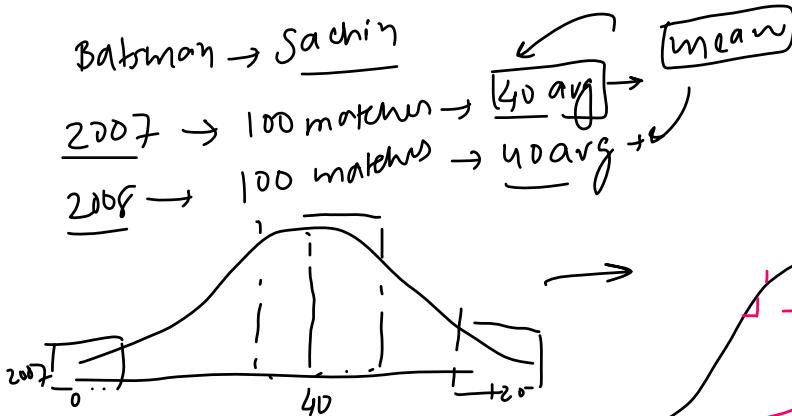
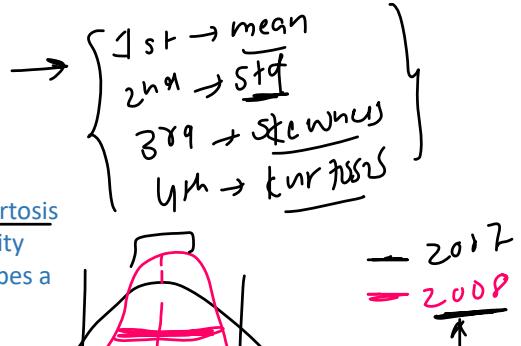


Kurtosis

23 March 2023 13:18

- What is Kurtosis?

{ Kurtosis is the 4th statistical moment. In probability theory and statistics, kurtosis (meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable. Like skewness, kurtosis describes a particular aspect of a probability distribution.



- False notation about Kurtosis

<https://en.wikipedia.org/wiki/Kurtosis>

- Formula

sample_kurtosis

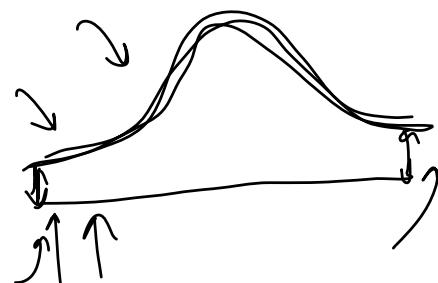
$$\left\{ \frac{n * (n+1)}{(n-1) * (n-2) * (n-3)} * \sum_i^n \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3 * (n-1)^2}{(n-2) * (n-3)}$$



- Practical Use-case

In finance, kurtosis risk refers to the risk associated with the possibility of extreme outcomes or "fat tails" in the distribution of returns of a particular asset or portfolio.

If a distribution has high kurtosis, it means that there is a higher likelihood of extreme events occurring, either positive or negative, compared to a normal distribution.



In finance, kurtosis risk is important to consider because it indicates that there is a greater probability of large losses or gains occurring, which can have significant implications for investors. As a result, investors may want to adjust their investment strategies to account for kurtosis risk.

- Excess Kurtosis & Types

Excess kurtosis is a measure of how much more peaked or flat a distribution is

Excess Kurtosis & Types

Excess kurtosis is a measure of how much more peaked or flat a distribution is compared to a normal distribution, which is considered to have a kurtosis of 0. It is calculated by subtracting 3 from the sample kurtosis coefficient.

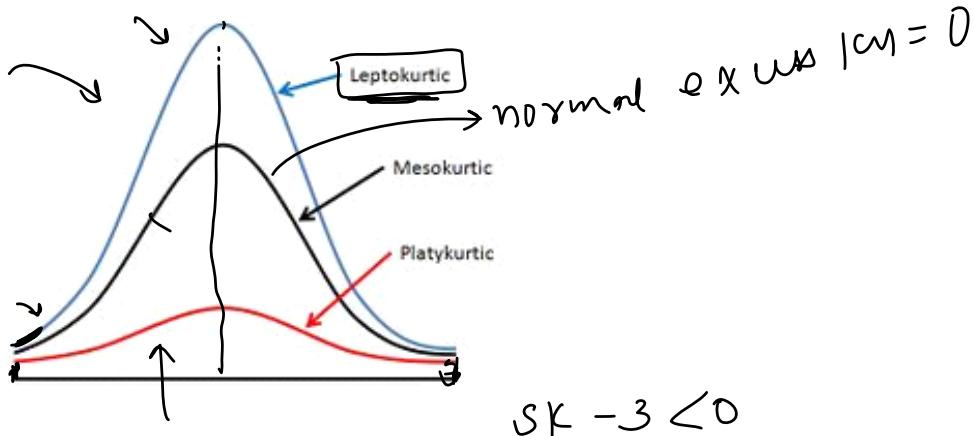
Types of Kurtosis

$$SK - 3 > 0$$

Leptokurtic

A distribution with positive excess kurtosis is called leptokurtic. "Lepto-" means "slender". In terms of shape, a leptokurtic distribution has fatter tails. This indicates that there are more extreme values or outliers in the distribution.

Example - Assets with positive excess kurtosis are riskier and more volatile than those with a normal distribution, and they may experience sudden price movements that can result in significant gains or losses.



Platykurtic

A distribution with negative excess kurtosis is called platykurtic. "Platy-" means "broad". In terms of shape, a platykurtic distribution has thinner tails. This indicates that there are fewer extreme values or outliers in the distribution.

Assets with negative excess kurtosis are less risky and less volatile than those with a normal distribution, and they may experience more gradual price movements that are less likely to result in large gains or losses.

Mesokurtic

$$(\mu, \sigma)$$

Distributions with zero excess kurtosis are called mesokurtic. The most prominent example of a mesokurtic distribution is the normal distribution family, regardless of the values of its parameters.

Mesokurtic is a term used to describe a distribution with a excess kurtosis of 0, indicating that it has the same degree of "peakedness" or "flatness" as a normal distribution.

Example -

In finance, a mesokurtic distribution is considered to be the ideal distribution for assets or portfolios, as it represents a balance between risk and return.

QQ Plot

23 March 2023 13:19

- How to find if a given distribution is normal or not?

o **Visual inspection**: One of the easiest ways to check for normality is to visually inspect a histogram or a density plot of the data. A normal distribution has a bell-shaped curve, which means that the majority of the data falls in the middle, and the tails taper off symmetrically. If the distribution looks approximately bell-shaped, it is likely to be normal.

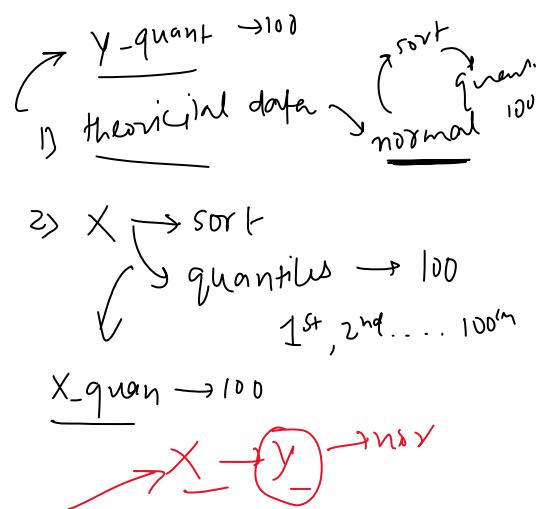
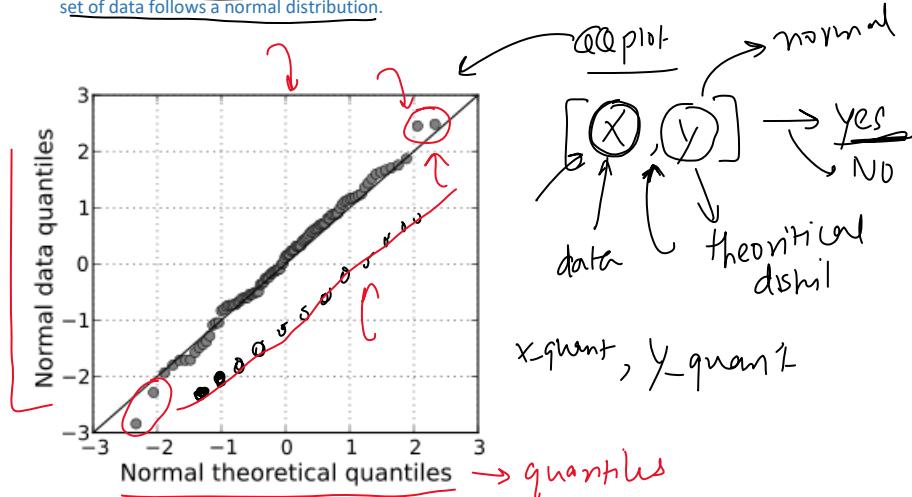


o **QQ Plot**: Another way to check for normality is to create a normal probability plot (also known as a Q-Q plot) of the data. A normal probability plot plots the observed data against the expected values of a normal distribution. If the data points fall along a straight line, the distribution is likely to be normal.

o **Statistical tests**: There are several statistical tests that can be used to test for normality, such as the Shapiro-Wilk test, the Anderson-Darling test, and the Kolmogorov-Smirnov test. These tests compare the observed data to the expected values of a normal distribution and provide a p-value that indicates whether the data is likely to be normal or not. A p-value less than the significance level (usually 0.05) suggests that the data is not normal.

- What is a QQ Plot and how is it plotted?

A QQ plot (quantile-quantile plot) is a graphical tool used to assess the similarity of the distribution of two sets of data. It is particularly useful for determining whether a set of data follows a normal distribution.

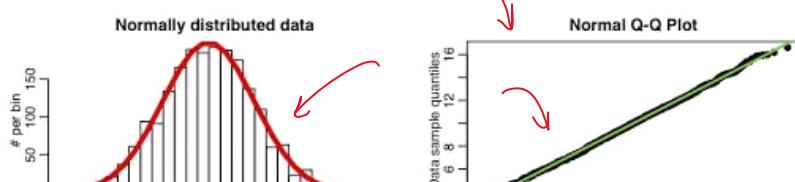


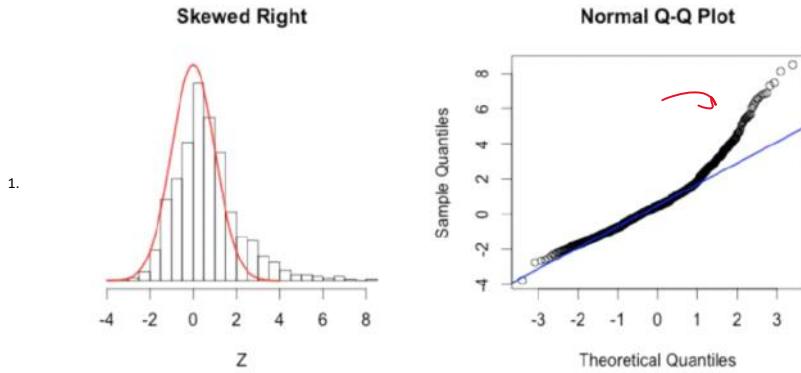
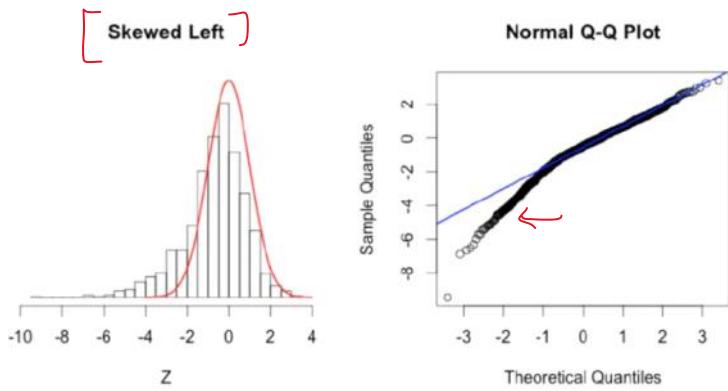
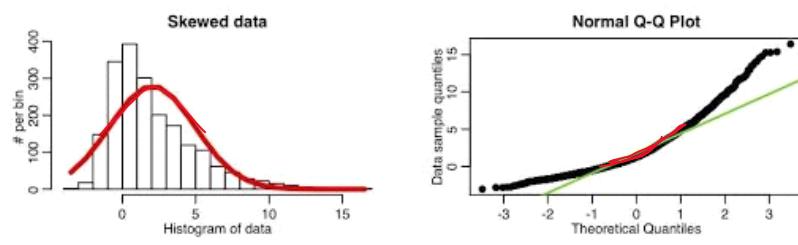
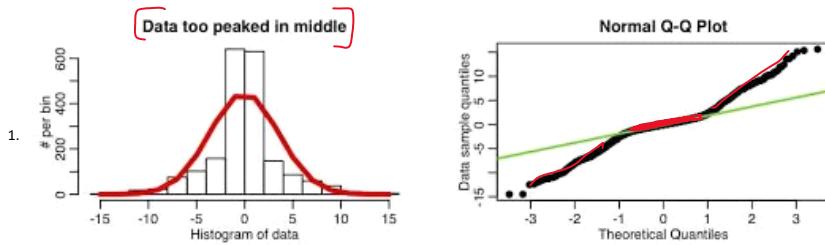
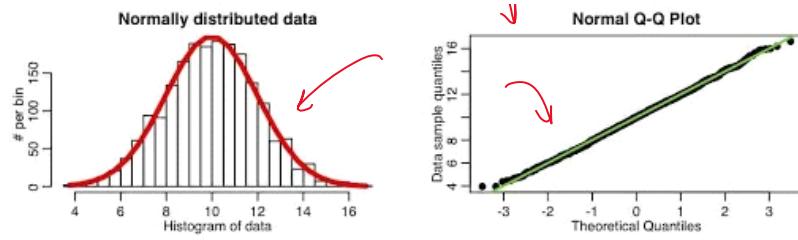
In a QQ plot, the quantiles of the two sets of data are plotted against each other. The quantiles of one set of data are plotted on the x-axis, while the quantiles of the other set of data are plotted on the y-axis. If the two sets of data have the same distribution, the points on the QQ plot will fall on a straight line. If the two sets of data do not have the same distribution, the points will deviate from the straight line.

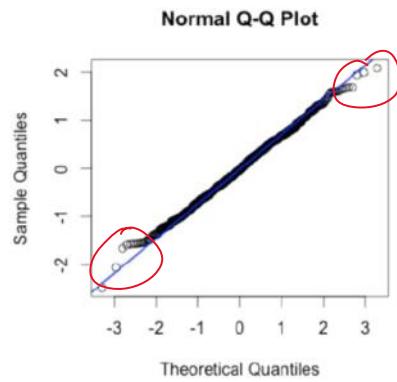
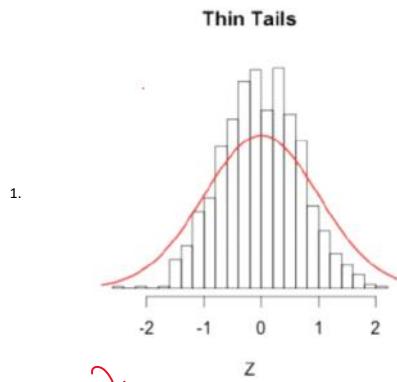
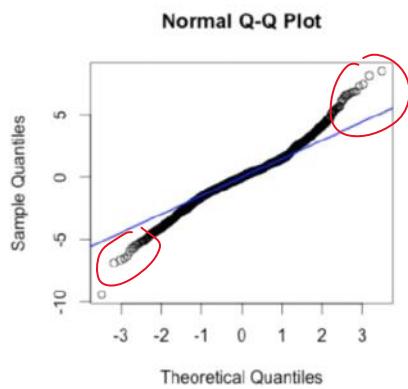
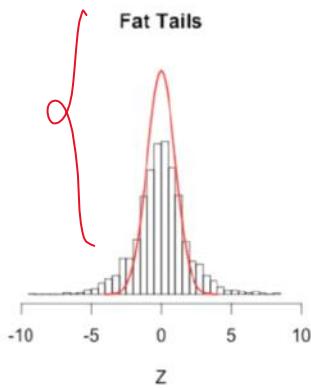
- Python example ✓

<https://www.statsmodels.org/dev/generated/statsmodels.graphics.gofplots.qqplot.html>

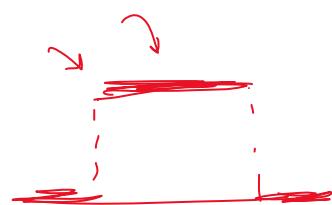
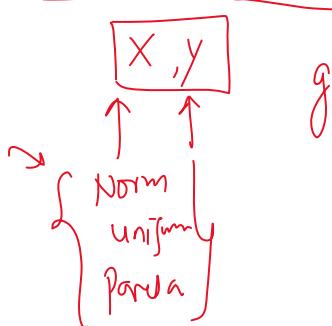
- How to interpret QQ plots







- Does QQ plot only detect normal distribution?



Uniform Distribution

23 March 2023 13:19

- What is Uniform Distribution and its types

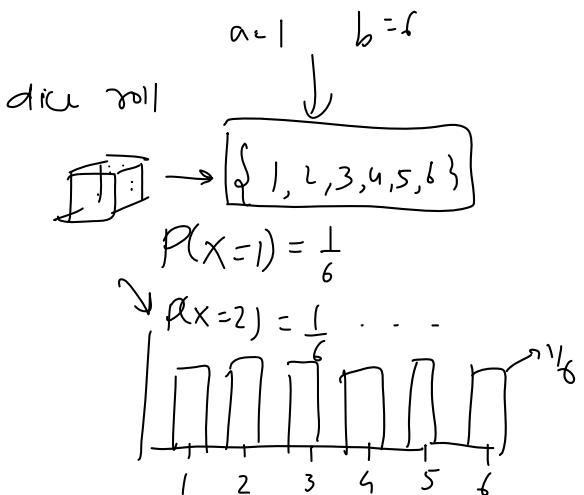
In probability theory and statistics, a uniform distribution is a probability distribution where all outcomes are equally likely within a given range. This means that if you were to select a random value from this range, any value would be as likely as any other value.

Types



Denoted as

$$X \sim U(a, b) \rightarrow \text{parameters } a \leftarrow \begin{matrix} \downarrow \\ \text{lower} \end{matrix} \quad b \leftarrow \begin{matrix} \uparrow \\ \text{upper} \end{matrix}$$

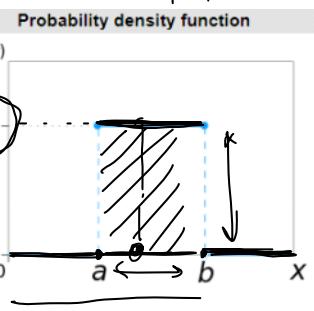


- Examples

- The height of a person randomly selected from a group of individuals whose heights range from 5'6" to 6'0" would follow a continuous uniform distribution.
- The time it takes for a machine to produce a product, where the production time ranges from 5 to 10 minutes, would follow a continuous uniform distribution.
- The distance that a randomly selected car travels on a tank of gas, where the distance ranges from 300 to 400 miles, would follow a continuous uniform distribution.
- The weight of a randomly selected apple from a basket of apples that weighs between 100 and 200 grams, would follow a continuous uniform distribution.

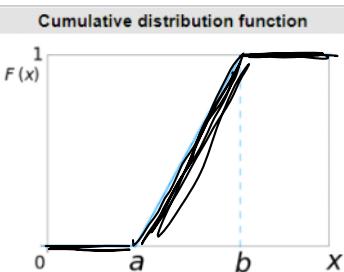
- PDF CDF and Graphs

PDF



$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

CDF

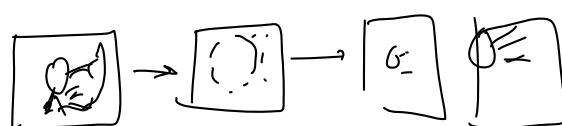


https://en.wikipedia.org/wiki/Continuous_uniform_distribution

- Skewness $\rightarrow 0$ \rightarrow Symmetric \rightarrow Normal

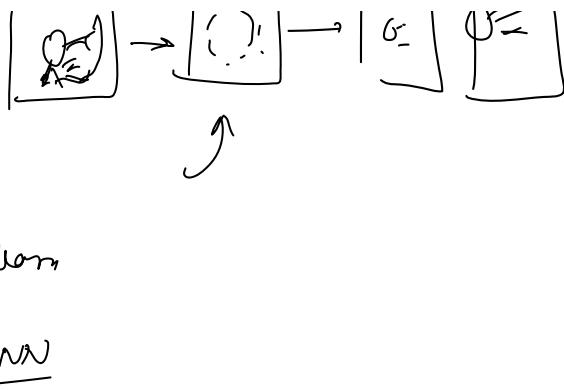
- Application in Machine learning and Data Science

- Random initialization: In many machine learning algorithms, such as neural networks and k-means clustering, the initial values of the parameters can have a significant impact on the final result. Uniform distribution is often used to randomly initialize the parameters, as it ensures that all values in the range have an equal probability of being selected.



In many machine learning algorithms, such as neural networks and k-means clustering, the initial values of the parameters can have a significant impact on the final result. Uniform distribution is often used to randomly initialize the parameters, as it ensures that all values in the range have an equal probability of being selected.

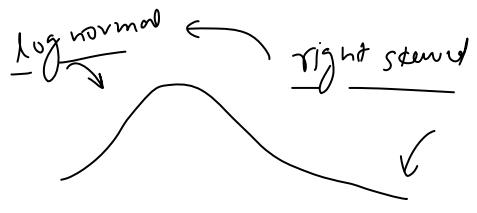
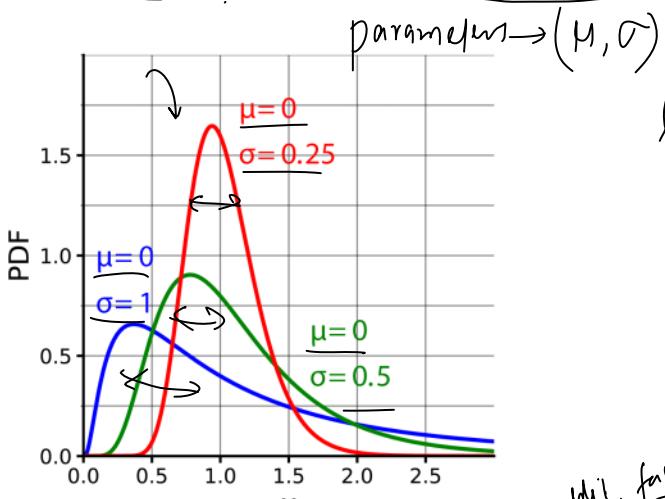
- b. **Sampling:** Uniform distribution can also be used for sampling. For example, if you have a dataset with an equal number of samples from each class, you can use uniform distribution to randomly select a subset of the data that is representative of all the classes.
- c. **Data augmentation:** In some cases, you may want to artificially increase the size of your dataset by generating new examples that are similar to the original data. Uniform distribution can be used to generate new data points that are within a specified range of the original data.
- d. **Hyperparameter tuning:** Uniform distribution can also be used in hyperparameter tuning, where you need to search for the best combination of hyperparameters for a machine learning model. By defining a uniform prior distribution for each hyperparameter, you can sample from the distribution to explore the hyperparameter space.



Log Normal Distribution

23 March 2023 13:19

In probability theory and statistics, a lognormal distribution is a heavy tailed continuous probability distribution of a random variable whose logarithm is normally distributed.



$$\log(x) \sim N(\mu, \sigma)$$

$$x \sim \text{log Normal}$$

$$\log(x) \sim N(\mu, \sigma) \quad \ln(27) \rightarrow -$$

$$\ln(28) \rightarrow -$$

$$\ln(29) \rightarrow -$$

$$\ln(30) \rightarrow -$$

$$\ln(31) \rightarrow -$$

$$\ln(32) \rightarrow -$$

$$\ln(33) \rightarrow -$$

$$\ln(34) \rightarrow -$$

$$\ln(35) \rightarrow -$$

$$\ln(36) \rightarrow -$$

$$\ln(37) \rightarrow -$$

$$\ln(38) \rightarrow -$$

$$\ln(39) \rightarrow -$$

$$\ln(40) \rightarrow -$$

$$\ln(41) \rightarrow -$$

$$\ln(42) \rightarrow -$$

$$\ln(43) \rightarrow -$$

$$\ln(44) \rightarrow -$$

$$\ln(45) \rightarrow -$$

$$\ln(46) \rightarrow -$$

$$\ln(47) \rightarrow -$$

$$\ln(48) \rightarrow -$$

$$\ln(49) \rightarrow -$$

$$\ln(50) \rightarrow -$$

$$\ln(51) \rightarrow -$$

$$\ln(52) \rightarrow -$$

$$\ln(53) \rightarrow -$$

$$\ln(54) \rightarrow -$$

$$\ln(55) \rightarrow -$$

$$\ln(56) \rightarrow -$$

$$\ln(57) \rightarrow -$$

$$\ln(58) \rightarrow -$$

$$\ln(59) \rightarrow -$$

$$\ln(60) \rightarrow -$$

$$\ln(61) \rightarrow -$$

$$\ln(62) \rightarrow -$$

$$\ln(63) \rightarrow -$$

$$\ln(64) \rightarrow -$$

$$\ln(65) \rightarrow -$$

$$\ln(66) \rightarrow -$$

$$\ln(67) \rightarrow -$$

$$\ln(68) \rightarrow -$$

$$\ln(69) \rightarrow -$$

$$\ln(70) \rightarrow -$$

$$\ln(71) \rightarrow -$$

$$\ln(72) \rightarrow -$$

$$\ln(73) \rightarrow -$$

$$\ln(74) \rightarrow -$$

$$\ln(75) \rightarrow -$$

$$\ln(76) \rightarrow -$$

$$\ln(77) \rightarrow -$$

$$\ln(78) \rightarrow -$$

$$\ln(79) \rightarrow -$$

$$\ln(80) \rightarrow -$$

$$\ln(81) \rightarrow -$$

$$\ln(82) \rightarrow -$$

$$\ln(83) \rightarrow -$$

$$\ln(84) \rightarrow -$$

$$\ln(85) \rightarrow -$$

$$\ln(86) \rightarrow -$$

$$\ln(87) \rightarrow -$$

$$\ln(88) \rightarrow -$$

$$\ln(89) \rightarrow -$$

$$\ln(90) \rightarrow -$$

$$\ln(91) \rightarrow -$$

$$\ln(92) \rightarrow -$$

$$\ln(93) \rightarrow -$$

$$\ln(94) \rightarrow -$$

$$\ln(95) \rightarrow -$$

$$\ln(96) \rightarrow -$$

$$\ln(97) \rightarrow -$$

$$\ln(98) \rightarrow -$$

$$\ln(99) \rightarrow -$$

$$\ln(100) \rightarrow -$$

$$\ln(101) \rightarrow -$$

$$\ln(102) \rightarrow -$$

$$\ln(103) \rightarrow -$$

$$\ln(104) \rightarrow -$$

$$\ln(105) \rightarrow -$$

$$\ln(106) \rightarrow -$$

$$\ln(107) \rightarrow -$$

$$\ln(108) \rightarrow -$$

$$\ln(109) \rightarrow -$$

$$\ln(110) \rightarrow -$$

$$\ln(111) \rightarrow -$$

$$\ln(112) \rightarrow -$$

$$\ln(113) \rightarrow -$$

$$\ln(114) \rightarrow -$$

$$\ln(115) \rightarrow -$$

$$\ln(116) \rightarrow -$$

$$\ln(117) \rightarrow -$$

$$\ln(118) \rightarrow -$$

$$\ln(119) \rightarrow -$$

$$\ln(120) \rightarrow -$$

$$\ln(121) \rightarrow -$$

$$\ln(122) \rightarrow -$$

$$\ln(123) \rightarrow -$$

$$\ln(124) \rightarrow -$$

$$\ln(125) \rightarrow -$$

$$\ln(126) \rightarrow -$$

$$\ln(127) \rightarrow -$$

$$\ln(128) \rightarrow -$$

$$\ln(129) \rightarrow -$$

$$\ln(130) \rightarrow -$$

$$\ln(131) \rightarrow -$$

$$\ln(132) \rightarrow -$$

$$\ln(133) \rightarrow -$$

$$\ln(134) \rightarrow -$$

$$\ln(135) \rightarrow -$$

$$\ln(136) \rightarrow -$$

$$\ln(137) \rightarrow -$$

$$\ln(138) \rightarrow -$$

$$\ln(139) \rightarrow -$$

$$\ln(140) \rightarrow -$$

$$\ln(141) \rightarrow -$$

$$\ln(142) \rightarrow -$$

$$\ln(143) \rightarrow -$$

$$\ln(144) \rightarrow -$$

$$\ln(145) \rightarrow -$$

$$\ln(146) \rightarrow -$$

$$\ln(147) \rightarrow -$$

$$\ln(148) \rightarrow -$$

$$\ln(149) \rightarrow -$$

$$\ln(150) \rightarrow -$$

$$\ln(151) \rightarrow -$$

$$\ln(152) \rightarrow -$$

$$\ln(153) \rightarrow -$$

$$\ln(154) \rightarrow -$$

$$\ln(155) \rightarrow -$$

$$\ln(156) \rightarrow -$$

$$\ln(157) \rightarrow -$$

$$\ln(158) \rightarrow -$$

$$\ln(159) \rightarrow -$$

$$\ln(160) \rightarrow -$$

$$\ln(161) \rightarrow -$$

$$\ln(162) \rightarrow -$$

$$\ln(163) \rightarrow -$$

$$\ln(164) \rightarrow -$$

$$\ln(165) \rightarrow -$$

$$\ln(166) \rightarrow -$$

$$\ln(167) \rightarrow -$$

$$\ln(168) \rightarrow -$$

$$\ln(169) \rightarrow -$$

$$\ln(170) \rightarrow -$$

$$\ln(171) \rightarrow -$$

$$\ln(172) \rightarrow -$$

$$\ln(173) \rightarrow -$$

$$\ln(174) \rightarrow -$$

$$\ln(175) \rightarrow -$$

$$\ln(176) \rightarrow -$$

$$\ln(177) \rightarrow -$$

$$\ln(178) \rightarrow -$$

$$\ln(179) \rightarrow -$$

$$\ln(180) \rightarrow -$$

$$\ln(181) \rightarrow -$$

$$\ln(182) \rightarrow -$$

$$\ln(183) \rightarrow -$$

$$\ln(184) \rightarrow -$$

$$\ln(185) \rightarrow -$$

$$\ln(186) \rightarrow -$$

$$\ln(187) \rightarrow -$$

$$\ln(188) \rightarrow -$$

$$\ln(189) \rightarrow -$$

$$\ln(190) \rightarrow -$$

$$\ln(191) \rightarrow -$$

$$\ln(192) \rightarrow -$$

$$\ln(193) \rightarrow -$$

$$\ln(194) \rightarrow -$$

$$\ln(195) \rightarrow -$$

$$\ln(196) \rightarrow -$$

$$\ln(197) \rightarrow -$$

$$\ln(198) \rightarrow -$$

$$\ln(199) \rightarrow -$$

$$\ln(200) \rightarrow -$$

$$\ln(201) \rightarrow -$$

$$\ln(202) \rightarrow -$$

$$\ln(203) \rightarrow -$$

$$\ln(204) \rightarrow -$$

$$\ln(205) \rightarrow -$$

$$\ln(206) \rightarrow -$$

$$\ln(207) \rightarrow -$$

$$\ln(208) \rightarrow -$$

$$\ln(209) \rightarrow -$$

$$\ln(210) \rightarrow -$$

$$\ln(211) \rightarrow -$$

$$\ln(212) \rightarrow -$$

$$\ln(213) \rightarrow -$$

$$\ln(214) \rightarrow -$$

$$\ln(215) \rightarrow -$$

$$\ln(216) \rightarrow -$$

$$\ln(217) \rightarrow -$$

$$\ln(218) \rightarrow -$$

$$\ln(219) \rightarrow -$$

$$\ln(220) \rightarrow -$$

$$\ln(221) \rightarrow -$$

$$\ln(222) \rightarrow -$$

$$\ln(223) \rightarrow -$$

$$\ln(224) \rightarrow -$$

$$\ln(225) \rightarrow -$$

$$\ln(226) \rightarrow -$$

$$\ln(227) \rightarrow -$$

$$\ln(228) \rightarrow -$$

$$\ln(229) \rightarrow -$$

$$\ln(230) \rightarrow -$$

$$\ln(231) \rightarrow -$$

$$\ln(232) \rightarrow -$$

$$\ln(233) \rightarrow -$$

$$\ln(234) \rightarrow -$$

$$\ln(235) \rightarrow -$$

$$\ln(236) \rightarrow -$$

$$\ln(237) \rightarrow -$$

$$\ln(238) \rightarrow -$$

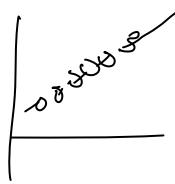
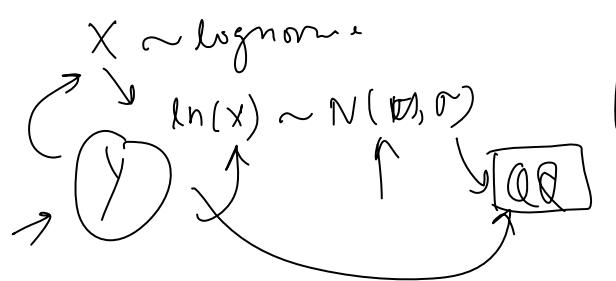
$$\ln(239) \rightarrow -$$

$$\ln(240) \rightarrow -$$

$$\ln(241) \rightarrow$$

skewness skewed

How to check if a random variable is log normally distributed?



Pareto Distribution

23 March 2023 13:19

Pareto Distribution

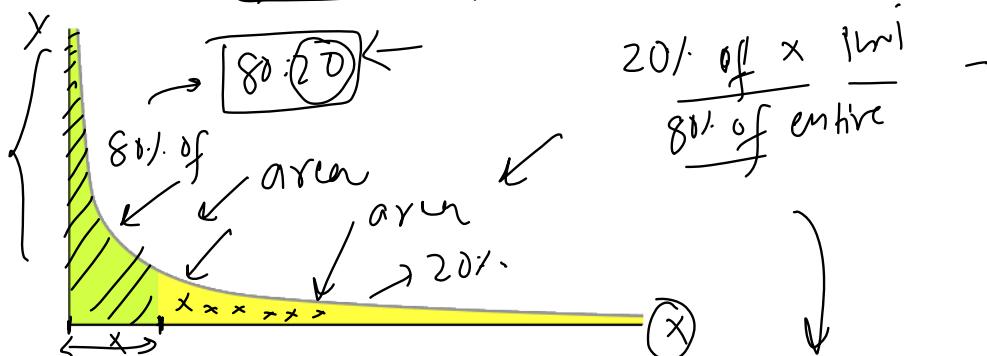
The Pareto distribution is a type of probability distribution that is commonly used to model the distribution of wealth, income, and other quantities that exhibit a similar power-law behaviour

What is Power Law

In mathematics, a power law is a functional relationship between two variables, where one variable is proportional to a power of the other. Specifically, if y and x are two variables related by a power law, then the relationship can be written as:

$$y = k * x^\alpha$$

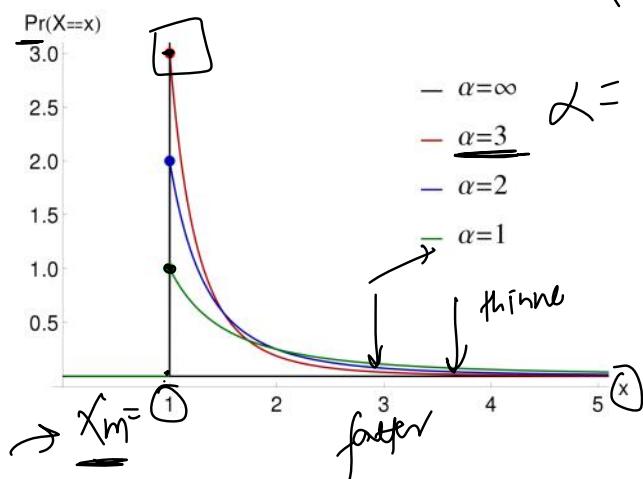
$$y = K x^\alpha$$



Vilfredo Pareto originally used this distribution to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. He also used it to describe distribution of income. This idea is sometimes expressed more simply as the Pareto principle or the "80-20 rule" which says that 20% of the population controls 80% of the wealth

↑ ↗ application ↘

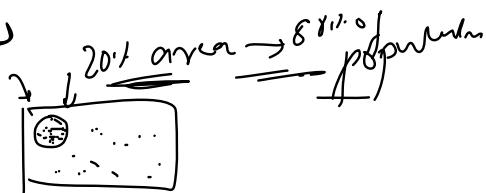
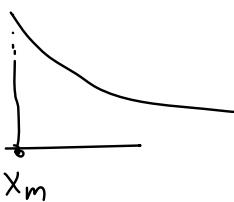
Graph & Parameters



$$X \sim \text{Pr}(x)$$

$$\alpha = 1.11$$

$$\text{pdf} = \frac{\alpha x^\alpha}{\alpha + 1} = y$$



80% of entire globe

20% of globe

20% of file

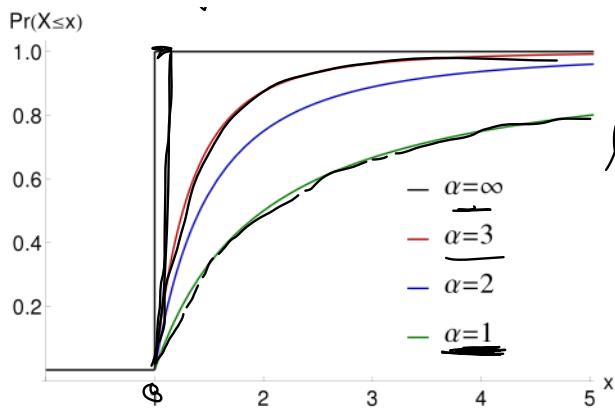
80%

Examples

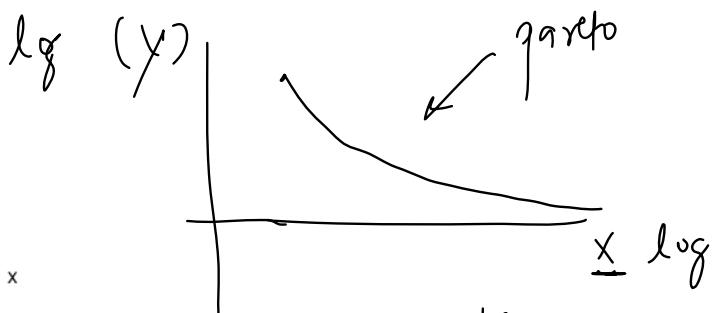
- The sizes of human settlements (few cities, many hamlets/villages)
- File size distribution of Internet traffic which uses the TCP protocol (many smaller files, few larger ones)

CDF

$$\text{Pr}(X < x)$$

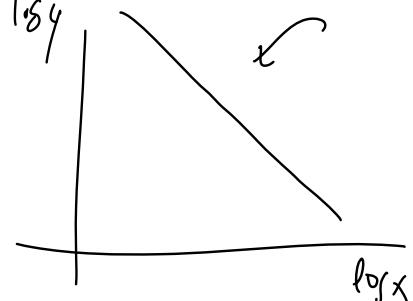


Skewness → skewed

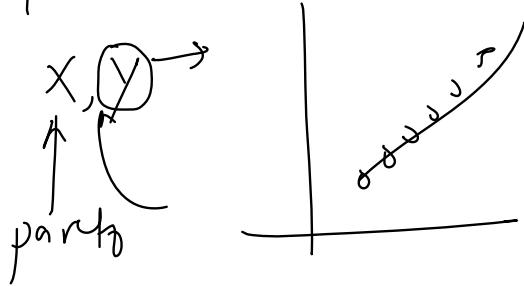


[How to detect if a distribution is Pareto Distribution?]

$$Y = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$$

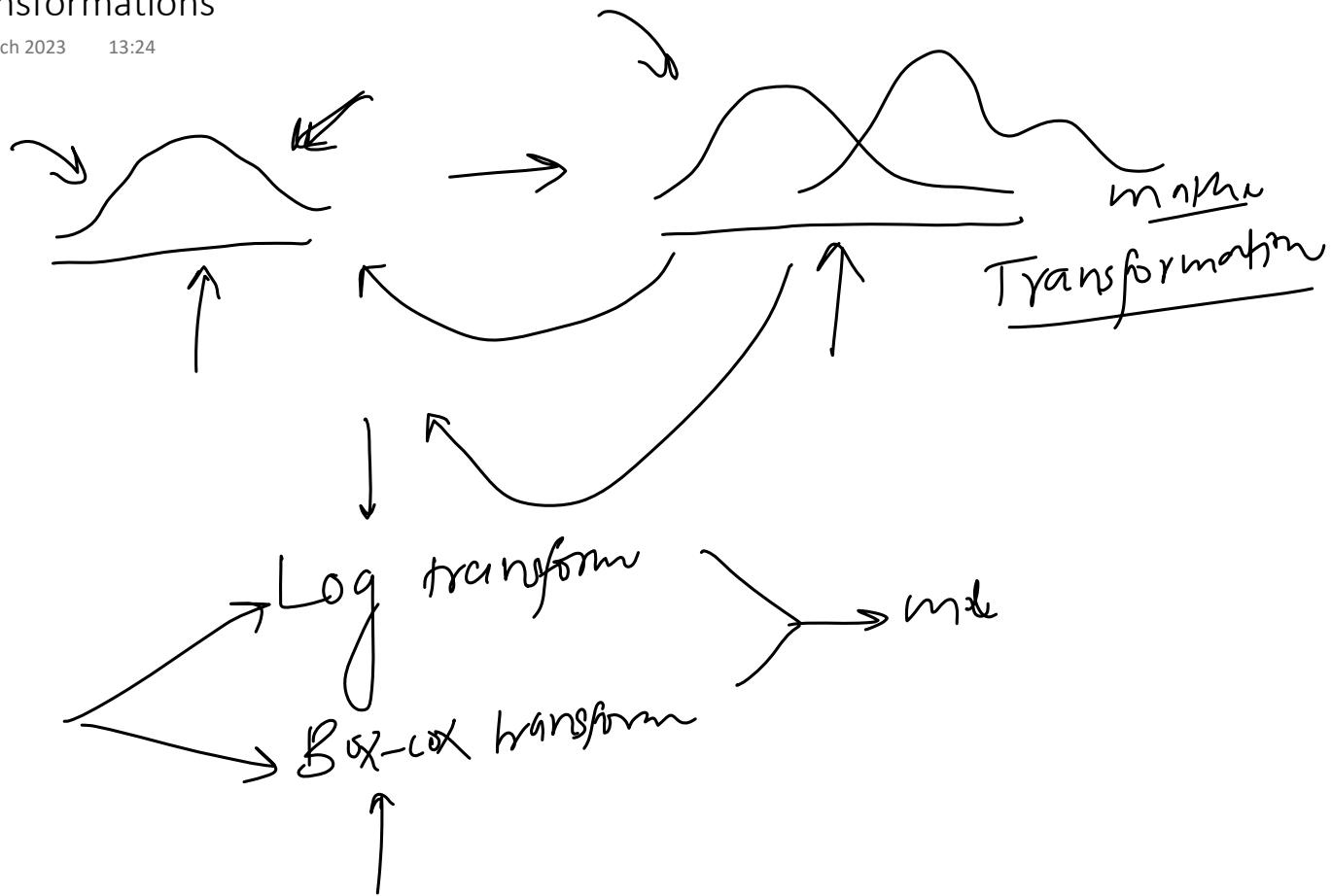


Q-Q plot



Transformations

23 March 2023 13:24



Views

24 March 2023 17:37



- What are Views?

In SQL, a view is a virtual table that does not store any data on its own but presents a customized view of one or more tables in a database. A view can be thought of as a pre-defined SELECT statement that retrieves data from one or more tables and returns a specific subset of data to the user.

So basically it is a logical table instead of a physical table

Once a view is created, it can be used in the same way as a table in SQL queries, and any changes made to the underlying tables will be reflected in the view. (Show)

Simple Views - Created from 1 single table

Complex Views - Created from multiple tables with the help of joins, subquery etc.

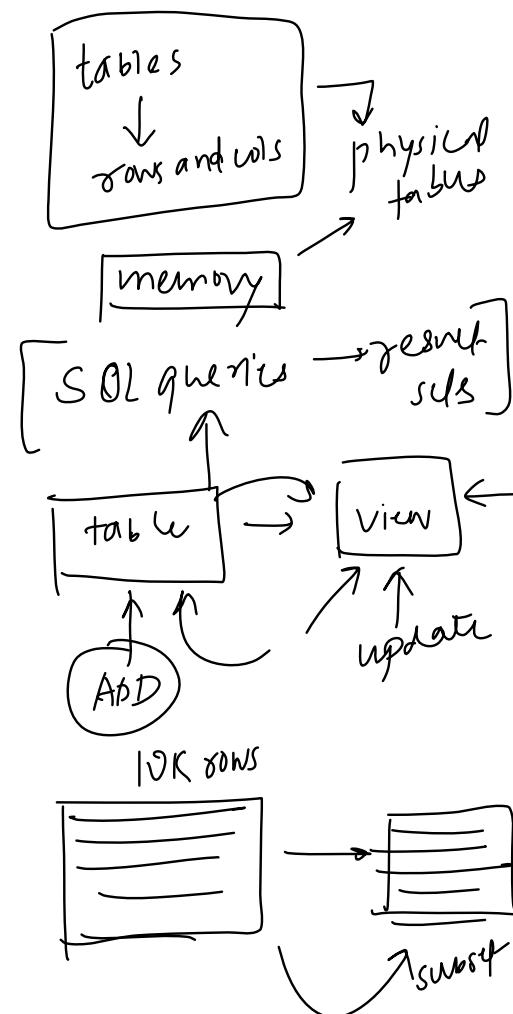
- Read only Vs Updatable Views

a. Read-only views: As the name suggests, read-only views are views that cannot be updated. They are used to simplify the process of querying data, but they cannot be used to modify or delete data in the underlying tables.

b. Updatable views: Updatable views are views that allow you to modify, insert or delete data in the underlying tables through the view. They behave like normal tables, but with restrictions.

To make a view updatable, certain conditions must be met. For example, the view must not contain any derived columns, subqueries, or aggregate functions.

Additionally, the view must be based on a single table or a join of tables with a unique one-to-one relationship.



- Materialized Views

A materialized view is a database object in SQL that contains the results of a query. Unlike regular views, which are just virtual tables that store SQL queries, materialized views are physical tables that store the results of a query. Materialized views are precomputed and stored on disk, which makes them much faster to access than regular views.

Benefit - Faster queries

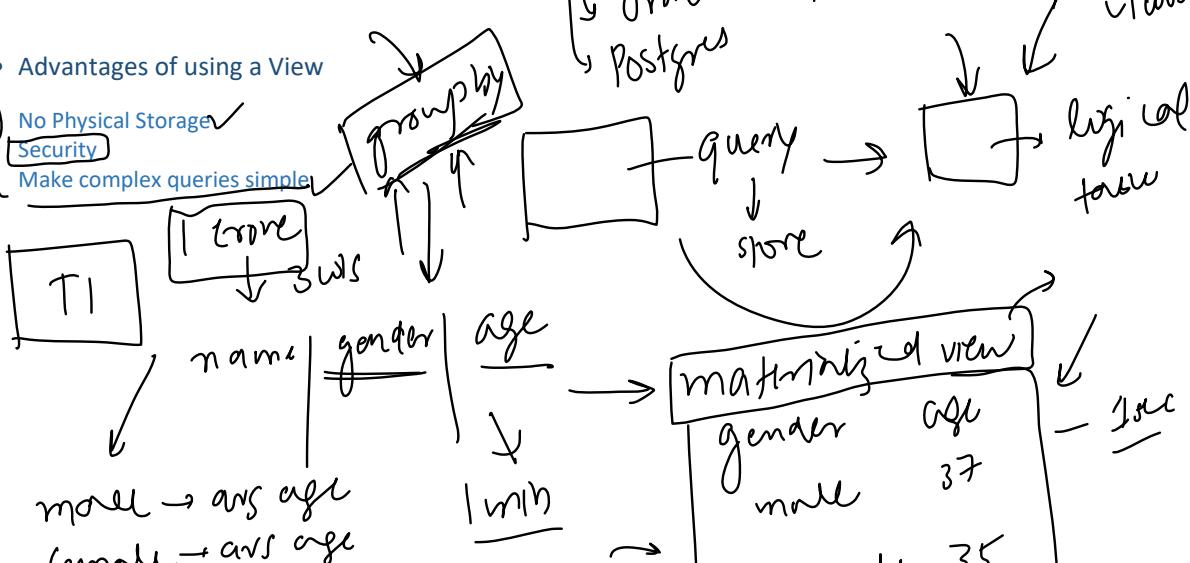
Disadvantage - Need to manually update the view

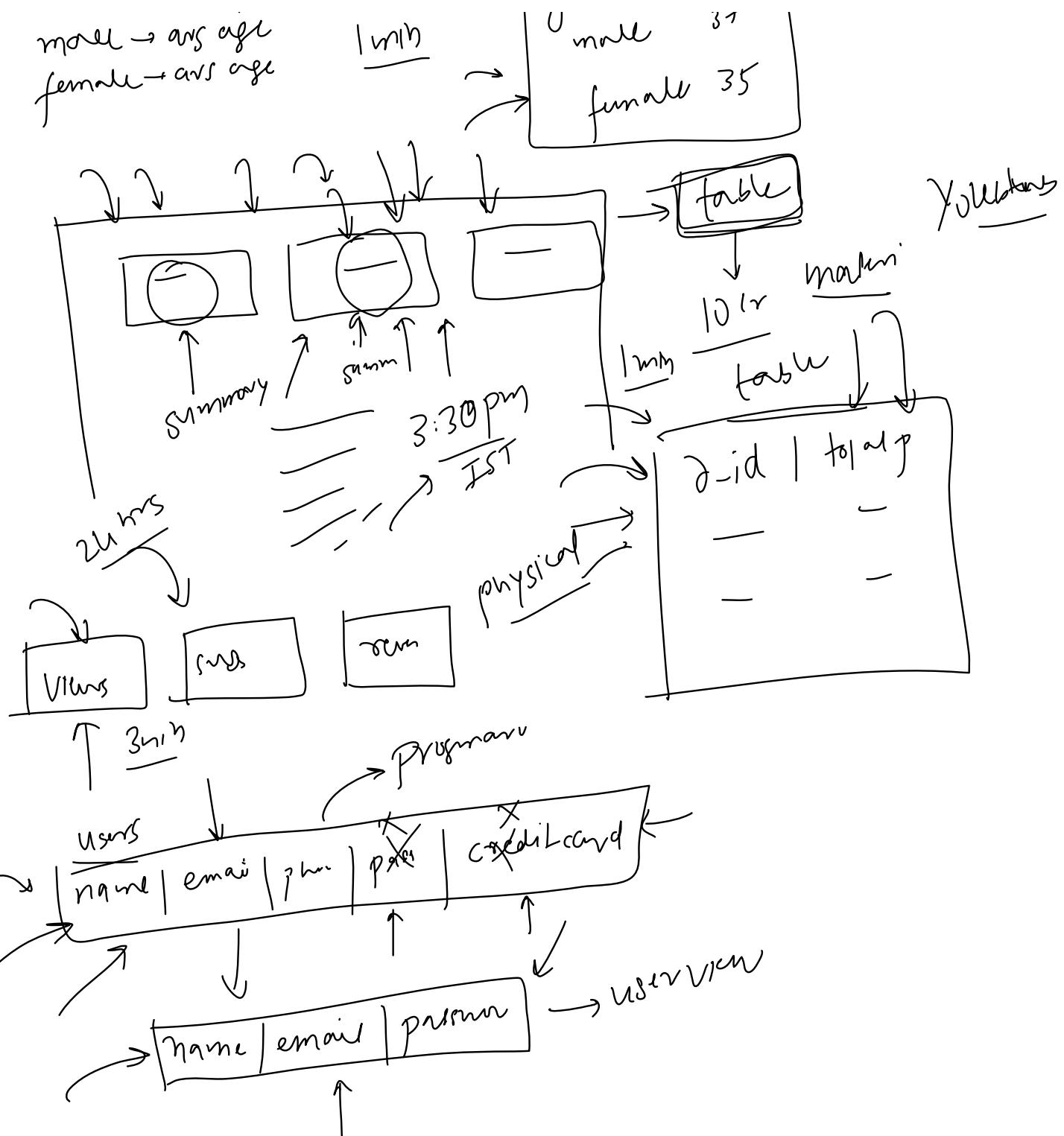
- Advantages of using a View

No Physical Storage

Security

Make complex queries simple





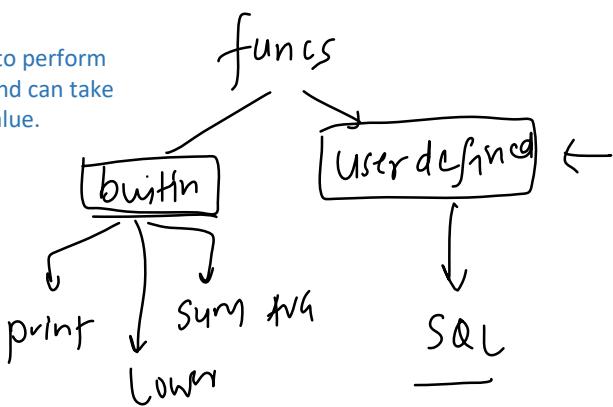
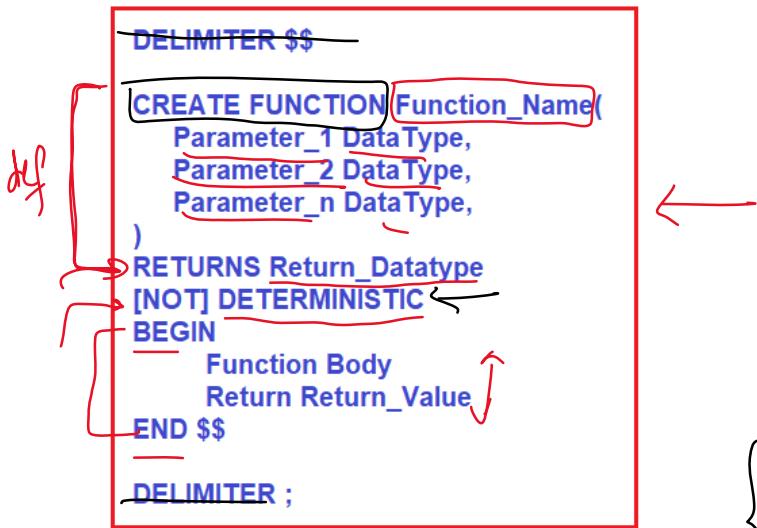
User Defined Functions

24 March 2023 18:58

- What are function and their advantages

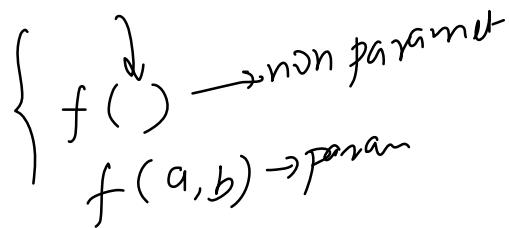
User-defined functions (UDFs) in SQL are functions that are created by users to perform specific tasks. These functions can be used just like built-in functions in SQL and can take parameters as input, perform some operations on them, and then return a value.

- Syntax

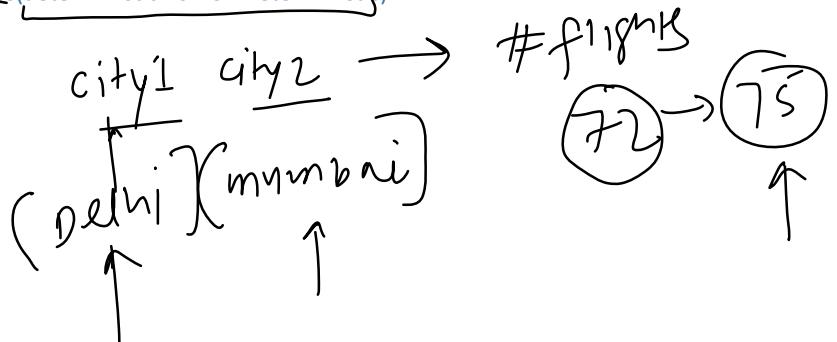
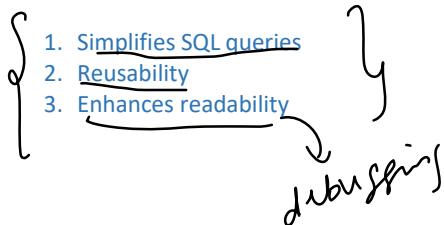


Examples

- hello world
- parameterized vs non parameterized
- Calculate age in years(for col)
- greet with name -> conditional title
- Date formatting and flights between 2 cities(deterministic Vs Non Deterministic)
- show all functions of a database
- Drop function



Benefits



Stored Procedures

25 March 2023 16:39

A stored procedure is a named block of SQL statements and procedural logic that is stored in a database and can be executed by a user or application.

Stored procedures are often used to encapsulate business logic and application logic, such as data validation, data processing, and database updates. By using stored procedures, developers can separate application logic from the presentation layer and simplify the application code.

- Create hello world stored procedure
- Create stored procedure to create a new user
- Show error message
- Create stored procedure to show orders placed by 1 single user
- Create a stored procedure to place an order

Some of the benefits of using stored procedures include:

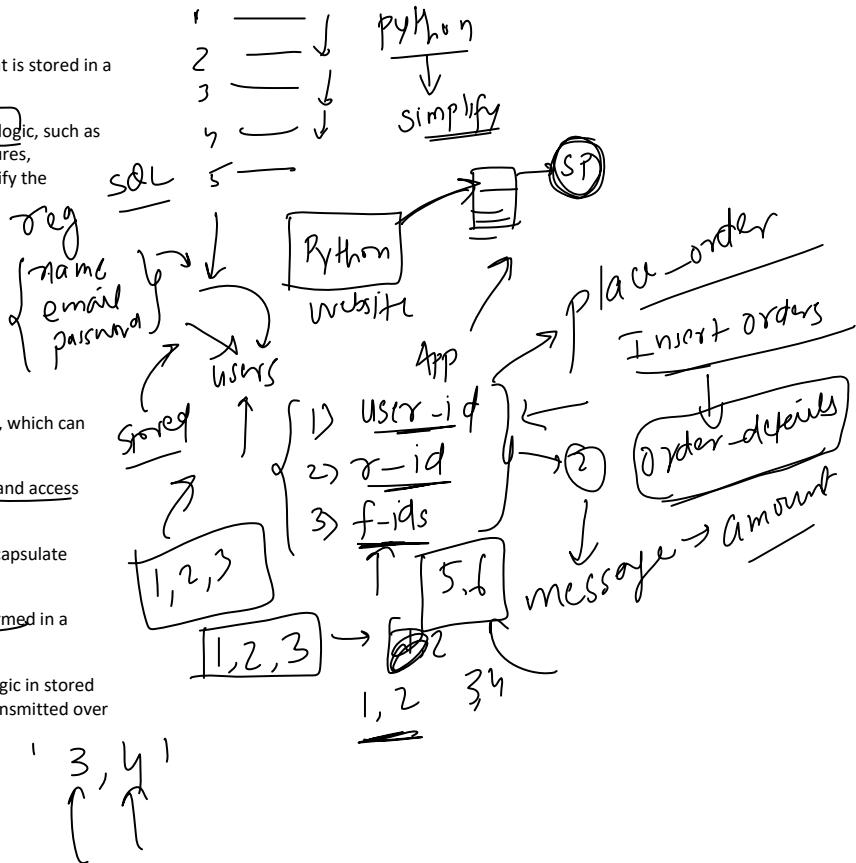
1. **Improved performance:** Stored procedures are precompiled and optimized, which can improve performance and reduce network traffic.
2. **Enhanced security:** Stored procedures can be granted specific permissions and access rights, which can improve security and limit access to sensitive data.
3. **Encapsulation of business logic:** Stored procedures allow developers to encapsulate complex business logic and make it easier to maintain and update.
4. **Consistency:** Stored procedures ensure that database operations are performed in a consistent manner, which can help to maintain data integrity.
5. **Reduced network traffic:** By encapsulating data access and manipulation logic in stored procedures, developers can reduce the amount of data that needs to be transmitted over the network.



'1,2'

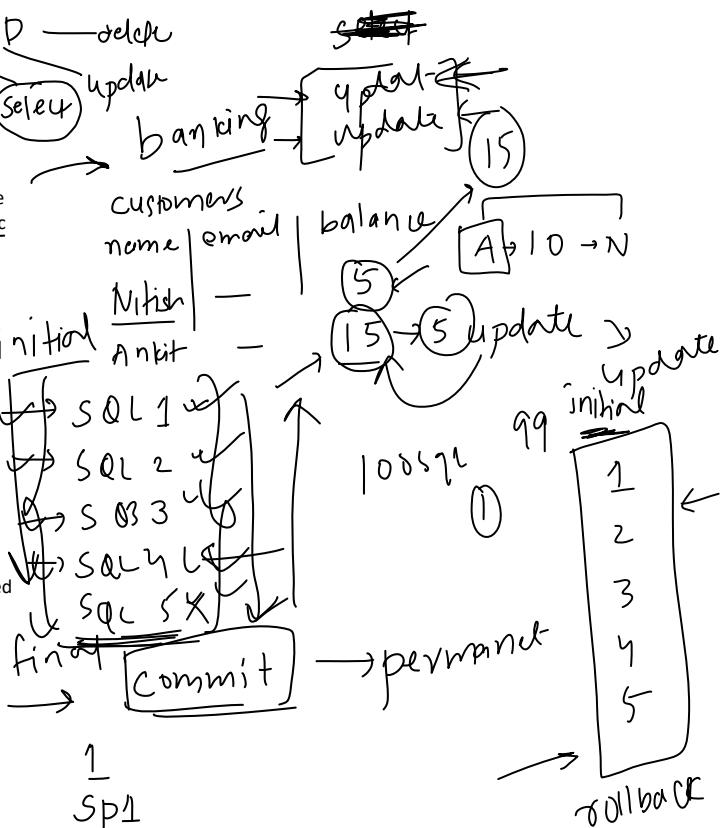
'1,2'

'3,4'



trans \rightarrow card(s)
Write (I, U, D)

CRUD
insert
Select
update
delete



What are Transactions?

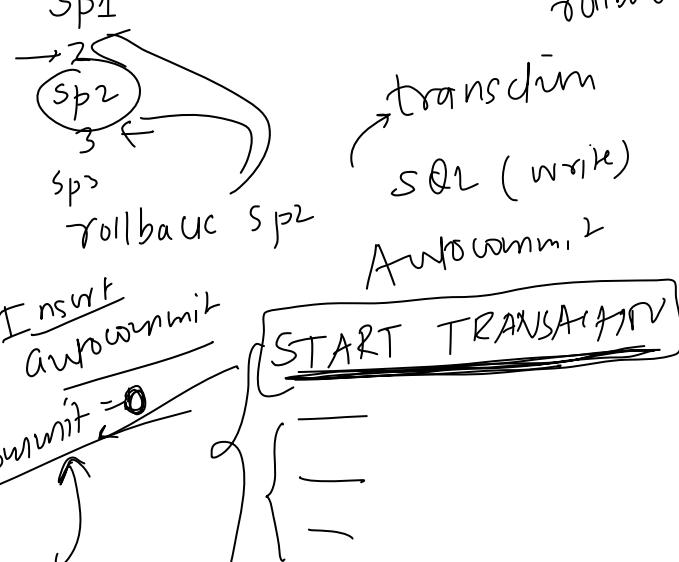
- A database transaction is a sequence of operations that are performed as a single logical unit of work in a database management system (DBMS). A transaction may consist of one or more database operations, such as inserts, updates, or deletes, which are treated as a single atomic operation by the DBMS.

It follows the principle of all or none.

What is Commit, Rollback and Savepoint?

In a database transaction, there are three main commands that are used to manage the transaction:

- Commit:** A commit command is used to permanently save the changes made by a transaction to the database. When a transaction is committed, all changes made by the transaction are made permanent and cannot be rolled back.
- Rollback:** A rollback command is used to undo the changes made by a transaction and return the database to its state before the transaction began. When a transaction is rolled back, all changes made by the transaction are discarded and the database is returned to its previous state.
- Savepoint:** A savepoint command is used to mark a specific point within a transaction where a rollback can be performed. This allows for partial rollbacks of a transaction, where only changes made after the savepoint are undone, while changes made before the savepoint are still committed to the database.



What is Autocommit?

Autocommit is a feature of database management systems (DBMS) that automatically commits each individual database transaction as soon as it is completed, rather than requiring an explicit commit command to be issued.

When Autocommit is enabled, each individual SQL statement issued against the database is treated as a separate transaction and is committed immediately after it is executed. This means that each SQL statement becomes a separate, independent transaction, and its effects are immediately visible to other users.

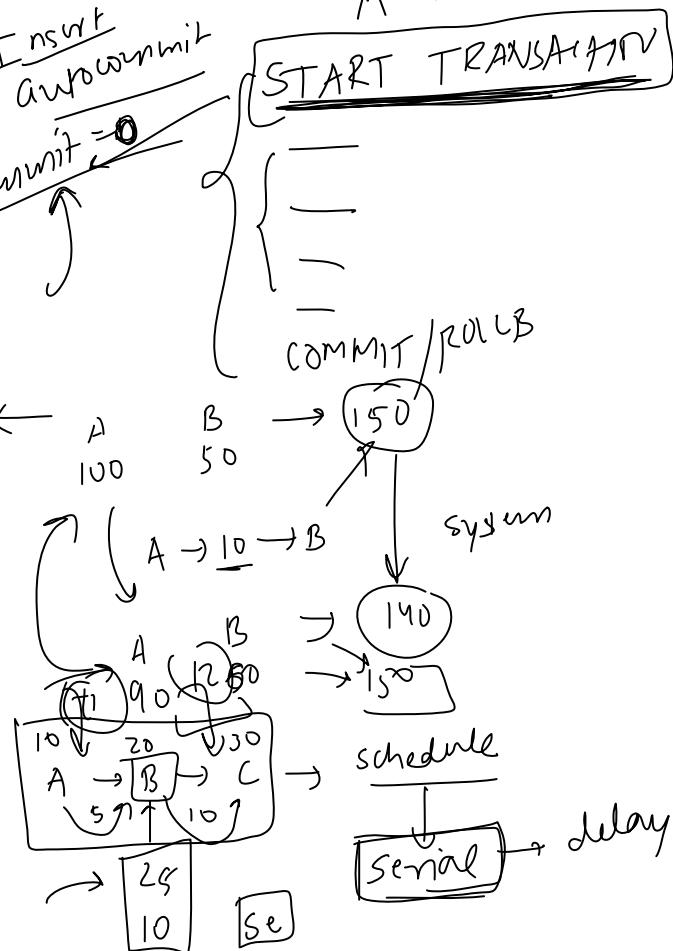
- Each SQL write statement is Autocommit (prove)
- set Autocommit = 0 and (show)
- START TRANSACTION -> show operations without committing
- START TRANSACTION -> with commit
- START TRANSACTION -> all or none with commit
- rollback
- rollback with savepoint
- rollback and commit together

What is ACID properties of a Transaction?

ACID is an acronym that stands for Atomicity, Consistency, Isolation, and Durability, which are a set of properties that ensure reliable database transactions:

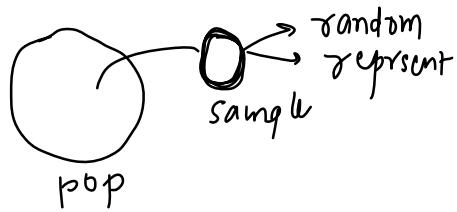
- Atomicity:** This property ensures that a transaction is treated as a single, indivisible unit of work. This means that either all of the changes made by a transaction are committed to the database, or none of them are. If any part of the transaction fails, the entire transaction is rolled back, and all changes are undone.
- Consistency:** This property ensures that a transaction takes the database from one valid state to another valid state. It requires that all data in the database must conform to a set of rules, or constraints, which ensure data integrity.
- Isolation:** This property ensures that concurrent transactions do not interfere with each other. It requires that each transaction executes as if it were the only transaction executing against the database, even if multiple transactions are executing at the same time.
- Durability:** This property ensures that once a transaction is committed, its changes are permanently stored in the database, even in the event of a system failure or power outage. This is typically achieved through the use of database backups, replication, or other forms of data redundancy.

Together, these properties ensure that database transactions are reliable, consistent, and accurate, and that the data stored in a database is both protected and available at all times. The ACID properties are essential for mission-critical applications that require high levels of data integrity and availability, such as banking, finance, and healthcare systems.



Some Terms

30 March 2023 07:09



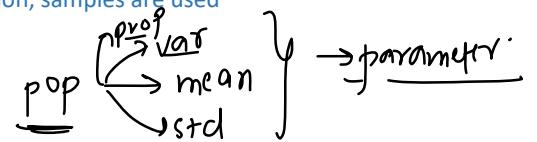
Population Vs Sample

Population: A population is the entire group or set of individuals, objects, or events that a researcher wants to study or draw conclusions about. It can be people, animals, plants, or even inanimate objects, depending on the context of the study. The population usually represents the complete set of possible data points or observations.

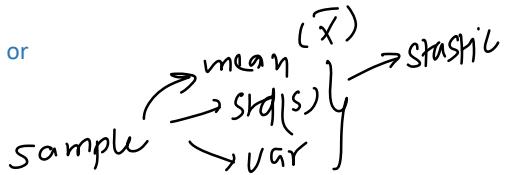
Sample: A sample is a subset of the population that is selected for study. It is a smaller group that is intended to be representative of the larger population. Researchers collect data from the sample and use it to make inferences about the population as a whole. Since it is often impractical or impossible to collect data from every member of a population, samples are used as an efficient and cost-effective way to gather information.

Parameter Vs Estimate

Parameter: A parameter is a numerical value that describes a characteristic of a population. Parameters are usually denoted using Greek letters, such as μ (mu) for the population mean or σ (sigma) for the population standard deviation. Since it is often difficult or impossible to obtain data from an entire population, parameters are usually unknown and must be estimated based on available sample data.



$$\bar{x} \rightarrow \mu$$



Statistic A statistic is a numerical value that describes a characteristic of a sample, which is a subset of the population. By using statistics calculated from a representative sample, researchers can make inferences about the unknown respective parameter of the population. Common statistics include the sample mean (denoted by \bar{x} , pronounced "x-bar"), the sample median, and the sample standard deviation (denoted by s).



Inferential Statistics

Inferential statistics is a branch of statistics that focuses on making predictions, estimations, or generalizations about a larger population based on a sample of data taken from that population. It involves the use of probability theory to make inferences and draw conclusions about the characteristics of a population by analysing a smaller subset or sample.

The key idea behind inferential statistics is that it is often impractical or impossible to collect data from every member of a population, so instead, we use a representative sample to make inferences about the entire group. Inferential statistical techniques include hypothesis testing, confidence intervals, and regression analysis, among others.

These methods help researchers answer questions like:

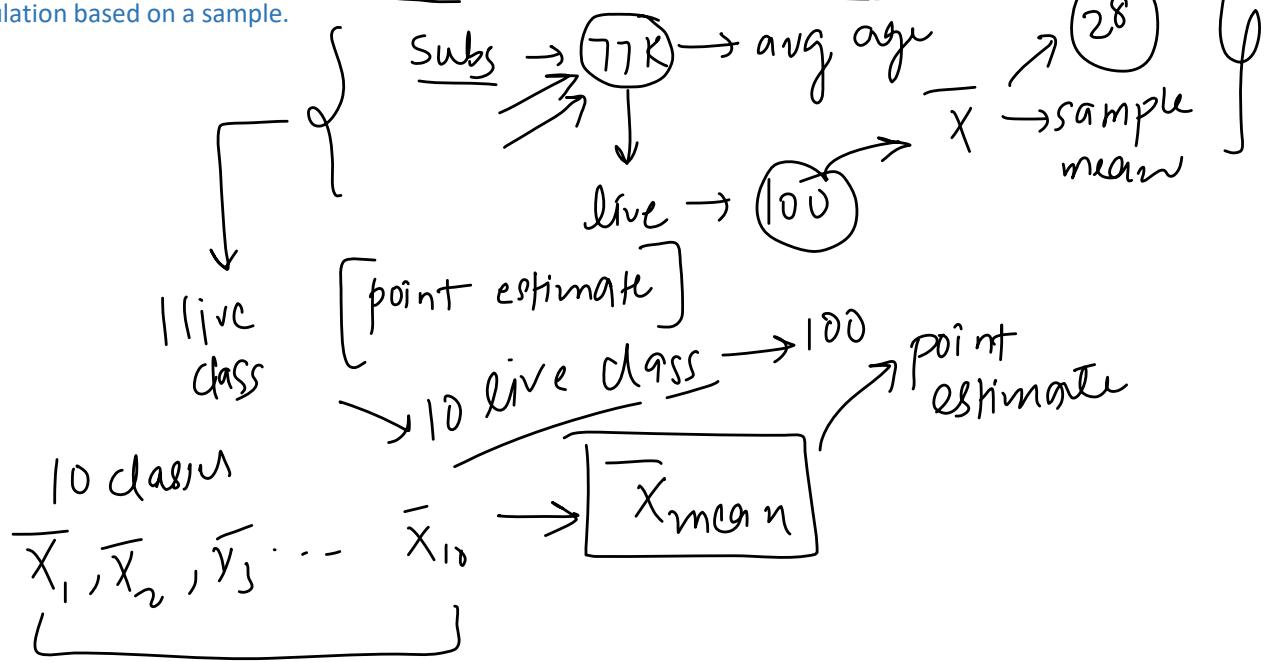
- a. Is there a significant difference between two groups?
- b. Can we predict the outcome of a variable based on the values of other variables?
- c. What is the relationship between two or more variables?

Inferential statistics are widely used in various fields, such as economics, social sciences, medicine, and natural sciences, to make informed decisions and guide policy based on limited data.

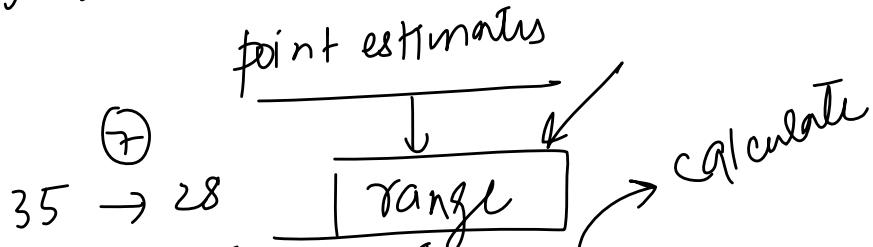
Point Estimate

30 March 2023 07:19

A point estimate is a single value calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.



28 → YT subs → avg age
yesterday → NO



MS Dhoni

- 1) exact → 25 → 25 → 1000000 (1cr)
- 2) ± 10 → 7500 (75 lac) ✓
- 3) ± 20 → 500 (50 lac) ✓

Confidence interval

Confidence Interval

30 March 2023 07:18

$$\mu \pm \sigma \rightarrow [25, 32] \leftarrow 95\% \text{ confident}$$

Confidence interval, in simple words, is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

Confidence level, usually expressed as a percentage like 95%, indicates how sure we are that the true value lies within the interval.

Confidence Interval = Point Estimate \pm Margin of Error
Ways to calculate CI:

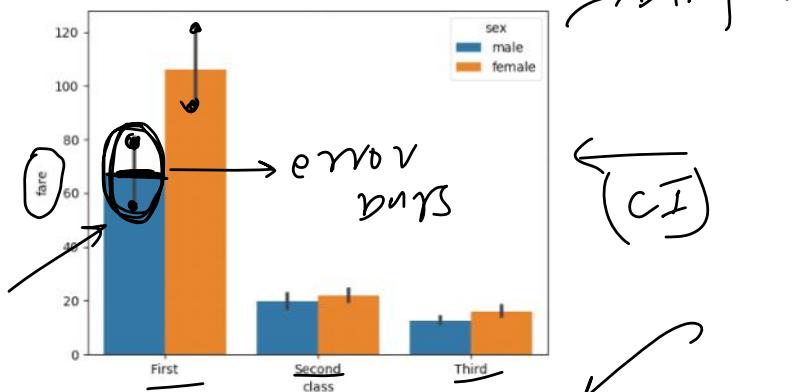
$$25 \pm 4 [21, 29]$$

Z procedure t procedure
 $\text{pop} \rightarrow \text{std available}$ (σ)

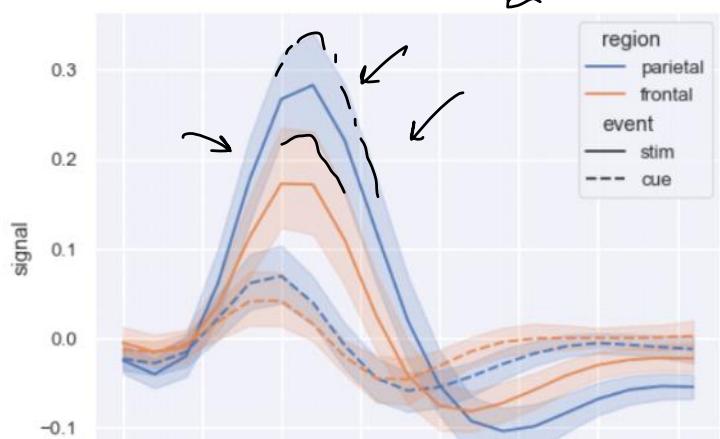
Confidence Interval is created for Parameters and not statistics. Statistics help us get the confidence interval for a parameter.

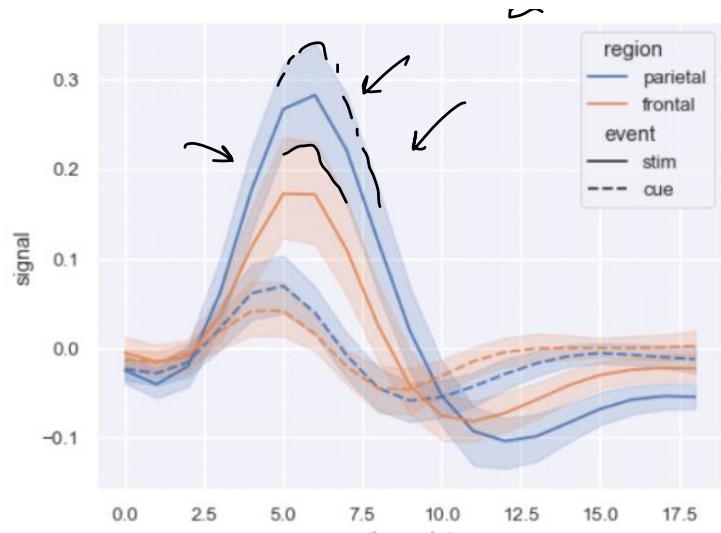
Examples of CT usage

seaborn



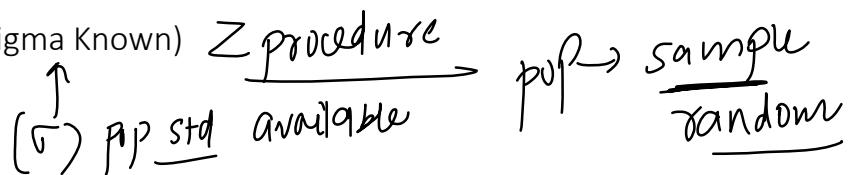
(CI)





Confidence Interval (Sigma Known)

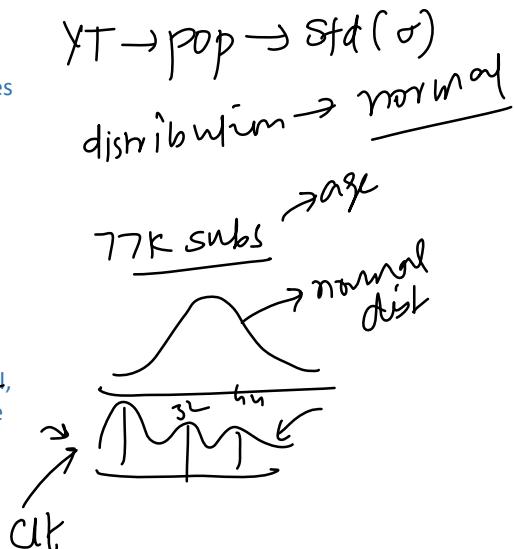
30 March 2023 07:13



Assumptions

- 1 **Random sampling:** The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
- 2 **Known population standard deviation:** The population standard deviation (σ) must be known or accurately estimated. In practice, the population standard deviation is often unknown, and the sample standard deviation (s) is used as an estimate. However, if the sample size is large enough, the sample standard deviation can provide a reasonably accurate approximation.
- 3 **Normal distribution or large sample size:** The Z-procedure assumes that the underlying population is normally distributed. However, if the population distribution is not normal, the Central Limit Theorem can be applied when the sample size is large (usually, sample size $n \geq 30$ is considered large enough). According to the Central Limit Theorem, the sampling distribution of the sample mean will approach a normal distribution as the sample size increases, regardless of the shape of the population distribution.

Sample size $n \geq 30$ → Z procedure



A $(1 - \alpha) * 100\%$ Confidence Interval for μ :

$Y_T \rightarrow \text{campus} \rightarrow 77K \rightarrow 28 \pm 14 \rightarrow$

$\rightarrow [16, 42] \leftarrow \text{confidence interval}$

$\text{Confidence level} \rightarrow 95\% \uparrow$

formula CI using Z procedure

$$\sigma = 15$$

$$CI = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Intuition

1) Intuition

2) $Z_{\alpha/2}$

$Z \rightarrow ?$

$(1 - \alpha) \rightarrow \text{confidence level}$

$(1 - \alpha) \rightarrow 95\%$

$\sigma \rightarrow \text{std pop}$

$n \rightarrow \text{sample size} \rightarrow 100$

Intuition
Point estimate $\tilde{x} \rightarrow CLT$

The diagram illustrates the sampling distribution of the sample mean. It starts with a population distribution labeled "normally dist" with mean μ and standard deviation σ . A sample of size n is drawn from this population, resulting in a sample mean \bar{X} . The sampling distribution of the sample mean is shown as a bell-shaped curve centered at μ , with standard deviation σ/\sqrt{n} . The area under this curve between two z-scores, $-z_{\alpha/2}$ and $z_{\alpha/2}$, is shaded and labeled $1 - \alpha$, representing the probability of the sample mean falling within this range. The z-score $z = (\bar{X} - \mu) / (\sigma/\sqrt{n})$ is also shown. A red box highlights the formula $95\% = (1 - \alpha)$ and the label "sure". Below the graph, the confidence interval formula is given as $CI = \bar{X} \pm (z_{\alpha/2}) \frac{\sigma}{\sqrt{n}}$.

$$P(Z_{\alpha_2} < Z < Z_{\alpha_1}) = 1 - \alpha$$

$$P\left(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1 - \alpha$$

(4) \rightarrow (1)

$$P\left(-Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < -\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P(-\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) \rightarrow 95\%$$

\bar{X} → Sample

$$P(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

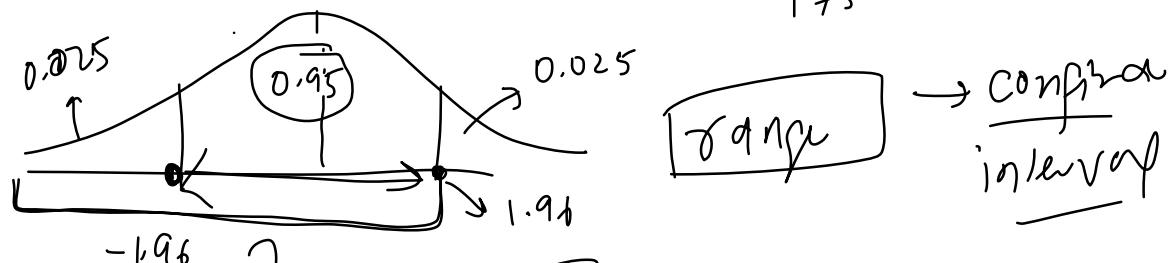
$1 - \alpha = 0.95 \quad \alpha = 0.05 \quad \frac{0.05}{2} = 0.025$

$$CI \quad \underline{\mu} = \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{confidence } (1 - \alpha) \rightarrow 95\%$$

$$\mu = \bar{X} \pm z_{0.025} \frac{\sigma}{\sqrt{n}}$$

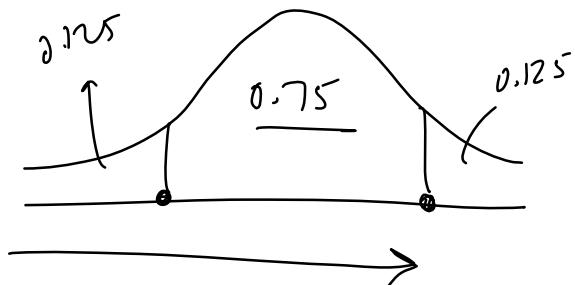
$z_{0.025} = 1.96$

$$\frac{95}{0.025} = 975$$



$$\boxed{\mu = \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}} \quad \text{CI with 95\% confidence level}$$

$$z_{\alpha/2} \rightarrow 1.96 \quad 50\% \quad 75\% \quad 99\% \quad 0.87 \quad R_{\alpha/2}$$



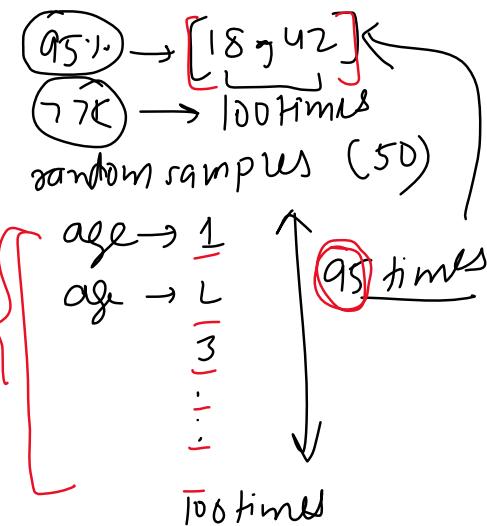
Interpreting Confidence Interval

30 March 2023 08:33

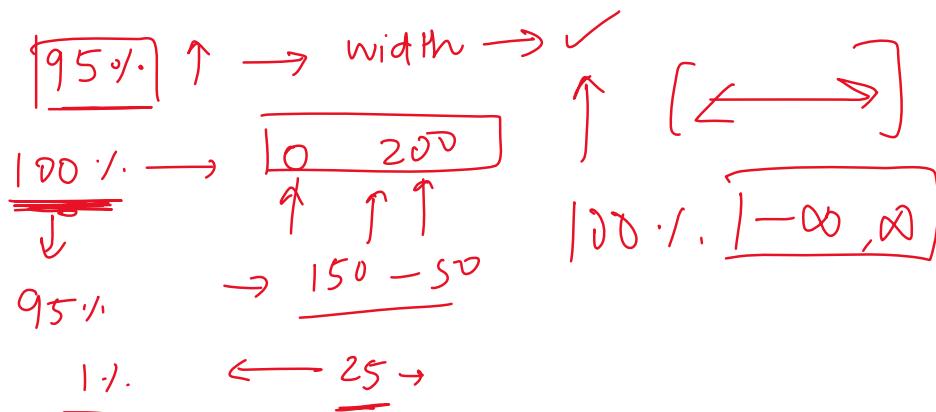
$$\text{pop} \rightarrow 45 \rightarrow [14 - 42] \leftarrow \text{fixed}$$

A confidence interval is a range of values within which a population parameter, such as the population mean, is estimated to lie with a certain level of confidence. The confidence interval provides an indication of the precision and uncertainty associated with the estimate. To interpret the confidence interval values, consider the following points:

- Confidence level:** The confidence level (commonly set at 90%, 95%, or 99%) represents the probability that the confidence interval will contain the true population parameter if the sampling and estimation process were repeated multiple times. For example, a 95% confidence interval means that if you were to draw 100 different samples from the population and calculate the confidence interval for each, approximately 95 of those intervals would contain the true population parameter.
- Interval range:** The width of the confidence interval gives an indication of the precision of the estimate. A narrower confidence interval suggests a more precise estimate of the population parameter, while a wider interval indicates greater uncertainty. The width of the interval depends on the sample size, variability in the data, and the desired level of confidence.
- Interpretation:** To interpret the confidence interval values, you can say that you are "X% confident that the true population parameter lies within the range (lower limit, upper limit)." Keep in mind that this statement is about the interval, not the specific point estimate, and it refers to the confidence level you chose when constructing the interval.



What is the trade-off



Factors Affecting Margin of Error

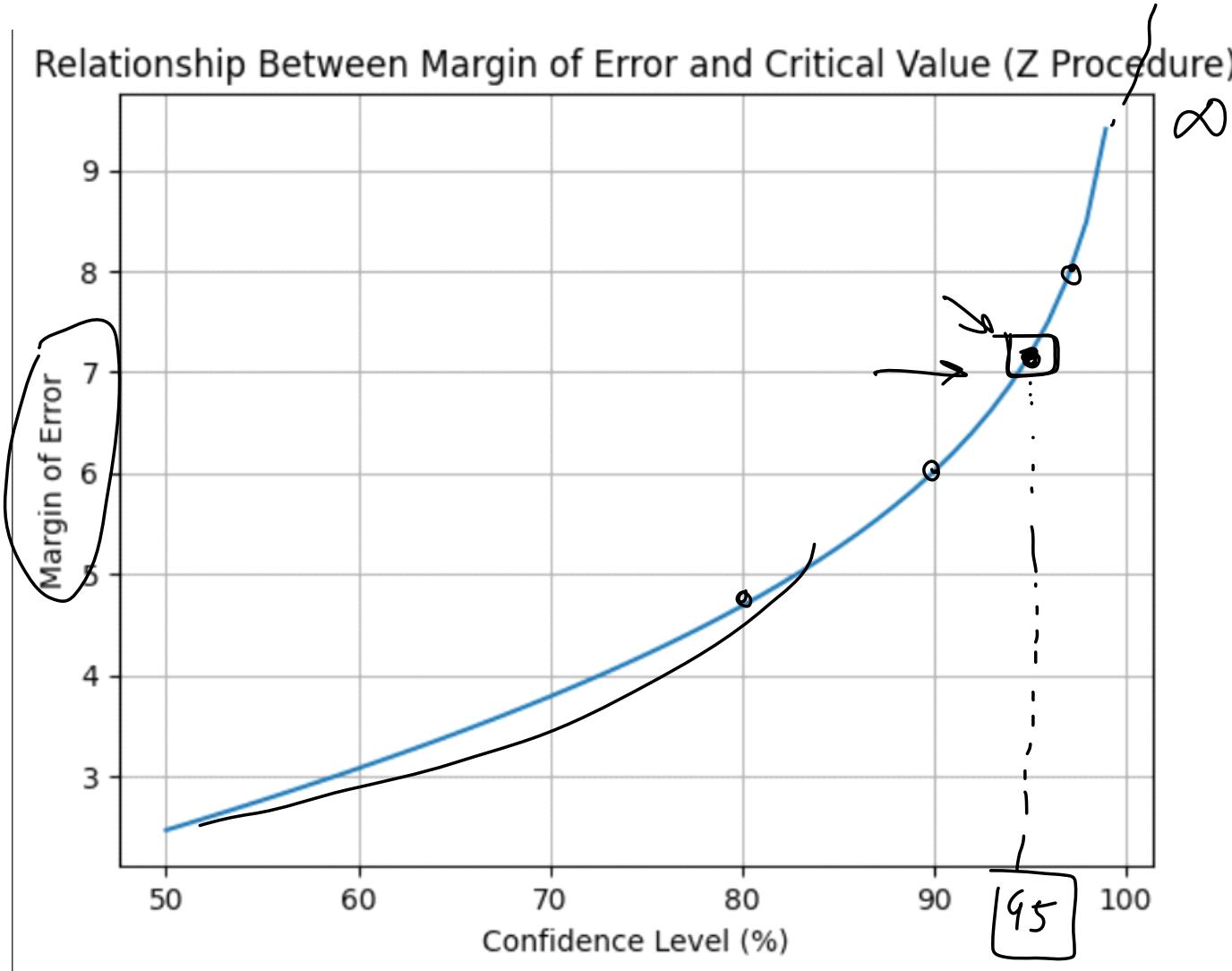
30 March 2023 07:15

1. Confidence Level (1-alpha)
2. Sample Size
3. Population Standard Deviation

$$\frac{\text{Upper} - \text{Lower}}{z} \rightarrow \text{margin}$$

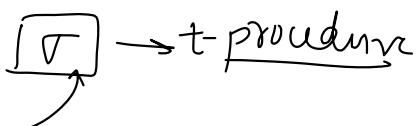
$$\begin{aligned} CI &= \text{point estimate} \pm \text{margin of error} \\ &= \boxed{\bar{x}} \pm z_{\alpha/2} \cdot \frac{\text{pop std}}{\sqrt{n}} \\ &\text{Sample mean} \quad \text{critical value} \quad \text{pop std} \quad \text{sample std} \end{aligned}$$

Relationship Between Margin of Error and Critical Value (Z Procedure)



Confidence Interval (Sigma not known)

30 March 2023 07:15



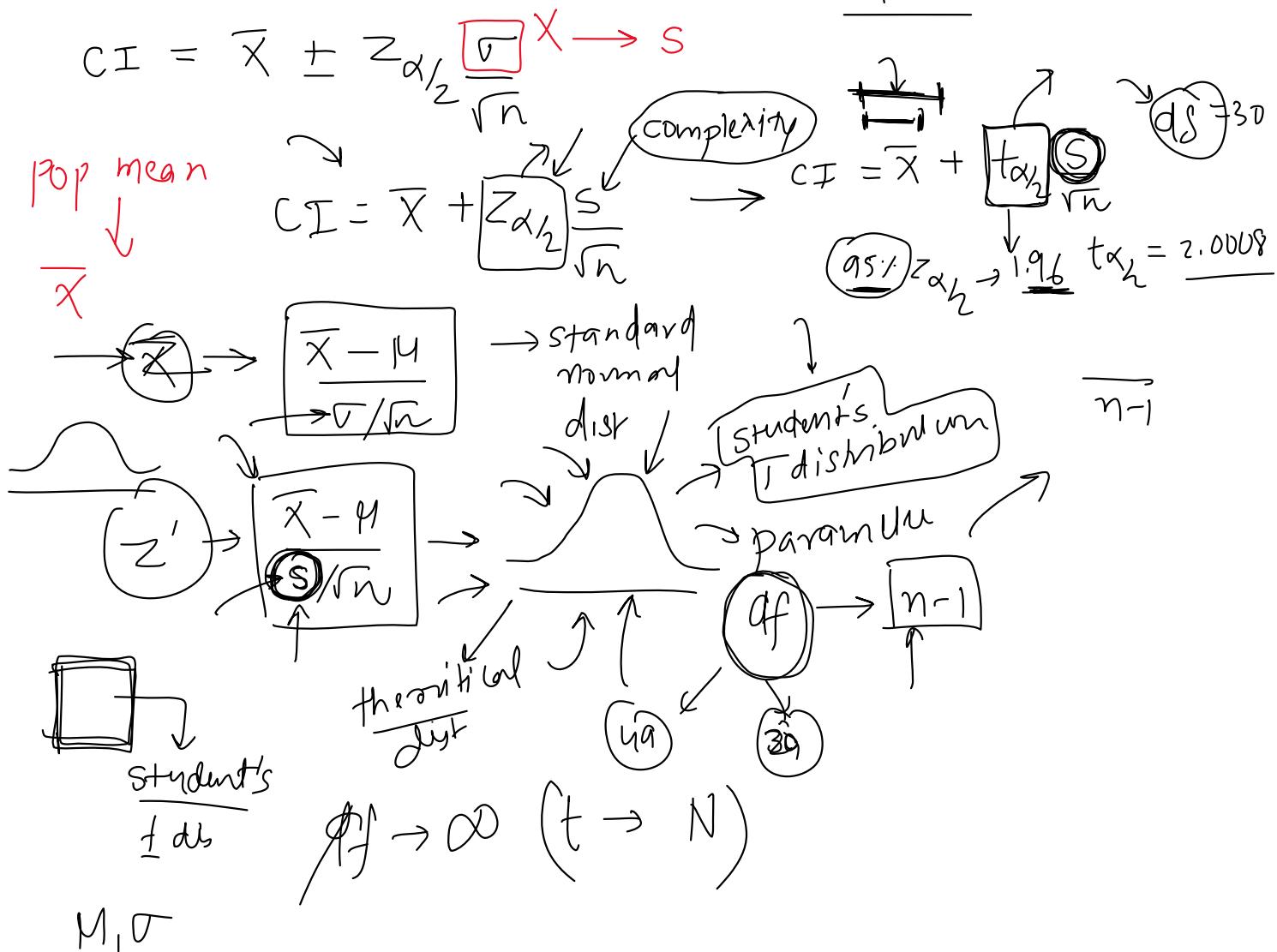
Using the t procedure

Assumptions

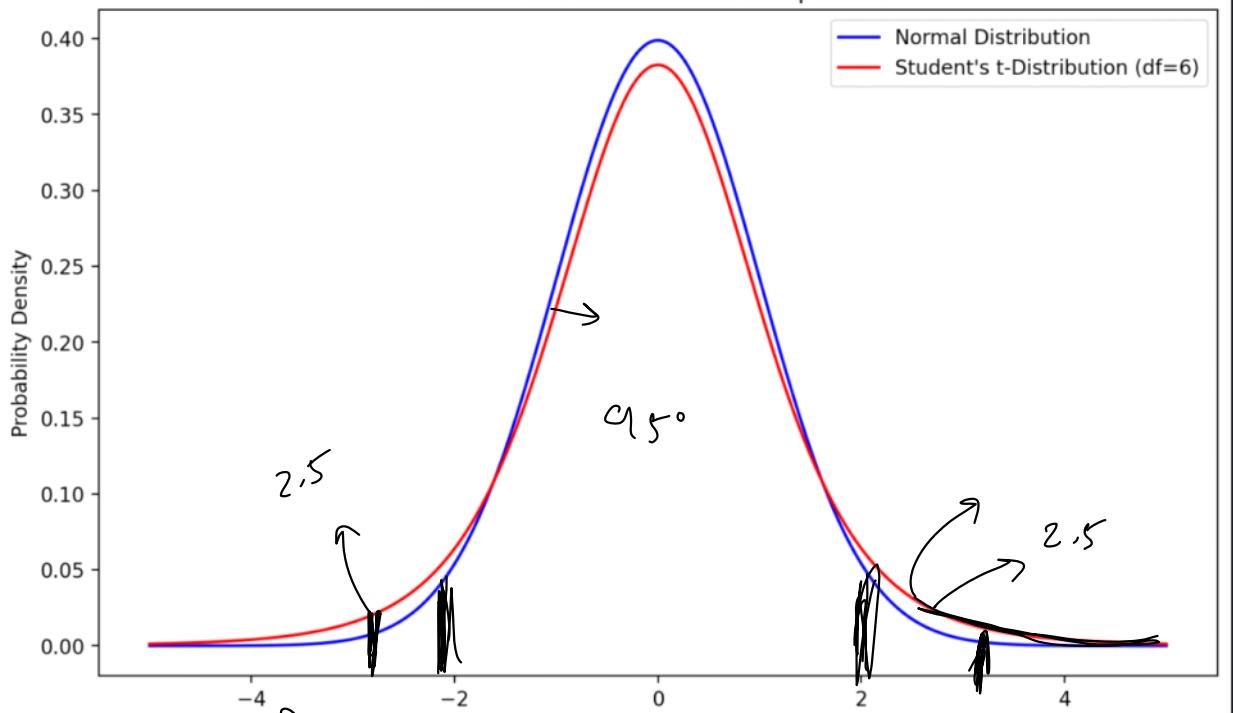
1. [Random sampling] The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. [Sample standard deviation] ^{may be} The population standard deviation (σ) is unknown, and the sample standard deviation (s) is used as an estimate. The t-distribution is specifically designed to account for the additional uncertainty introduced by using the sample standard deviation instead of the population standard deviation.
3. [Approximately normal distribution] The t-procedure assumes that the underlying population is approximately normally distributed, or the sample size is large enough for the Central Limit Theorem to apply. If the population distribution is heavily skewed or has extreme outliers, the t-procedure may not be accurate, and non-parametric methods should be considered.
4. [Independent observations] The observations in the sample should be independent of each other. In other words, the value of one observation should not influence the value of another observation. This is particularly important when working with time series data or data with inherent dependencies.

$n > 30 \rightarrow$ if it's not normal
 normal \rightarrow small sample
 sum \rightarrow CLT
 $t_{\alpha/2} > z_{\alpha/2}$
 $t_{\alpha/2} \approx z_{\alpha/2}$

Z-table



Normal and t-Distribution Comparison



$$CI = \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Diagram showing two normal distributions. The left one has mean \bar{x} , standard deviation s , and sample size n . The right one has mean μ and standard deviation σ .

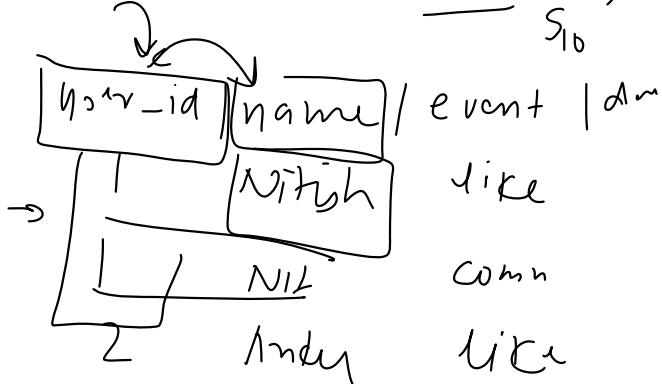
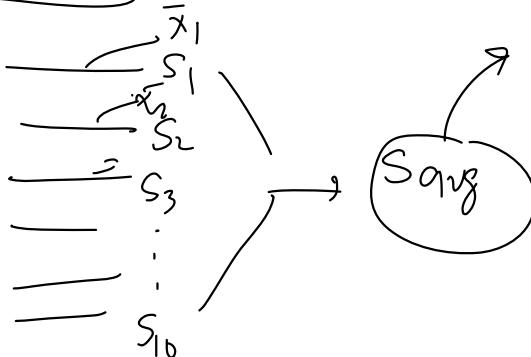
$$= \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$CI = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

1 N i h i z C L T

2 A n k u l



user_wml

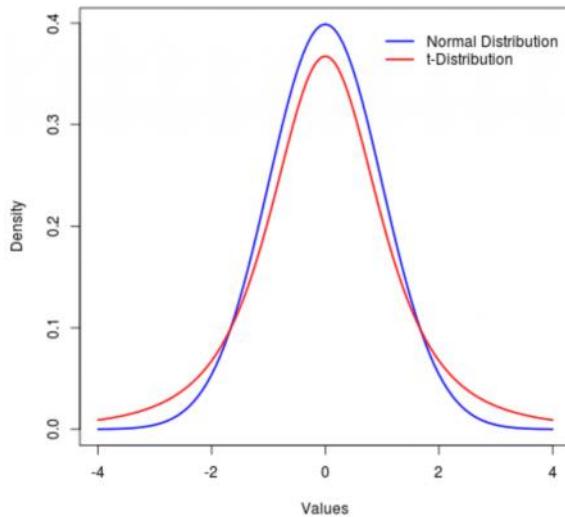
1
2

2

1

Student's T Distribution

30 March 2023 07:16



Student's t-distribution, or simply the t-distribution, is a probability distribution that arises when estimating the mean of a normally distributed population when the sample size is small and the population standard deviation is unknown. It was introduced by William Sealy Gosset, who published under the pseudonym "Student."

The t-distribution is similar to the normal distribution (also known as the Gaussian distribution or the bell curve) but has heavier tails. The shape of the t-distribution is determined by the degrees of freedom, which is closely related to the sample size (degrees of freedom = sample size - 1). As the degrees of freedom increase (i.e., as the sample size increases), the t-distribution approaches the normal distribution.

In hypothesis testing and confidence interval estimation, the t-distribution is used in place of the normal distribution when the sample size is small (usually less than 30) and the population standard deviation is unknown. The t-distribution accounts for the additional uncertainty that arises from estimating the population standard deviation using the sample standard deviation.

To use the t-distribution in practice, you look up critical t-values from a t-distribution table, which provides values corresponding to specific degrees of freedom and confidence levels (e.g., 95% confidence). These critical t-values are then used to calculate confidence intervals or perform hypothesis tests.

Titanic Case Study

31 March 2023 18:00

$$\begin{array}{l} \text{Pop} \rightarrow 1360 \\ \xrightarrow{\quad} \\ \mu \rightarrow X \\ \sigma \rightarrow X \\ \text{CLT} \rightarrow \text{10 times} \rightarrow \underline{30} \text{ s.e} \\ \text{95% confidence level} \\ \text{inference} \end{array}$$

1. Why Probability Distributions are important?

01 April 2023 16:42

2. Why divide by n-1?

01 April 2023 16:43

1. Demo of this happening
2. Mathematical Derivation
3. Geometric Intuition

The "sample standard deviation" is often called "the standard deviation of the sample," but that's not technically correct. Really, it's an *estimate*, derived from the sample, of the standard deviation of *the population*.

\bar{x} is estimated from the values x_i , the points x_i are a little closer to \bar{x} , on average than to μ . Therefore, the above estimate would underestimate the true population standard deviation!

When we calculate the sample variance, we use the sample mean as an approximation for the population mean. This approximation usually results in an underestimation of the actual population variance because the sample mean is affected by the sample itself, and it's closer to the sample data points than the true population mean.

4. Why divide by n-1 (Bessel's correction)
5. But what is n-1 and why n-1 only

<https://colab.research.google.com/drive/1BXNWprKm3ig6AwXP3JHjkdNpKmuUROxj?usp=sharing>

3. Sampling Techniques

01 April 2023 16:43

Probability sampling is a method of selecting samples from a population in a way that each member of the population has a known, non-zero chance of being selected. This approach ensures that the sample is representative of the population and reduces sampling bias. There are several types of probability sampling techniques, including:

1. Simple Random Sampling: Every member of the population has an equal chance of being selected. This can be done with or without replacement.
2. Systematic Sampling: Members of the population are chosen at regular intervals, usually using a random starting point. For example, selecting every 10th individual in a list.
3. Stratified Sampling: The population is divided into homogeneous subgroups (strata) based on specific characteristics, such as age or income. Then, simple random sampling is used to select a proportional number of individuals from each stratum.
4. Cluster Sampling: The population is divided into clusters, usually based on geographic regions or natural groupings. A random sample of clusters is selected, and then all individuals within the chosen clusters are included in the sample.
5. Multistage Sampling: This method combines two or more sampling techniques, often using a hierarchical approach. For example, one might first use cluster sampling to select specific regions, and then use stratified sampling within those regions.

When to use Systematic sampling over SRS

Systematic sampling can be advantageous over simple random sampling (SRS) in certain situations. Here are some cases where you might prefer to use systematic sampling over SRS:

- a. When you need a more evenly distributed sample: Systematic sampling ensures that the sample is spread out evenly across the entire population. This can be particularly useful when there is a pattern or trend in the population, and you want to make sure that the sample covers the entire range of the population.
- b. When the population is ordered or sequential: In cases where the population is ordered or has a natural sequence (e.g., items on a production line, patients in a waiting list, or data points in time series), systematic sampling can be more practical and efficient than SRS. It helps to capture periodic patterns or trends that might be present in the population.
- c. When data collection is easier with systematic sampling: In some situations, the logistics of data collection may be more manageable with systematic sampling. For example, if you are surveying households in a neighbourhood, it may be more practical to interview every nth household rather than randomly selecting households from a list. This can reduce travel time and effort for the survey team.
- d. When simple random sampling is difficult or impossible: In cases where the population list is not readily available or it is difficult to generate random numbers (e.g., in remote areas with limited resources), systematic sampling can be a more feasible alternative to SRS.

Likelihood Vs Probability

01 April 2023 16:43

Expected Value Vs Expected Variance

01 April 2023 16:44

Statistical Moments

01 April 2023 16:44

Spearman Correlation Coefficient

01 April 2023 16:44

Discrete Uniform Distribution

01 April 2023 16:45

Poisson Distribution

01 April 2023 16:45

What is Degree of Freedom?

01 April 2023 17:26

Bernoulli Distribution

27 March 2023 16:06

$p \rightarrow$ prob of success
 $1-p \rightarrow$ prob of failure

$\rightarrow P$

Experiment

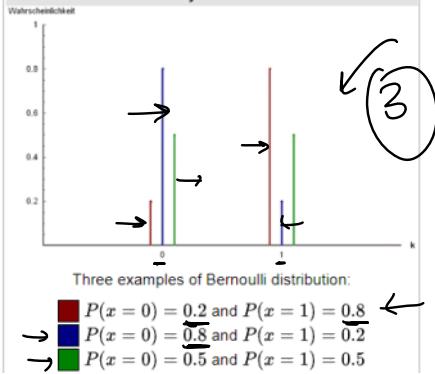
Bernoulli distribution is a probability distribution that models a binary outcome, where the outcome can be either success (represented by the value 1) or failure (represented by the value 0). The Bernoulli distribution is named after the Swiss mathematician Jacob Bernoulli, who first introduced it in the late 1600s.

The Bernoulli distribution is characterized by a single parameter, which is the probability of success, denoted by p . The probability mass function (PMF) of the Bernoulli distribution is:

$$\boxed{\text{pmf}} = p(X=x) = \boxed{p^x (1-p)^{1-x}}$$

Bernoulli distribution

Probability mass function



machine learning

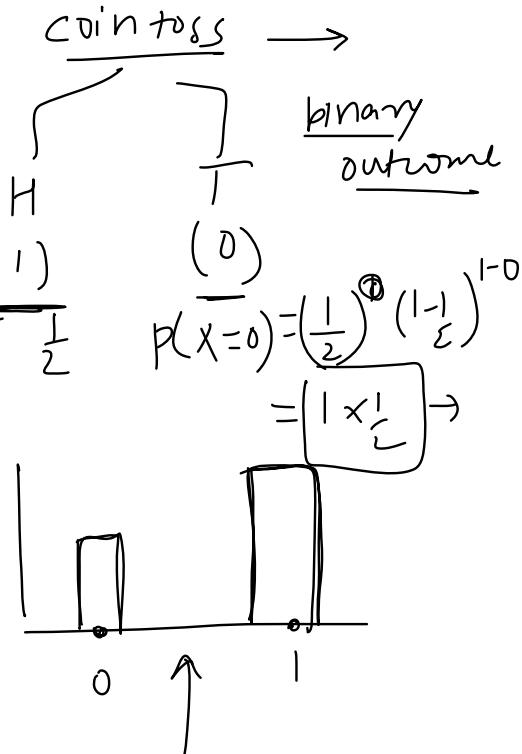
$$P(X=1) = \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^0 = \frac{1}{2}$$

$$P(X=0) = \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^1 = \frac{1}{2}$$

rolling a dice
getting a 5

$$\left(\frac{1}{6}\right)$$

$$\left(\frac{5}{6}\right)$$



The Bernoulli distribution is commonly used in machine learning for modelling binary outcomes, such as whether a customer will make a purchase or not, whether an email is spam or not, or whether a patient will have a certain disease or not.

Binomial distribution

Binomial Distribution

27 March 2023 16:36

$$(p) \quad \boxed{(n, p)}$$

$n=1 \rightarrow$ binomial
Bernoulli

Bernoulli trial

n times

\circlearrowleft Binomial

Binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials with two possible outcomes (often called "success" and "failure"), where the probability of success is constant for each trial. The binomial distribution is characterized by two parameters: the number of trials n and the probability of success p .

$$0.5$$

Feedback

$\rightarrow 10$ students

\rightarrow T Y Y X

2 Y Y N ✓

3 Y N Y ✓

4 Y N N →

5 N Y Y ↗

6 N Y N ↗

7 N N Y ↗

8 N N N ↗

The Probability of anyone watching this lecture in the future and then liking it is 0.5. What is the probability that:

1. No-one out of 3 people will like it

$$\frac{1}{8}$$

1. 1 out of 3 people will like it

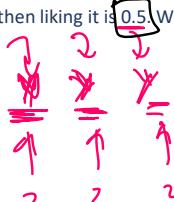
$$\frac{3}{8}$$

1. 2 out of 3 people will like it

$$\frac{3}{8}$$

1. 3 out of 3 people will like it

$$\frac{1}{8}$$



Binomial

$$P(X=x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$n \rightarrow$ # of trials

$p \rightarrow$ prob of success

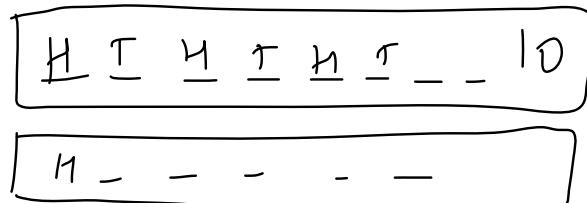
$x \rightarrow$ desired result.

PDF Formula:

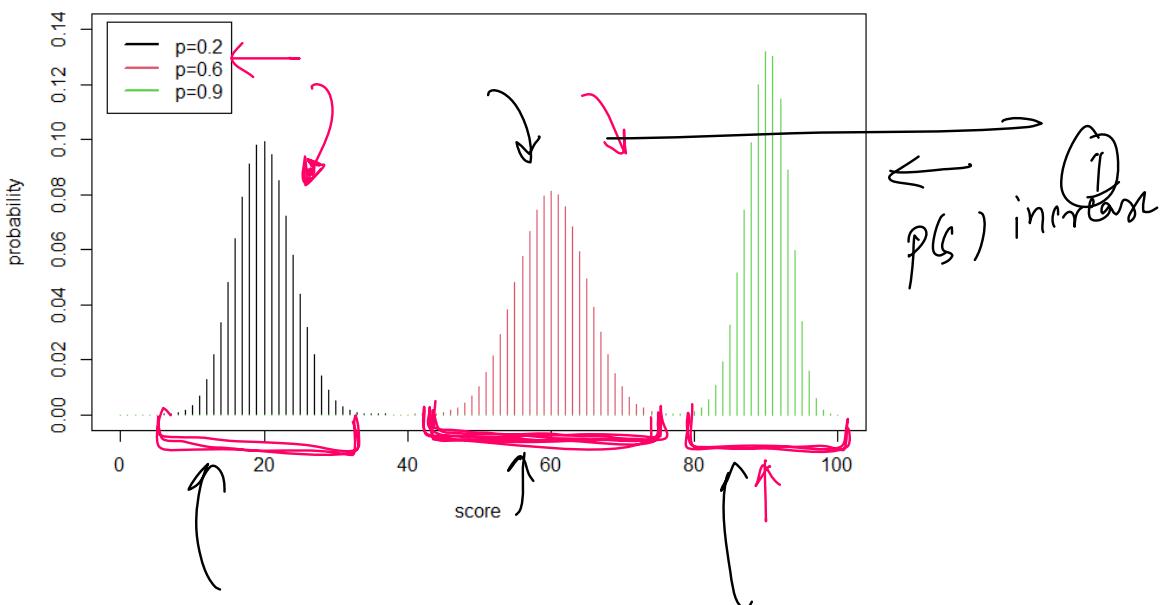
$$\frac{3!}{2!1!} \times \frac{1}{8} = \frac{3}{8}$$

$$3 \times \frac{1}{8} = \frac{3}{8}$$

Graph of PDF:



Binomial distribution with different probabilities of success



Criteria:

Criteria:

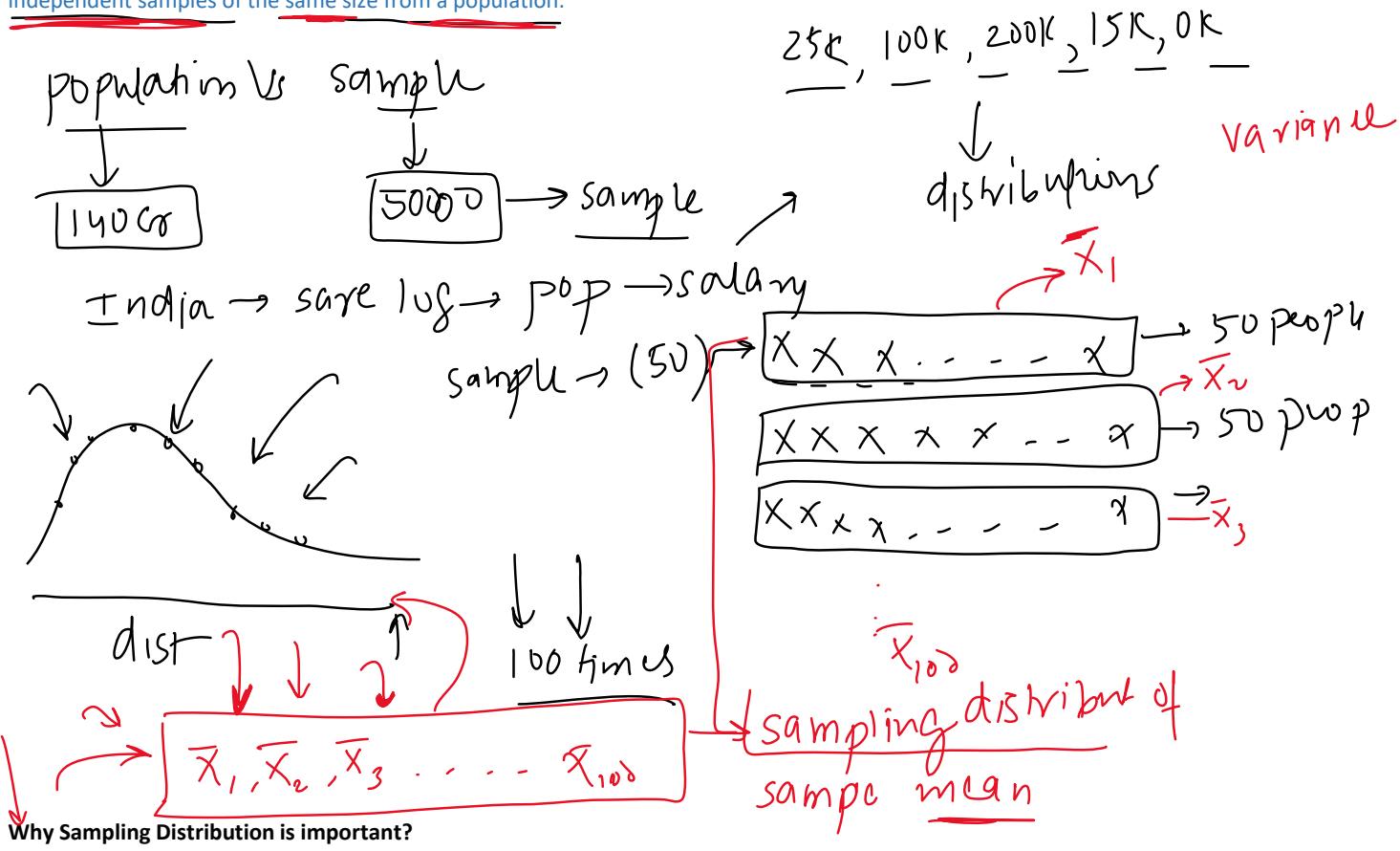
1. The process consists of n trials
2. Only 2 exclusive outcomes are possible, a success and a failure.
3. $P(\text{success}) = p$ and $P(\text{failure}) = 1-p$ and it is fixed from trial to trial
4. The trials are independent.

1. **Binary classification problems:** In binary classification problems, we often model the probability of an event happening as a binomial distribution. For example, in a spam detection system, we may model the probability of an email being spam or not spam using a binomial distribution.
2. **Hypothesis testing:** In statistical hypothesis testing, we use the binomial distribution to calculate the probability of observing a certain number of successes in a given number of trials, assuming a null hypothesis is true. This can be used to make decisions about whether a certain hypothesis is supported by the data or not.
3. **Logistic regression:** Logistic regression is a popular machine learning algorithm used for classification problems. It models the probability of an event happening as a logistic function of the input variables. Since the logistic function can be viewed as a transformation of a linear combination of inputs, the output of logistic regression can be thought of as a binomial distribution.
4. **A/B testing:** A/B testing is a common technique used to compare two different versions of a product, web page, or marketing campaign. In A/B testing, we randomly assign individuals to one of two groups and compare the outcomes of interest between the groups. Since the outcomes are often binary (e.g., click-through rate or conversion rate), the binomial distribution can be used to model the distribution of outcomes and test for differences between the groups.

Sampling Distribution

27 March 2023 17:10

Sampling distribution is a probability distribution that describes the statistical properties of a sample statistic (such as the sample mean or sample proportion) computed from multiple independent samples of the same size from a population.

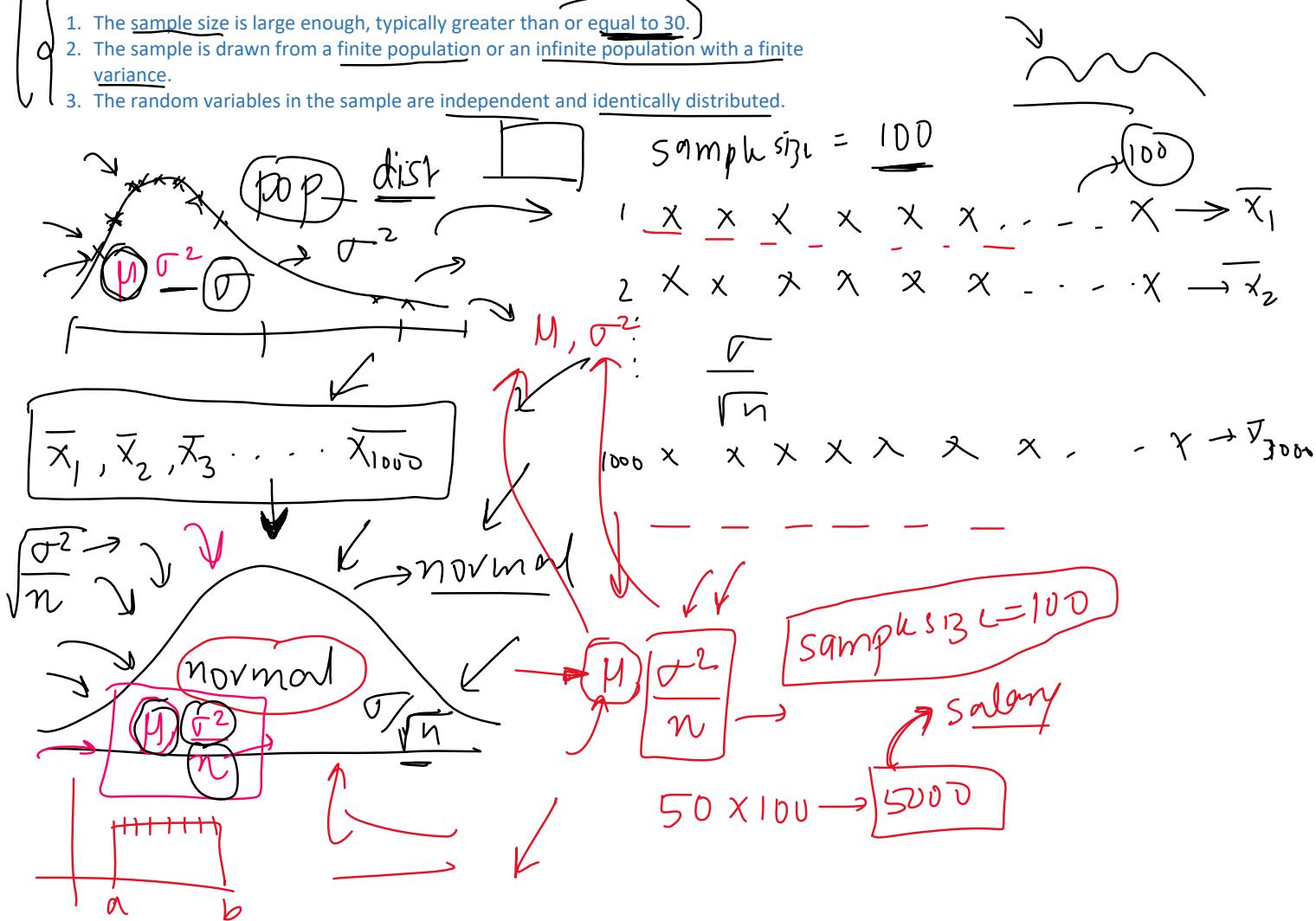


interval $\boxed{\mu = 32.584} \rightarrow$ fare $\mu = 32.584$

The Central Limit Theorem (CLT) states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of the variables.

The conditions required for the CLT to hold are:

1. The sample size is large enough, typically greater than or equal to 30.
2. The sample is drawn from a finite population or an infinite population with a finite variance.
3. The random variables in the sample are independent and identically distributed.

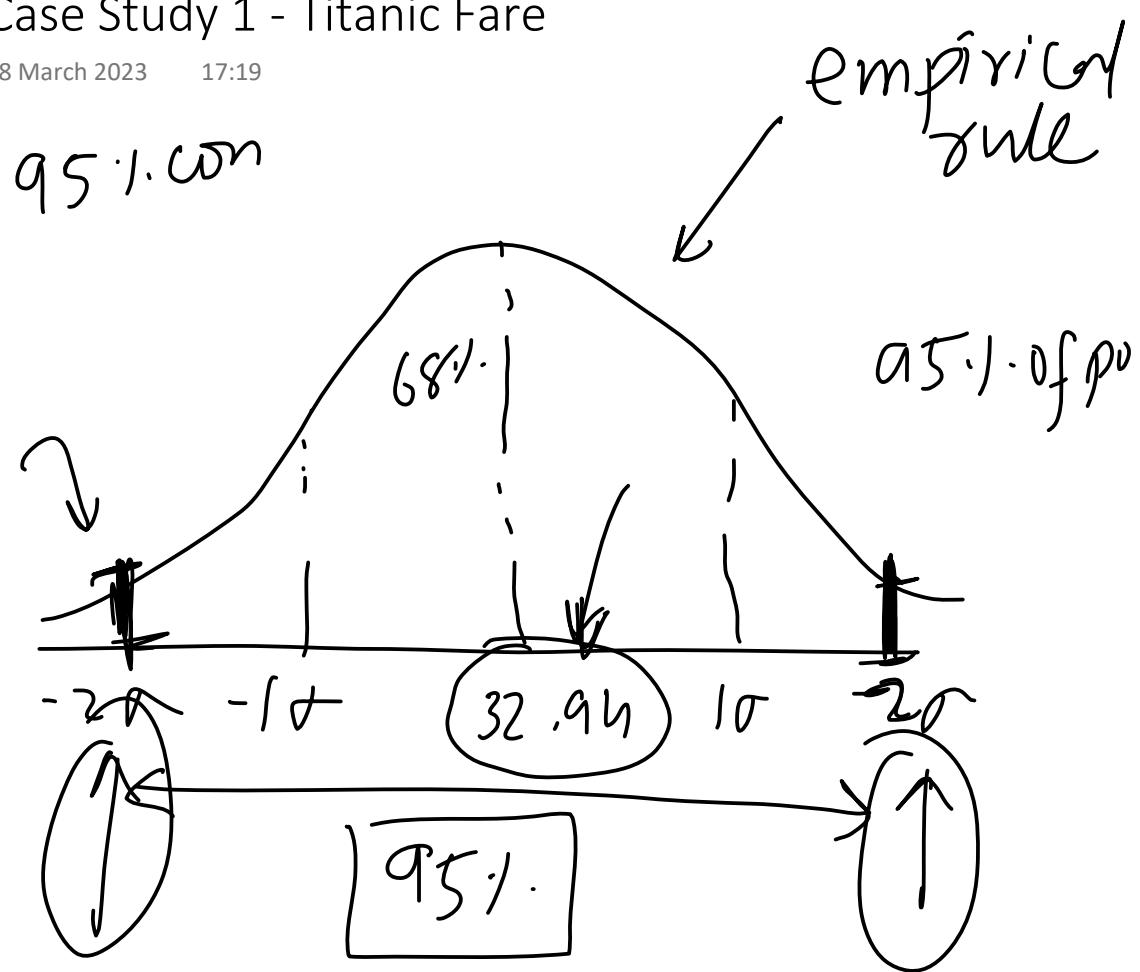


The CLT is important in statistics and machine learning because it allows us to make probabilistic inferences about a population based on a sample of data. For example, we can use the CLT to construct confidence intervals, perform hypothesis tests, and make predictions about the population mean based on the sample data. The CLT also provides a theoretical justification for many commonly used statistical techniques, such as t-tests, ANOVA, and linear regression.

Case Study 1 - Titanic Fare

28 March 2023

17:19



Case Study - What is the average income of Indians

28 March 2023 15:49

Step-by-step process:

1. Collect multiple random samples of salaries from a representative group of Indians. Each sample should be large enough (usually, $n > 30$) to ensure the CLT holds. Make sure the samples are representative and unbiased to avoid skewed results.
2. Calculate the sample mean (average salary) and sample standard deviation for each sample.
3. Calculate the average of the sample means. This value will be your best estimate of the population mean (average salary of all Indians).
4. Calculate the standard error of the sample means, which is the standard deviation of the sample means divided by the square root of the number of samples.
5. Calculate the confidence interval around the average of the sample means to get a range within which the true population mean likely falls. For a 95% confidence interval:

```
lower_limit = average_sample_means - 1.96 * standard_error  
upper_limit = average_sample_means + 1.96 * standard_error
```

6. Report the estimated average salary and the confidence interval.

Python code

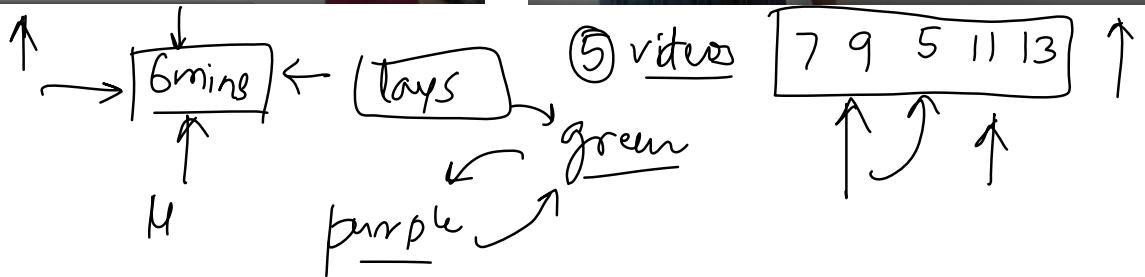
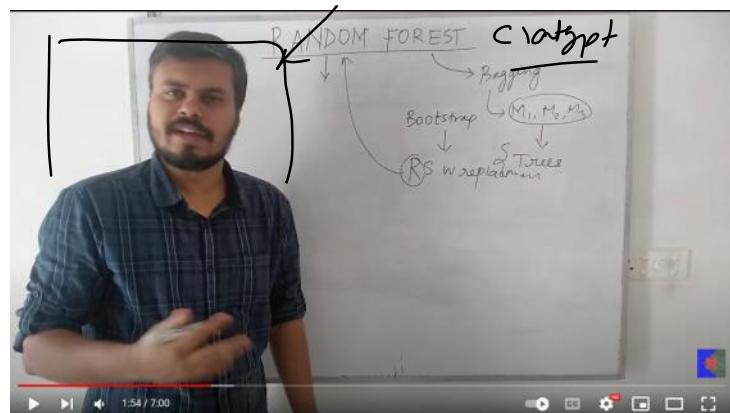
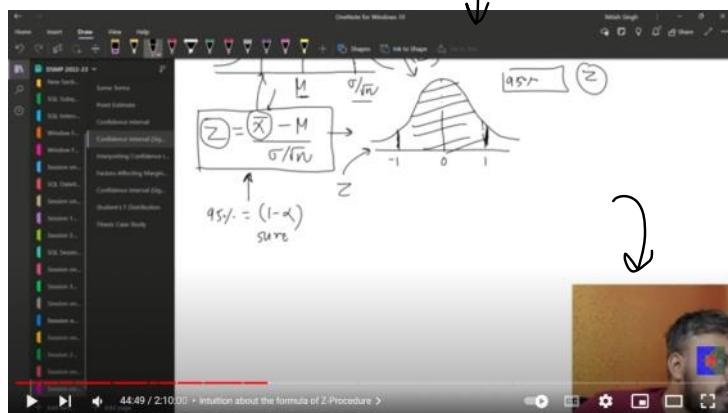
Remember that the validity of your results depends on the quality of your data and the representativeness of your samples. To obtain accurate results, it's crucial to ensure that your samples are unbiased and representative.

Hypothesis Testing

04 April 2023 07:04

$$\boxed{a+b=8}$$

1 video → 13 mins



A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.

Null and Alternate Hypothesis

04 April 2023 07:09

H_0

new shooting
more avg view
duration

$H_0: \mu = 6\text{min}$

$H_1: \mu > 6\text{ mins}$

1. Null hypothesis (H_0):

In simple terms, the null hypothesis is a statement that assumes there is no significant effect or relationship between the variables being studied. It serves as the starting point for hypothesis testing and represents the **status quo** or the assumption of **no effect until proven otherwise**. The purpose of hypothesis testing is to gather evidence (data) to either reject or fail to reject the null hypothesis in favour of the alternative hypothesis, which claims there is a significant effect or relationship.

Null →

reject →

H_1, H_a

2. Alternative hypothesis (H_1 or H_a):

The alternative hypothesis, is a statement that contradicts the null hypothesis and claims there is a significant effect or relationship between the variables being studied. It represents the **research hypothesis** or the claim that the researcher wants to support through statistical analysis.

$H_0:$ ✓

Important Points

- How to decide what will be **Null hypothesis** and what will be **Alternate Hypothesis** (Typically the Null hypothesis says nothing new is happening)
- We try to gather evidence to **reject the null hypothesis**
- It's important to note that failing to reject the null hypothesis doesn't necessarily mean that the **null hypothesis is true**; it just means that there isn't enough evidence to support the alternative hypothesis.

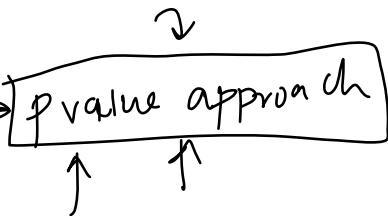
Hypothesis tests are similar to jury trials, in a sense. In a jury trial, H_0 is similar to the not-guilty verdict, and H_a is the guilty verdict. You assume in a jury trial that the defendant isn't guilty unless the prosecution can show beyond a reasonable doubt that he or she is guilty. If the jury says the evidence is beyond a reasonable doubt, they reject H_0 , not guilty, in favour of H_a , guilty.

Steps involved in Hypothesis Testing

04 April 2023 13:25

✓
Rejection Region Approach

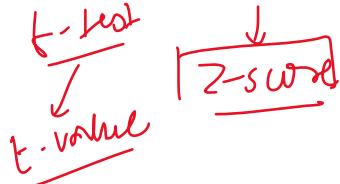
basil



$$H_0: \mu = 6 \text{ mins} \rightarrow \text{p-value}$$

$$H_a: \mu > 6 \text{ mins} \rightarrow \square$$

- 1. Formulate a Null and Alternate hypothesis
- 2. Select a significance level (This is the probability of rejecting the null hypothesis when it is actually true, usually set at 0.05 or 0.01)
- 3. Check assumptions (example distribution) $\sigma \leftarrow$
- 4. Decide which test is appropriate (Z-test, T-test, Chi-square test, ANOVA)
- 5. State the relevant test statistic
- 6. Conduct the test
- 7. Reject or not reject the Null Hypothesis.
- 8. Interpret the result



0.05 or 0.01
 $\rightarrow 5\% \rightarrow 1\%$
 normally distributed
 σ given \leftarrow t-test
 z-test

Performing a Z test Example 1

04 April 2023 07:15

$$\begin{array}{l} \mu = 50 \quad \sigma = 5 \\ n = 30 \quad \bar{x} = 53 \end{array}$$

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was 50 units per day with a known population standard deviation of 5 units. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day. The company wants to know if the new training program has significantly increased productivity.

1) $H_0: \mu = 50 \quad H_a: \mu > 50$

2) $\alpha = 0.05 \rightarrow 5\%$

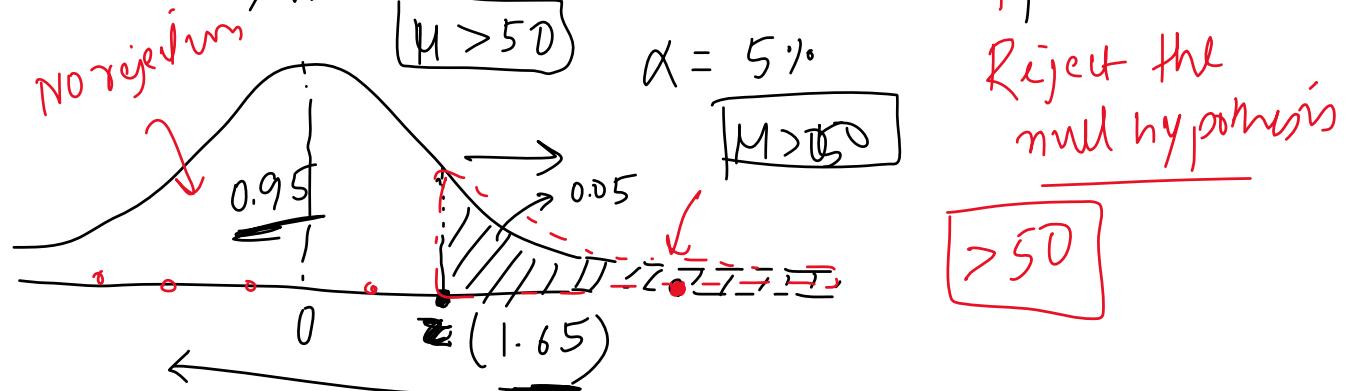
3) normality valid / pop std (σ) known

4) Z test

5) Z

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{53 - 50}{5/\sqrt{30}} = \frac{3}{5/\sqrt{30}} = 3.28$$

Rejection



Example 2

04 April 2023 16:06

Z test

{ Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a known population standard deviation of 4 grams.

$$\mu = 50 \quad n = 40 \quad \bar{x} = 49 \quad \sigma = 4$$

1) $H_0: \mu = 50$ $H_a: \mu \neq 50$

2) $\alpha = 0.05$

3) Normality ✓ \rightarrow Z test

4) Z test

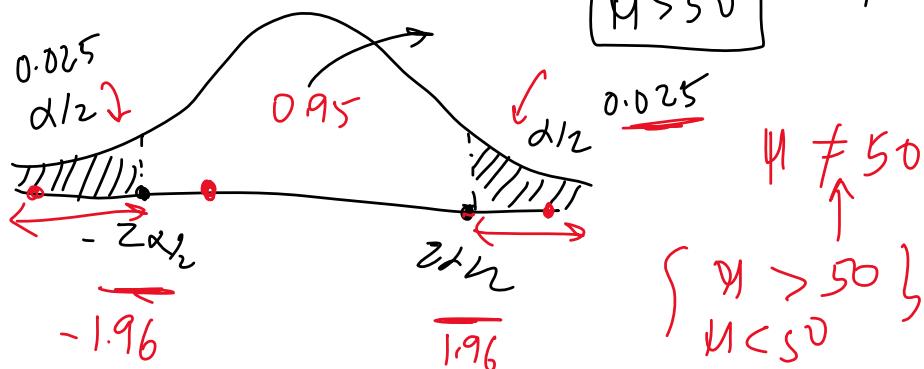
5) Z

$$6) Z = \frac{49 - 50}{4/\sqrt{40}} = \frac{-\sqrt{40}}{4} = [-1.58]$$

$\mu \neq 50$

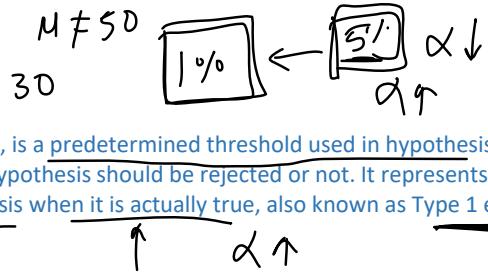
can't reject the null hypothesis

$$\alpha = 5\%$$



Rejection Region

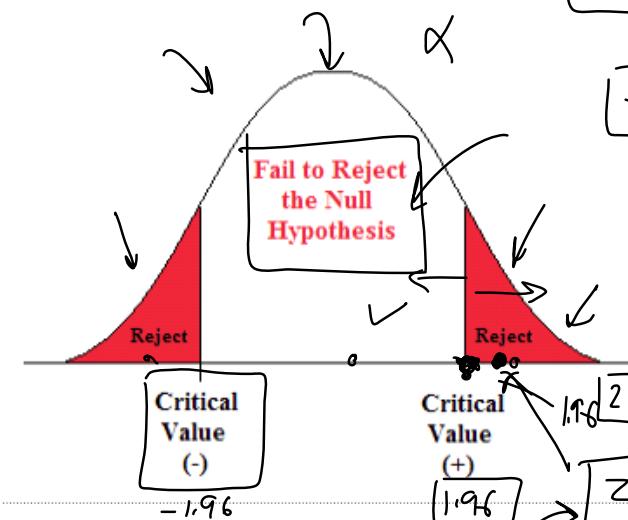
04 April 2023 16:21



Significance level - denoted as α (alpha), is a predetermined threshold used in hypothesis testing to determine whether the null hypothesis should be rejected or not. It represents the probability of rejecting the null hypothesis when it is actually true, also known as Type 1 error.

rejection

The critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.



evidentum
Strength

[z]

[0.0]

[1.95]

Null reg cut
[P-value]

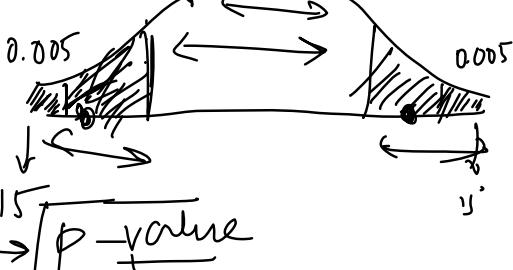
[z]

[1.97]

[z = 15]

[z = 2.00]
[z = 15]

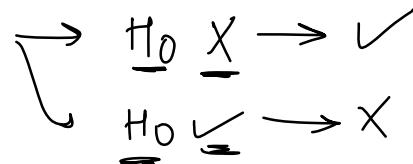
Problem with Rejection Region Approach



Type 1 vs Type 2 Error

04 April 2023 13:29

hypothesis



In hypothesis testing, there are two types of errors that can occur when making a decision about the null hypothesis: Type I error and Type II error.

Type-I (False Positive) error occurs when the sample results lead to the rejection of the null hypothesis when it is in fact true.

In other words, it's the mistake of finding a significant effect or relationship when there is none. The probability of committing a Type I error is denoted by α (alpha), which is also known as the significance level. By choosing a significance level, researchers can control the risk of making a Type I error.

Type-II (False Negative) error occurs when based on the sample results, the null hypothesis is not rejected when it is in fact false.

This means that the researcher fails to detect a significant effect or relationship when one actually exists. The probability of committing a Type II error is denoted by β (beta).

Trade-off between Type 1 and Type 2 errors

$$\alpha = 0.05$$

$H_0 \rightarrow H_0$ Crime

Truth about the population

H_0 true H_0 false

significant

H_0 true

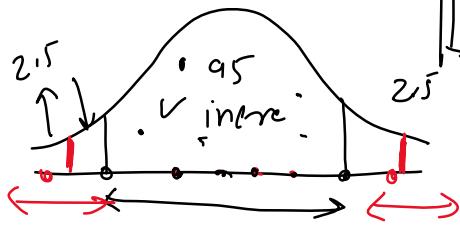
Reject H_0

Decision based on sample

Accept H_0

Type I error	Correct decision
Correct decision	Type II error

$\alpha = 5\%$



$$\alpha$$

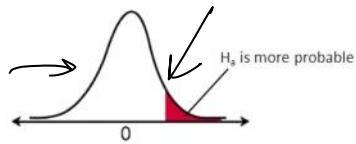
One sided vs two sided test

04 April 2023 13:29

one-sided

One-sided (one-tailed) test: A one-sided test is used when the researcher is interested in testing the effect in a specific direction (either greater than or less than the value specified in the null hypothesis). The alternative hypothesis in a one-sided test contains an inequality (either " $>$ " or " $<$ ").

Example: A researcher wants to test whether a new medication increases the average recovery rate compared to the existing medication.

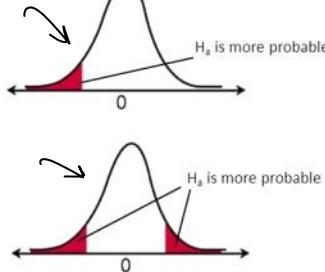


Right-tail test

$H_a: \mu > \text{value}$

Two-sided (two-tailed) test: A two-sided test is used when the researcher is interested in testing the effect in both directions (i.e., whether the value specified in the null hypothesis is different, either greater or lesser). The alternative hypothesis in a two-sided test contains a "not equal to" sign (\neq).

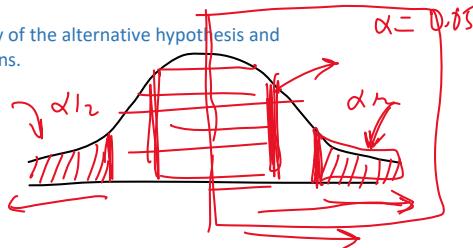
Example: A researcher wants to test whether a new medication has a different average recovery rate compared to the existing medication.



Two-tail test

$H_a: \mu \neq \text{value}$

The main difference between them lies in the directionality of the alternative hypothesis and how the significance level is distributed in the critical regions.



Advantages and Disadvantages?

Two-tailed test (two-sided):

Advantages:

1. Detects effects in both directions: Two-tailed tests can detect effects in both directions, which makes them suitable for situations where the direction of the effect is uncertain or when researchers want to test for any difference between the groups or variables.
2. More conservative: Two-tailed tests are more conservative because the significance level (α) is split between both tails of the distribution. This reduces the risk of Type I errors in cases where the direction of the effect is uncertain.

power = $1 - \beta$

Disadvantages:

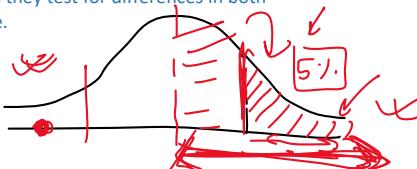
1. Less powerful: Two-tailed tests are generally less powerful than one-tailed tests because the significance level (α) is divided between both tails of the distribution. This means the test requires a larger effect size to reject the null hypothesis, which could lead to a higher risk of Type II errors (failing to reject the null hypothesis when it is false).
2. Not appropriate for directional hypotheses: Two-tailed tests are not ideal for cases where the research question or hypothesis is directional, as they test for differences in both directions, which may not be of interest or relevance.

MF100gms

One-tailed test (one-sided):

Advantages:

1. More powerful: One-tailed tests are generally more powerful than two-tailed tests, as the entire significance level (α) is allocated to one tail of the distribution. This means that the test is more likely to detect an effect in the specified direction, assuming the effect exists.
2. Directional hypothesis: One-tailed tests are appropriate when there is a strong theoretical or practical reason to test for an effect in a specific direction.



Disadvantages:

1. Missed effects: One-tailed tests can miss effects in the opposite direction of the specified alternative hypothesis. If an effect exists in the opposite direction, the test will not be able to detect it, which could lead to incorrect conclusions.
2. Increased risk of Type I error: One-tailed tests can be more prone to Type I errors if the effect is actually in the opposite direction than the one specified in the alternative hypothesis.

Where can be Hypothesis Testing Applied?

04 April 2023 07:15

1. Testing the effectiveness of interventions or treatments: Hypothesis testing can be used to determine whether a new drug, therapy, or educational intervention has a significant effect compared to a control group or an existing treatment.
 2. Comparing means or proportions: Hypothesis testing can be used to compare means or proportions between two or more groups to determine if there's a significant difference. This can be applied to compare average customer satisfaction scores, conversion rates, or employee performance across different groups.
 3. Analysing relationships between variables: Hypothesis testing can be used to evaluate the association between variables, such as the correlation between age and income or the relationship between advertising spend and sales.
 4. Evaluating the goodness of fit: Hypothesis testing can help assess if a particular theoretical distribution (e.g., normal, binomial, or Poisson) is a good fit for the observed data.
 5. Testing the independence of categorical variables: Hypothesis testing can be used to determine if two categorical variables are independent or if there's a significant association between them. For example, it can be used to test if there's a relationship between the type of product and the likelihood of it being returned by a customer.
 6. A/B testing: In marketing, product development, and website design, hypothesis testing is often used to compare the performance of two different versions (A and B) to determine which one is more effective in terms of conversion rates, user engagement, or other metrics.

Hypothesis testing can be used to determine if a significant effect exists to compare means or proportions, such as a significant difference in scores, conversion rates, or proportions.

Correlation can be used to evaluate the relationship between variables like age and income or the relationship between gender and purchase history.

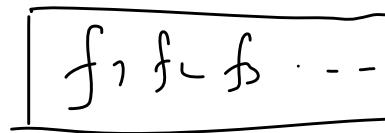
Chi-square test is used to assess if a particular theoretical distribution fits the observed data.

Z test, t test can be used to determine if there's a significant association or relationship between the variables being compared.

	gender	purchase history
M	1	0
F	0	1

Hypothesis Testing ML Applications

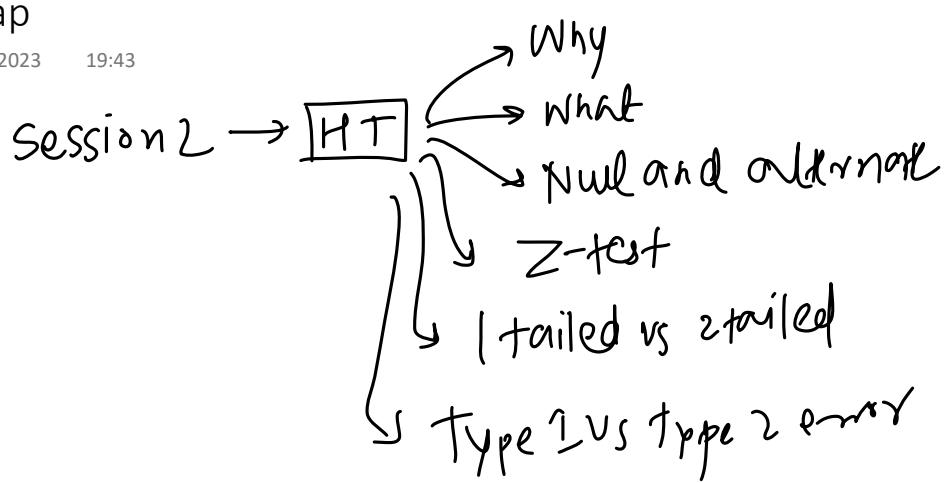
04 April 2023 16:50



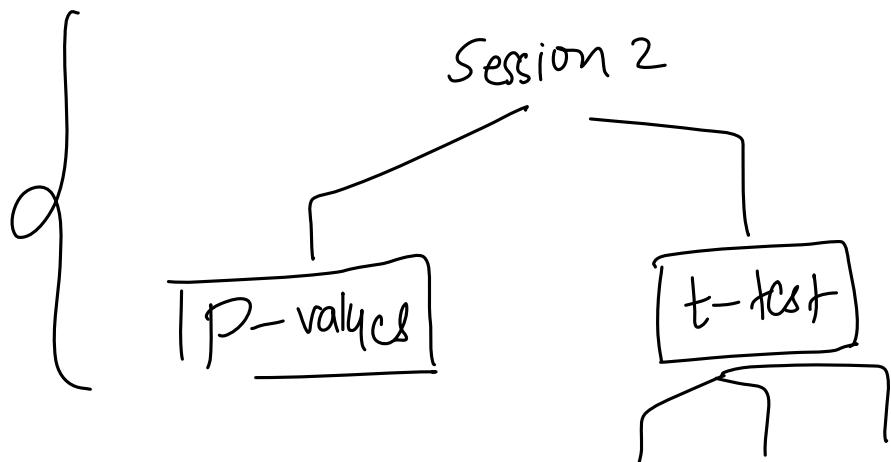
1. **Model comparison:** Hypothesis testing can be used to compare the performance of different machine learning models or algorithms on a given dataset. For example, you can use a paired t-test to compare the accuracy or error rate of two models on multiple cross-validation folds to determine if one model performs significantly better than the other.
2. **Feature selection:** Hypothesis testing can help identify which features are significantly related to the target variable or contribute meaningfully to the model's performance. For example, you can use a t-test, chi-square test, or ANOVA to test the relationship between individual features and the target variable. Features with significant relationships can be selected for building the model, while non-significant features may be excluded.
3. **Hyperparameter tuning:** Hypothesis testing can be used to evaluate the performance of a model trained with different hyperparameter settings. By comparing the performance of models with different hyperparameters, you can determine if one set of hyperparameters leads to significantly better performance.
4. **Assessing model assumptions:** In some cases, machine learning models rely on certain statistical assumptions, such as linearity or normality of residuals in linear regression. Hypothesis testing can help assess whether these assumptions are met, allowing you to determine if the model is appropriate for the data.

Recap

06 April 2023 19:43



significance level (α)



$\text{of } 53 \rightarrow 53 \text{ head}$

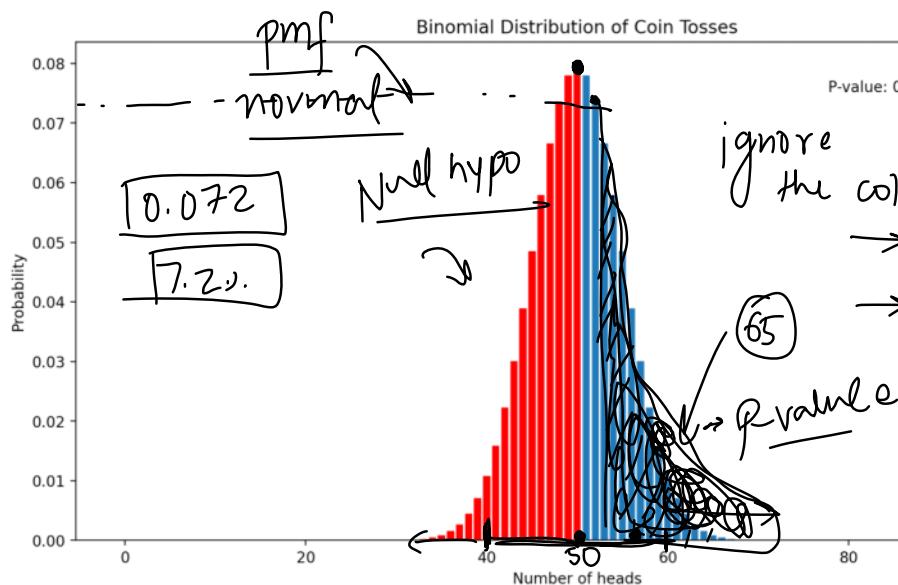
$$P(H > 53) | p, n$$

p-value

experiment

↓
1 coin100 times
toss

P-value is the probability of getting a sample as or more extreme having more evidence against H_0 than our own sample given the Null Hypothesis (H_0) is true.



P-value: 0.

← ignore the colors

$$\rightarrow H_0: P(H) = P(T)$$

$$\rightarrow H_a: P(H) > P(T)$$

binomial
distribution
← #heads

exp → 100 times → 53 heads

exp → 100 times

30 times

In simple words p-value is a measure of the strength of the evidence against the Null Hypothesis that is provided by our sample data.

Null hyp
100 cap → 80 times → 0

2 times

$P = 0.3$

53 A

Interpreting p-value

06 April 2023 08:25

reject your H_0

With significance value

$$\alpha = 0.05 / 0.01 \rightarrow p\text{-value} \leq \alpha$$

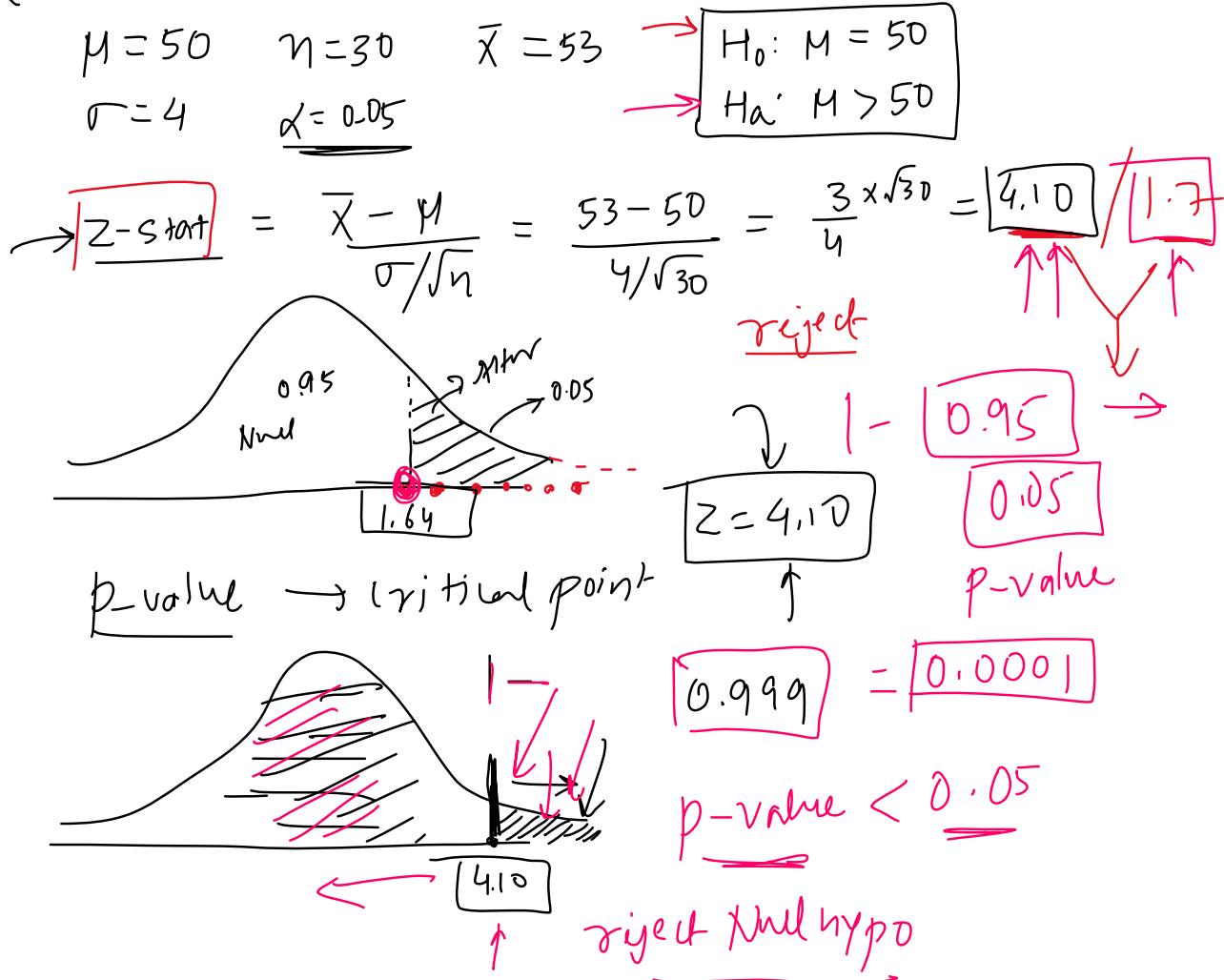
Without significance value

$$\alpha \rightarrow 0.05$$

1. Very small p-values (e.g., $p < 0.01$) indicate strong evidence against the null hypothesis, suggesting that the observed effect or difference is unlikely to have occurred by chance alone.
2. Small p-values (e.g., $0.01 \leq p < 0.05$) indicate moderate evidence against the null hypothesis, suggesting that the observed effect or difference is less likely to have occurred by chance alone.
3. Large p-values (e.g., $0.05 \leq p < 0.1$) indicate weak evidence against the null hypothesis, suggesting that the observed effect or difference might have occurred by chance alone, but there is still some level of uncertainty.
4. Very large p-values (e.g., $p \geq 0.1$) indicate weak or no evidence against the null hypothesis, suggesting that the observed effect or difference is likely to have occurred by chance alone.

rejection region approach p-value approach

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was 50 units per day. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day and the pop std is 4. The company wants to know if the new training program has significantly increased productivity.



Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a pop standard deviation of 5 grams.

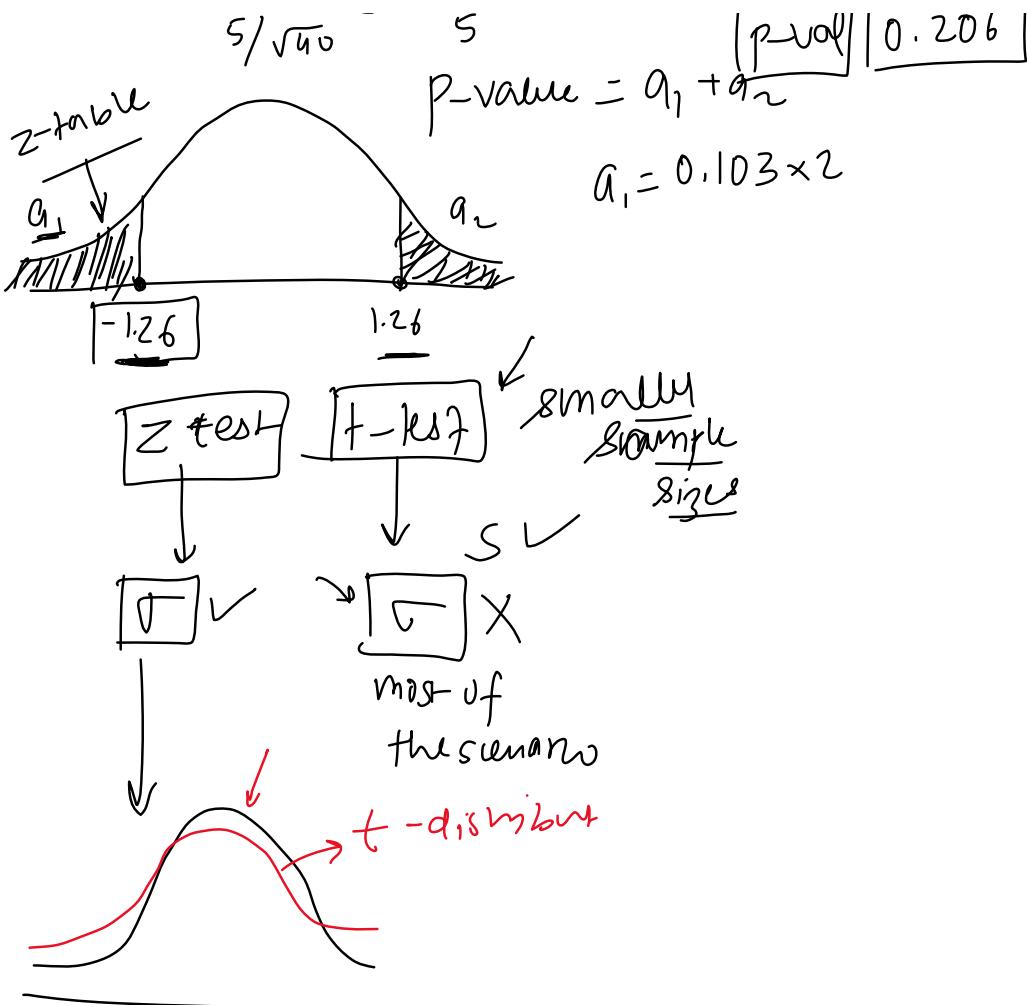
$$\begin{array}{lll} \mu = 50 & n = 40 & \bar{x} = 49 \\ \sigma = 5 & \alpha = 0.05 & \end{array}$$

$H_0: \mu = 50 \quad H_a: \mu \neq 50$

p-value 2tailed $0.206 > 0.05$

$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{49 - 50}{5/\sqrt{40}} = -\frac{\sqrt{10}}{5} = -1.26$

p-value $= \alpha_1 + \alpha_2$ 0.206



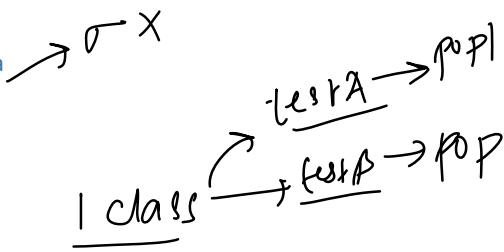
T-tests

06 April 2023 14:14

A t-test is a statistical test used in hypothesis testing to compare the means of two samples or to compare a sample mean to a known population mean. The t-test is based on the t-distribution, which is used when the population standard deviation is unknown and the sample size is small.

There are three main types of t-tests:

$$| \text{sample} \rightarrow \bar{x} \rightarrow M$$



One-sample t-test: The one-sample t-test is used to compare the mean of a single sample to a known population mean. The null hypothesis states that there is no significant difference between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.

Independent two-sample t-test: The independent two-sample t-test is used to compare the means of two independent samples. The null hypothesis states that there is no significant difference between the means of the two samples, while the alternative hypothesis states that there is a significant difference.

Paired t-test (dependent two-sample t-test): The paired t-test is used to compare the means of two samples that are dependent or paired, such as pre-test and post-test scores for the same group of subjects or measurements taken on the same subjects under two different conditions. The null hypothesis states that there is no significant difference between the means of the paired differences, while the alternative hypothesis states that there is a significant difference.

Single Sample t-test

06 April 2023 14:14

t-test

$t \times s \sqrt{n}$

A one-sample t-test checks whether a sample mean differs from the population mean.

Assumptions for a single sample t-test

1. Normality - Population from which the sample is drawn is normally distributed
2. Independence - The observations in the sample must be independent, which means that the value of one observation should not influence the value of another observation.
3. Random Sampling - The sample must be a random and representative subset of the population.
4. Unknown population std - The population std is not known.

sample normally
distrib

Suppose a manufacturer claims that the average weight of their new chocolate bars is 50 grams, we highly doubt that and want to check this so we drew out a sample of 25 chocolate bars and measured their weight, the sample mean came out to be 49.7 grams and the sample std deviation was 1.2 grams. Consider the significance level to be 0.05

$$\mu = 50 \quad n = 25$$

$$\bar{x} = 49.7 \quad \alpha = 0.05$$

$$s = 1.2$$



$$H_0: \mu = 50$$

$$H_a: \mu \neq 50$$

assuming it is normal

$$\frac{t}{t} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{49.7 - 50}{\frac{1.2}{\sqrt{25}}} = \frac{-0.3 \times 5}{1.2} = \frac{-1.5}{1.2} = -1.25$$

$$df = n - 1 = 24$$

$$H_0: \mu = 35 \rightarrow$$

$$H_a: \mu < 35$$

$$\mu = 35$$

$$\bar{x}, s, \alpha = 0.05$$

25 sample \rightarrow normal
age

t-test
Shapiro-Wilk
tstat

p-value < 0.05 not normal
p-value > 0.05 normal

Python Case Study 1

06 April 2023 17:27

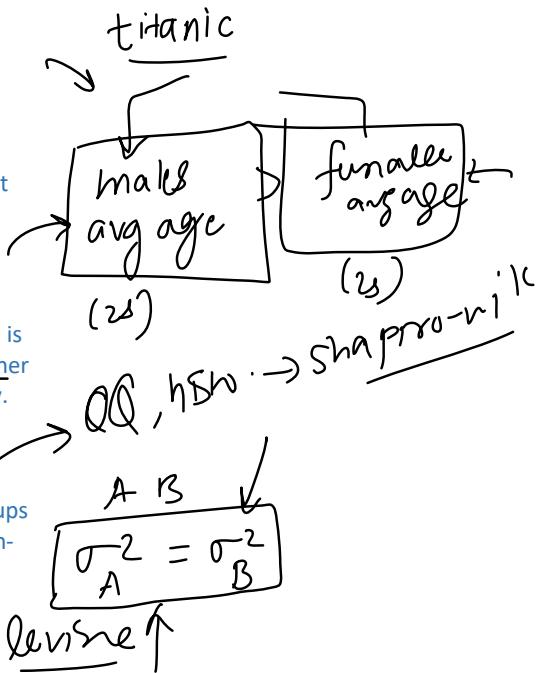
Independent 2 sample t-test

06 April 2023 14:15

An independent two-sample t-test, also known as an unpaired t-test, is a statistical method used to compare the means of two independent groups to determine if there is a significant difference between them.

Assumptions for the test:

1. Independence of observations: The two samples must be independent, meaning there is no relationship between the observations in one group and the observations in the other group. The subjects in the two groups should be selected randomly and independently.
2. Normality: The data in each of the two groups should be approximately normally distributed. The t-test is considered robust to mild violations of normality, especially when the sample sizes are large (typically $n \geq 30$) and the sample sizes of the two groups are similar. If the data is highly skewed or has substantial outliers, consider using a non-parametric test, such as the Mann-Whitney U test.
3. Equal variances (Homoscedasticity): The variances of the two populations should be approximately equal. This assumption can be checked using F-test for equality of variances. If this assumption is not met, you can use Welch's t-test, which does not require equal variances.
4. Random sampling: The data should be collected using a random sampling method from the respective populations. This ensures that the sample is representative of the population and reduces the risk of selection bias.



p-value < 0.05

$\sigma_A^2 \neq \sigma_B^2$

p-value > 0.05

$\sigma_A^2 = \sigma_B^2$

reject my H_0

Suppose a website owner claims that there is no difference in the average time spent on their website between desktop and mobile users. To test this claim, we collect data from 30 desktop users and 30 mobile users regarding the time spent on the website in minutes. The sample statistics are as follows:

desktop users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14] → avg-time

mobile_users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11]

Desktop users:

- o Sample size (n_1): 30
- o Sample mean (mean1): 18.5 minutes
- o Sample standard deviation (std_dev1): 3.5 minutes

Mobile users:

- o Sample size (n_2): 30
- o Sample mean (mean2): 14.3 minutes
- o Sample standard deviation (std_dev2): 2.7 minutes

We will use a significance level (α) of 0.05 for the hypothesis test.

avg-time

$$H_0: \mu_d = \mu_m$$

$$H_a: \mu_d \neq \mu_m$$

check assumptions

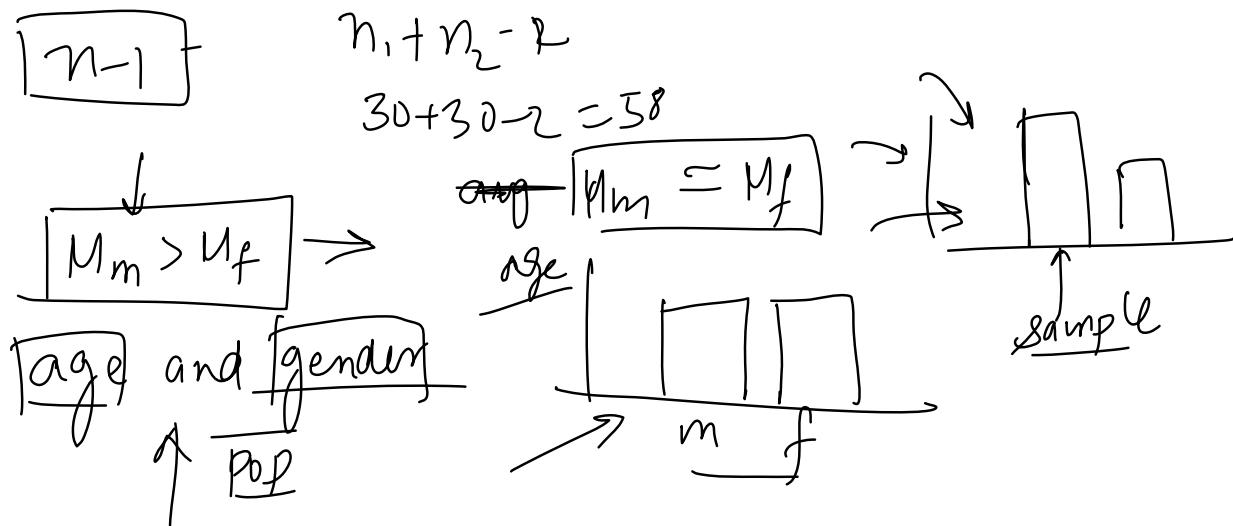
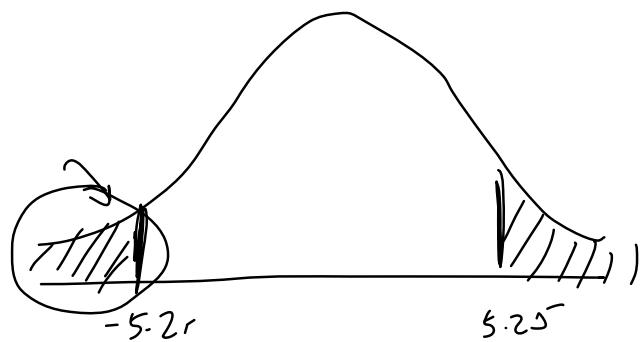
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-statistic

$$t = \frac{18.5 - 14.3}{\sqrt{\frac{3.5^2}{30} + \frac{2.7^2}{30}}} = 5.25$$

$$t = \frac{\overline{X} - \mu}{\sqrt{\frac{(3.5)^2}{30} + \frac{(2.7)^2}{30}}} = \frac{4.2}{\sqrt{\frac{19.5}{30}}} = 0.154$$



Python Case Study 2

06 April 2023 17:27

Paired 2 sample t-test

06 April 2023 14:21

A paired two-sample t-test, also known as a dependent or paired-samples t-test, is a statistical test used to compare the means of two related or dependent groups.

Common scenarios where a paired two-sample t-test is used include:

1. Before-and-after studies: Comparing the performance of a group before and after an intervention or treatment.
2. Matched or correlated groups: Comparing the performance of two groups that are matched or correlated in some way, such as siblings or pairs of individuals with similar characteristics.

Assumptions

1. Paired observations: The two sets of observations must be related or paired in some way, such as before-and-after measurements on the same subjects or observations from matched or correlated groups.
2. Normality: The differences between the paired observations should be approximately normally distributed. This assumption can be checked using graphical methods (e.g., histograms, Q-Q plots) or statistical tests for normality (e.g., Shapiro-Wilk test). Note that the t-test is generally robust to moderate violations of this assumption when the sample size is large.
3. Independence of pairs: Each pair of observations should be independent of other pairs. In other words, the outcome of one pair should not affect the outcome of another pair. This assumption is generally satisfied by appropriate study design and random sampling.

	I	II	d
A	50	55	-5
B	60	60	0
C	76	66	10
D	40	60	-20
E	25	100	-75

normal

Let's assume that a fitness center is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants' weight.

Before the program:

[80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91]

After the program:

[78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]

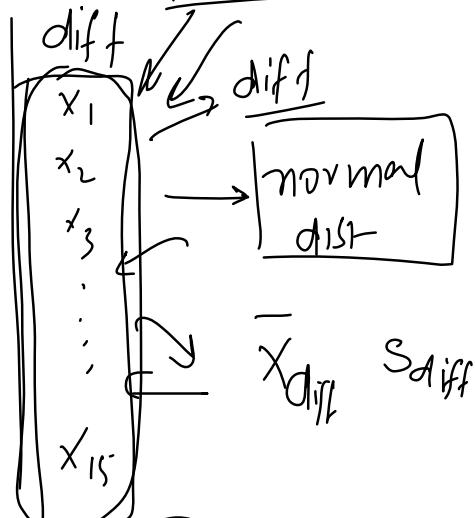
Significance level (α) = 0.05

$$H_0: \mu_{\text{before}} - \mu_{\text{after}} = 0$$

$$H_1: \mu_{\text{before}} > \mu_{\text{after}}$$

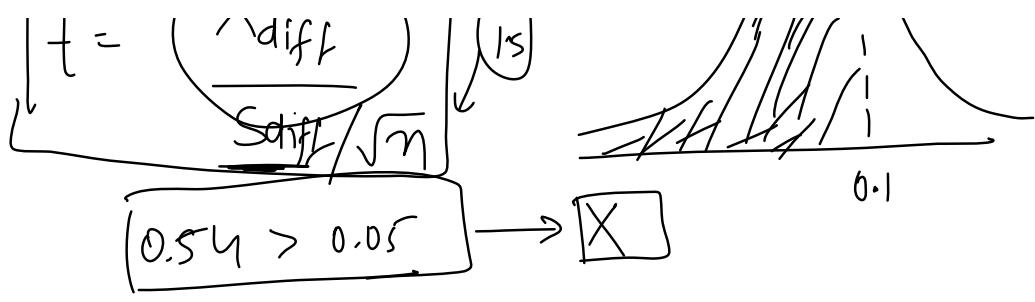
> 0.05

name	wt before	wt after
A	80	78
B	92	93
C	75	71
.	71	71
;	71	71
K	91	88



$$t = \frac{\bar{x}_{\text{diff}}}{s_{\text{diff}}}$$

ls



Chi Square Distribution

07 April 2023 10:29

→ distribution → continuous pd

$$X \sim N(0,1)$$

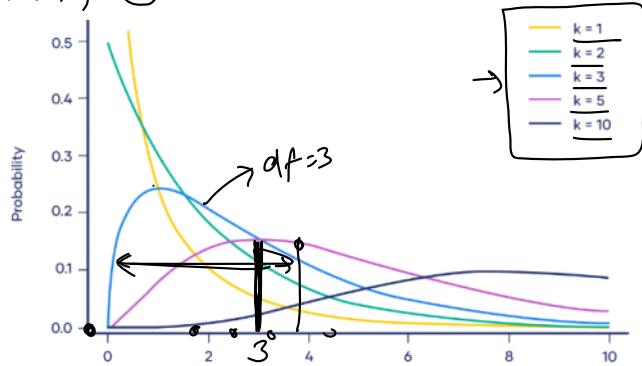
The Chi-Square distribution, also written as χ^2 distribution, is a continuous probability distribution that is widely used in statistical hypothesis testing, particularly in the context of goodness-of-fit tests and tests for independence in contingency tables. It arises when the sum of the squares of independent standard normal random variables follows this distribution.

The Chi-Square distribution has a single parameter, the degrees of freedom (df), which influences the shape and spread of the distribution. The degrees of freedom are typically associated with the number of independent variables or constraints in a statistical problem.

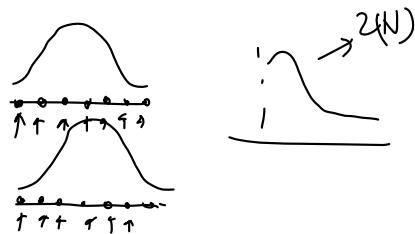
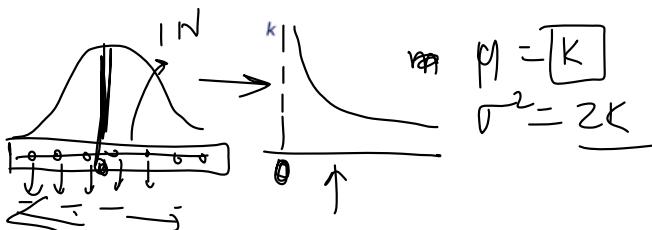
Some key properties of the Chi-Square distribution are:

- It is a continuous distribution, defined for non-negative values.
- It is positively skewed, with the degree of skewness decreasing as the degrees of freedom increase.
- The mean of the Chi-Square distribution is equal to its degrees of freedom, and its variance is equal to twice the degrees of freedom.
- As the degrees of freedom increase, the Chi-Square distribution approaches the normal distribution in shape.

The Chi-Square distribution is used in various statistical tests, such as the Chi-Square goodness-of-fit test which evaluates whether an observed frequency distribution fits an expected theoretical distribution, and the Chi-Square test for independence which checks the association between categorical variables in a contingency table.



$$\sqrt{z}$$



$$\chi^2 = |Z^2| \rightarrow \text{degree of freedom} \\ df = 1$$

$$\chi^2 = Z_1^2 + Z_2^2 \rightarrow df = 2$$

$$\chi^2 = Z_1^2 + Z_2^2 + Z_3^2 \rightarrow df = 3$$

$$\chi^2 = \sum_{i=1}^k Z_i^2 \quad \boxed{df = k}$$

$df \uparrow$

Chi Square Test

07 April 2023 15:03

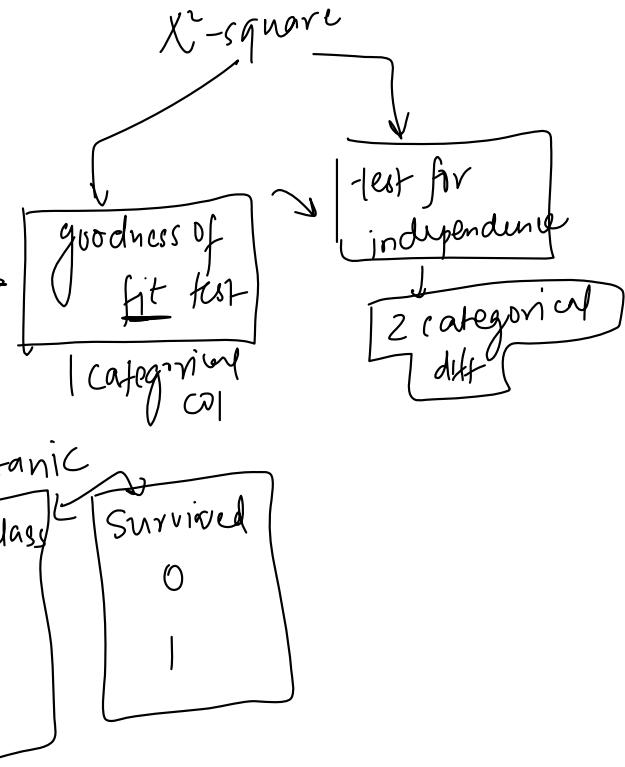
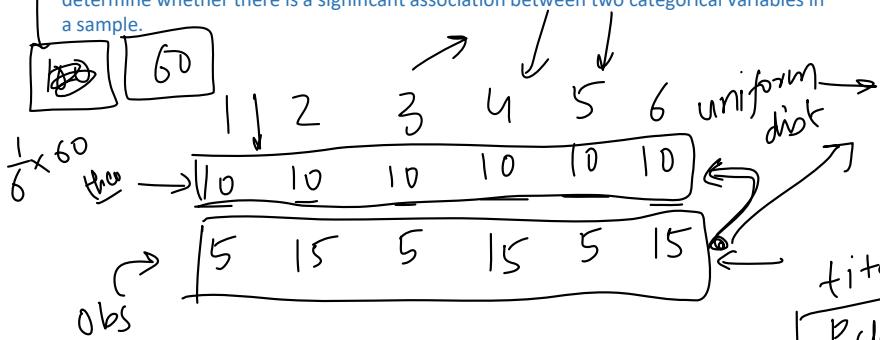
→ Categorical ←

Z -test t -test

continuous

The Chi-Square test is a statistical hypothesis test used to determine if there is a significant association between categorical variables or if an observed distribution of categorical data differs from an expected theoretical distribution. It is based on the Chi-Square (χ^2) distribution, and it is commonly applied in two main scenarios:

1. Chi-Square Goodness-of-Fit Test: This test is used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution. It is often applied to check if the data follows a specific probability distribution, such as the uniform or binomial distribution.
2. Chi-Square Test for Independence (Chi-Square Test for Association): This test is used to determine whether there is a significant association between two categorical variables in a sample.



P class	Survived
1	0
2	1
3	

Goodness of Fit Test

07 April 2023 10:29

The Chi-Square Goodness-of-Fit test is a statistical hypothesis test used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution. It helps to evaluate whether the data follows a specific probability distribution, such as uniform, binomial, or Poisson distribution, among others. This test is particularly useful when you want to assess if the sample data is consistent with an assumed distribution or if there are significant deviations from the expected pattern.

Steps

The Chi-Square Goodness-of-Fit test involves the following steps:

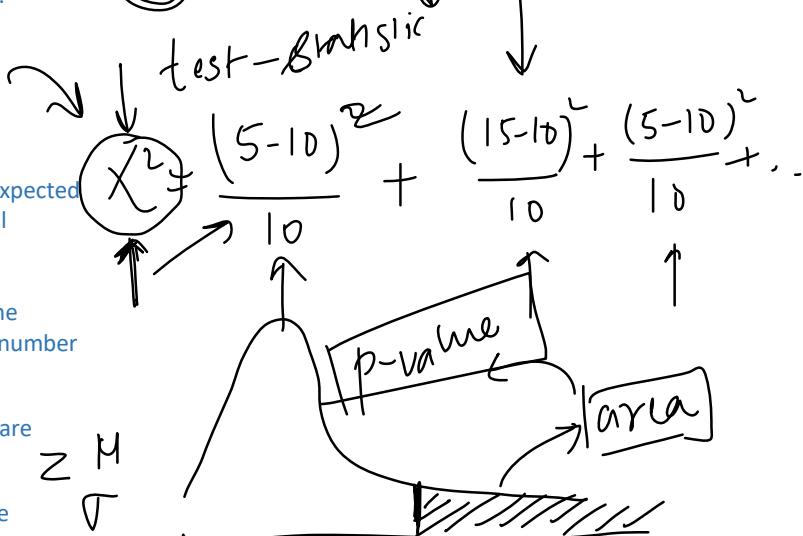
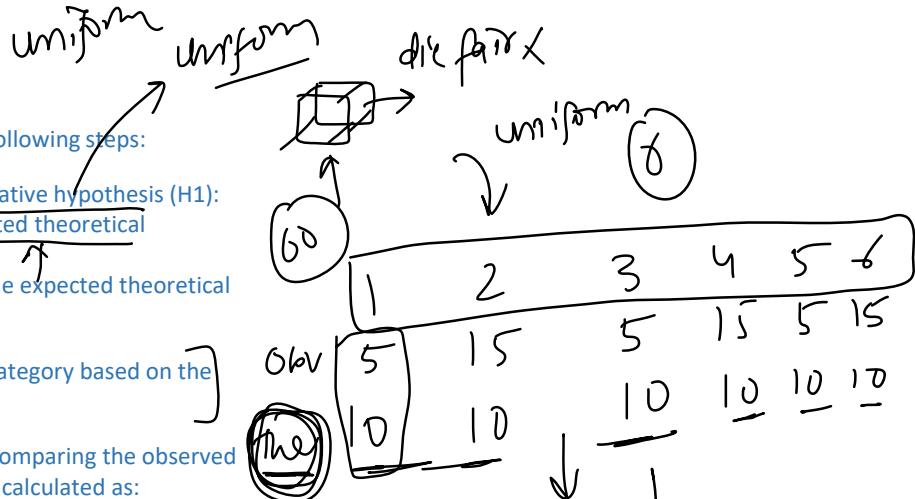
- Define the null hypothesis (H_0) and the alternative hypothesis (H_1):
 - H_0 : The observed data follows the expected theoretical distribution.
 - H_1 : The observed data does not follow the expected theoretical distribution.
- Calculate the expected frequencies for each category based on the theoretical distribution and the sample size.
- Compute the Chi-Square test statistic (χ^2) by comparing the observed and expected frequencies. The test statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- where O_i is the observed frequency in category i , E_i is the expected frequency in category i , and the summation is taken over all categories.
- Determine the degrees of freedom (df), which is typically the number of categories minus one ($df = k - 1$), where k is the number of categories.
- Calculate the p-value for the test statistic using the Chi-Square distribution with the calculated degrees of freedom.
- Compare the test statistic to the critical value or the p-value

Assumptions

- Independence:** The observations in the sample must be independent of each other. This means that the outcome of one observation should not influence the outcome of another observation.
- Categorical data:** The variable being analysed must be categorical, not continuous or ordinal. The data should be divided into mutually exclusive and exhaustive categories.
- Expected frequency:** Each category should have an expected frequency of at least 5. This guideline helps ensure that the Chi-Square distribution is a reasonable approximation for the distribution of the test statistic. Having small expected frequencies can lead to an inaccurate estimation of the Chi-Square distribution, potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).



$$t - \text{test} \quad t = \bar{x}$$

$$df = n - 1 = 5 \quad \alpha = 0.05$$

$$0.06$$

Potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).

4. Fixed distribution: The theoretical distribution being compared to the observed data should be specified before the test is conducted. It is essential to avoid choosing a distribution based on the observed data, as doing so can lead to biased results.

parametric or
non-parametric

The Chi-Square Goodness-of-Fit test is a non-parametric test. Non-parametric tests do not assume that the data comes from a specific probability distribution or make any assumptions about population parameters like the mean or standard deviation.

In the Chi-Square Goodness-of-Fit test, we compare the observed frequencies of the categorical data to the expected frequencies based on a hypothesized distribution. The test doesn't rely on any assumptions about the underlying distribution's parameters. Instead, it focuses on comparing observed counts to expected counts, making it a non-parametric test.

Example 1

07 April 2023 16:26

Suppose we have a six-sided fair die, and we want to test if the die is indeed fair. We roll the die 60 times and record the number of times each side comes up. We'll use the Chi-Square Goodness-of-Fit test to determine if the observed frequencies are consistent with a fair die (i.e., a uniform distribution of the sides).

Observed frequencies:

- Side 1: 12 times
- Side 2: 8 times
- Side 3: 11 times
- Side 4: 9 times
- Side 5: 10 times
- Side 6: 10 times

H_0 : die is fair \rightarrow uniform

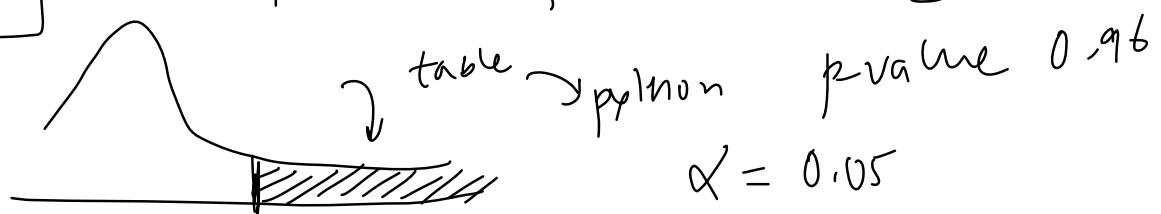
H_1 : die is not fair

expected

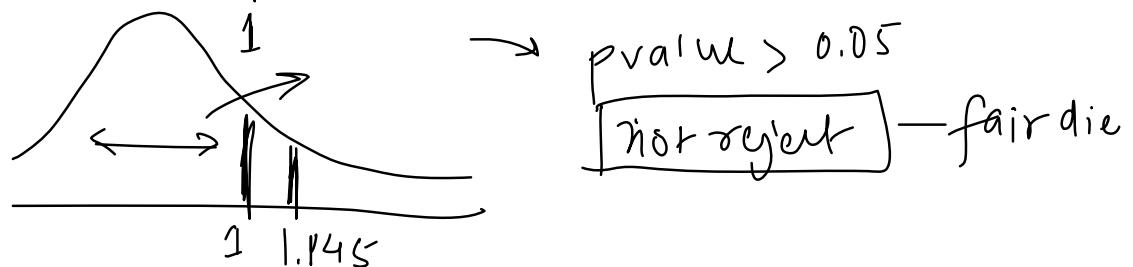
	1	2	3	4	5	6	
Obs	12	8	11	9	10	10	$\rightarrow 60$
	10	10	10	10	10	10	

$$\chi^2 = \frac{(12-10)^2 + (8-10)^2 + (11-10)^2 + (9-10)^2}{10} = \frac{4+4+1+1}{10} = 1$$

$$\boxed{\chi^2 = 1} \rightarrow \text{chisquare} \quad df = n-1 = 6-1 = 5$$



$$\alpha = 0.05$$



Example 2

07 April 2023 15:26

uniform

Suppose a marketing team at a retail company wants to understand the distribution of visits to their website by day of the week. They have a hypothesis that visits are uniformly distributed across all days of the week, meaning they expect an equal number of visits on each day. They collected data on website visits for four weeks and want to test if the observed distribution matches the expected uniform distribution.

Observed frequencies (number of website visits per day of the week for four weeks):

- Monday: 420
- Tuesday: 380
- Wednesday: 410
- Thursday: 400
- Friday: 410
- Saturday: 430
- Sunday: 390

	mon	tu e	wed	thu	fri	sat	SUN
Obs	420	380	410	400	410	430	390
EXP	405	405	405	405	405	405	405

H_0 : Uniform dis

H_a : Not uniform

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(420 - 405)^2 + (380 - 405)^2 + (410 - 405)^2 + (400 - 405)^2 + (410 - 405)^2 + (430 - 405)^2 + (390 - 405)^2}{405}$$

$$\chi^2 = [LB]$$

$$df = n - 1 = 7 - 1 = 6$$



[0.08] → p-value

0.08 < α →

Example 3

07 April 2023 15:27

survey \rightarrow village \rightarrow 800 families

A survey of 800 families in a village with 4 children each revealed the following distribution:

# girls	4	3	2	1	0
# boys	0	1	2	3	4
	↑1	2	3	4	5
# families	32	178	290	236	64
theoretical					

binomial

4 children

150h

75%

$$P(S) = P(d) = \frac{1}{2}$$

$$\left\{ \begin{array}{l} H_0: P(m) = P(f) = \frac{1}{2} \\ H_a: P(m) \neq P(f) \end{array} \right\} \xrightarrow{\text{binomial}} p = \frac{1}{2}$$

$$P_g = 1 - p$$

$$\rightarrow n=4, p=\frac{1}{2}$$

$${}^n C_x p^x (1-p)^{n-x}$$

$$P(0) = {}^4 C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \frac{1}{16} \times 800 = 50$$

$$P(1) = {}^4 C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \frac{4!}{3!} \times \frac{1}{16} = \frac{1}{4} \times 800 = 200$$

$$P(2) = {}^4 C_2 \left(\frac{1}{2}\right)^2 = \frac{4!}{2!2!} \times \frac{1}{16} = \frac{6}{16} \times 800$$

$$P(3) = {}^4 C_3 \left(\frac{1}{2}\right)^3 = \frac{4!}{1!3!} \times \frac{1}{16} = \frac{4 \times 3}{16} \times \frac{1}{16} \times 800$$

$$\frac{6}{16} \times 800$$

	0	1	2	3	4
Obs	32	178	290	236	64
Theo	50	200	300	200	50

$$\chi^2 = \frac{(32-50)^2}{50} + \frac{(178-200)^2}{200} + \frac{(290-300)^2}{300} + \frac{(236-200)^2}{200} + \frac{(64-50)^2}{50}$$

$$= \frac{324}{50} + \frac{484}{200} + \frac{100}{300} + \frac{1296}{200} + \frac{196}{50}$$

$$= 6.2 + 2.3 + 0.33 + 6.2 + 3.9$$

$$\chi^2 = \frac{18.93}{\uparrow} \quad df = 5-1 = 4$$

$$0.00081 < \alpha(0.05)$$

reject the Null hypothesis

Python Case Study

07 April 2023 15:27

$$\begin{array}{cccc} 1 & 2 & 3 & \\ \overbrace{216} & \underline{184} & \underline{491} & \text{obs} \\ \overline{299} & \cancel{299} & 270 & \text{exp} \\ x = & & & \end{array}$$

Test for Independence

07 April 2023 10:29

The Chi-Square test for independence, also known as the Chi-Square test for association, is a statistical test used to determine whether there is a significant association between two categorical variables in a sample. It helps to identify if the occurrence of one variable is dependent on the occurrence of the other variable, or if they are independent of each other.

The test is based on comparing the observed frequencies in a contingency table (a table that displays the frequency distribution of the variables) with the frequencies that would be expected under the assumption of independence between the two variables.

Steps

1. State the null hypothesis (H_0) and alternative hypothesis (H_1):

- H_0 : There is no association between the two categorical variables (they are independent).
- H_1 : There is an association between the two categorical variables (they are dependent).

2. Create a contingency table with the observed frequencies for each combination of the categories of the two variables.

3. Calculate the expected frequencies for each cell in the contingency table assuming that the null hypothesis is true (i.e., the variables are independent).

4. Compute the Chi-Square test statistic:

$$\chi^2 = \sum [(O_{ij} - E_{ij})^2 / E_{ij}]$$

where O_{ij} is the observed frequency in each cell and E_{ij} is the expected frequency.

$$(2-1)(3-1) \\ 1 \times 2 = 2 \\ df$$

5. Determine the degrees of freedom: $df = (\text{number of rows} - 1) * (\text{number of columns} - 1)$

6. Obtain the critical value or p-value using the Chi-Square distribution table or a statistical software/calculator with the given degrees of freedom and significance level (commonly $\alpha = 0.05$).

7. Compare the test statistic to the critical value or the p-value to the significance level to decide whether to reject or fail to reject the null hypothesis. If the test statistic is greater than the critical value, or if the p-value is less than the significance level, we reject the null hypothesis and conclude that there is a significant association between the two variables.

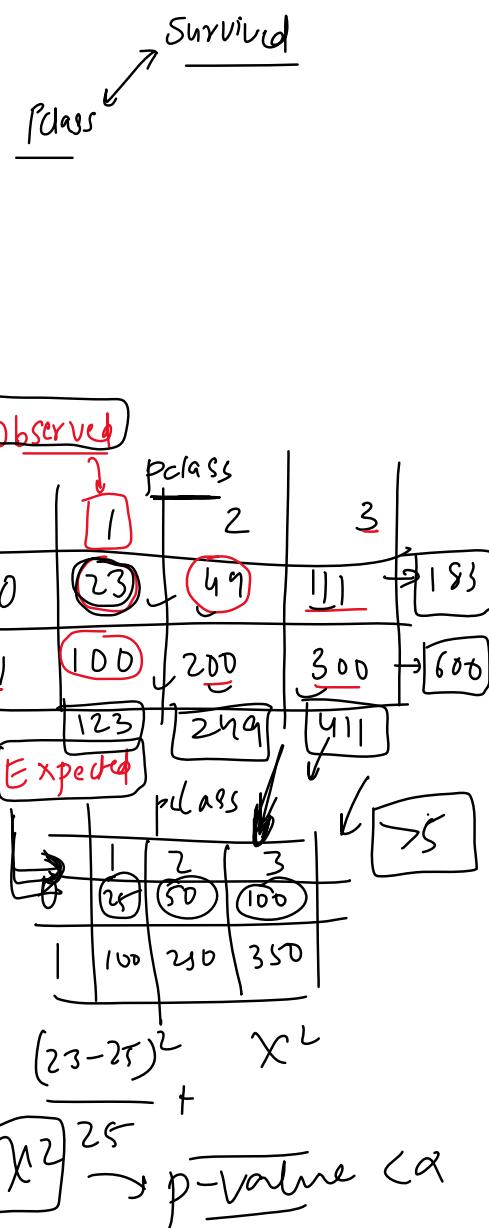
Assumptions

1. Independence of observations: The observations in the sample should be independent of each other. This means that the occurrence of one observation should not affect the occurrence of another observation. In practice, this usually implies that the data should be collected using a simple random sampling method.

2. Categorical variables: Both variables being tested must be categorical, either ordinal or nominal. The Chi-Square test for independence is not appropriate for continuous variables.

3. Adequate sample size: The sample size should be large enough to ensure that the expected frequency for each cell in the contingency table is sufficient. A common rule of thumb is that the expected frequency for each cell should be at least 5. If some cells have expected frequencies less than 5, the test may not be valid, and other methods like Fisher's exact test may be more appropriate.

4. Fixed marginal totals: The marginal totals (the row and column sums of the contingency table) should be fixed before the data is collected. This is because the Chi-Square test for independence assesses the association between the two variables under the assumption that the marginal totals are fixed and not influenced by the relationship between the variables.



Example 1

07 April 2023 17:50

A researcher wants to investigate if there is an association between the level of education (categorical variable) and the preference for a particular type of exercise (categorical variable) among a group of 150 individuals. The researcher collects data and creates the following contingency table

Education	Exercise Type			Total
	Yoga	Running	Swimming	
High School	15	20	10	45
Bachelor's	20	30	15	65
Master's or PhD	5	15	20	40
Total	40	65	45	150

Observed ✓

edu \leftrightarrow exercise (independ)

H_0 : they are independent X

H_1 : they are associated

need a criterion

$$\frac{45 \times 40}{150} \times \frac{65 \times 45}{150} \times \frac{45 \times 45}{150}$$

expected

	Yoga	Run	swim
High	12	19	13.5
Bach	17	28	14
phd	10	17	12

$$\frac{(15-12)^2}{12} + \frac{(20-19)^2}{20} + \frac{(10-13.5)^2}{10}$$

p-value 0.04 (< α)

$$\chi^2 = 9.95$$

reject null df = $(3-1)(3-1) = 4$

Python Case Study

07 April 2023 15:06

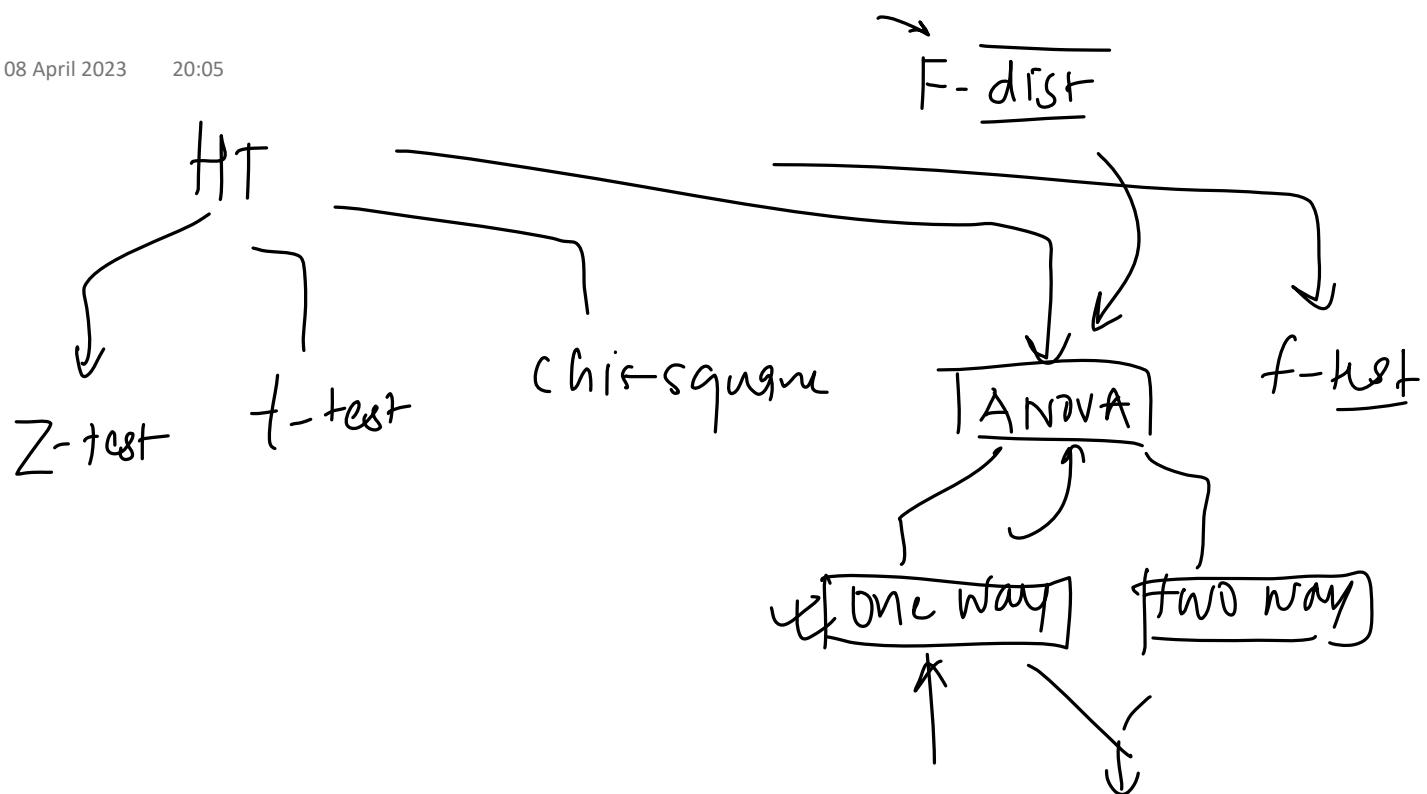
Applications in Machine Learning

07 April 2023 17:30



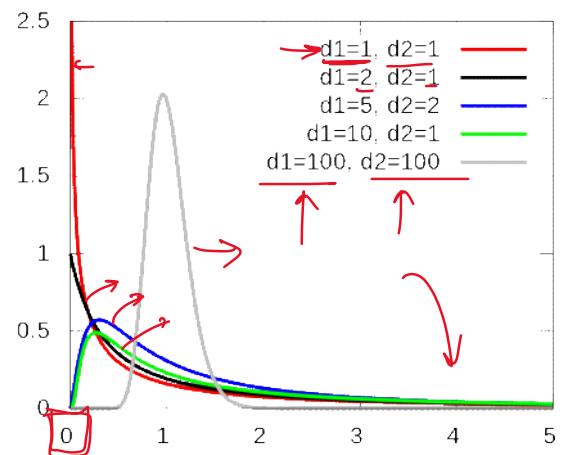
- 1 Feature selection: Chi-Square test can be used as a filter-based feature selection method to rank and select the most relevant categorical features in a dataset. By measuring the association between each categorical feature and the target variable, you can eliminate irrelevant or redundant features, which can help improve the performance and efficiency of machine learning models.
- 2 Evaluation of classification models: For multi-class classification problems, the Chi-Square test can be used to compare the observed and expected class frequencies in the confusion matrix. This can help assess the goodness of fit of the classification model, indicating how well the model's predictions align with the actual class distributions.
- 3 Analysing relationships between categorical features: In exploratory data analysis, the Chi-Square test for independence can be applied to identify relationships between pairs of categorical features. Understanding these relationships can help inform feature engineering and provide insights into the underlying structure of the data.
- 4 Discretization of continuous variables: When converting continuous variables into categorical variables (binning), the Chi-Square test can be used to determine the optimal number of bins or intervals that best represent the relationship between the continuous variable and the target variable.
- 5 Variable selection in decision trees: Some decision tree algorithms, such as the CHAID (Chi-squared Automatic Interaction Detection) algorithm, use the Chi-Square test to determine the most significant splitting variables at each node in the tree. This helps construct more effective and interpretable decision trees.

AGF



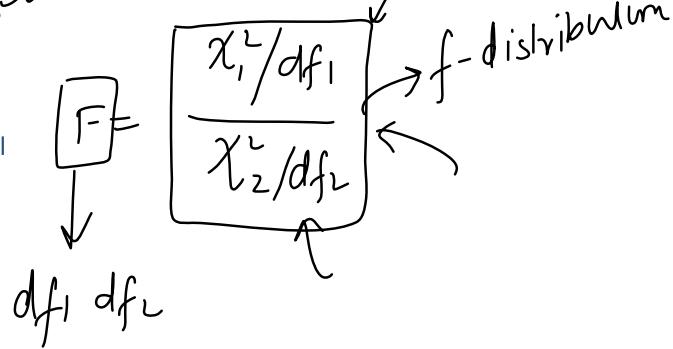
→ chi-square → (normal)²

- Continuous probability distribution:** The F-distribution is a continuous probability distribution used in statistical hypothesis testing and analysis of variance (ANOVA).
- Fisher-Snedecor distribution:** It is also known as the Fisher-Snedecor distribution, named after Ronald Fisher and George Snedecor, two prominent statisticians.
- Degrees of freedom:** The F-distribution is defined by two parameters - the degrees of freedom for the numerator (df_1) and the degrees of freedom for the denominator (df_2).
- Positively skewed and bounded:** The shape of the F-distribution is positively skewed, with its left bound at zero. The distribution's shape depends on the values of the degrees of freedom.
- Testing equality of variances:** The F-distribution is commonly used to test hypotheses about the equality of two variances in different samples or populations.
- Comparing statistical models:** The F-distribution is also used to compare the fit of different statistical models, particularly in the context of ANOVA.
- F-statistic:** The F-statistic is calculated by dividing the ratio of two sample variances or mean squares from an ANOVA table. This value is then compared to critical values from the F-distribution to determine statistical significance.
- Applications:** The F-distribution is widely used in various fields of research, including psychology, education, economics, and the natural and social sciences, for hypothesis testing and model comparison.



SSB $\xrightarrow{\chi^2} df_1$

SSW $\xrightarrow{\chi^2} df_2$

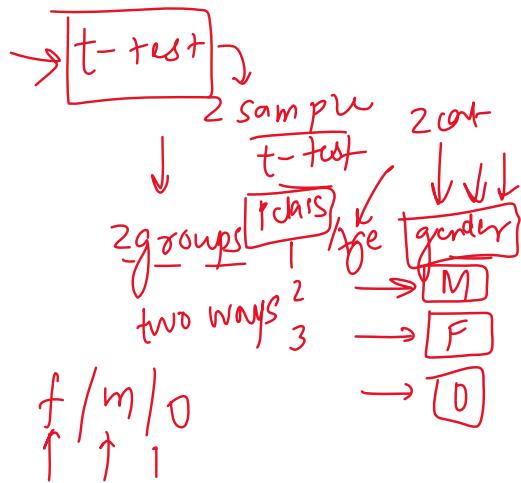


One way ANOVA test

08 April 2023 13:12

One-way ANOVA [Analysis of Variance] is a statistical method used to compare the means of three or more independent groups to determine if there are any significant differences between them. It is an extension of the t-test, which is used for comparing the means of two independent groups. The term "one-way" refers to the fact that there is only one independent variable (factor) with multiple levels (groups) in this analysis.

The primary purpose of one-way ANOVA is to test the null hypothesis that all the group means are equal. The alternative hypothesis is that at least one group mean is significantly different from the others.



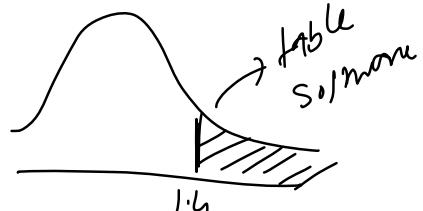
Steps

- Define the null and alternative hypotheses.
- Calculate the overall mean (grand mean) of all the groups combined and mean of all the groups individually.
- Calculate the "between-group" and "within-group" sum of squares (SS).
- Find the between group and within group degree of freedoms
- Calculate the "between-group" and "within-group" mean squares (MS) by dividing their respective sum of squares by their degrees of freedom.
- Calculate the F-statistic by dividing the "between-group" mean square by the "within-group" mean square.

ANOVA table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS) (This is SS Divided by d.f.) and is an Estimation of Variance to be Used in F-ratio	F-ratio	t est statistic
Between samples or categories	$\sum n_i (\bar{X}_i - \bar{X})^2$	$(k-1)$	$\frac{SS \text{ between}}{(k-1)}$	$f = \frac{\text{MS between}}{\text{MS within}}$	$\frac{SSB}{df_{ssb}}$
Within samples or categories	$\sum (X_{ij} - \bar{X}_i)^2 + \dots + \sum (X_{jk} - \bar{X}_k)^2$	$(n-k)$	$\frac{SS \text{ within}}{(n-k)}$	$f = \frac{SSW}{df_{ssw}}$	$\frac{SSW}{df_{ssw}}$
Total	$\sum (X_{ij} - \bar{X})^2$	$(n-1)$			$f\text{-dist}$

- Calculate the p-value associated with the calculated F-statistic using the F-distribution and the appropriate degrees of freedom. The p-value represents the probability of obtaining an F-statistic as extreme or more extreme than the calculated value, assuming the null hypothesis is true.
 - Choose a significance level (α), typically 0.05.
 - Compare the calculated p-value with the chosen significance level (α).
- a. If the p-value is less than or equal to α , reject the null hypothesis in favour of the alternative hypothesis, concluding that there is a significant difference between at least one pair of group means.
- b. If the p-value is greater than α , fail to reject the null hypothesis, concluding that there is not enough evidence to suggest a significant difference between the group means.



It's important to note that one-way ANOVA only determines if there is a significant difference between the group means; it does not identify which specific groups have significant differences. To determine which pairs of groups are significantly different, post-hoc tests, such as Tukey's HSD or Bonferroni, are conducted after a significant ANOVA result.

Example 1

08 April 2023 13:22

A	B	C
3	1	8
6	8	6
3	9	10

number count

ANOVA

$$H_0: \mu_A = \mu_B = \mu_C$$

H_1 : at least 1 of mean is significant

$$n = 9 \quad k = 3$$

$$\bar{x} = 6 \quad \bar{x}_A = 4 \quad \bar{x}_B = 6 \quad \bar{x}_C = 8$$

$$[SST] \rightarrow \text{sum of square Total} \rightarrow df = n-1 = 9-1 = 8$$

$$(6-3)^2 + (6-6)^2 + (6-3)^2 + (6-1)^2 + (6-8)^2 + (6-8)^2 + (6-6)^2 + (6-10)^2$$

$$9 + 0 + 9 + 25 + 4 + 9 + 4 + 16 = 76$$

$$[SSW] \rightarrow \text{sum of squares within} \quad \downarrow df = 6 \quad n-k = 9-3 = 6$$

A	B	C
3	1	8
6	8	6
3	9	10

$$(4-3)^2 + (4-6)^2 + (4-3)^2 +$$

$$(6-1)^2 + (6-8)^2 + (6-9)^2 +$$

$$(8-8)^2 + (8-1)^2 + (8-10)^2$$

$$SSW = 52$$

$$1 + 9 + 1 + 25 + 4 + 9 + 4 + 4$$

A	B	C
3	1	8
6	8	6
3	9	10

$$SSB \rightarrow df = 2$$

$$3 \times \underline{(6-4)^2} + 3 \times (6-6)^2 + 3 \times (6-8)^2$$

$$3 \times 4 + 0 + 3 \times 4 = 24 = SSB$$

quantity

error

value

71.

df

12

$$F = \frac{24}{2} = 12$$

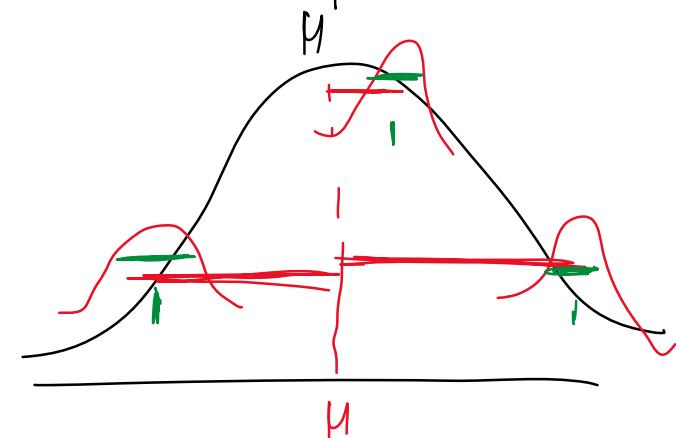
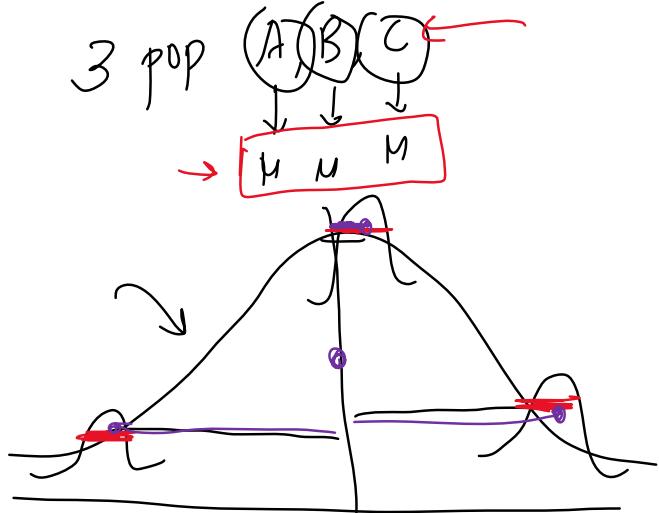
quantity

$$\rightarrow SSB$$

$$\rightarrow \underline{SSW}$$

$$\rightarrow \boxed{SST}$$

$$SSW = \frac{24}{52} = \frac{72}{12 \times 6}$$



$$IT$$

$$\begin{array}{|c|} \hline 2 \\ \hline 6 \\ \hline \end{array}$$

$$F = \frac{24}{\frac{2}{52}} = \frac{12 \times 6}{6} = \frac{72}{6}$$

$$f = 1.4$$

interval variance groups (SSW)

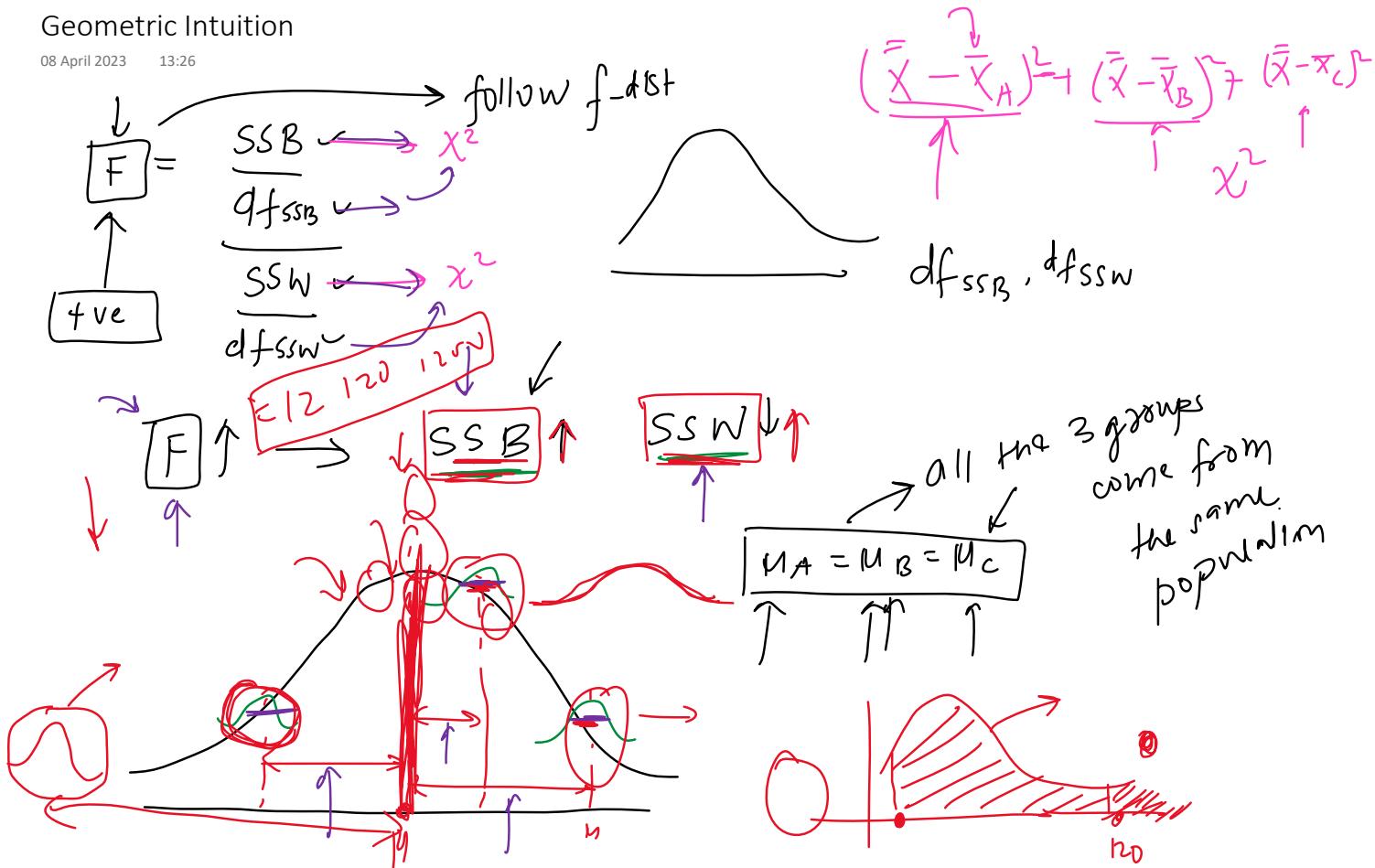
p-value = 0.31 > $\alpha (0.05)$

Null hy can't reject

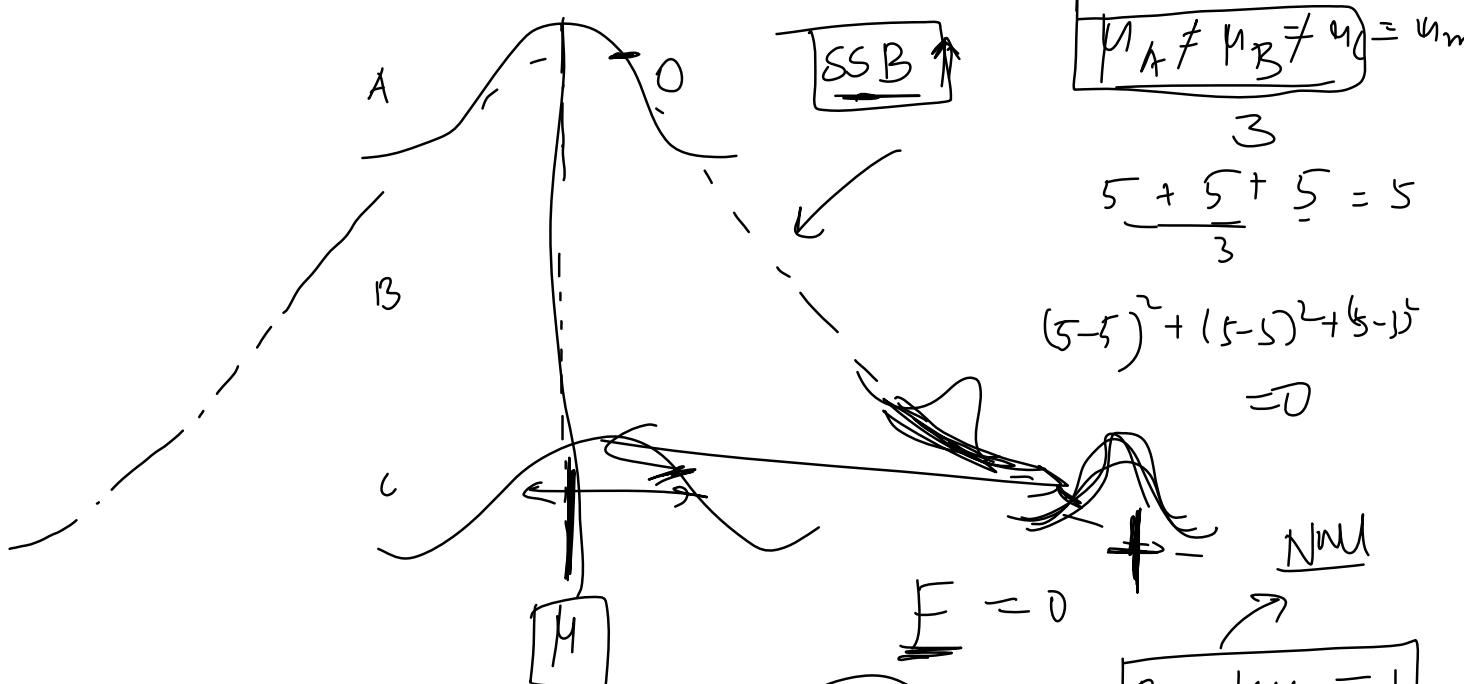
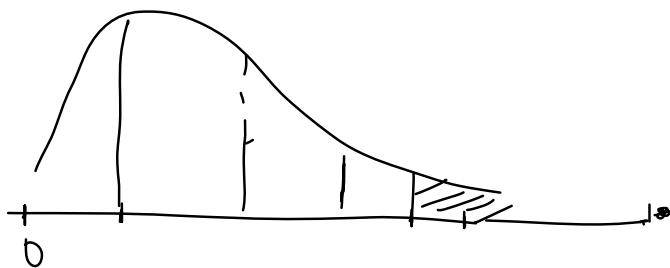
$$\boxed{M_A = M_B = M_C}$$

Geometric Intuition

08 April 2023 13:26



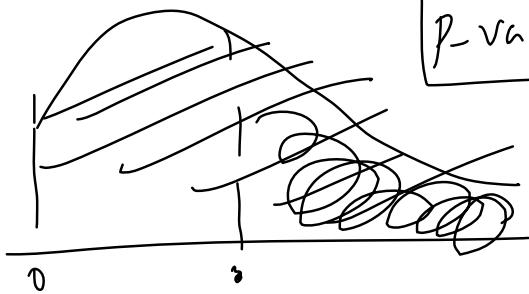
$F \uparrow \rightarrow SSB \uparrow \rightarrow p \text{ value} \rightarrow \text{small } p\text{-value} <$



μ

$t = 0$

p-value = 1



Assumptions

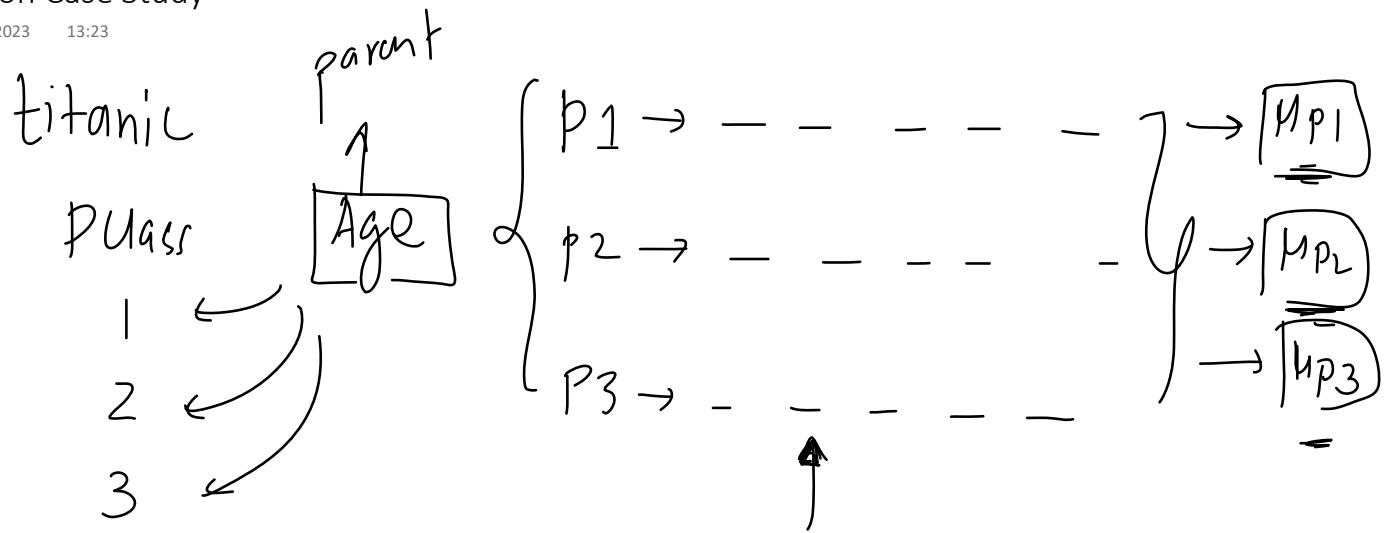
08 April 2023 16:48

Assumptions

1. **Independence:** The observations within and between groups should be independent of each other. This means that the outcome of one observation should not influence the outcome of another. Independence is typically achieved through random sampling or random assignment of subjects to groups.
2. **Normality:** The data within each group should be approximately normally distributed. While one-way ANOVA is considered to be robust to moderate violations of normality, severe deviations may affect the accuracy of the test results. If normality is in doubt, non-parametric alternatives like the Shapiro-wilk test can be considered.
3. **Homogeneity of variances:** The variances of the populations from which the samples are drawn should be equal, or at least approximately so. This assumption is known as homoscedasticity. If the variances are substantially different, the accuracy of the test results may be compromised. Levene's test or Bartlett's test can be used to assess the homogeneity of variances. If this assumption is violated, alternative tests such as Welch's ANOVA can be used.

Python Case Study

08 April 2023 13:23



Post-hoc Test

08 April 2023 13:23

1 - 39 2 - 29 - 3 → 24

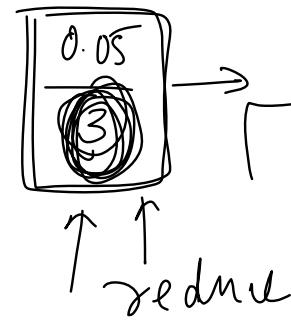
3 groups

Post hoc tests, also known as post hoc pairwise comparisons or multiple comparison tests, are used in the context of ANOVA when the overall test indicates a significant difference among the group means. These tests are performed after the initial one-way ANOVA to determine which specific groups or pairs of groups have significantly different means.

The main purpose of post hoc tests is to control the family-wise error rate (FWER) and adjust the significance level for multiple comparisons to avoid inflated Type I errors. There are several post hoc tests available, each with different characteristics and assumptions. Some common post hoc tests include:

- Bonferroni correction:** This method adjusts the significance level (α) by dividing it by the number of comparisons being made. It is a conservative method that can be applied when making multiple comparisons, but it may have lower statistical power when a large number of comparisons are involved.
- Tukey's HSD (Honestly Significant Difference) test:** This test controls the FWER and is used when the sample sizes are equal and the variances are assumed to be equal across the groups. It is one of the most commonly used post hoc tests.

When performing post hoc tests, it is essential to choose a test that aligns with the assumptions of your data (e.g., equal variances, equal sample sizes) and provides an appropriate balance between controlling Type I errors and maintaining statistical power.

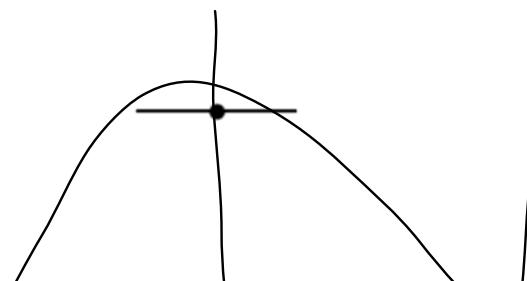
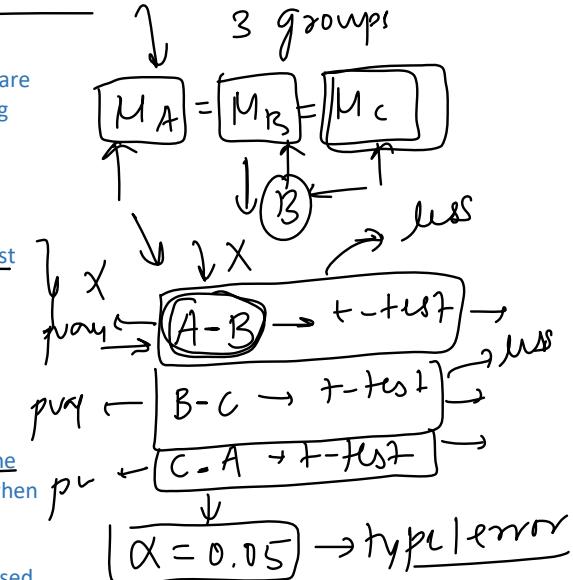


$$0.05 \times 0.05 \times 0.05$$

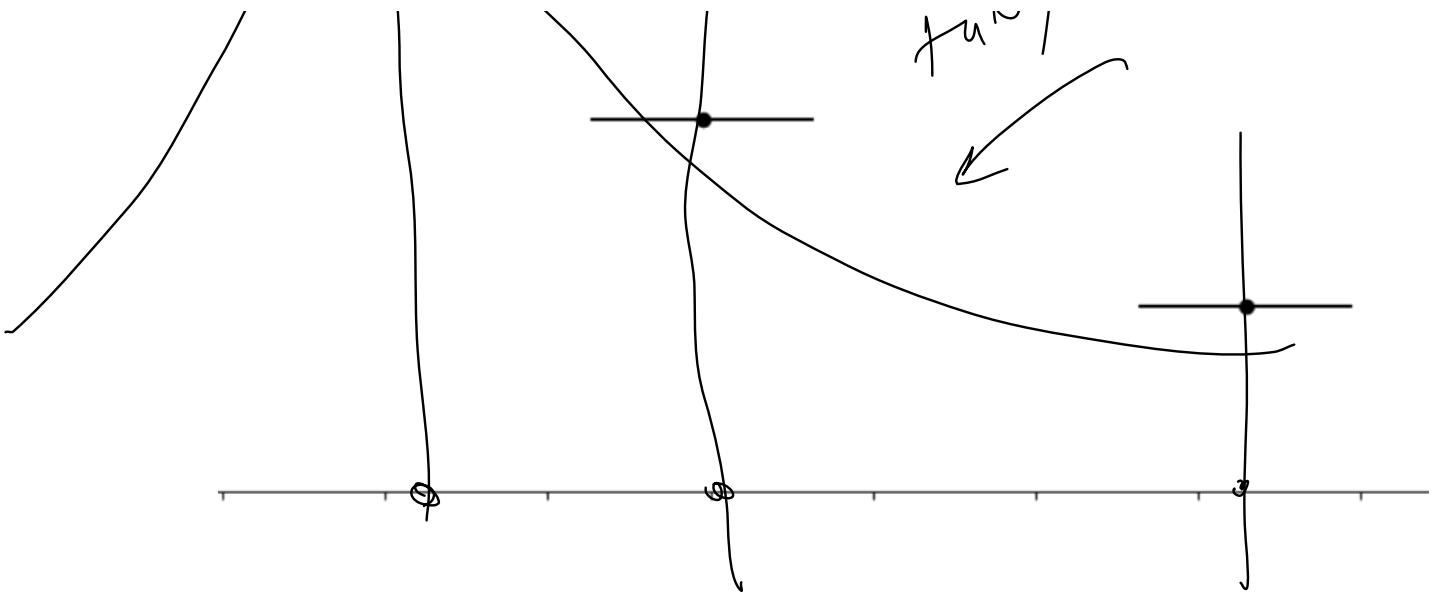
$$\frac{0.15}{3} = 0.05$$

$$1 - 0.85 = 0.15$$

$$1 - 0.95 \times 0.95 \times 0.95$$



ANOVA
Tukey

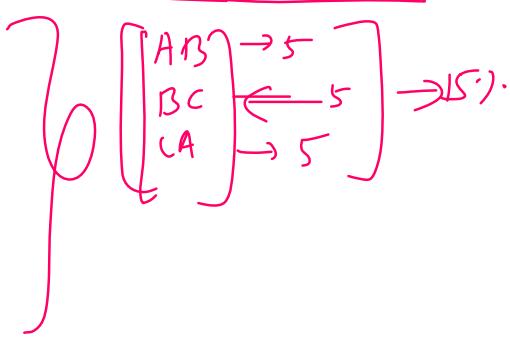


Why t-test is not used for more than 3 categories?

08 April 2023 13:23

1. **Increased Type I error:** When you perform multiple comparisons using individual t-tests, the probability of making a Type I error (false positive) increases. The more tests you perform, the higher the chance that you will incorrectly reject the null hypothesis in at least one of the tests, even if the null hypothesis is true for all groups.
2. **Difficulty in interpreting results:** When comparing multiple groups using multiple t-tests, the interpretation of the results can become complicated. For example, if you have 4 groups and you perform 6 pairwise t-tests, it can be challenging to interpret and summarize the overall pattern of differences among the groups.
3. **Inefficiency:** Using multiple t-tests is less efficient than using a single test that accounts for all groups, such as one-way ANOVA. One-way ANOVA uses the information from all the groups simultaneously to estimate the variability within and between the groups, which can lead to more accurate conclusions.

2 sample indep



Applications in Machine Learning

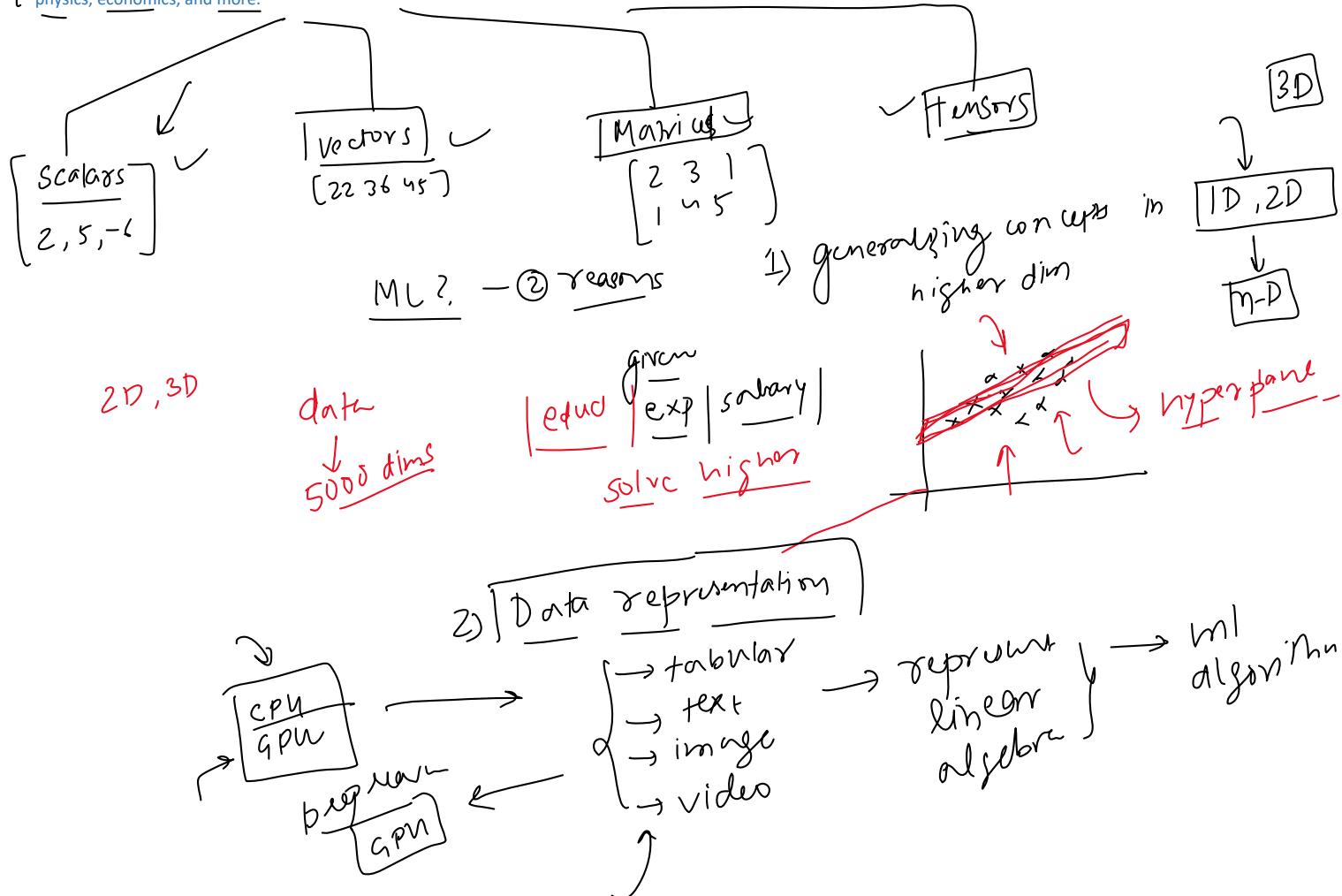
08 April 2023 13:27

- 
1. **Hyperparameter tuning:** When selecting the best hyperparameters for a machine learning model, one-way ANOVA can be used to compare the performance of models with different hyperparameter settings. By treating each hyperparameter setting as a group, you can perform one-way ANOVA to determine if there are any significant differences in performance across the various settings.
 2. **Feature selection:** One-way ANOVA can be used as a univariate feature selection method to identify features that are significantly associated with the target variable, especially when the target variable is categorical with more than two levels. In this context, the one-way ANOVA is performed for each feature, and features with low p-values are considered to be more relevant for prediction.
 3. **Algorithm comparison:** When comparing the performance of different machine learning algorithms, one way ANOVA can be used to determine if there are any significant differences in their performance metrics (e.g., accuracy, F1 score, etc.) across multiple runs or cross-validation folds. This can help you decide which algorithm is the most suitable for a specific problem.
 4. **Model stability assessment:** One-way ANOVA can be used to assess the stability of a machine learning model by comparing its performance across different random seeds or initializations. If the model's performance varies significantly between different initializations, it may indicate that the model is unstable or highly sensitive to the choice of initial conditions.

Linear Algebra

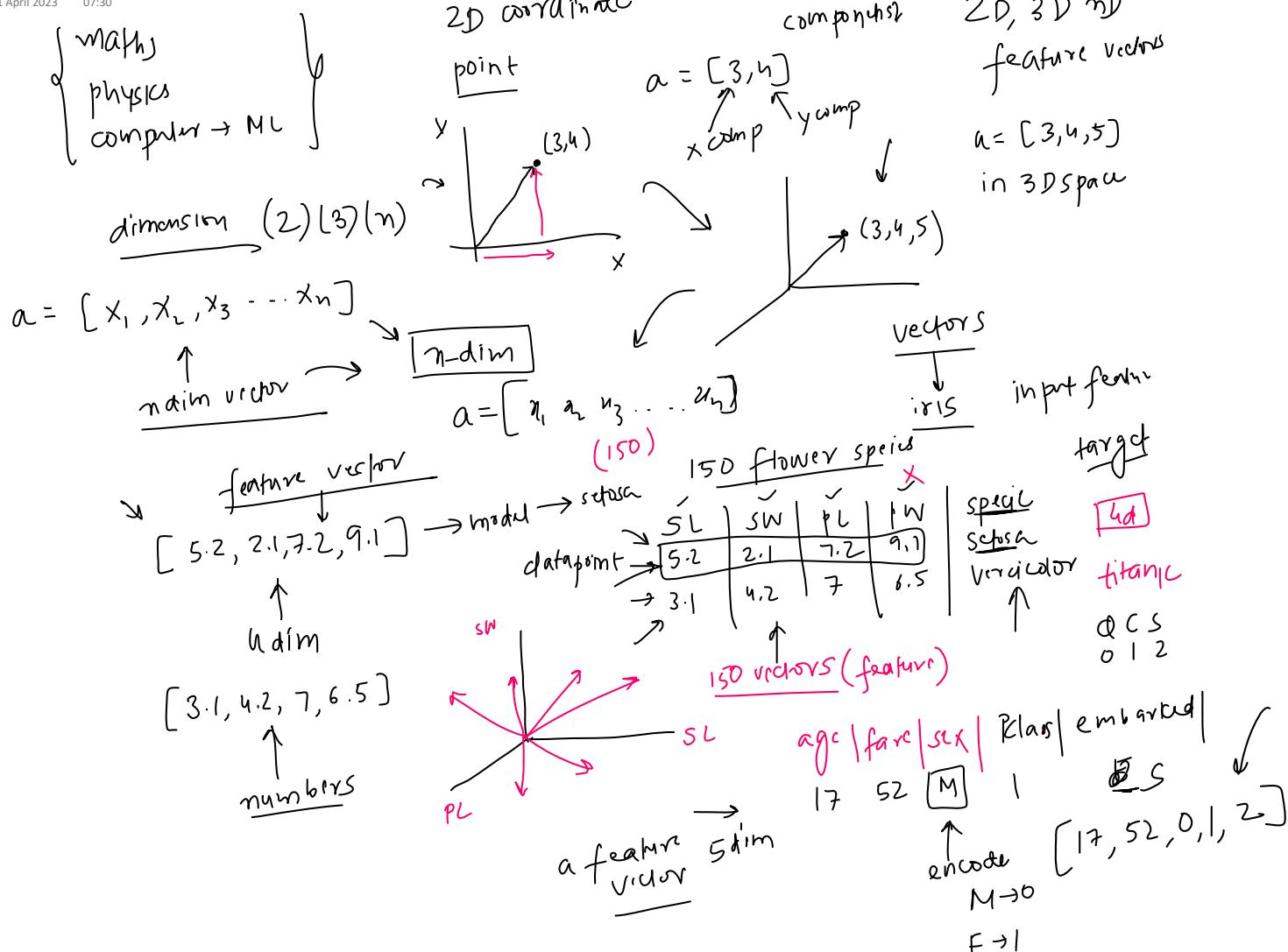
11 April 2023 07:29

{ Linear algebra is a branch of mathematics that deals with the study of linear systems, which are sets of equations involving linear functions of variables. It is a foundational subject in mathematics and has applications in many areas, including computer science, engineering, physics, economics, and more.



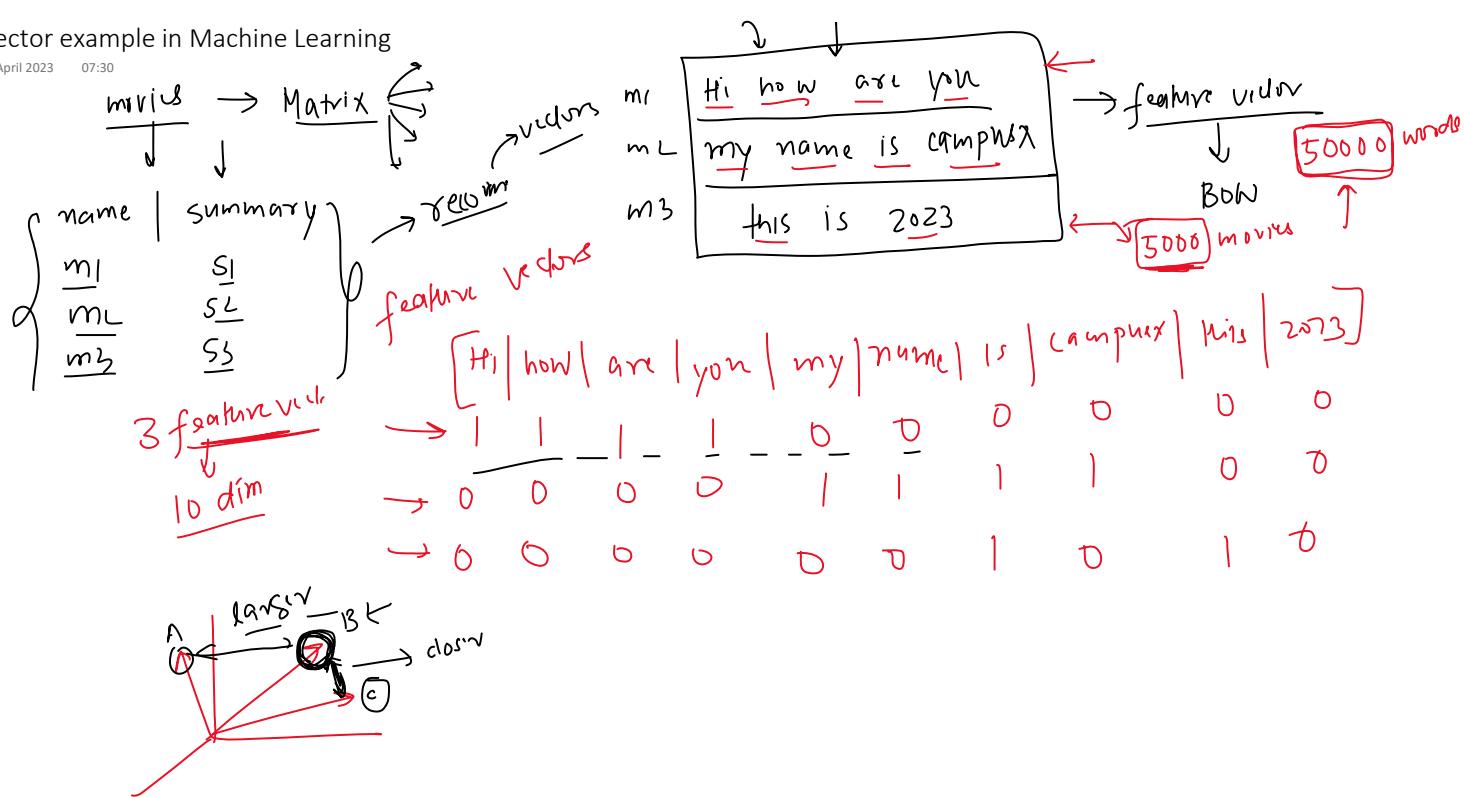
What are Vectors

11 April 2023 07:30



Vector example in Machine Learning

11 April 2023 07:30



Row and Column Vector

11 April 2023 07:30

$$a = [x_1, x_2, x_3, \dots, x_n]$$

rows *cols* *$n \times 1$*

→ *row vector*

a is a row vector

<i>SL</i>	<i>SW</i>	<i>PL</i>	<i>PW</i>	<i>species</i>
—	—	—	—	
—	—	—	—	⋮
—	—	—	—	⋮
—	—	—	—	—

col vector
150x1

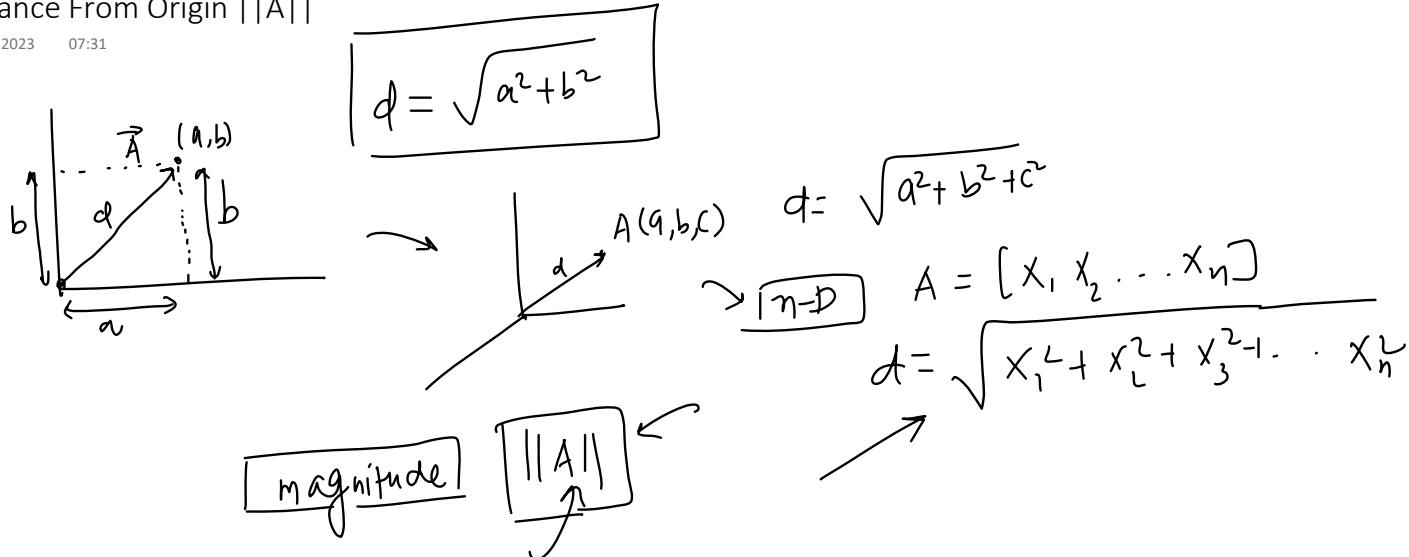
$$b = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

rows *cols* *$n \times n$*

→ *row vector*
 $(1 \times n)$

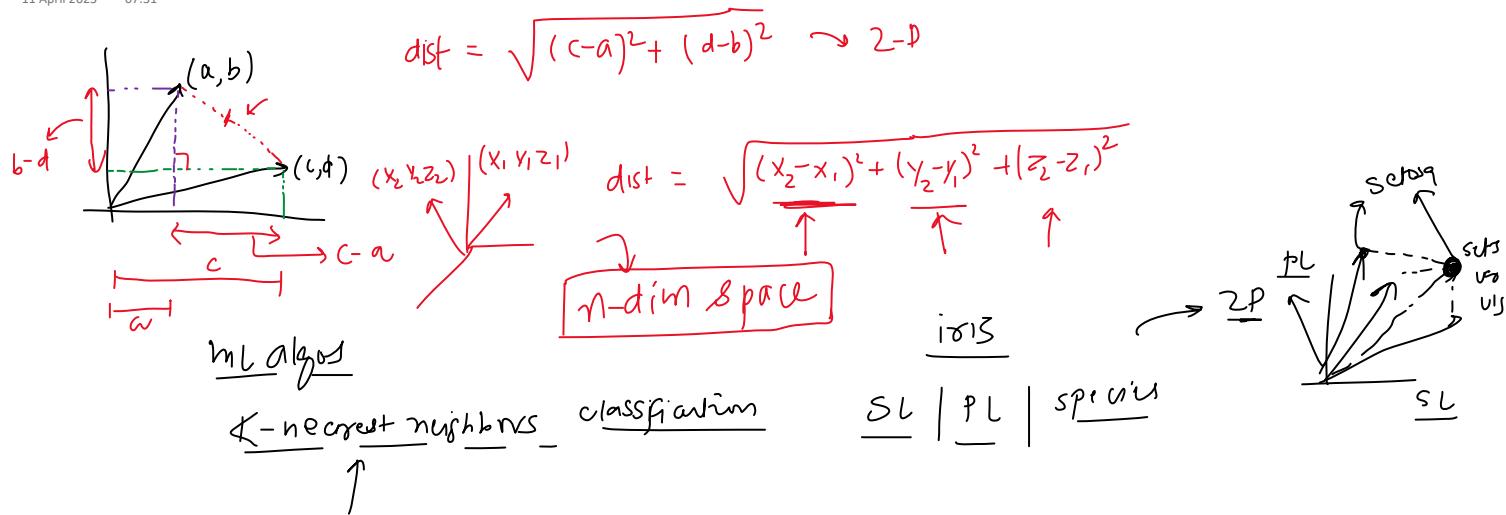
Distance From Origin $\|A\|$

11 April 2023 07:31



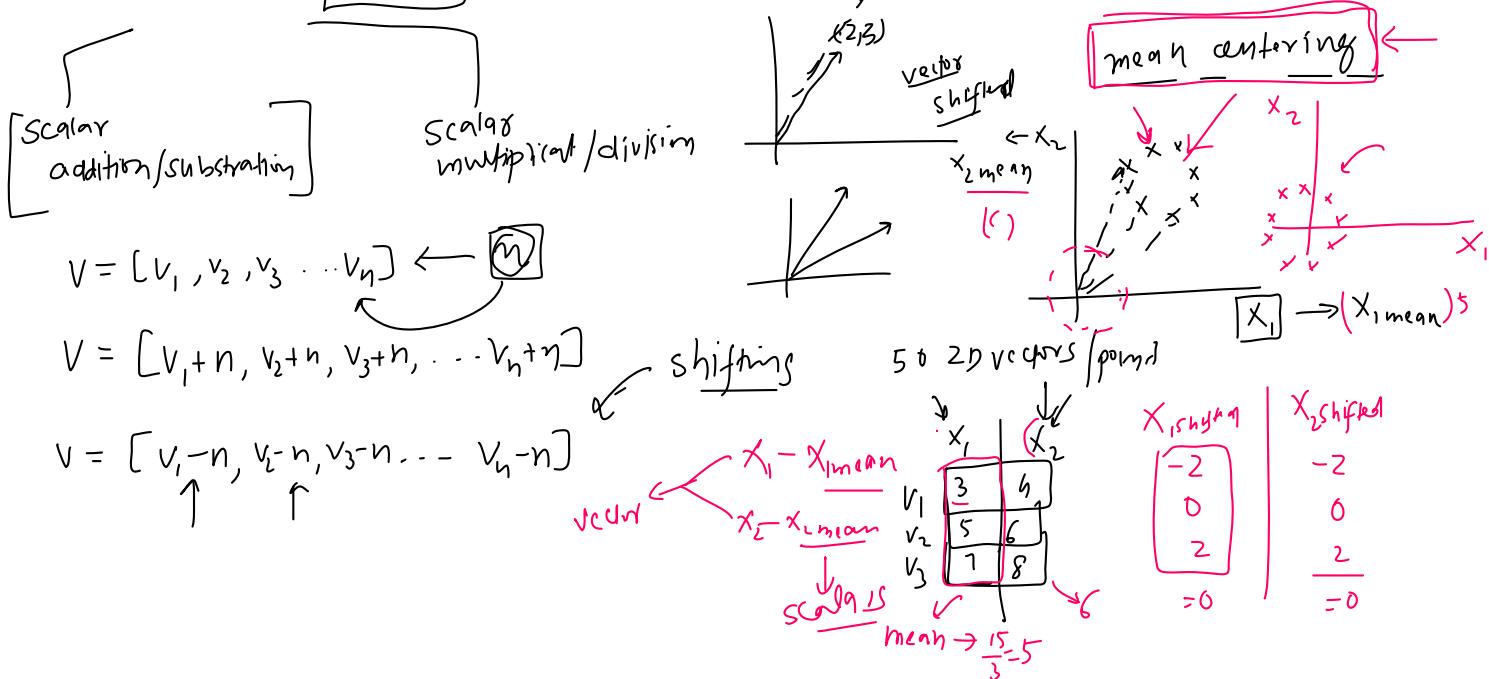
Euclidean Distance

11 April 2023 07:31



Scalar Addition/Subtraction (Shifting)

11 April 2023 07:31



Mean centering is a useful pre-processing technique in various machine learning applications. It can improve the performance, convergence, and interpretability of the model. Some practical examples where mean centering is applied include:

1. **Principal Component Analysis (PCA):** PCA is a dimensionality reduction technique that transforms the data into a new coordinate system by identifying the directions (principal components) with the highest variance. Before applying PCA, it is essential to mean center the data to ensure that the first principal component represents the direction with the highest variance in the dataset, rather than being influenced by the location of the data in the coordinate system.
2. **Linear regression:** In linear regression, mean centering can help improve the interpretability of the model coefficients by making them directly comparable. When the features are mean-centered, the intercept term represents the expected value of the dependent variable when all independent variables are at their mean values. Additionally, mean centering can help with multicollinearity issues, especially when there are interaction terms in the model.
3. **Gradient-based optimization algorithms:** Some machine learning algorithms, such as gradient descent, can converge faster when the input features are mean-centered. This is because mean centering can lead to better conditioning of the optimization problem, allowing the gradient descent algorithm to take larger, more consistent steps towards the optimal solution.
4. **Clustering algorithms:** Mean centering can help improve the performance of clustering algorithms like k-means by ensuring that the initial cluster centroids are not heavily influenced by the location of the data in the coordinate system. This can lead to faster convergence and better clustering results.
5. **Regularization:** In machine learning models that use regularization techniques, such as ridge regression or LASSO, mean centering can help ensure that the regularization term has a consistent effect across all features. By mean centering the features, the model is less likely to penalize the intercept term, which can lead to better generalization.

Scalar Multiplication/Division [Scaling]

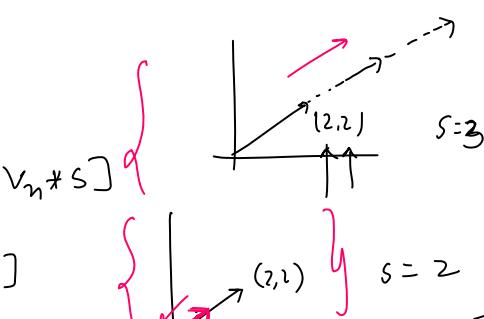
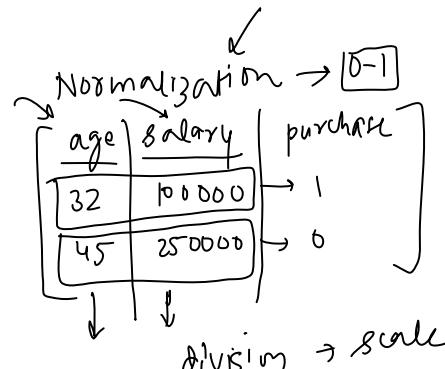
11 April 2023 08:10

$$v = [v_1, v_2, v_3, \dots, v_n]$$

s

$$v \times s = [v_1 \times s, v_2 \times s, v_3 \times s, \dots, v_n \times s]$$

$$\frac{v}{s} = [v_1/s, v_2/s, v_3/s, \dots, v_n/s]$$

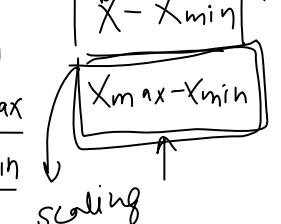
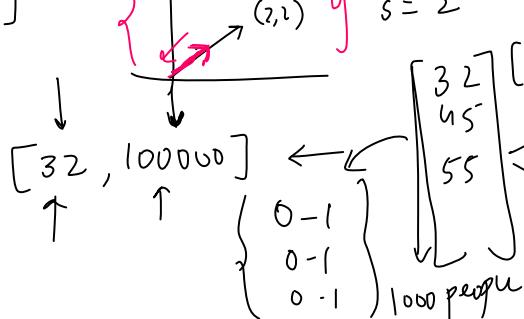


direction

magnitude

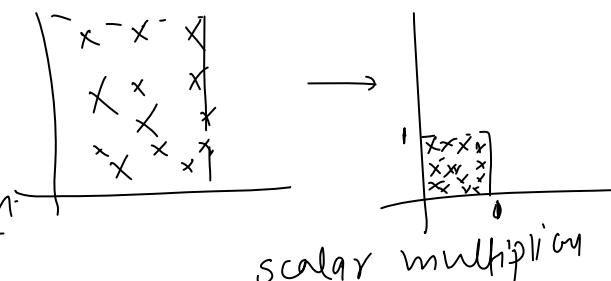
scale \rightarrow scalar

vector



gradient descent

w ₁	w ₂	w ₃	w _n
age	iq	10th marks	12th marks
18	120	90	95
\rightarrow 10			



scalar multiplication

linear reg

$$\frac{\partial L}{\partial w}$$

$$\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \frac{\partial L}{\partial w_3}, \dots, \frac{\partial L}{\partial w_n}$$

$$w_{new} = w_{old} - \eta \frac{\Delta L}{\Delta w}$$

scale your vector

1000

linear reg

$\frac{\partial L}{\partial w}$

$\frac{\Delta L}{\Delta w}$

$\frac{\partial L}{\partial w_1}$

$\frac{\partial L}{\partial w_2}$

$\frac{\partial L}{\partial w_3}$

$\frac{\Delta L}{\Delta w_1}$

$\frac{\partial L}{\partial w_n}$

$\frac{\partial L}{\partial w_n}$

$\frac{\Delta L}{\Delta w_n}$

$\frac{\partial L}{\partial w}$

$\frac{\partial L}{\partial w_1}$

$\frac{\partial L}{\partial w_2}$

$\frac{\Delta L}{\Delta w_1}$

$\frac{\partial L}{\partial w_n}$

$\frac{\partial L}{\partial w_n}$

$\frac{\partial L}{\partial w_n}$

$\frac{\Delta L}{\Delta w_n}$

$\frac{\partial L}{\partial w}$

$\frac{\partial L}{\partial w_1}$

$\frac{\partial L}{\partial w_2}$

$\frac{\Delta L}{\Delta w_1}$

$\frac{\partial L}{\partial w_n}$

$\frac{\partial L}{\partial w_n}$

$\frac{\partial L}{\partial w_n}$

$\frac{\Delta L}{\Delta w_n}$

Vector Addition/Subtraction

11 April 2023 08:08

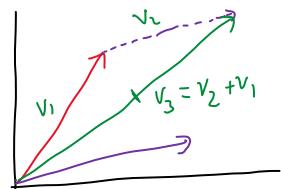
$$\boxed{v_1 \ v_2 \ \dots \ v_n}$$

$$v_1 = [a_1, a_2, a_3, \dots, a_n]$$

$$v_2 = [b_1, b_2, b_3, \dots, b_n]$$

$$v_1 + v_2 = [a_1 + b_1, a_2 + b_2, a_3 + b_3, \dots, a_n + b_n]$$

resultant

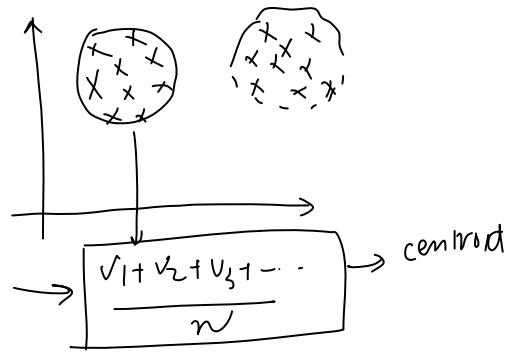


$$v_1 = [a_1, a_2, a_3, \dots, a_n]$$

$$v_2 = [b_1, b_2, \dots, b_n]$$

$$v_1 - v_2 = [a_1 - b_1, a_2 - b_2, \dots, a_n - b_n]$$

$$v_1 + v_2 = v_2 + v_1$$



gradient descent
weight = $[w_1, w_2, w_3, \dots, w_n]$

$$\text{dev} = \left[\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \frac{\partial L}{\partial w_3}, \dots, \frac{\partial L}{\partial w_n} \right]$$

$$w_n = w_0 - \eta \text{dev}$$

Rules

→ greater 2 vector $v_1, v_2, v_3 = v_1 + v_2 + v_3$

→ dimension should be same $\begin{matrix} 2D \text{ vector} \\ 3D \text{ vector} \end{matrix}$

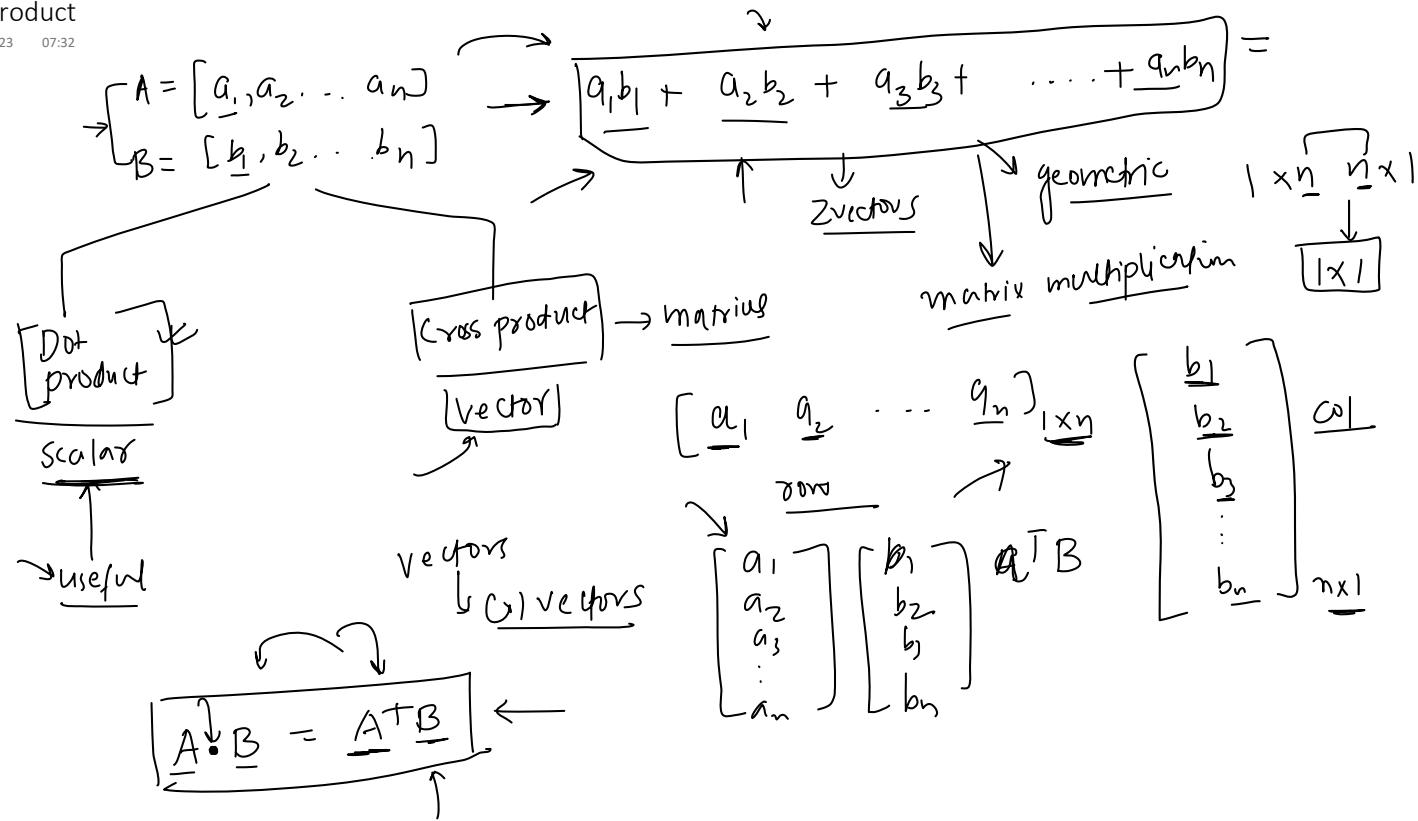
$$\rightarrow A + B = B + A$$

y

$$\rightarrow (A + B) + C = A + (B + C)$$

Dot Product

11 April 2023 07:32



Rules

1) Commutative

$$A \cdot B = B \cdot A$$

$n-D$

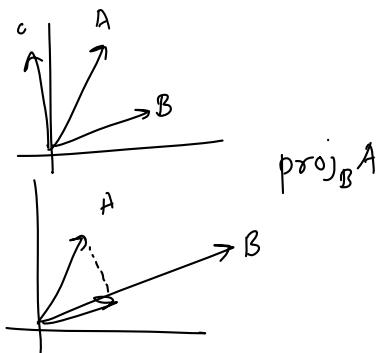
2) Distributive

$$A \cdot (B+C) = A \cdot B + A \cdot C$$

Use

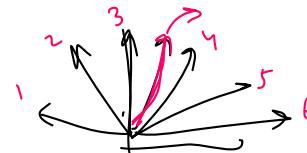
- Compute similarity between 2 vectors
- projections
- perform matrix multiplication

dot product



Machine

→ movies | Summary



Recommend

cosine similarity

→ Deep learning → matrix → dot product

Angle between 2 vectors

11 April 2023 07:32

$A \cdot B = ||A|| ||B|| \cos \theta$

θ is the angle between A and B .

A, B both non-zero \rightarrow SVM

$\Rightarrow A \cdot B = 0 \rightarrow ||A|| ||B|| \cos \theta = 0$

$\theta = 90^\circ$ \rightarrow $A \perp B$ \rightarrow perpendicular

$\theta = 90^\circ$ \rightarrow orthogonal

Cosine similarity \leftarrow

$\cos \theta = \frac{A \cdot B}{||A|| ||B||}$

ML similarity measure

$-1 \rightarrow 1$

$\theta = 0^\circ$ acute \rightarrow same direction

$\theta = 180^\circ$ opposite \rightarrow similar

$\theta = 90^\circ$ orthogonal

more recommended

$m \times s \rightarrow$ vectors

of recommendation

ML

angle small

$\theta = 0^\circ$

$0-30-60-90$

cosine similarity \rightarrow dot product

Unit Vector

11 April 2023 14:31

Projection of a vector

11 April 2023 07:31

Equation of line in n-D

11 April 2023 07:33

Vector Norms

11 April 2023 07:33

linear reg → OLS [this week]
gradient descent [next week]

ML → mistakes → live

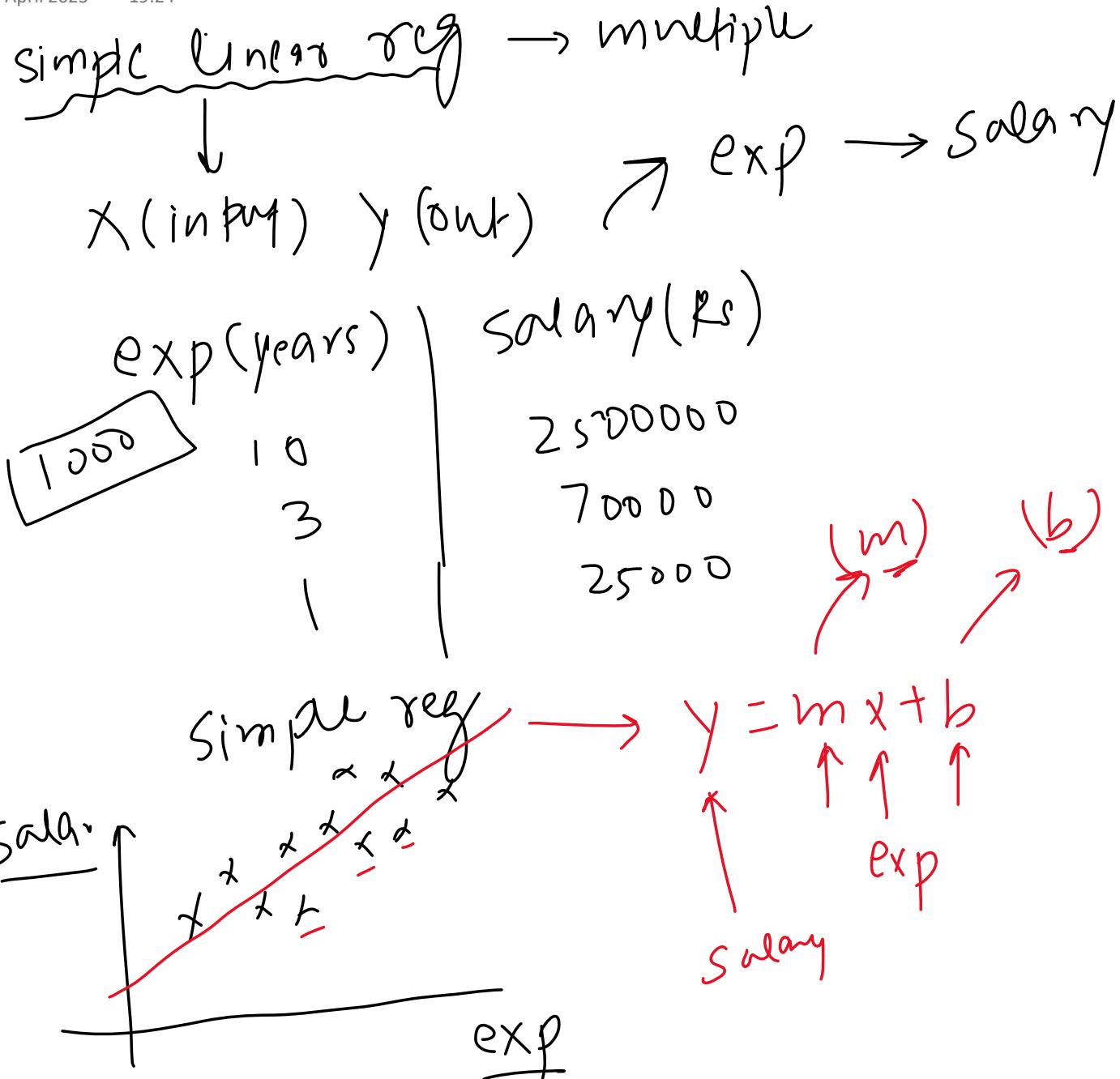
↳ live sessions

↳ linear reg
gradient desc
PCA

→ ✓
↳ rest
algo Y

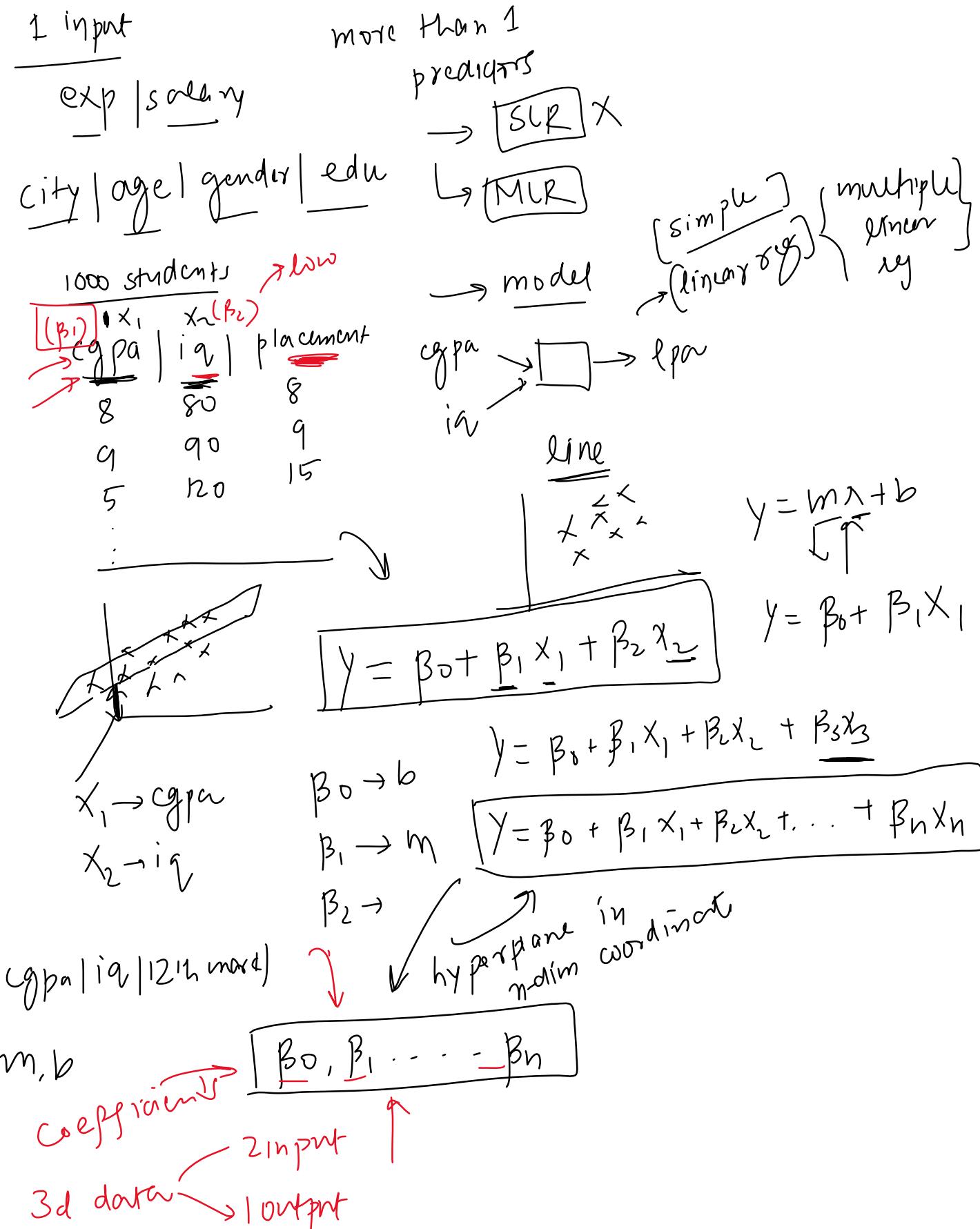
Recap

19 April 2023 19:24



What is Multiple Linear Regression

19 April 2023 19:24



Python Code

19 April 2023 19:24

(B) m cols $\rightarrow n$ students

x_1	x_2	y
Cgpa	ia	placement
(β_0)	(β_1)	
8	80	8
7	70	7
5	120	15
x_{11}	x_{12}	
x_{21}	x_{22}	
x_{31}	x_{32}	

$$y_1 = 8 \quad y_2 = 7 \quad y_3 = 15$$

$$\hat{y}_1 = ? \quad \hat{y}_2 = ? \quad \hat{y}_3 = ?$$

$$\begin{cases} \hat{y}_1 = \beta_0 + \beta_1 8 + \beta_2 80 \\ \hat{y}_2 = \beta_0 + \beta_1 7 + \beta_2 70 \\ \hat{y}_3 = \beta_0 + \beta_1 5 + \beta_2 120 \end{cases}$$

$$\begin{aligned} \hat{y}_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14} + \dots + \beta_m x_{1m} \\ \hat{y}_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m} \\ \hat{y}_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \dots + \beta_m x_{3m} \\ \vdots \\ \hat{y}_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm} \end{aligned}$$

$$\begin{aligned} \hat{y} &= \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14} + \dots + \beta_m x_{1m} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m} \\ \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \dots + \beta_m x_{3m} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm} \end{bmatrix} \\ &\quad \text{with } \hat{y} = X\beta \end{aligned}$$

$$= \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

$$\hat{y} = X\beta \quad \text{where } X \in \mathbb{R}^{n \times (m+1)}$$

$$\hat{y} = X\beta$$

$$d_1^2 + d_2^2 + d_3^2 + \dots + d_n^2$$

$$E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \begin{array}{l} \text{minimize} \\ \text{matrix form} \end{array}$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} \quad n \times 1$$

$$e = y - \hat{y} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

$$e = y - \hat{y} = \begin{bmatrix} 1 \\ y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}_{n \times 1}$$

$$e^T e = \left[\begin{array}{cccc} y_1 - \hat{y}_1 & y_2 - \hat{y}_2 & \dots & y_n - \hat{y}_n \end{array} \right]_{1 \times n} \left[\begin{array}{c} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{array} \right]_{n \times 1} =$$

$$e^T e = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots + (y_n - \hat{y}_n)^2$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\boxed{E = e^T e}$$

$$\uparrow_{\text{miniz}} E = (y - \hat{y})^T (y - \hat{y}) = (y^T - \hat{y}^T) (y - \hat{y})$$

$$E = y^T y - \boxed{y^T \hat{y} - \hat{y}^T y} + \hat{y}^T \hat{y}$$

$$\boxed{E = y^T y - 2 y^T \hat{y} + \hat{y}^T \hat{y}}$$

$$\rightarrow \text{eqn } 3 \quad \hat{y} = X\beta$$

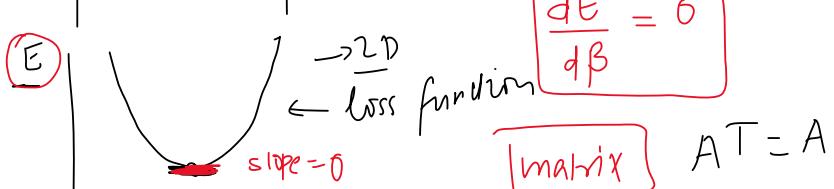
$$\textcircled{Y} = f(x) \rightarrow \textcircled{x}$$

$$E = y^T y - 2 y^T X\beta + (X\beta)^T (X\beta)$$

$$\boxed{E = y^T y - 2 y^T X\beta + \boxed{\beta^T X^T X\beta}} \quad \text{eqn } 4$$

$E(\beta)$ = find such value of β matrix for which E is min

$$\frac{dE}{d\beta} = 0 - 2 y^T X + 2 \beta^T X^T X = 0$$



$$\cancel{\beta^T X^T X} = \cancel{\beta^T X}$$

$$\boxed{\beta^T X^T X = \beta^T X}$$

$$d \cancel{\beta^T X^T X \beta} \quad A^T = A \quad A \text{ is symmetric} \quad \boxed{\beta^T X^T X \underbrace{(X^T X)^{-1}}_{= Y^T X (X^T X)^{-1}} = Y^T X}$$

$y \rightarrow \text{data output}$
 $x \rightarrow \text{data input}$

$$\boxed{\beta^T X^T X \beta}$$

$$f(x) = x^2 \rightarrow \boxed{\frac{d}{dx} y}$$

$$\frac{d}{d\beta} \underbrace{\beta^T X^T X \beta}_{X^T A X} = \boxed{2 \beta^T X^T X}$$

\cancel{P}

A is symmetric

$$(X^T X)^T = X^T X$$

$A^T = A$

$$\beta = [(X^T X)^{-1}]^T (Y^T X)^T$$

$$\beta = [(X^T X)^{-1}]^T X^T Y$$

$$\boxed{\beta = (X^T X)^{-1} X^T Y} \quad \text{eq ⑤ OLS}$$

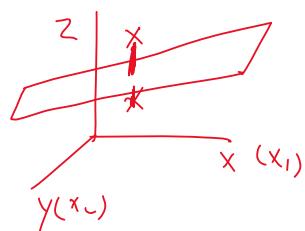
$$\beta = \text{values} \quad (m+1 \times 1)$$

$(X^T X)^{-1} \quad | \quad (m+1) \times 1$

$m+1 \times n \quad n \times (m+1)$

$(m+1) \times (m+1) \quad (m+1) \times n$

$(m+1) \times n \quad n \times 1$



$$\exp \left| 12^m \max \right| \text{salon}^1$$

$$\text{salon} = \exp \left| \begin{matrix} \beta_1 \\ 0 \end{matrix} \right| + 12^m \times \beta_2 + \boxed{\beta_0}$$

$$\beta_0 = 0$$

$$\underbrace{\beta^T X^T X}_{\beta^T I} (\underbrace{X^T X}_{\text{symmetric}})^{-1} = Y^T X (X^T X)^{-1}$$

$$\beta^T = Y^T X (X^T X)^{-1}$$

$$\boxed{(\beta^T)^T = \left[\underbrace{Y^T X}_{A} \underbrace{(X^T X)^{-1}}_{B} \right]^T}$$

$$\underbrace{(X^T X)^{-1}}_{\text{symmetric}} \quad \boxed{[(X^T X)^{-1}]^T = (X^T X)^{-1}}$$

$$\underbrace{X^T X}_{A} = A \quad \boxed{(X^T X)^T = X^T X \quad A^T = A}$$

$$AA^{-1} = I$$

$$(AA^{-1})^T = I^T$$

$$(A^{-1})^T A^T = I$$

$$(A^{-1})^T A = I$$

$$(A^{-1})^T A A^{-1} = I A^{-1}$$

$$(A^{-1})^T I = A^{-1}$$

$$(A^{-1})^T = \boxed{A^{-1}}$$

$$\left(\begin{matrix} X^T X & X^T Y \end{matrix} \right)$$

$$(X^T X)^{-1} X^T Y$$



$$\beta_0 = 0$$

$$(X^T X)^{-1} X^T Y$$

$$\begin{array}{|c|c|c|c|} \hline & x_1 & x_L & y \\ \hline | & | & | & | \\ | & & & \\ | & & & \\ \hline \end{array}$$

$$Y = A \quad \hat{Y} = B$$

$$(AB)^T = B^T A^T$$

$$\underline{A^T B} = \underline{(A+B)^T}$$

$$\boxed{C = C^T}$$

$$(A^T)^T = A$$

$$\boxed{A^T + B = C}$$

symmetric matrix

$$\underline{A^T B} = \underline{B^T A} \checkmark$$

$$(A^T B)^T = B^T A \checkmark$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \leftarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\hat{Y} = X\beta$$

$n \times (m+1)$

$A^T B$ is symmetric

$$\cancel{Y^T \hat{Y}} = Y^T \cancel{X\beta}$$

$1 \times 1 \rightarrow$ scalar

(1) [2]

$$(n \times n) \quad n \times (m+1) \quad (m+1) \times 1$$

$A^T B$ is sym

$$\begin{bmatrix} 1 \times n & n \times 1 \\ \checkmark & \end{bmatrix}$$

Code From Scratch

19 April 2023 19:25

Problem with OLS solution

19 April 2023 19:25

linear reg

OLS (does)
gradient descent

10 input OLS $\rightarrow \sqrt{(\underline{\underline{X^T X}})^{-1}}$ $X \rightarrow n \times (m+1)$

O(n^3) inverse $(m+1) \times n \quad n \times (m+1)$

\boxed{X} $\rightarrow n \text{ row } (m+1)$ $(m+1) \times (m+1)$
 $n \times (m+1)$ 11×11

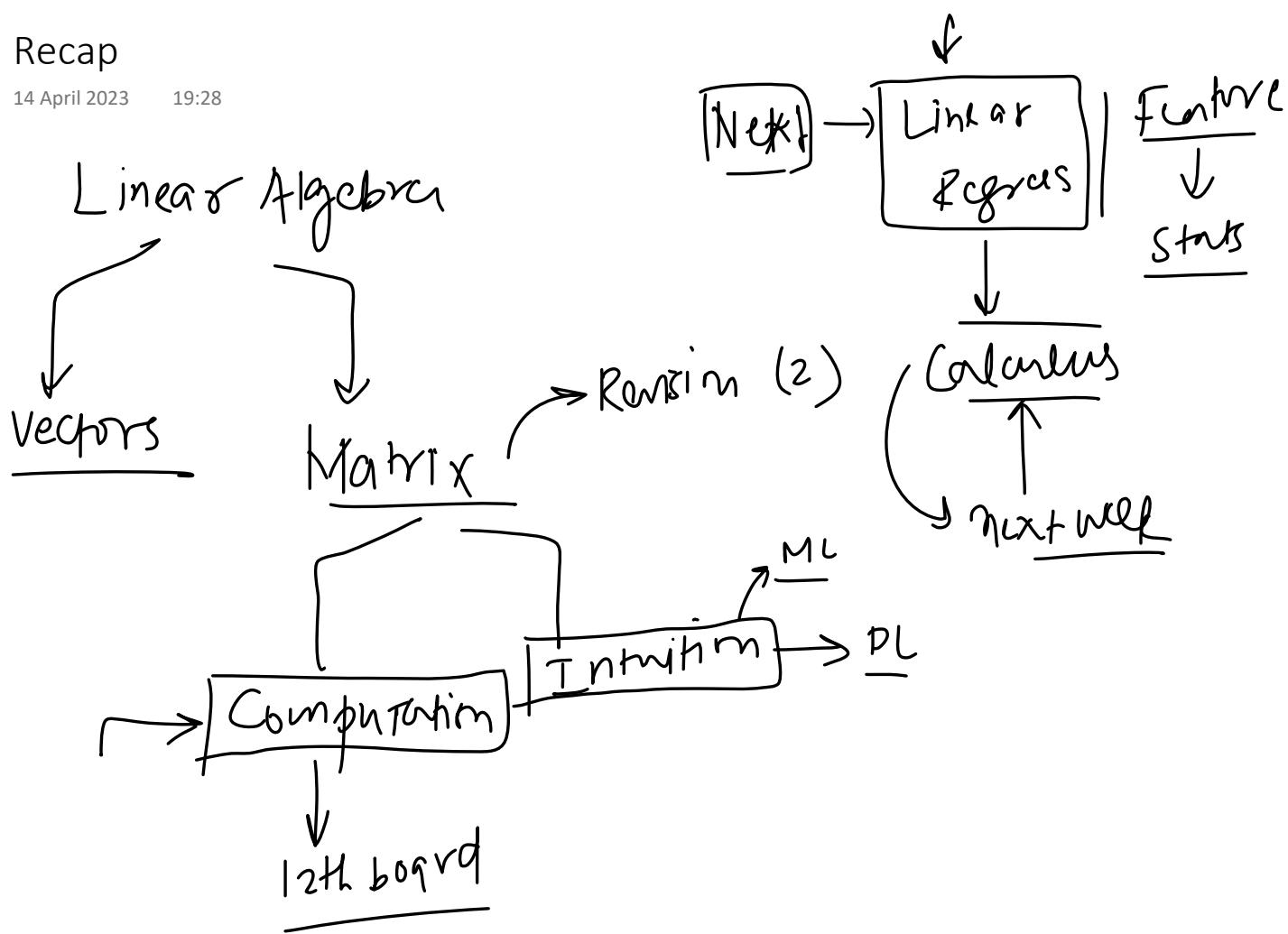
$(m+1) \times n$ $n \times (m+1)$ $(100)^3$

$(m+1) \quad (m+1)$ (100×100) $[1000000]$

$(100 \times 100)^3$ $(100000)^3$

Recap

14 April 2023 19:28



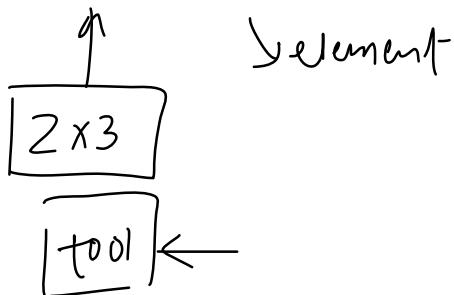
What are Matrices

14 April 2023 14:48

2×2

A matrix is a rectangular array of numbers, symbols, or expressions arranged in rows and columns. The numbers, symbols, or expressions are called the elements of the matrix.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

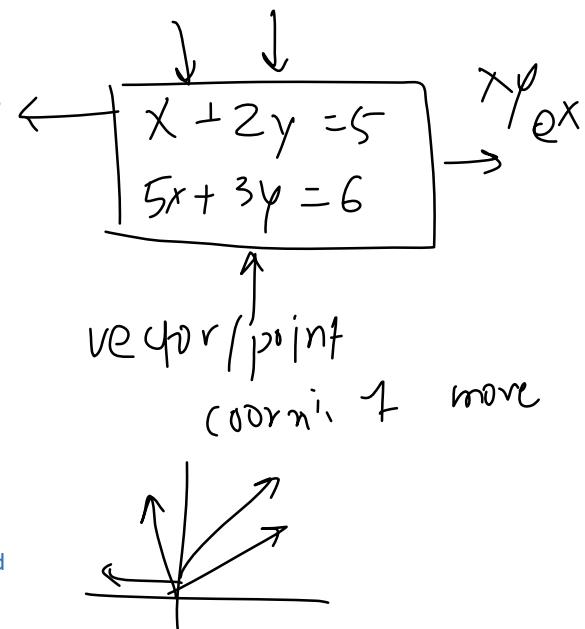
What are Matrices
Order of a matrix
Notation
Uses and Application Areas

order of matrix

rows \times # cols

shape $\leftarrow [2D]$

1. **Linear Systems:** Matrices can be used to represent and solve systems of linear equations. A system of linear equations can be written in matrix form as $Ax = b$, where A is the matrix of coefficients, x is the column vector of unknowns, and b is the column vector of constants. Methods such as Gaussian elimination, LU decomposition, and matrix inversion can be employed to find the solutions to the system.
2. **Linear Transformations:** Matrices are used to represent linear transformations between vector spaces. A matrix can define a linear transformation that maps vectors from one space to another while preserving the operations of vector addition and scalar multiplication. For example, rotation, scaling, and reflection transformations in geometry can be represented using matrices.
3. **Eigenvalues and Eigenvectors:** Matrices are used in the study of eigenvalues and eigenvectors, which are essential in various applications such as differential equations, stability analysis, and diagonalization of matrices. An eigenvalue-eigenvector pair (λ, v) of a square matrix A satisfies the equation $Av = \lambda v$.
4. **Graph Theory:** In graph theory, matrices can be used to represent graphs through adjacency matrices, incidence matrices, and Laplacian matrices. These matrix representations provide a convenient way to analyze the properties of graphs and perform operations on them.
5. **Markov Chains:** Matrices are used in the study of Markov chains, which are stochastic processes that undergo transitions from one state to another according to certain probabilistic rules. Transition matrices describe the probabilities of transitioning between different states in a Markov chain and can be used to analyze the long-term behavior of the system.
6. **Computer Graphics:** Matrices are used extensively in computer graphics to represent transformations such as translation, rotation, scaling, and projection. These transformations are applied to 2D or 3D models to manipulate their position, orientation, and size in a virtual environment.
7. **Control Theory:** In control theory, matrices are used to represent and analyze linear systems, such as state-space models and transfer functions. The use of matrices in control theory allows for the design and analysis of control strategies for complex systems.



- 7. **Control theory:** In control theory, matrices are used to represent and analyze linear systems, such as state-space models and transfer functions. The use of matrices in control theory allows for the design and analysis of control strategies for complex systems.
- 8. **Optimization:** In optimization problems, matrices can be used to represent constraints, objectives, and variables. Techniques such as linear programming, quadratic programming, and semidefinite programming rely on matrices and matrix operations to find optimal solutions.

Types of Matrices

14 April 2023 14:49

$$\begin{bmatrix} 2 & 1 & 3 \\ 4 & 1 & 5 \end{bmatrix} \leftarrow \text{normal}$$

special

- Row Matrix
- Col Matrix
- Square matrix(diagonal) and Non-square Matrix
- Diagonal and scalar Matrix
- Identity Matrix
- Zero Matrix

1) Row matrix → row vector

$$\begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}_{1 \times 4} \text{ row}$$

2) Col matrix

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}_{4 \times 1}$$

3) Square #rows = #cols

$$\begin{bmatrix} a_{11} & 2 \\ 1 & a_{22} \\ 3 & 4 \end{bmatrix}_{2 \times 2} \boxed{i=j}$$

i = rows

j = cols

$$\begin{bmatrix} a_{11} & 2 & 3 \\ 1 & a_{22} & 6 \\ 4 & 5 & a_{33} \\ 7 & 8 & 9 \end{bmatrix}_{3 \times 3}$$

4) Diagonal → square

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}_{3 \times 3} \begin{array}{l} \text{non dia} \\ i \neq j = 0 \end{array}$$

i = j ≠ 0

5) Scalar matrix → diagonal matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3 \times 3}$$

3x3

b) Identity $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \rightarrow \text{order } \rightarrow \textcircled{3}$

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow \text{order } \textcircled{2}$

7) Zero $\rightarrow \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

Matrix Equality

14 April 2023 15:45

2×2

2×2

$$\underline{A} = B$$

1) order should be same ✓

2) $\boxed{A_{ij} = B_{ij}}$

A

$$\begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}$$

2×2

B

$$\begin{bmatrix} \frac{1}{3} \\ \frac{2}{5} \end{bmatrix}$$

2×2

Scalar Operation

14 April 2023 14:50

single 2, 3, -5 $\xrightarrow{\text{add}}$ multiply

$$\boxed{K+A} = K=2 \quad A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$\rightarrow K+A = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}$$

$$\rightarrow KA = 2 \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}$$

negative

$$A = -A$$

$$A = \begin{bmatrix} 1 & -2 \\ 3 & -4 \end{bmatrix}$$

$$K = -1$$

$$KA = -1 \times \begin{bmatrix} 1 & -2 \\ 3 & -4 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ -3 & 4 \end{bmatrix}$$

$$\boxed{-4}$$

$$\boxed{A, B} \xrightarrow{K} (A+B) = \frac{KA}{\text{mat}} + \frac{KB}{\text{mat}}$$

- \rightarrow Scalar Addition
 - \rightarrow Scalar Multiplication
 - \rightarrow Negative of a Matrix
 - \rightarrow Rules
- $K(A+B) = KA+KB$

Matrix Addition and Subtraction

14 April 2023 14:51

$\begin{cases} \rightarrow \text{add} \\ \rightarrow \text{sub} \\ \rightarrow \text{multi} \\ \rightarrow \text{div?} \end{cases}$

criteria

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

A

(2×2)

\rightarrow order same

$$\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

B

(2×2)

\rightarrow Matrix Addition
 \rightarrow Matrix Subtraction
 Rules
 $\rightarrow A+B=B+A$
 $\rightarrow (A+B)+C=A+(B+C)$
 \rightarrow Additive Identity
 \rightarrow Additive Inverse

$$\begin{bmatrix} 1+5 \end{bmatrix}$$

100×200

100×200

$$\boxed{A+B = B+A}$$

order

Associative

$$\boxed{(A+B)+C = A+(B+C)}$$

Subtraction

\uparrow

$A - B$

\downarrow scalar multi / matrix add

$$\boxed{A + (-1)B}$$

negative of B

$-B$

1) Additive identity

$$A + \boxed{X} = A$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} + \boxed{X} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$$

\uparrow

$$X = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

zero matrix

$$A = \boxed{-A}$$

additive inverse

$$A - \boxed{X} = 0$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} - \boxed{X} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

additive
Inverses

$$\begin{bmatrix} 4 & 5 \\ -A \end{bmatrix} \xrightarrow{\quad} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Matrix Multiplication

14 April 2023

14:51

$$\begin{bmatrix} A \\ -1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} B \\ 4 & 5 \\ 6 & 7 \end{bmatrix} = \boxed{\quad}$$

matrix

$$2 \times 2 \leftrightarrow 2 \times 2$$

$$1 \times 3 \neq 1 \times 3$$

geometric

$$\begin{bmatrix} \text{row vector} \\ \begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{matrix} \end{bmatrix} \underset{2 \times 3}{=} \boxed{\quad}$$

$$2 \times 3 = 3 \times 3$$

$$\begin{bmatrix} 2 \times 3 \\ \text{Col vector} \\ \begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{matrix} \end{bmatrix} = \boxed{\quad}$$

Multiplying Matrix Rules
 -> $A \cdot B = B \cdot A$
 -> $(AB)C = A(BC)$
 -> $A(B+C) = AB+AC$
 -> Multiplicative Identity

$$\boxed{A \cdot B \neq B \cdot A}$$

$$\begin{bmatrix} 2 \times 3 & 3 \times 3 \\ A & B \\ B & A \\ 3 \times 3 & 2 \times 3 \end{bmatrix}$$

$$\begin{array}{c} \text{ASSOCIATIVITY} \\ (AB)C = A(BC) \\ \hline \boxed{A(B+C) = AB+AC} \end{array}$$

$$\begin{bmatrix} 30 & 36 & 42 \\ 66 & 71 & 100 \end{bmatrix}$$

multiplicative identity

$$A \times = A \quad \boxed{I}$$

$$A I = A$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} \boxed{x} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$$

$$\overbrace{\quad}^A \quad \overbrace{\quad}^m \quad \overbrace{\quad}^n \quad \overbrace{\quad}^{m \times n}$$

$$\begin{array}{c}
 \underline{A} \\
 \left[\begin{array}{cc} 2 & 3 \\ 4 & 5 \end{array} \right] \quad \left[\begin{array}{c} 1 \\ 0 \end{array} \right] \quad \left[\begin{array}{c} 0 \\ 1 \end{array} \right]
 \end{array}$$
$$\left[\begin{array}{cc} 2 & 3 \\ 4 & 5 \end{array} \right] \leftarrow$$

Transpose of a Matrix

14 April 2023 14:53

Solving eq⑩

Rules

Transpose

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \leftarrow$$

$$C = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 2 & 4 \end{bmatrix}_{2 \times 3}$$

$$C^T = \begin{bmatrix} 2 & 1 \\ 3 & 2 \\ 4 & 4 \end{bmatrix}_{3 \times 2}$$

Symmetric matrix

$$\boxed{A = A^T} \leftarrow$$

↑
Skew symmetric

$$\boxed{A^T = -A} \leftarrow$$

$$\rightarrow (A^T)^T = A$$

$$\rightarrow (A+B)^T = A^T + B^T$$

$$\rightarrow (AB)^T = B^T \cdot A^T$$

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Matrix Transpose
→ Symmetric Matrix
→ Skew Symmetric

Rules

- > $A^T \cdot A^T = A$
- > $(A+B)^T = A^T + B^T$
- > $(AB)^T = B^T \cdot A^T$

$$(A^T)^T = A$$

$$(m \times n)^T$$

$$(n \times m)^T = \underline{(m \times n)}$$

$$\boxed{B^T} = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \quad \boxed{(A+B)^T} = A^T + B^T$$

$$\underline{m \times n} = n \times m$$

$$\boxed{(AB)^T = B^T \cdot A^T}$$

geometrically

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad A^T = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

Determinant

14 April 2023 14:53

$2, 3, -11, -15$ calculate $\text{inv}(A)$

The determinant is a scalar value computed from a square matrix (a matrix with the same number of rows and columns) that carries important information about the matrix. It has several uses in linear algebra, including determining the invertibility of a matrix, finding the solution to systems of linear equations, and calculating the volume scaling factor for linear transformations.

What is Determinant
 $1 \times 1 \rightarrow 2 \times 2 \rightarrow \text{Det}(3 \times 3)$
 Rules
 $\rightarrow \det(A) = \det(A')$
 Singular Matrix

$$AX = B$$

$$\begin{bmatrix} 1 & 2 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 11 \end{bmatrix}$$

$$X = \frac{B}{A} = A^{-1} \cdot B$$

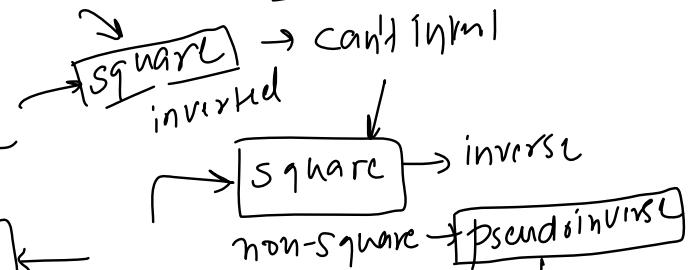
division X multiplication

$$\frac{1}{A} = A^{-1}$$

$$A^{-1} = A$$

A^{-1} given it is invertible

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$



Determinant

1 Minor → 2 Cofactor → 3 Adjoint

Inverso

1×1 2×2 3×3

$$A = [1]_{1 \times 1}$$

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$\det(A)[1]$$

$$B = \begin{bmatrix} -2 & 5 \\ -6 & 7 \end{bmatrix}$$

$$1 \sim 5 \mid -14 - (-30) \quad \checkmark$$

$$\det(A) = \Delta = \begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 1 \times 4 - 3 \times 2 = -2$$

$$\det(B) = \Delta = \begin{vmatrix} -6 & 7 \\ -2 & 5 \\ -6 & 7 \end{vmatrix} = \frac{-14 - (-30)}{30 - 14} = \boxed{16}$$

How to decide if inv of a matrix is possible
 $A \quad A^{-1} \quad A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

$\det(A) = 0$
 \downarrow
 singular

- 1) square ✓
 $\rightarrow 2) \det(A) \neq 0$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\boxed{ad - bc = 0} \text{ inverse } X$$

non-singular
matrices inverse

$\boxed{3 \times 3}$

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad 3 \times 3$$

$\boxed{1 \times 1}$

2×2

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\det(A) = \Delta =$$

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix}$$

$$a_{12} = 4 \quad a_{13} = 12 \quad a_{21} = 3$$

$$(-1)^{1+2} = -1 \quad (-1)^{1+3} = 1$$

$$= +1 \times \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} - 2 \times \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} + 3 \times \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix}$$

$$= -3 + 12 - 9 = 0$$

$\det(A) = 0 \rightarrow$ singular matrix

$$\text{in } A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

\downarrow
 inverse X

cofactor \rightarrow minor

Minor

14 April 2023 17:27

Minor of an element a_{ij} of a Determinant is the determinant obtained by deleting its i th row and j th col. It is denoted by M_{ij}

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{det}(A) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$$\underline{M_{11}} = a_{22} \quad \underline{M_{12}} = a_{21}$$

$$\underline{M_{21}} = a_{12} \quad \underline{M_{22}} = a_{11}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \boxed{a_{13}} \\ a_{21} & a_{22} & \underline{\underline{a_{23}}} \\ a_{31} & a_{32} & \underline{\underline{a_{33}}} \end{bmatrix}$$

$$\boxed{M_{11}} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = a_{22}a_{33} - a_{23}a_{32} \quad M_{12} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} \\ = a_{21}a_{33} - a_{23}a_{31}$$

$$\text{B} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad M$$

\downarrow
Det = Sum of the product of elements of any row(or col) with Their corresponding cofactors.

Cofactor of an element of a_{ij} of a determinant is defined by

$$A_{ij} = (-1)^{i+j} M_{ij} \text{ where } M_{ij} \text{ is the minor of } a_{ij}$$

$$A_{11} = (-1)^{1+1} M_{11} \leftarrow A_{21} = (-1)^{2+1} M_{21} = -M_{21}$$

$$A_{11}$$

$i=1$

$j=1$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$A_{32} = (-1)^{3+2} M_{32}$$

$$\det = \boxed{a_{11} A_{11} + a_{12} A_{12} + a_{13} A_{13}}$$

Adjoint

14 April 2023 17:27

$$\boxed{A \cdot A^{-1} = I}$$

$$\boxed{A^{-1}}$$

The adjugate of a matrix, also known as the classical adjoint, is a matrix formed by replacing each element in the original matrix with its corresponding cofactor and then taking the transpose of the resulting matrix. The adjugate of matrix A is denoted as $\text{adj}(A)$.

$$A = \begin{bmatrix} \underline{c_{11}} & \underline{a_{12}} & \underline{a_{13}} \\ \underline{a_{21}} & \underline{a_{22}} & \underline{a_{23}} \\ \underline{a_{31}} & \underline{a_{32}} & \underline{a_{33}} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}^T$$

↓

$$\frac{5}{7} \rightarrow \boxed{5 \times 7 - 1} \rightarrow \text{adj}(A) = \begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{bmatrix}$$

Inverse of Matrix

14 April 2023 14:54

An inverse matrix is a matrix that, when multiplied by the original matrix, results in the identity matrix. The inverse matrix is defined only for square matrices (matrices with the same number of rows and columns) and not all square matrices have an inverse.

A matrix is invertible (has an inverse) if and only if it is non-singular, meaning its determinant is non-zero. If the determinant of A is zero, A is called a singular matrix, and it does not have an inverse.

Inverse matrices play a crucial role in linear algebra and have many applications, such as solving systems of linear equations, finding the solution to a matrix equation, and performing various matrix operations. There are several methods for finding the inverse of a matrix, including Gaussian elimination, the adjugate method, and LU decomposition.

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

$\neq 0$ non-singular
invertible

$$(AB)^{-1} = B^{-1} A^{-1}$$

$$\begin{aligned} & (AB)^{-1} = B^{-1} A^{-1} \\ & A \cdot A^{-1} = I \\ & A X = B \\ & A^{-1} A X = A^{-1} B \\ & I X = A^{-1} B \\ & X = A^{-1} B \end{aligned}$$

Solving a system of linear equations

14 April 2023 14:54

$$\begin{array}{l} \xrightarrow{\quad} \left[\begin{array}{l} x+y=5 \\ 4x+3y=15 \end{array} \right] \quad \boxed{x,y} \\ \xrightarrow{\quad} \left[\begin{array}{cc|c} 1 & 1 & 5 \\ 4 & 3 & 15 \end{array} \right] \xrightarrow{\quad} \left[\begin{array}{cc|c} 1 & 1 & 5 \\ 0 & -1 & 15 \end{array} \right] \xrightarrow{\quad} \left[\begin{array}{cc|c} 1 & 1 & 5 \\ 0 & 1 & -15 \end{array} \right] \xrightarrow{\quad} \left[\begin{array}{l} x+y=5 \\ y=-15 \end{array} \right] \xrightarrow{\quad} \left[\begin{array}{l} x+y=5 \\ y=5 \end{array} \right] \xrightarrow{\quad} \left[\begin{array}{l} x=0 \\ y=5 \end{array} \right] \\ \text{computation} \\ \text{intuition} \end{array}$$

$$\boxed{A^{-1} \cdot B}$$

$$A = \left[\begin{array}{cc} 1 & 1 \\ 4 & 3 \end{array} \right]$$

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

$$= \frac{1}{-1} \left[\begin{array}{cc} 3 & -1 \\ -4 & 1 \end{array} \right]$$

$$A^{-1} = \left[\begin{array}{cc} -3 & 1 \\ 4 & -1 \end{array} \right] \cdot \left[\begin{array}{c} 5 \\ 15 \end{array} \right]$$

$$= \left[\begin{array}{c} 0 \\ 1 \end{array} \right] = \left[\begin{array}{c} x \\ y \end{array} \right]$$

$$\left[\begin{array}{l} x=0 \\ y=5 \end{array} \right]$$

$$A^{-1} A X = A^{-1} B$$

$$I_X = A^{-1} B$$

$$\rightarrow X = A^{-1} B$$

$$3 - 4 = -1 \neq 0 \quad \begin{matrix} (-1)^{1+1} 3 \\ (-1)^3 4 \end{matrix} \quad \begin{matrix} (-1)^{1+1} 1 \\ (-1)^3 1 \end{matrix}$$

$$\left[\begin{array}{cc} 3 & -4 \\ -1 & 1 \end{array} \right]^T = \left[\begin{array}{cc} 3 & -1 \\ -4 & 1 \end{array} \right]$$

$$\text{adj} \left[\begin{array}{cc} 3 & -1 \\ -4 & 1 \end{array} \right]$$

$$\left[\begin{array}{cc} -3 & 1 \\ 4 & -1 \end{array} \right] \cdot \left[\begin{array}{c} 5 \\ 15 \end{array} \right]$$

$$= \left[\begin{array}{c} 0 \\ 1 \end{array} \right] = \left[\begin{array}{c} x \\ y \end{array} \right]$$

$$= \begin{bmatrix} 0 \\ 5 \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} - \boxed{\begin{bmatrix} x \\ y \end{bmatrix}}$$

$$x + 3y + 4z = 5$$

$$6x + 4y + z = 16$$

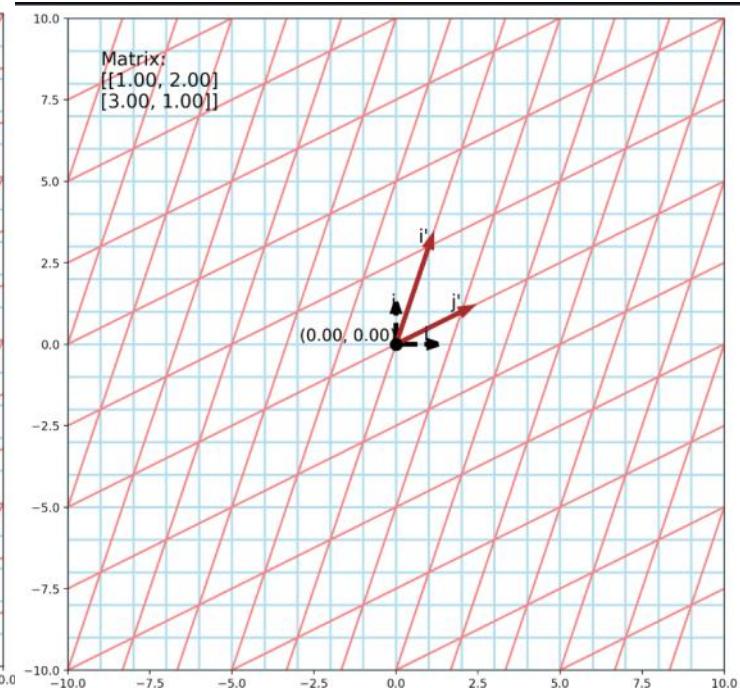
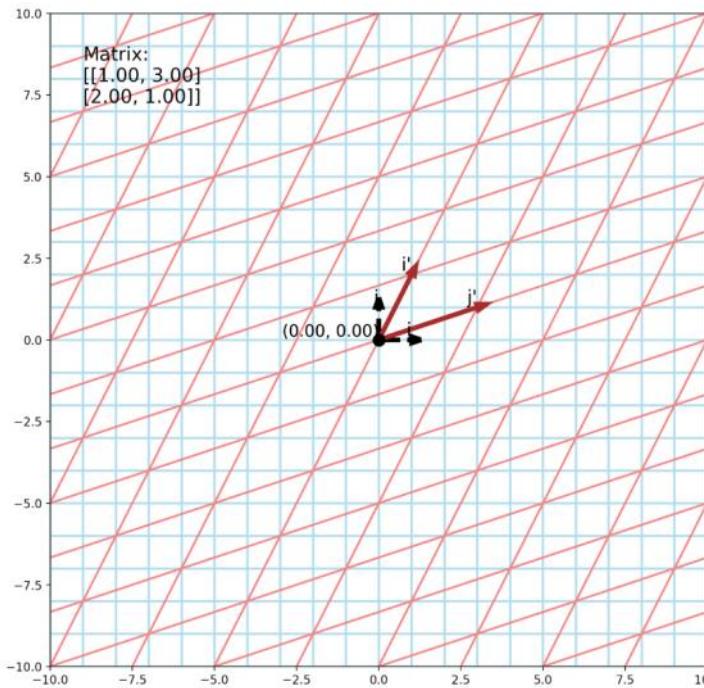
$$2x + 2y + 2z = 11$$

$$\rightarrow \begin{bmatrix} 1 & 3 & 4 \\ 6 & 4 & 1 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5 \\ 16 \\ 11 \end{bmatrix} \leftarrow$$

A X B

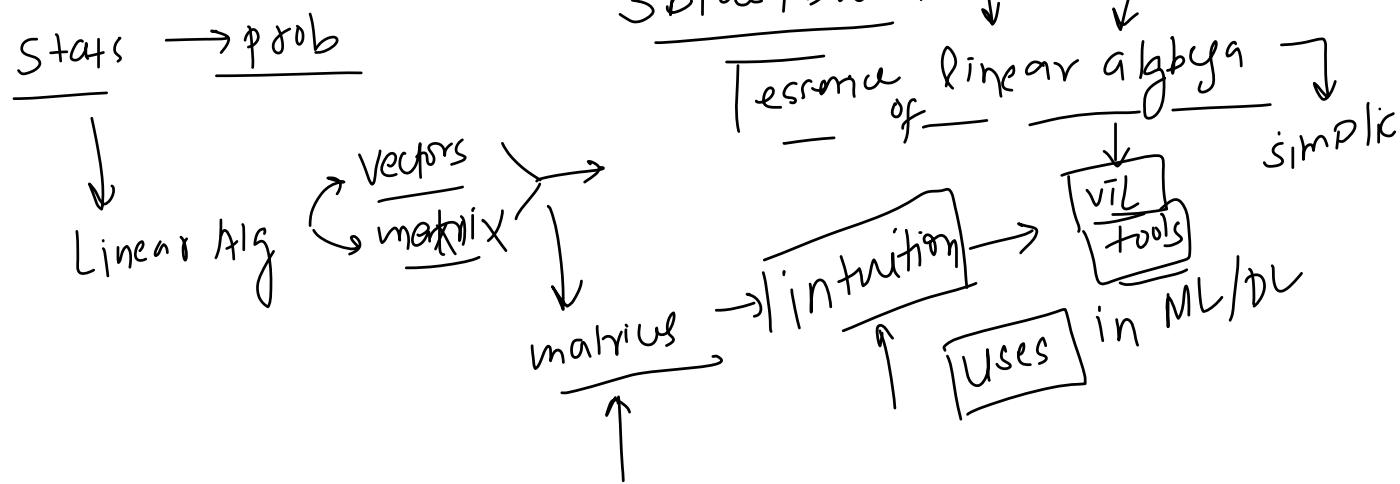
$$\boxed{AX = B}$$

$$\boxed{X = A^{-1}B}$$

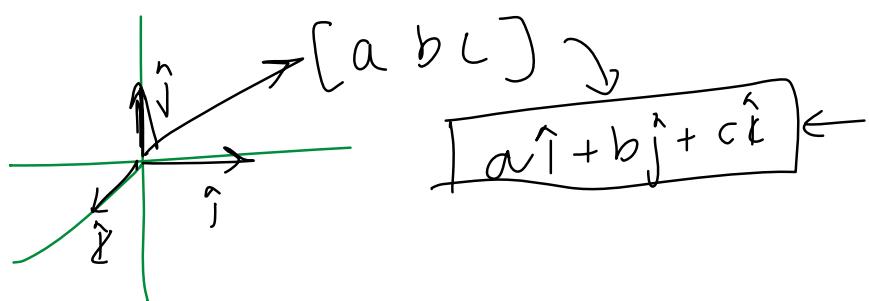
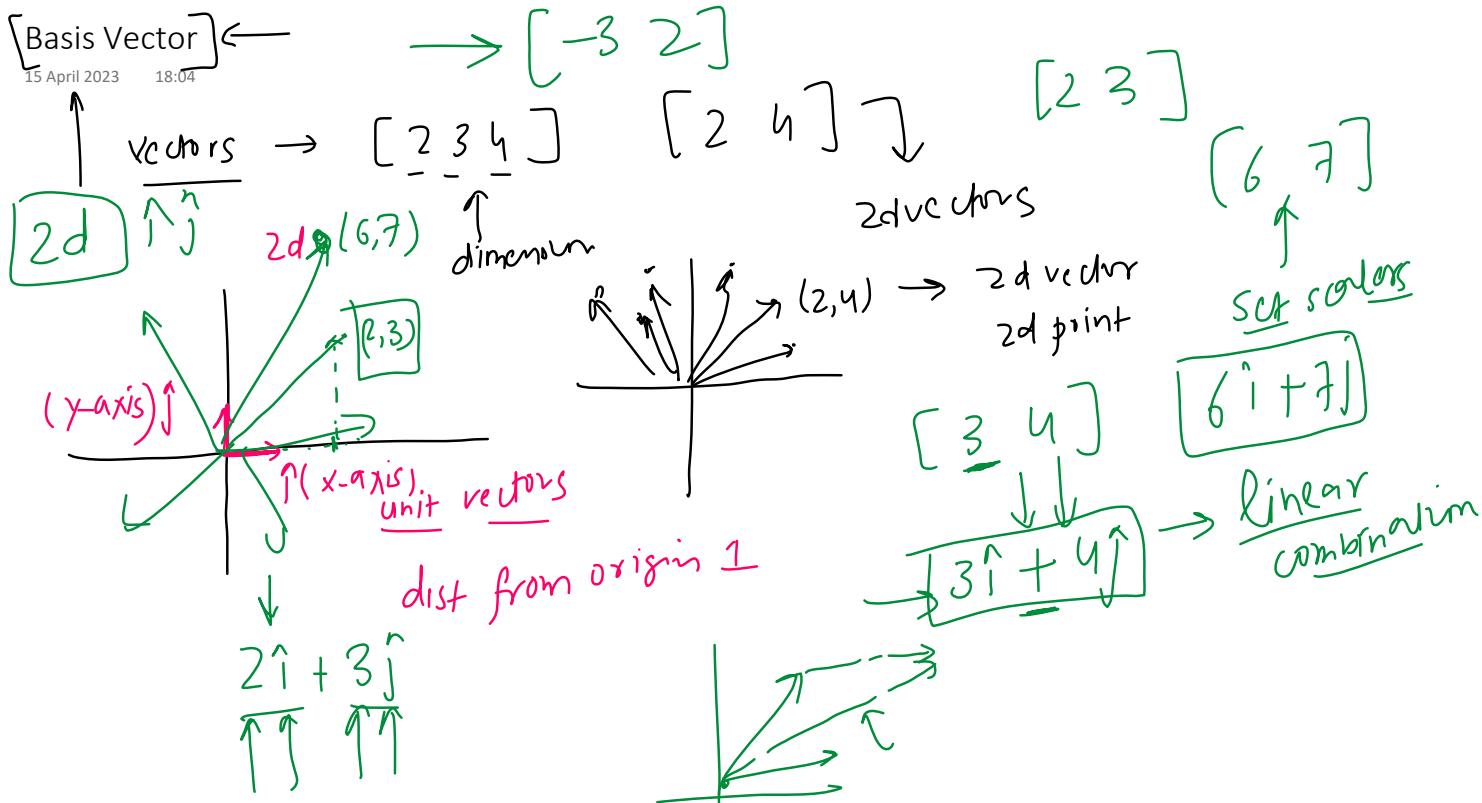


Recap

15 April 2023 18:04

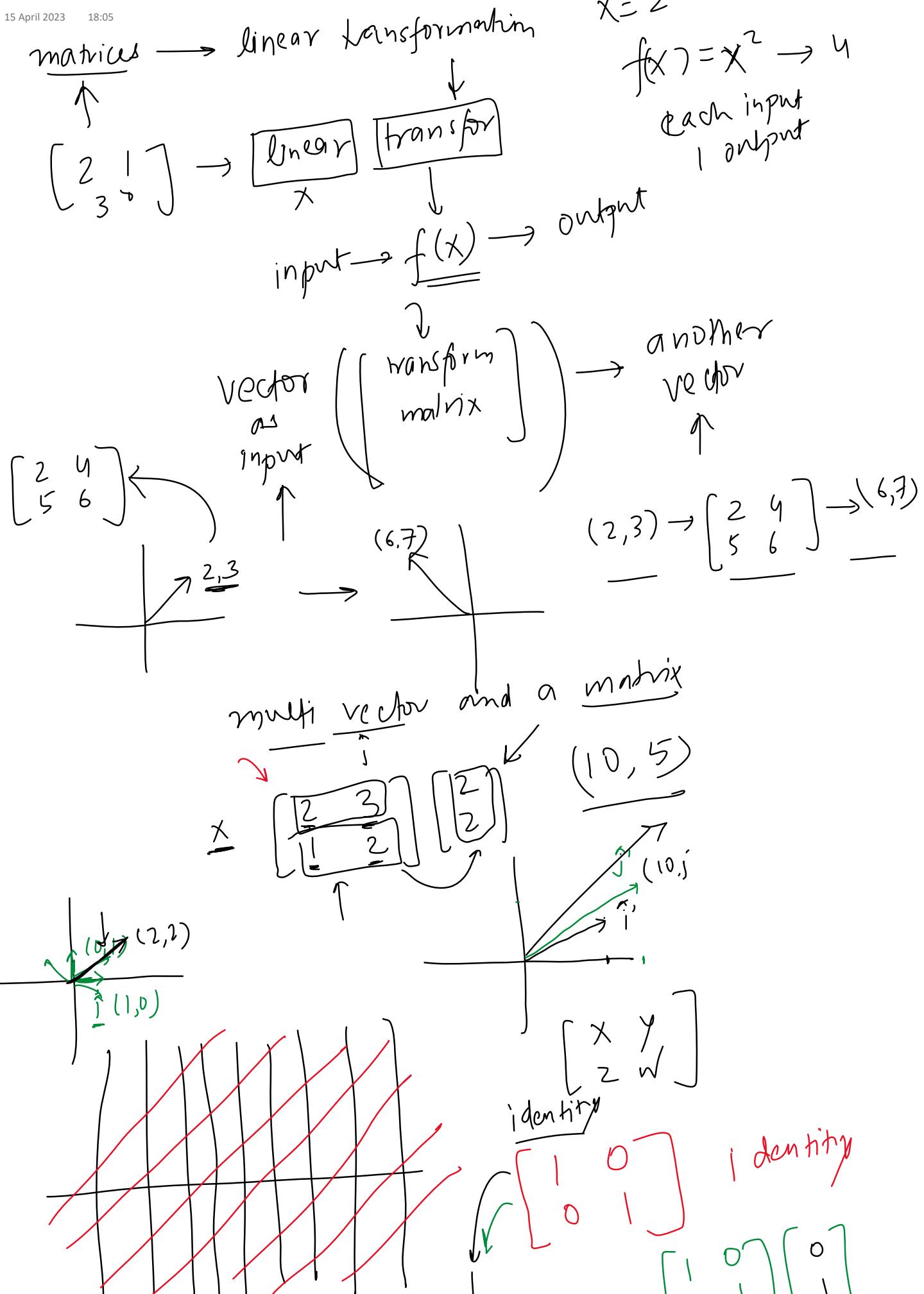


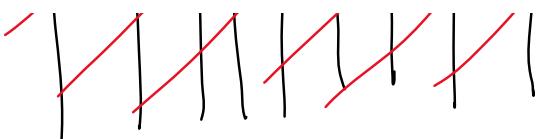
Basis Vector
15 April 2023 18:04



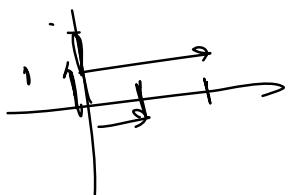
Linear Transformations

15 April 2023 18:05

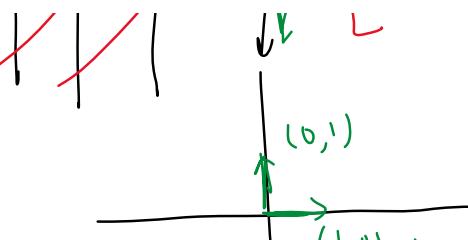




$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

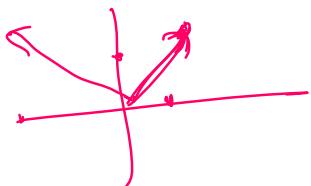
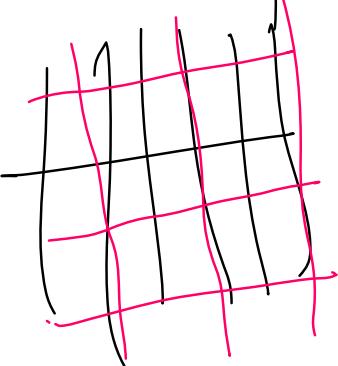


$$\begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

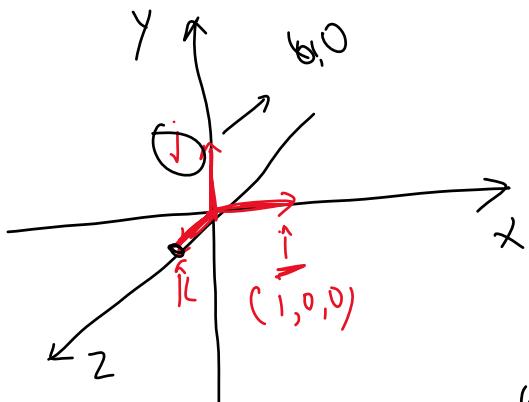
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$



$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

Linear Transformation in 3d

15 April 2023 18:16



$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = [2, 5, 8]$$

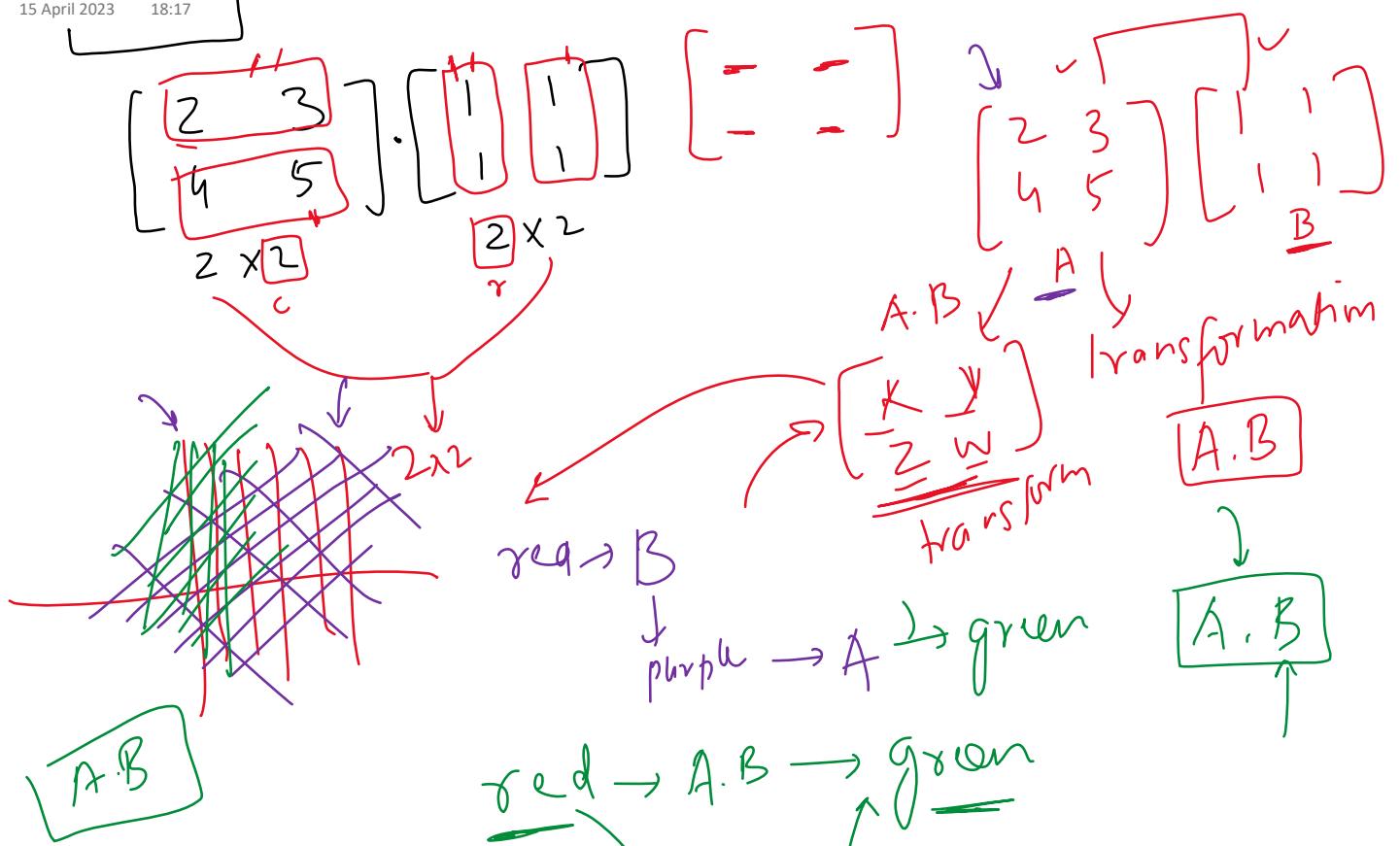
linear



$$(3, 0, 9)$$

Matrix Multiplication as Composition

15 April 2023 18:17



$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad B = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$$

$\underline{c} \rightarrow A \cdot B \rightarrow \underline{c}''$

$\underline{c} \rightarrow A \cdot B \rightarrow \underline{c}' \rightarrow A \rightarrow \underline{c}''$

3rd matrix
coordinate space

$\underline{c} \rightarrow B \rightarrow \underline{c}' \rightarrow A \rightarrow \underline{c}''$ composition of transformations

$$A \cdot B \rightarrow \underline{c}''$$

intuition

$$\begin{bmatrix} A \cdot B \\ X \\ Y \end{bmatrix} \rightarrow \begin{bmatrix} A \cdot B \\ X' \\ Y' \end{bmatrix}$$

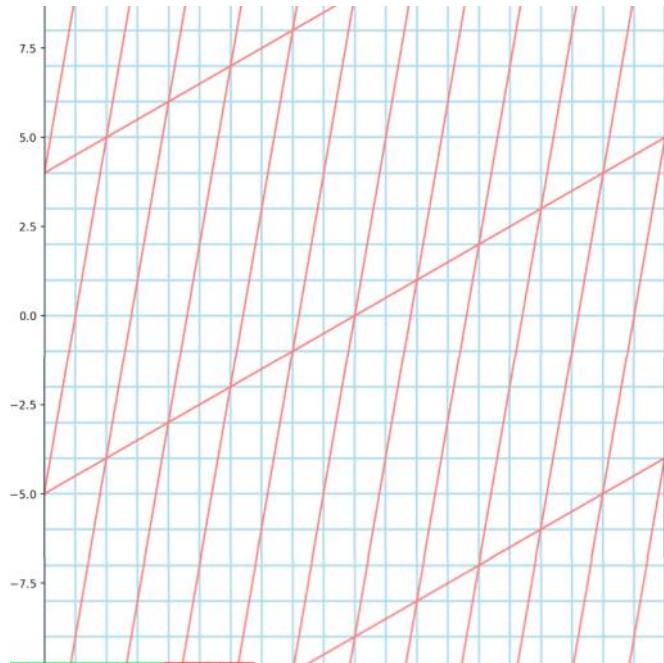
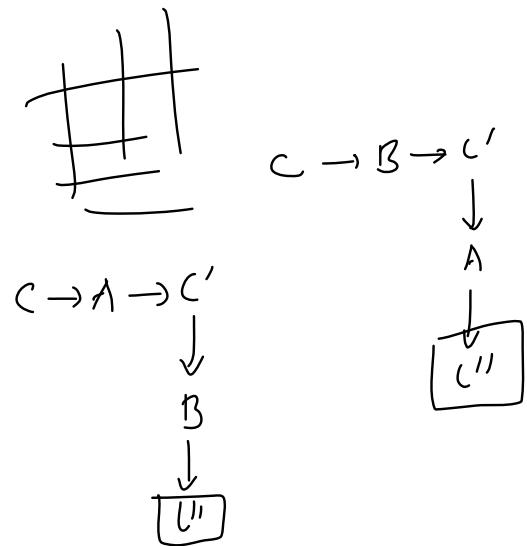
$$A \cdot B \begin{bmatrix} X \\ Y \end{bmatrix} \rightarrow \begin{bmatrix} X'' \\ Y'' \end{bmatrix}$$

$$A \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x' \\ y' \end{pmatrix}$$

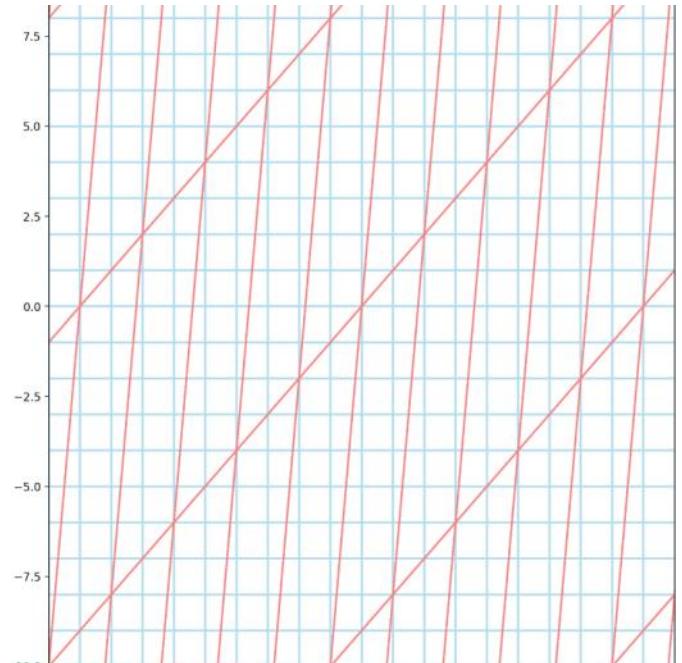
Test of Commutative Law

15 April 2023 18:17

$$\begin{array}{c} \cancel{A \cdot B} \neq \cancel{B \cdot A} \\ \checkmark A \cdot (B \cdot C) = (A \cdot B) \cdot C \end{array}$$



$B \cdot A \rightarrow A$ first then B



$A \cdot B \rightarrow B$ first then A

Determinant

15 April 2023 18:29

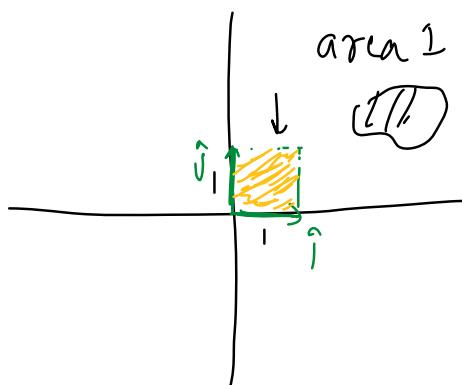
$$\begin{array}{c} |x| \quad 2 \times 2 \quad 3 \times 3 \\ \hline \Delta = \left| \begin{array}{cc} 2 & 3 \\ 4 & 5 \end{array} \right| = 10 - 12 = -2 \end{array}$$

non-square
transform

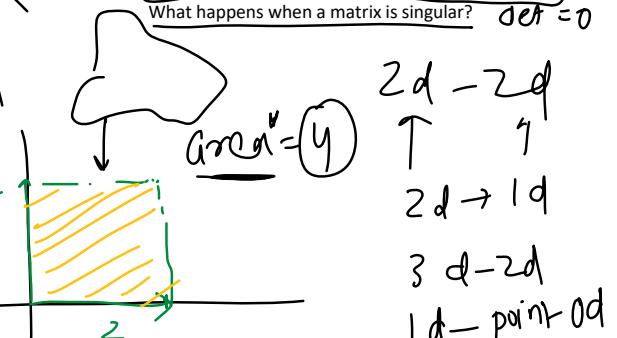
Why determinant is possible only for square matrices.
The interpretation of the determinant as a scaling factor is only meaningful for square matrices because the input and output spaces must have the same dimension for this concept to be applicable.

What does it mean to have a negative determinant?
What happens when a matrix is singular? $\det = 0$

$$\left[\begin{array}{cc} 2 & 0 \\ 0 & 2 \end{array} \right] = A$$

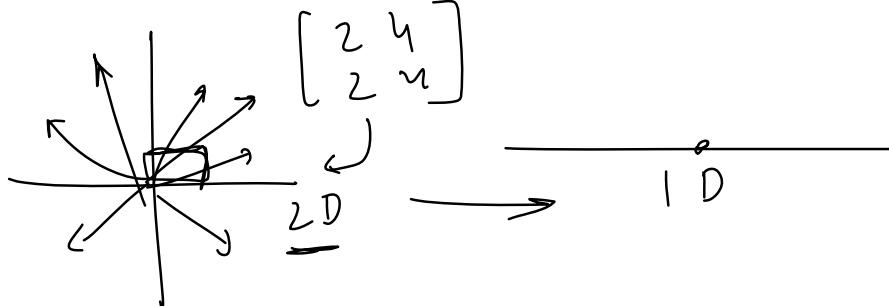


$$\det(A) =$$



$$\frac{\text{Area}'}{\text{Area}} = \frac{4}{1} = 4$$

$$\left| \begin{array}{cc} 2 & 4 \\ 2 & 4 \end{array} \right| = 2 \cdot 4 - 2 \cdot 4 = 0 \rightarrow \boxed{\text{inversc}}$$



Inverse

15 April 2023 18:32

$$A \rightarrow A^{-1}$$

$$A^{-1}A = I$$

$$A X = B$$

↑
some vector X

$$A X = B$$

$$\begin{array}{l} A A^{-1} = I \\ \uparrow \quad \uparrow \\ x + 2y = 5 \\ 3x + 5y = 6 \\ \rightarrow \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \end{bmatrix} \\ \text{---} \quad \text{---} \\ \text{---} \quad \text{---} \end{array}$$

$$A^{-1}A X = X^{-1}B$$

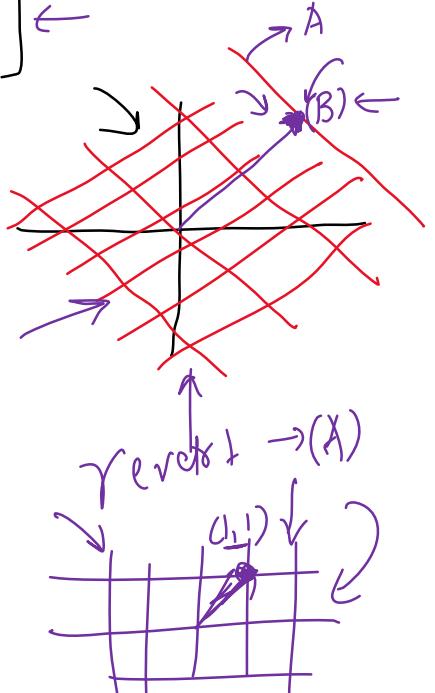
$$\begin{array}{l} I X = A^{-1}B \\ \downarrow \\ X = A^{-1}B \end{array}$$

reversing
the transform

$$A A^{-1} = I$$

Why $A \cdot A^{-1} = I$?
Why inverse is possible for square matrix only

$$\begin{array}{l} A^{-1} \cdot B \\ \rightarrow X, Y \end{array}$$



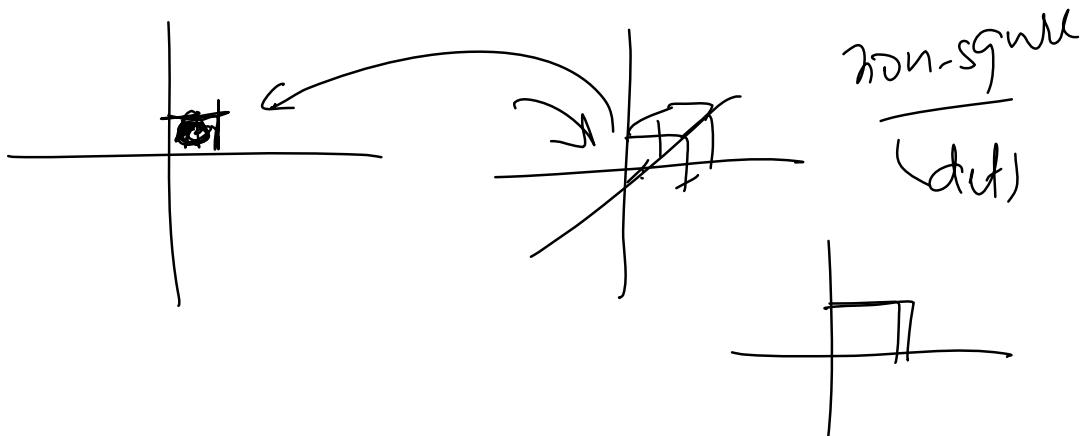
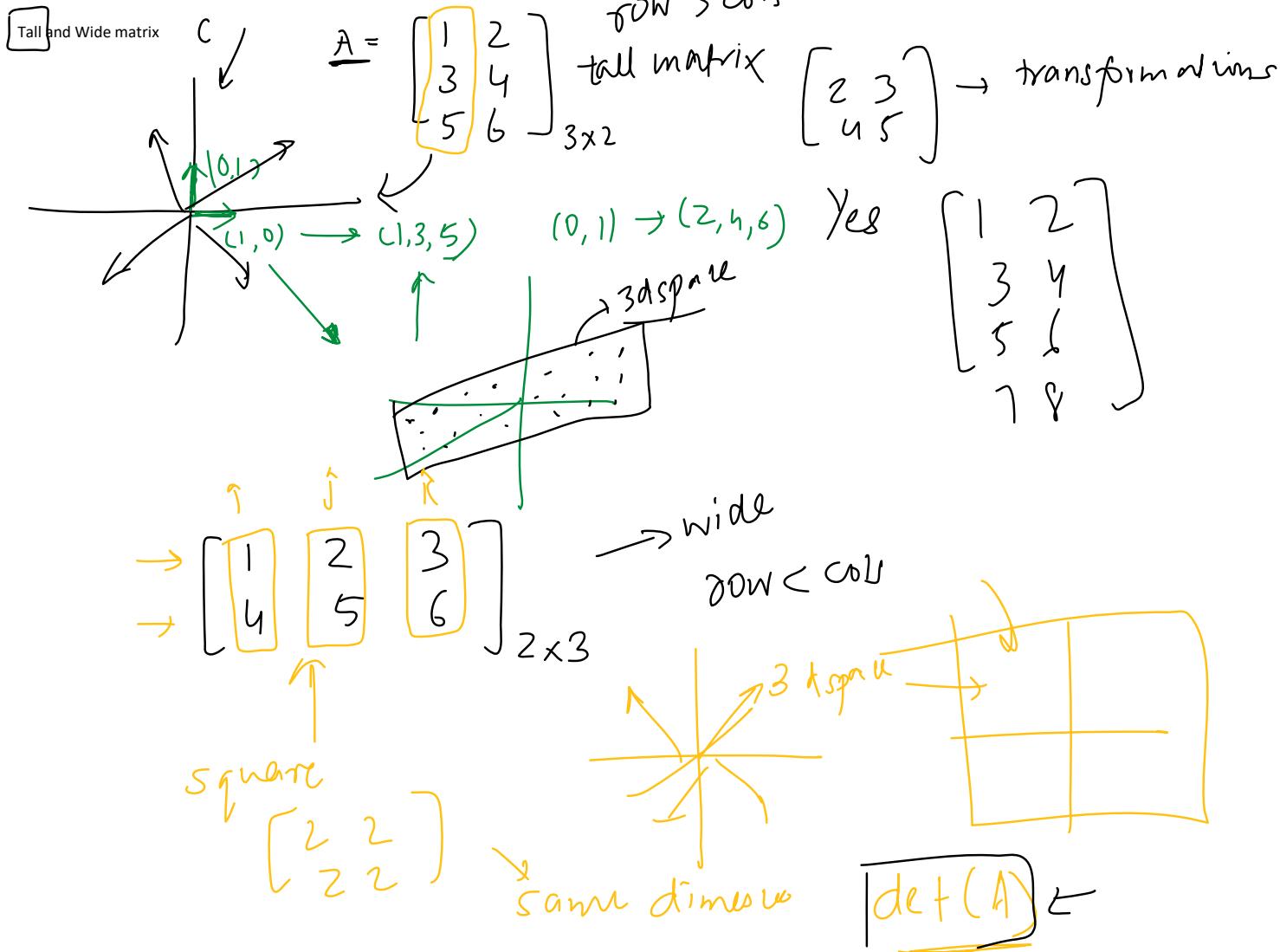
$$A A^{-1} = I$$

transformation for non square matrix?

15 April 2023 18:33

Square matrices ($n \times n$) represent linear transformations where the domain and codomain vector spaces have the same dimensions, i.e., $T: V \rightarrow V$. In these cases, the transformation maps a vector space onto itself. Non-square matrices can also represent linear transformations between vector spaces with different dimensions.

Square matrices



Why only square matrix has inverse

15 April 2023 19:11

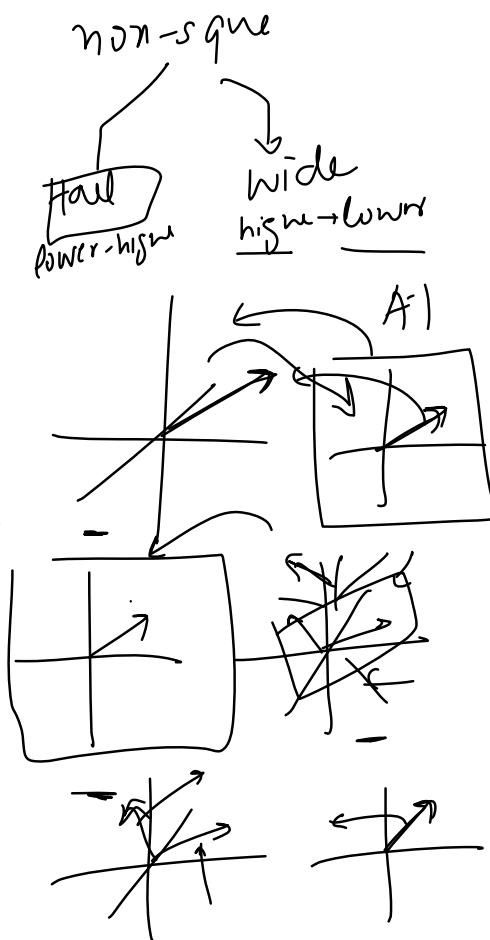
Non-square inv

An inverse is possible only for square matrices because it is related to the concept of a matrix being a bijective linear transformation, which implies both injectivity (one-to-one) and surjectivity (onto). A square matrix represents a linear transformation between vector spaces of the same dimension, where the domain and codomain are the same. When a square matrix is invertible, its linear transformation is bijective, meaning that it has a unique inverse transformation.

Let's consider why non-square matrices cannot have inverses:

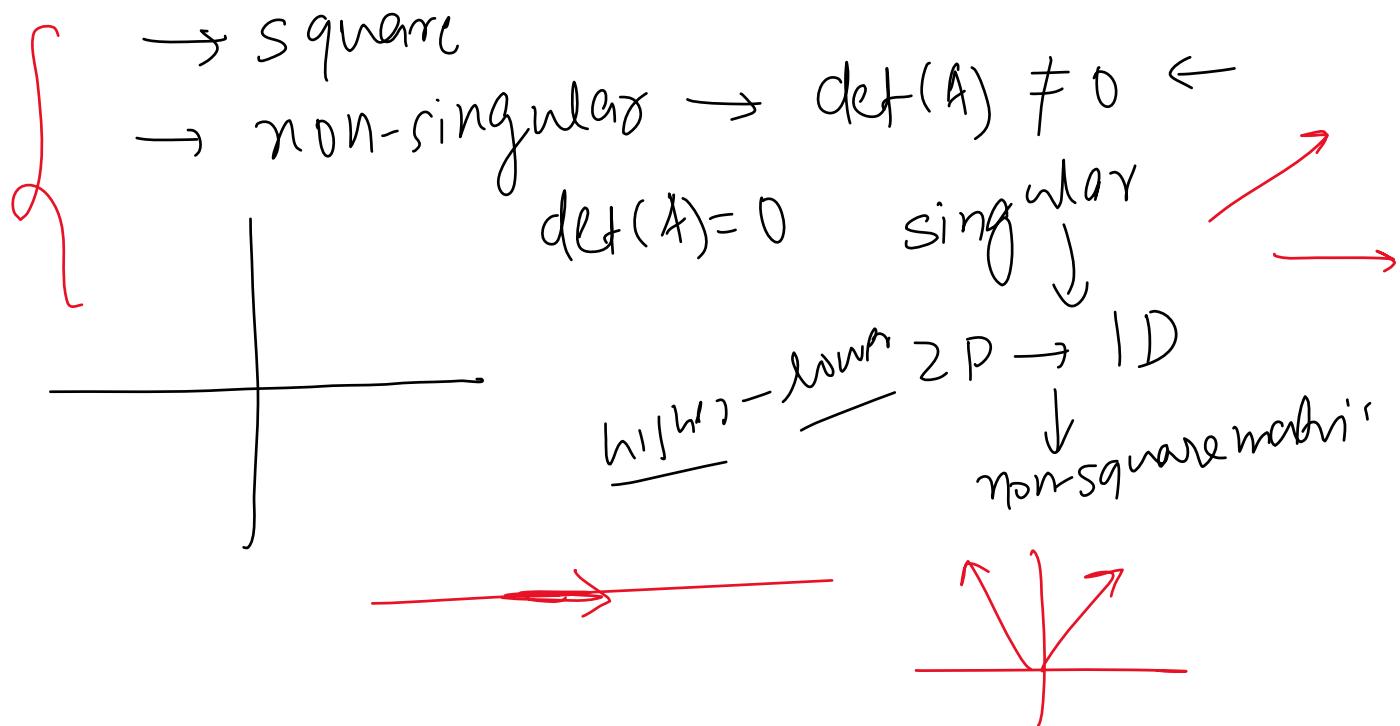
1. If a matrix A has more rows than columns ($m > n$), i.e., a tall matrix, the linear transformation it represents is from a lower-dimensional space to a higher-dimensional space. In this case, the transformation is generally not surjective (onto), as there are output vectors in the higher-dimensional space that have no corresponding input vector. Consequently, there is no inverse transformation that can map every output vector back to an input vector.
2. If a matrix A has more columns than rows ($m < n$), i.e., a wide matrix, the linear transformation it represents is from a higher-dimensional space to a lower-dimensional space (dimension reduction). In this case, the transformation is generally not injective (one-to-one), as multiple input vectors can map to the same output vector. Consequently, there is no unique inverse transformation that can map each output vector back to a unique input vector.

Again, the inverse of a matrix is possible only for square matrices because these matrices represent linear transformations between vector spaces of the same dimension. Only in these cases can a matrix potentially satisfy the conditions of being a bijective transformation, i.e., both injective and surjective, which allows the existence of a unique inverse transformation. However, not all square matrices have inverses; only those that are non-singular (with a non-zero determinant) have an inverse.



Why inverse is possible for non-singular matrices only

15 April 2023 19:50



Data matrix -> representation

15 April 2023 18:33

Matrix multiplication

15 April 2023 18:33

Hadamard product

15 April 2023 18:33

The Hadamard product, also known as the element-wise product or Schur product, is a binary operation that takes two matrices of the same dimensions and produces a third matrix where each element is the product of the corresponding elements of the input matrices. Specifically, given two matrices A and B of the same size $m \times n$, their Hadamard product C is also an $m \times n$ matrix, where each element is defined as:

$$C[i, j] = A[i, j] * B[i, j]$$

for all $i = 1, \dots, m$ and $j = 1, \dots, n$.

<https://ezyang.github.io/convolution-visualizer/>

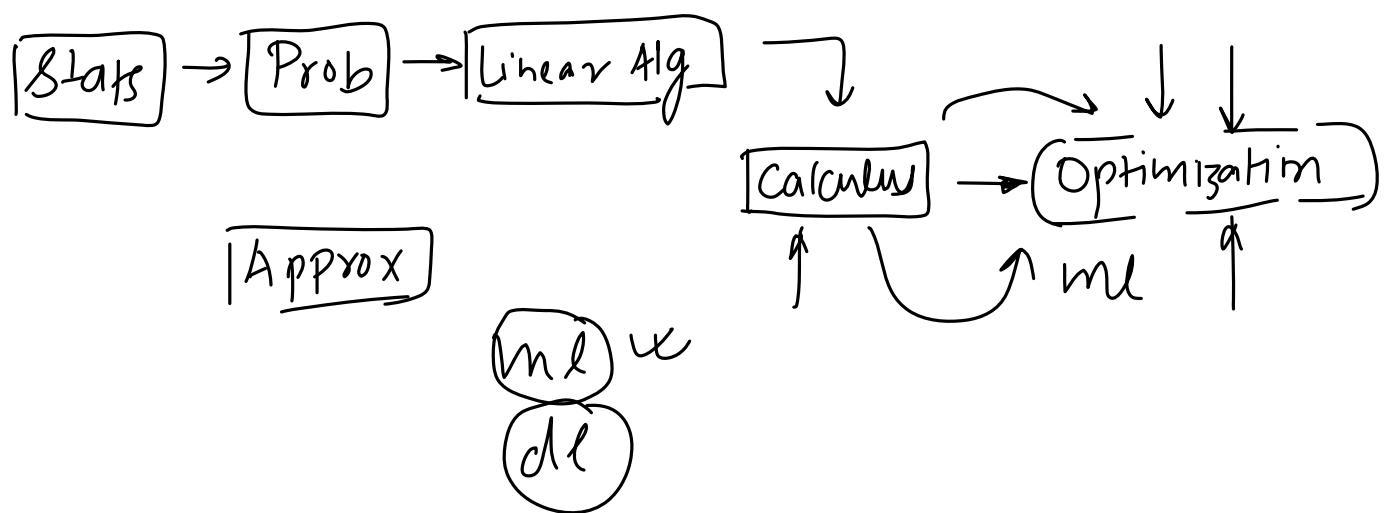
$S \cup R \rightarrow$

python →

maths →

python code?

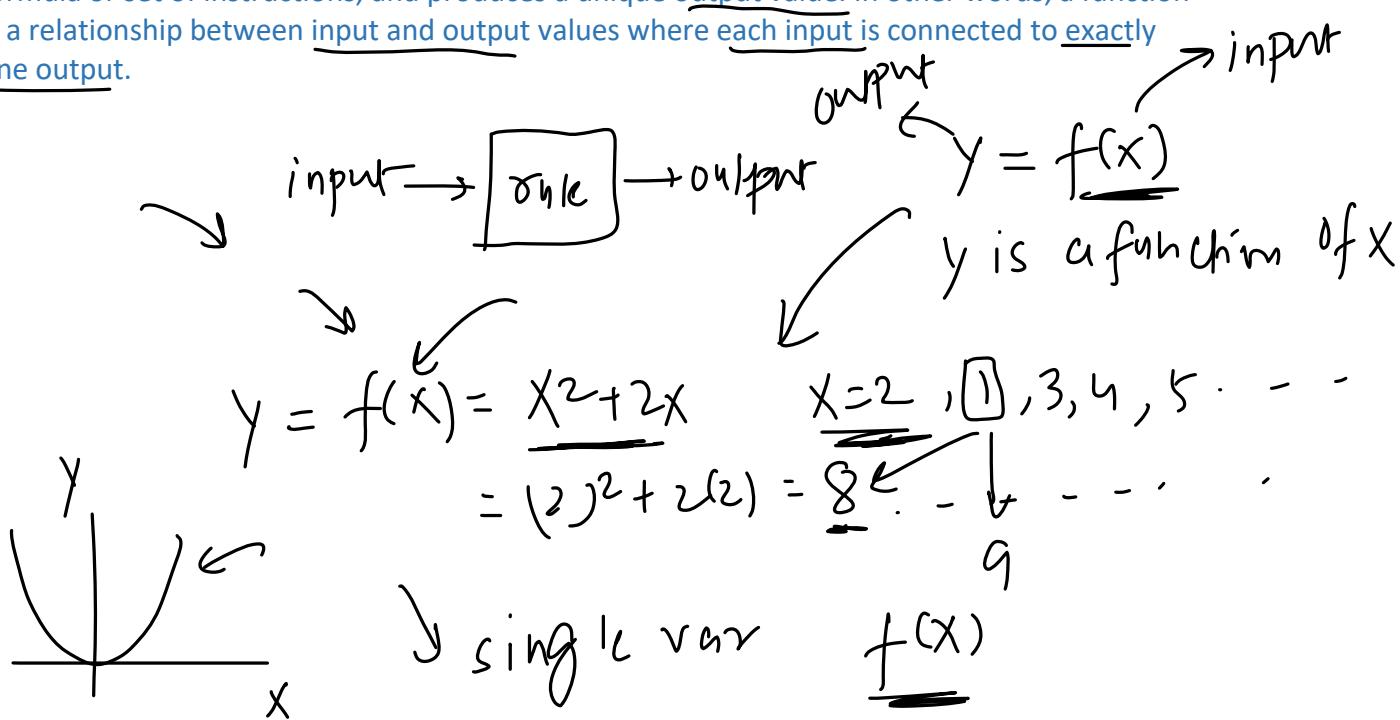
Regression includes



Functions

21 April 2023 14:28

A function is a mathematical rule that takes an input value, processes it according to a specific formula or set of instructions, and produces a unique output value. In other words, a function is a relationship between input and output values where each input is connected to exactly one output.

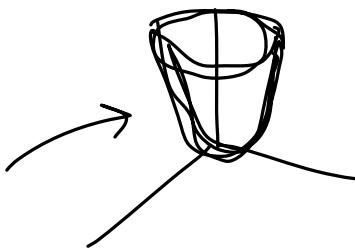


Multivariable Functions

21 April 2023 16:35

$$z = f(x, y) = x^2 + y^2$$

↑ ↑ ↑
output input Z+1



$$f(\lambda) \quad 1+1 \rightarrow 2$$

n+1 dim

$$x=1 \quad y=1 \rightarrow z=2$$

$$f(\underline{x_1, x_2, x_3, \dots, x_n})$$

$$\underline{x=1} \quad \underline{y=1} \rightarrow \underline{z \neq 2}$$

Parameters in a Function

21 April 2023 16:36

$$\underline{ax^2 + bx + c}$$

In mathematics, parameters of a function are the variables that are used to define the behaviour of the function. The parameters influence the function's output by determining how the input values are processed.

The parameters are the constants or coefficients that appear in the function's formula. For example, in the quadratic function $f(x) = ax^2 + bx + c$, 'a', 'b', and 'c' are the parameters of the function. By changing the values of these parameters, you can modify the shape and position of the parabola represented by the function.

$$y = f(x) = 5x^2$$

↑
parameter input

$$6x^2$$

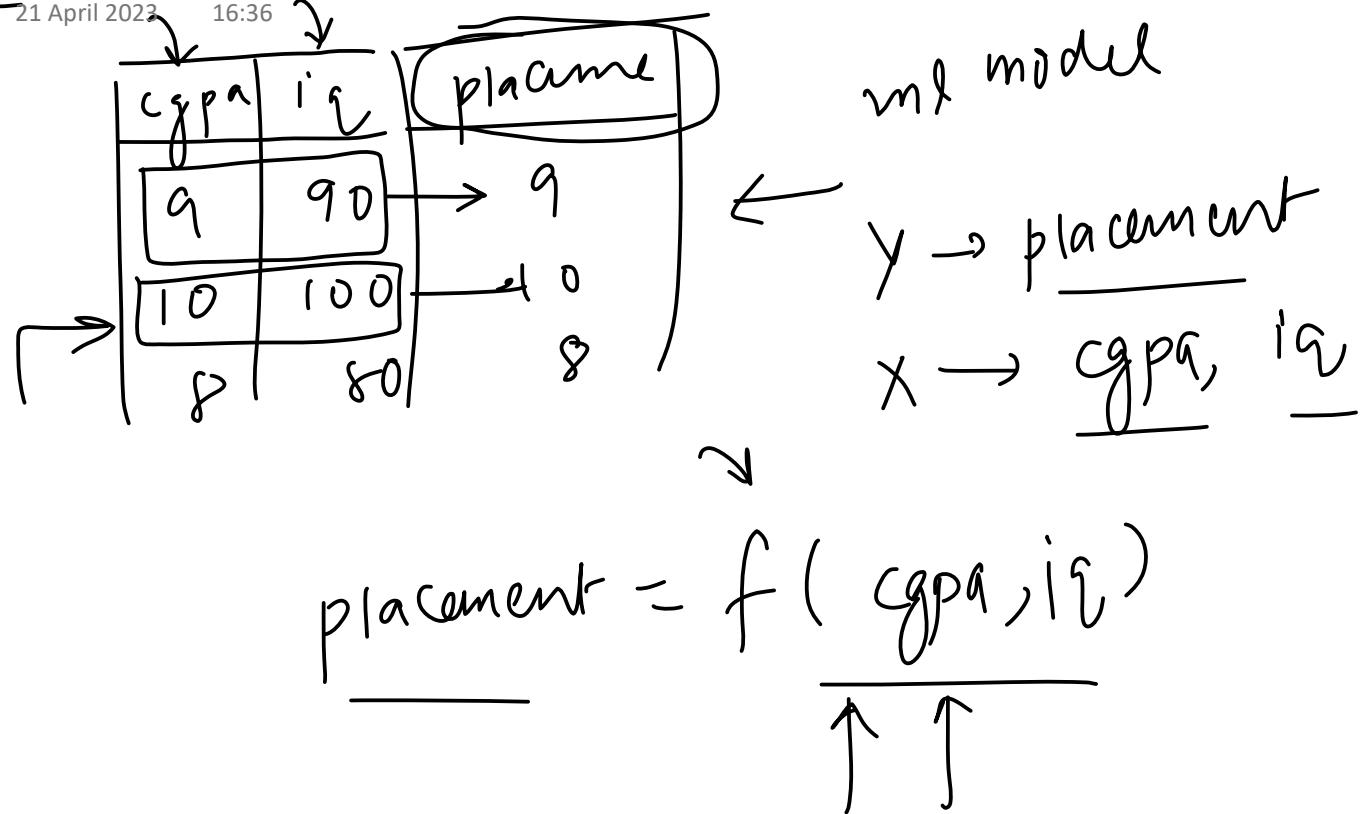
$$y = f(x) = ax^2$$

↑
parameter

[ML models] as Mathematical Function

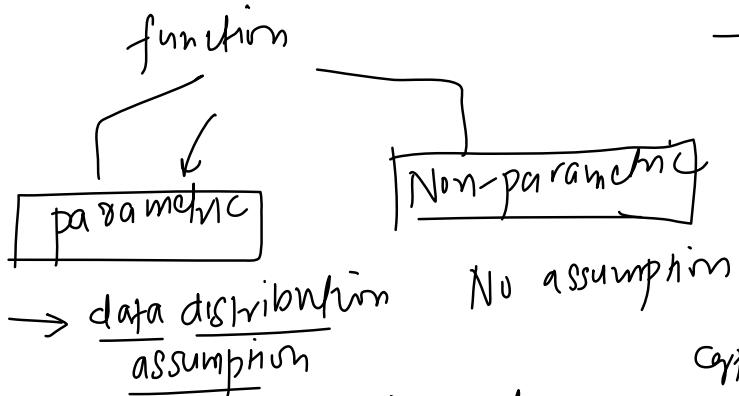
21 April 2023

16:36



Parametric Vs Non-Parametric ML models

21 April 2023 16:36



→ {fixed number
parameters
resp. their
rows}

-17
parametric and
cgpia | place

$$y = mx + b$$

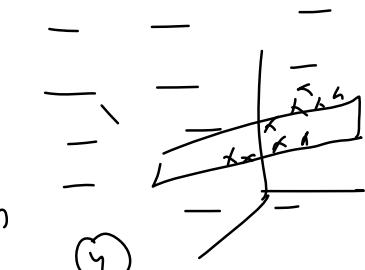
placement

→ cgpa / iq

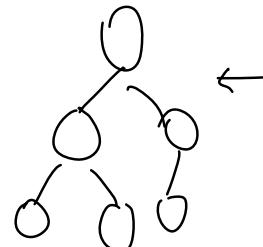
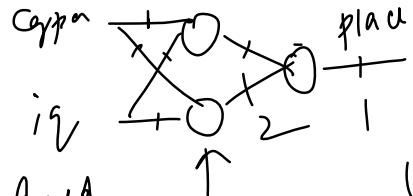
given

$$\text{placement} = f(\text{cgpa}, \text{iq})$$

cgpa iq place



$$y = \beta_0 + \beta_1 \text{cgpa} + \beta_2 \text{iq}$$



ml model

linear reg

8, 80

$$y = 1 + 2 \times 8 + 3 \times 80$$

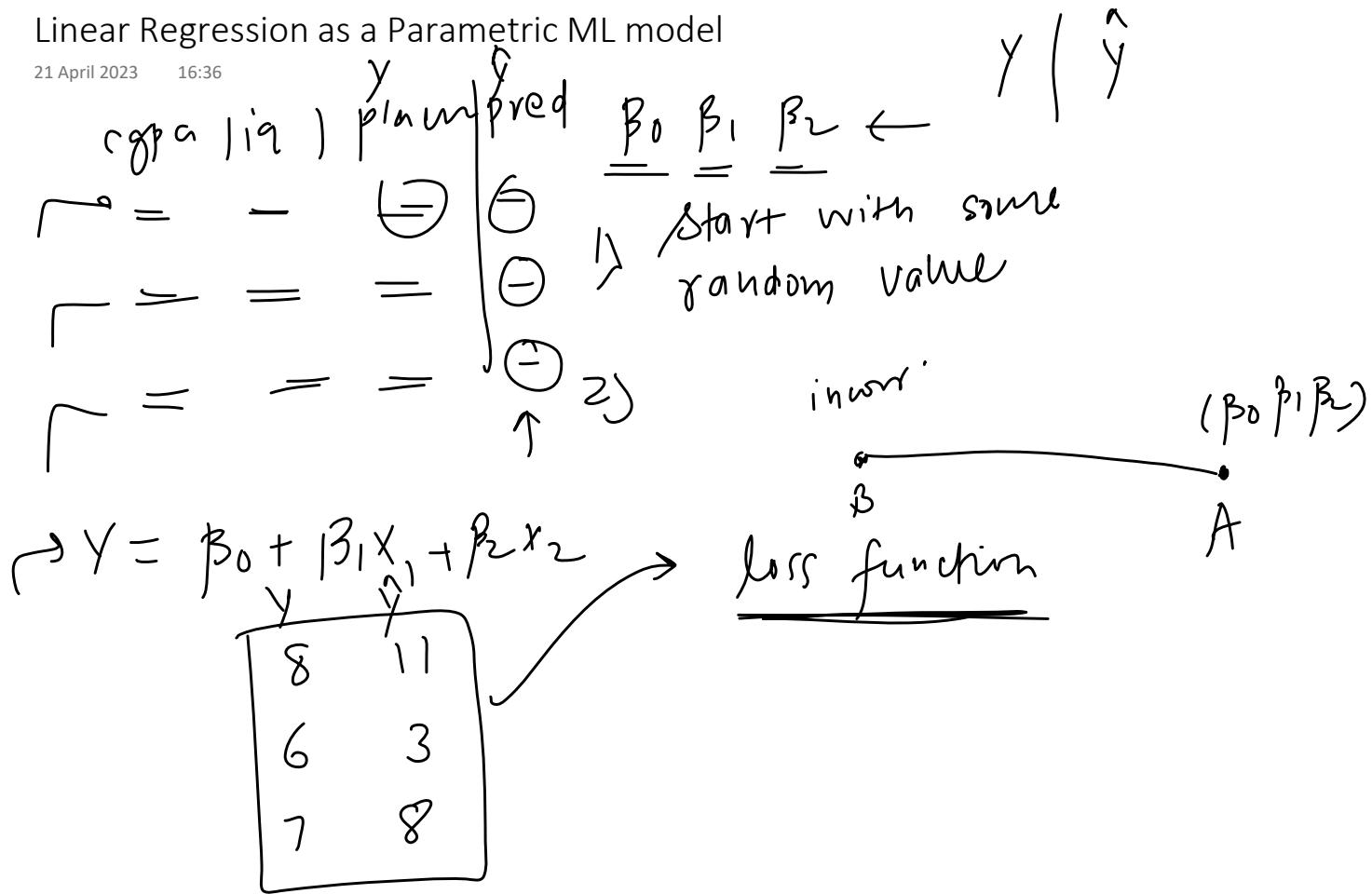
1, 2, 3

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\rightarrow \beta_0 \beta_1 \beta_2 \rightarrow \text{parameters}$$

Linear Regression as a Parametric ML model

21 April 2023 16:36

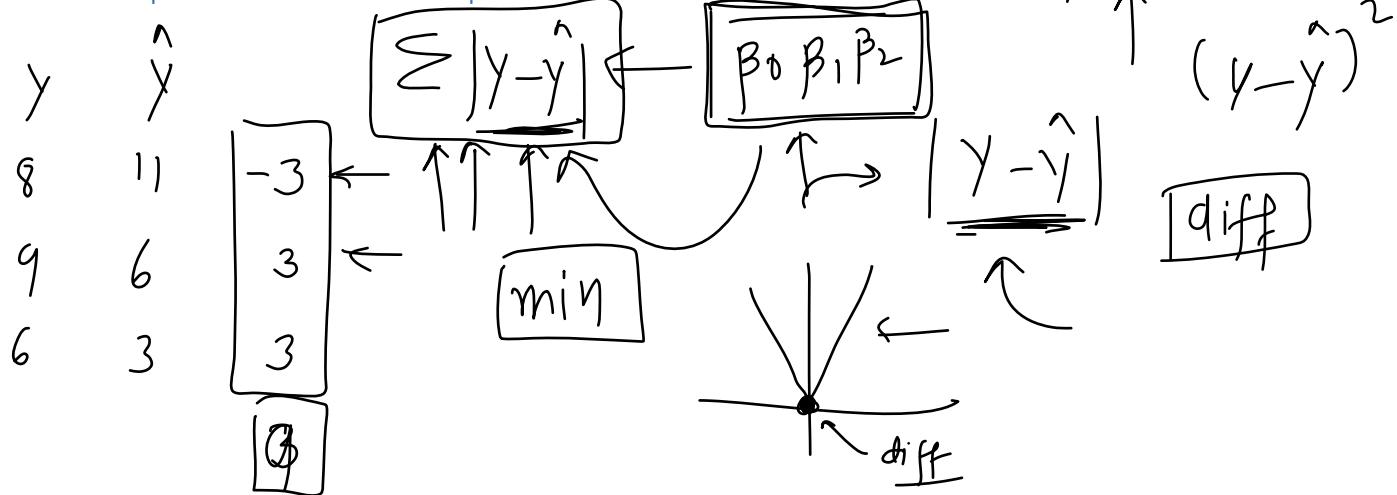


Loss Function

21 April 2023 16:37

A loss function, also known as a cost function or objective function, is a mathematical function that measures the difference between the predicted output and the actual target values in a machine learning model. The primary goal of training a machine learning model is to minimize the value of the loss function, which corresponds to improving the model's performance on the given task.

Loss functions play a crucial role in the optimization process, guiding the learning algorithm to adjust the model's parameters to achieve better predictions.



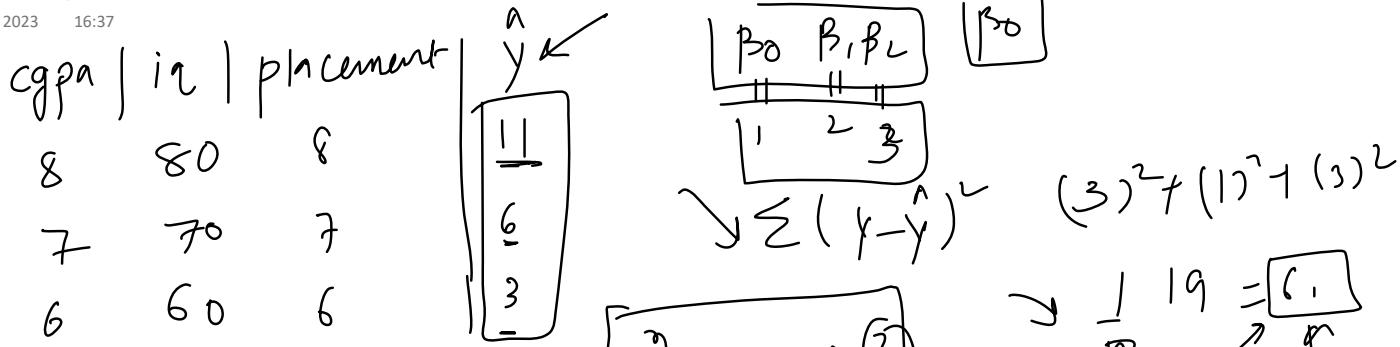
How to select a good Loss Function

21 April 2023 16:37

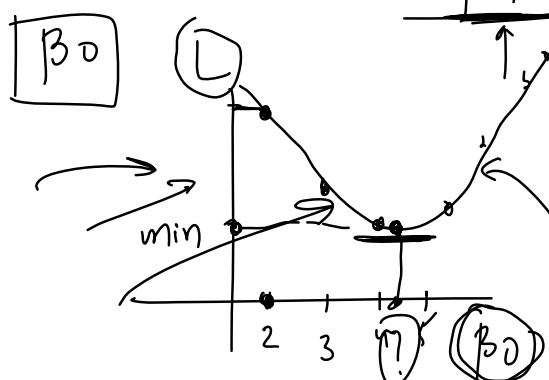
1. **Problem type:** The choice of a loss function depends on the type of problem you are solving. For example, in regression tasks, mean squared error (MSE) or mean absolute error (MAE) are commonly used. For binary classification, cross-entropy loss or hinge loss can be employed. For multi-class classification, categorical cross-entropy or multi-class hinge loss can be used. Choose a loss function that aligns with the objectives of the specific problem you are addressing.
2. **Robustness to outliers:** Some loss functions, like mean squared error, are more sensitive to outliers, which can lead to a model that is overly influenced by extreme values. If your dataset contains outliers or is prone to noise, consider using a loss function that is more robust to outliers, such as mean absolute error (MAE) or Huber loss.
3. **Interpretability and ease of use:** A good loss function should be interpretable and easy to implement. Simple loss functions like mean squared error or cross-entropy loss are widely used because they are easy to understand, compute, and differentiate. When possible, opt for a loss function that is easy to work with and can be easily incorporated into your optimization process.
4. **Differentiability:** Most optimization algorithms, like gradient descent, require the loss function to be differentiable. Choose a loss function that has continuous first-order derivatives, which makes it easier to compute the gradients needed for optimization.
5. **Compatibility with the model:** Ensure that the chosen loss function is compatible with the model architecture you are using. Some models have specific requirements or assumptions about the loss function. For example, linear regression assumes a Gaussian noise distribution, which is why mean squared error is a suitable loss function in that case.

Calculating Parameters From a Loss Function (the easy way and $\beta_1=1$ $\beta_2=2$ problem)

21 April 2023 16:37



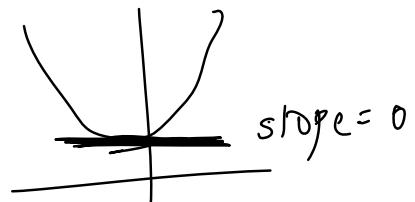
$$L(\beta_0, \beta_1, \beta_2) = \underset{\beta_0, \beta_1, \beta_2}{\operatorname{argmin}}$$



$$L \rightarrow \beta_0$$

parabolic

optimization
activation function



$$\frac{dL}{d\beta_0} = 0$$

$$L = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

linear
OLS

$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

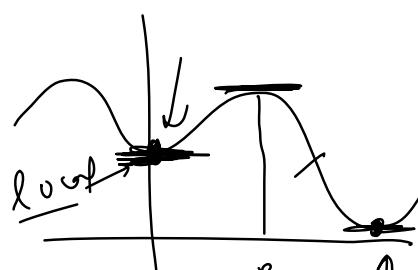
$$\frac{dL}{d\beta_0} = \sum_{i=1}^n (y_i - \hat{y}_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2)^2$$

$$y_i = \beta_0 + x_1 + 2x_2$$

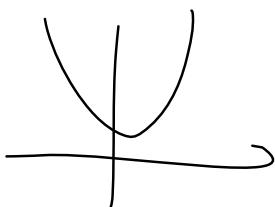
$$-2(y_i - \beta_0 - x_1 - 2x_2) = 0$$

$$\text{slope} = 0$$

$$(y_i - \beta_0 - x_1 - 2x_2) = \beta_0$$



$$\frac{dL}{d\beta_0} = 0$$



$$\left| \frac{dL}{d\beta_0} = 0 \right. \sim \text{minimized summing}$$

$$\boxed{\frac{\partial L}{\partial \beta_0} = 0} \rightarrow \begin{matrix} \text{gauge} \\ \text{limited summation} \end{matrix} \leftarrow$$

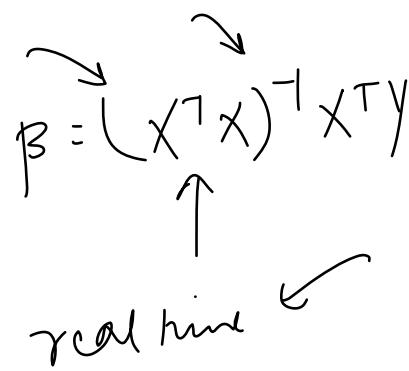
Problem with the easy way

21 April 2023 16:44

1. Non-convexity: The loss function may not always be convex, meaning that it might have multiple local minima and maxima. In such cases, setting the gradient to zero might lead to a local minimum or maximum, which is not necessarily the global minimum (the optimal solution).

$$\beta = (X^T X)^{-1} X^T Y$$

real time



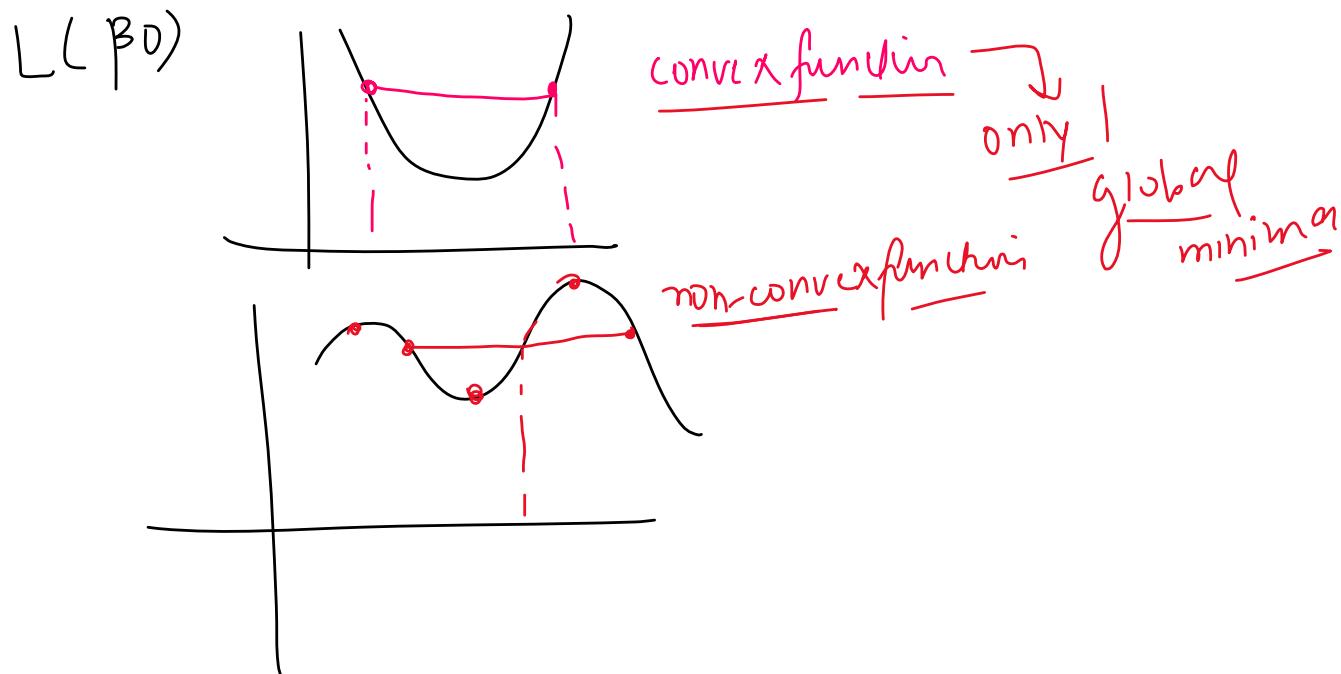
2. Complexity: For some models, the loss function can be highly complex, and finding the analytical solution by setting the gradient to zero might be computationally expensive or even impossible. This is particularly true for deep learning models, where the loss functions involve a large number of parameters and complex relationships between them.

3. Scalability: In large-scale machine learning problems with massive amounts of data or high-dimensional feature spaces, computing the analytical solution by setting the gradient to zero can be computationally prohibitive due to the high cost of processing and storing the data.

4. Online learning and streaming data: In some applications, the data is not available all at once but arrives in a continuous stream. In these scenarios, models need to be updated incrementally as new data arrives, and an analytical solution would not be practical. Gradient descent and its variants, such as stochastic gradient descent, are well-suited for online learning and handling streaming data.

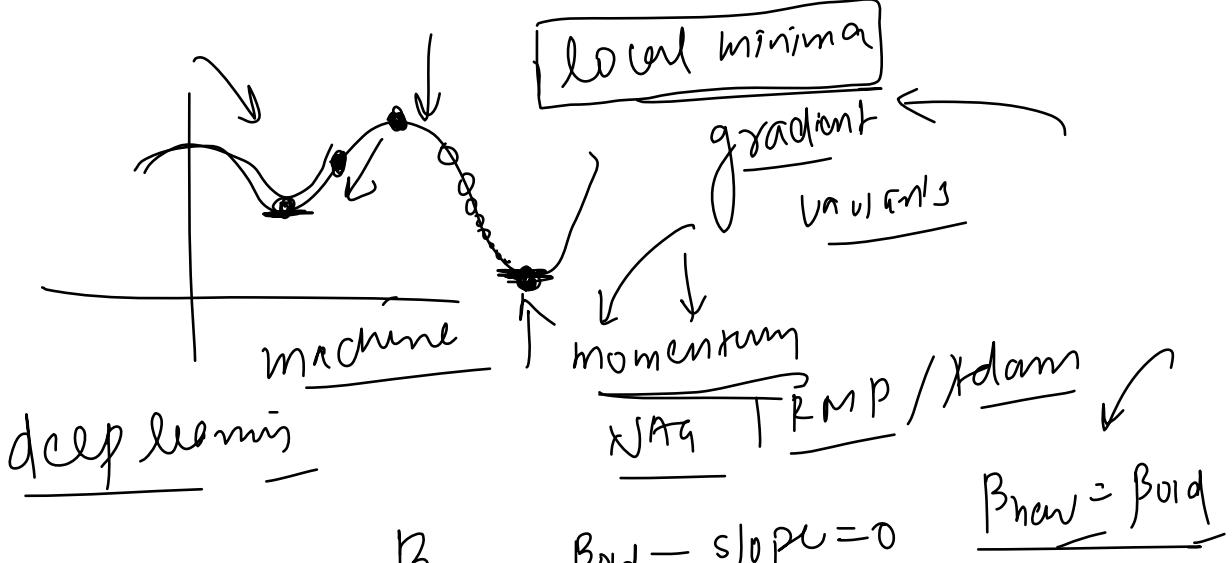
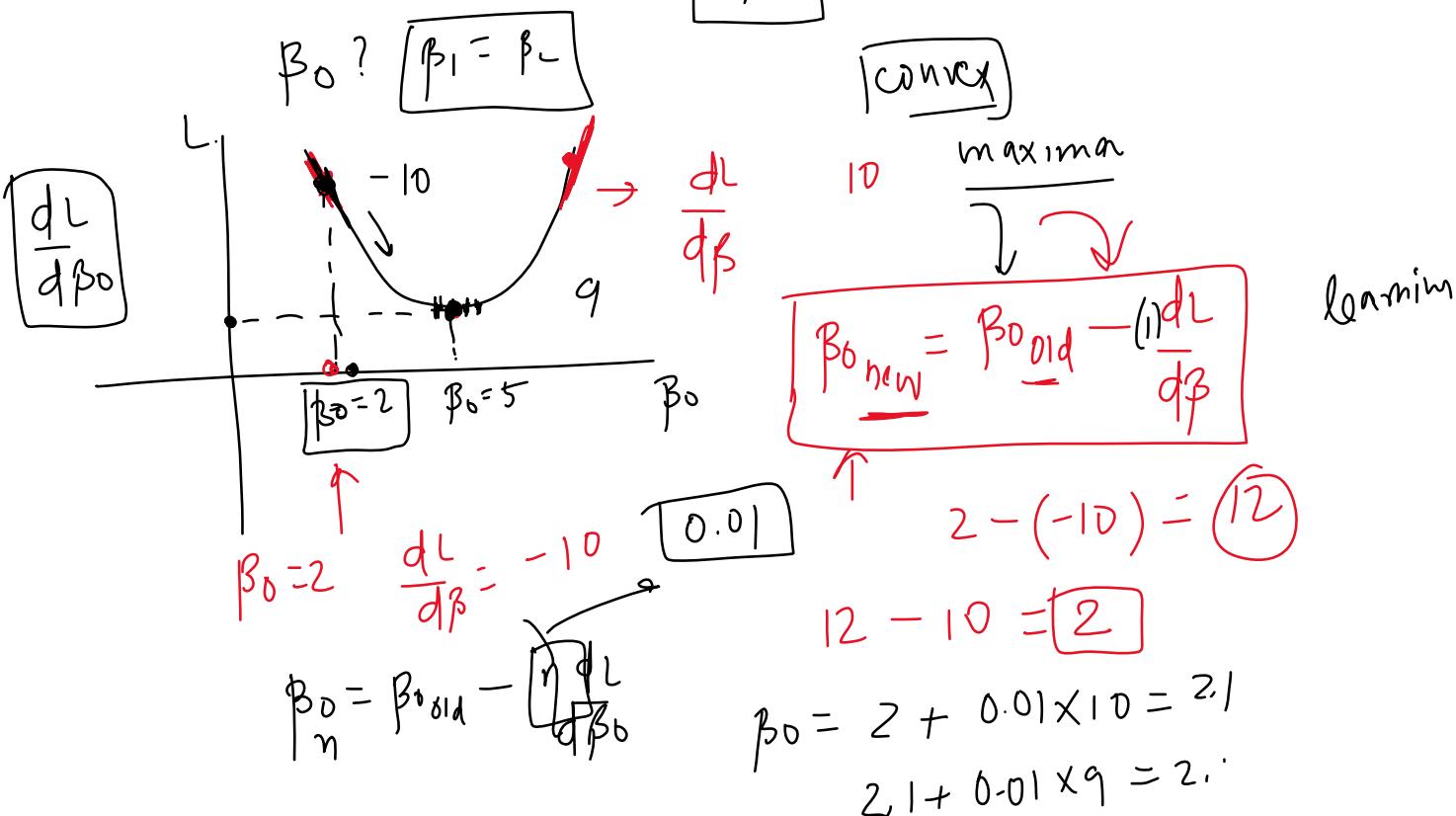
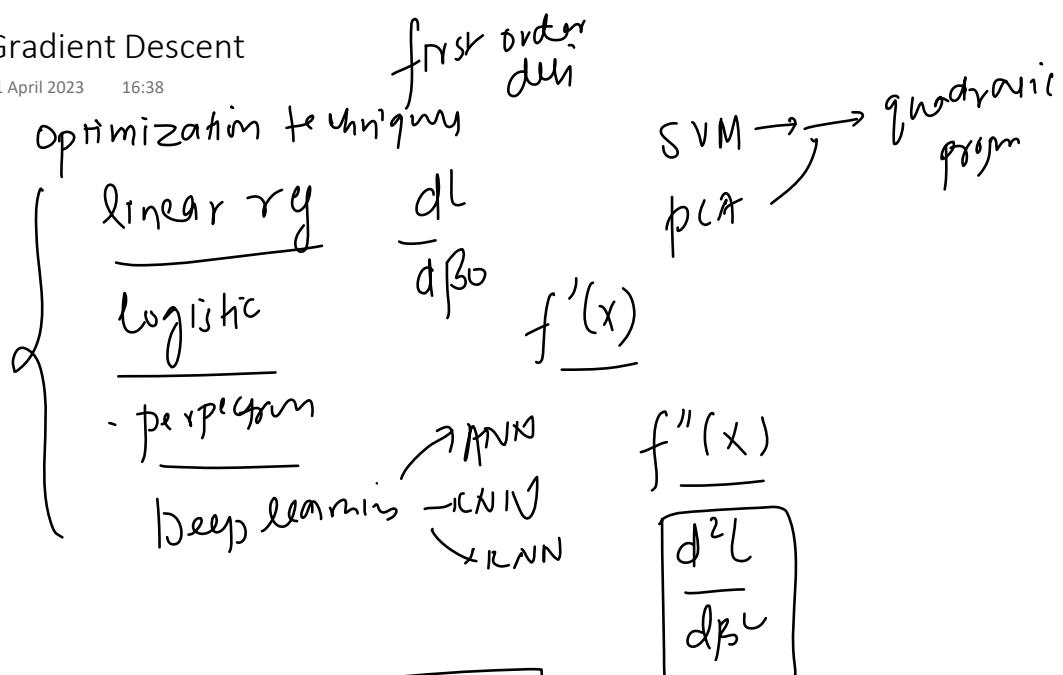
Convex And Non Convex Loss Functions

21 April 2023 16:38



Gradient Descent

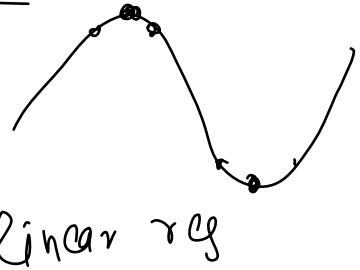
21 April 2023 16:38



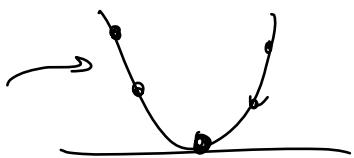
$$\beta_0_{\text{new}} = \beta_{\text{old}} - \frac{\text{slope}}{} = 0$$

$\beta_{\text{new}} = \beta_{\text{old}}$

epochs



linear reg



convex
1 minima

$$\boxed{\beta = 5} \quad \min \quad \max$$

$$f(\beta) = \beta^2 + 2 \quad \beta = 4.9 \\ = 27 >$$

epochs $(4.9)^2 + 2 = \boxed{26}$

$$\boxed{\beta_n = \beta_0 - \eta \frac{\partial L}{\partial \beta}}$$

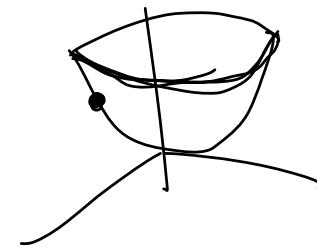
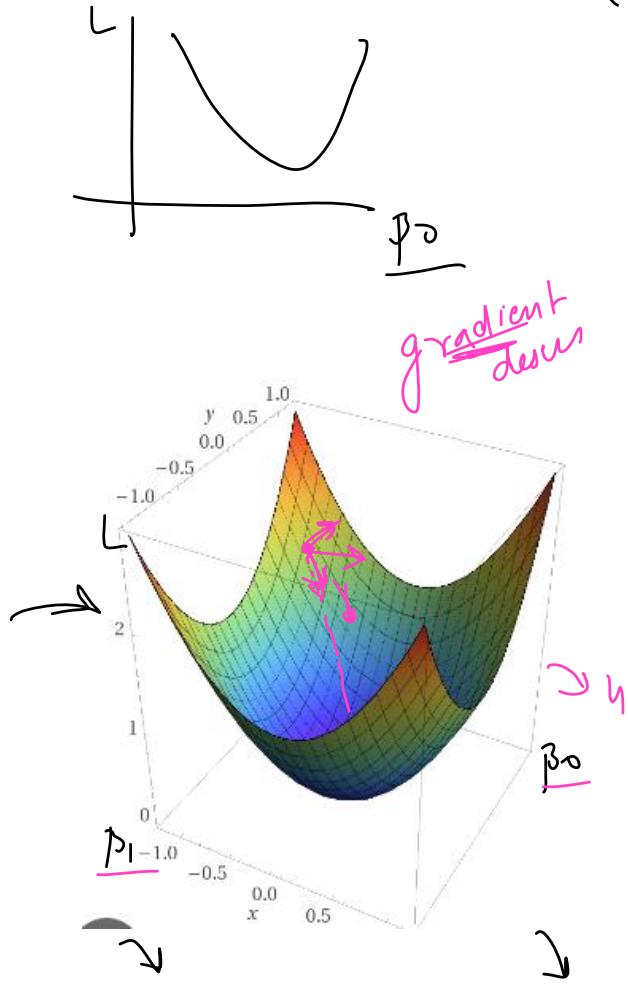
Gradient Descent with multiple Parameters

21 April 2023 16:39

$$L(\beta_0) \rightarrow \text{cols } x_1 \ x_2 \ \dots \ x_n \}$$

$$\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_n$$

$$(n+1) \quad L \ \underline{\beta_0} \ \underline{\beta_1}$$

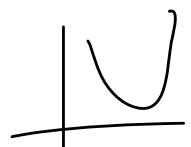


$$\frac{\partial L}{\partial \beta_1} \quad \frac{\partial L}{\partial \beta_0} \quad \beta_0 \ \beta_1 \ \beta_2$$

$$\boxed{\begin{aligned}\beta_0 &= \beta_0 - \eta \frac{\partial L}{\partial \beta_0} \\ \beta_1 &= \beta_1 - \eta \frac{\partial L}{\partial \beta_1}\end{aligned}}$$

$$\beta_2 = \beta_2 - \eta \frac{\partial L}{\partial \beta_2}$$

$$\nabla L(\underline{\beta_0 \dots \beta_n}) = \left(\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}, \dots, \frac{\partial L}{\partial \beta_n} \right)$$

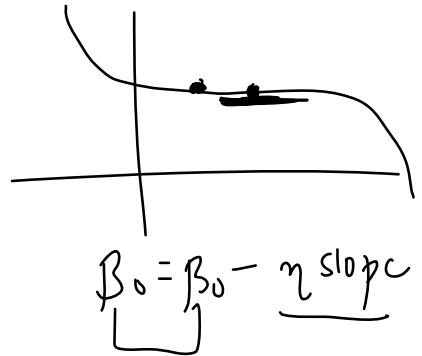


Problems faced in Optimization $\rightarrow \text{GD}$

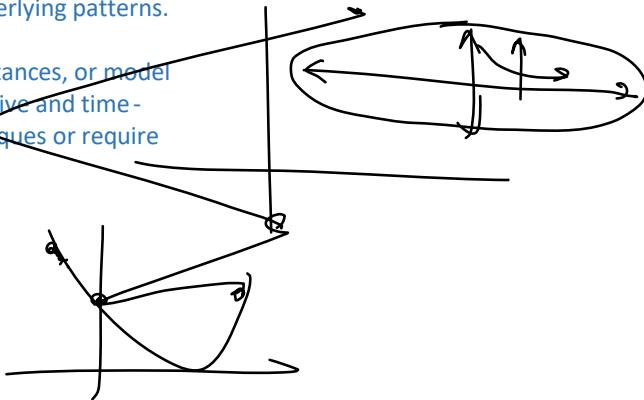
21 April 2023 16:39



1. **Non-convexity:** For many machine learning models, such as artificial neural networks, the loss function is non-convex, which means it has a complex landscape with multiple local minima, maxima, and saddle points. This makes it difficult for optimization algorithms to find the global minimum and can result in suboptimal solutions.
2. **Ill-conditioning:** The loss function may be ill-conditioned, meaning the gradients in some dimensions are much larger than in others. This can cause gradient-based optimization algorithms, such as gradient descent, to oscillate and converge slowly.
3. **Vanishing and exploding gradients:** In deep neural networks, the gradients can become very small (vanish) or very large (explode) as they propagate through the layers. This can lead to slow convergence or unstable training dynamics, making it difficult to optimize the loss function.
4. **Overfitting:** When optimizing the loss function, the algorithm may overfit the training data, resulting in a model that performs poorly on unseen data. This occurs when the model is too complex and learns the noise in the training data instead of the underlying patterns.
5. **Scalability:** For large-scale problems with a high number of features, instances, or model parameters, optimizing the loss function can be computationally expensive and time-consuming. This can limit the applicability of certain optimization techniques or require significant computational resources.



\checkmark $\frac{dL}{d\beta} \rightarrow \text{small} \quad \text{big}$
 neuron
 $(m, b) \rightarrow \boxed{10 \text{ thm}} \rightarrow \text{diff}$
 \downarrow
60 billion



Other optimization techniques

21 April 2023 16:39

gradient \rightarrow ds

Argmin β^0

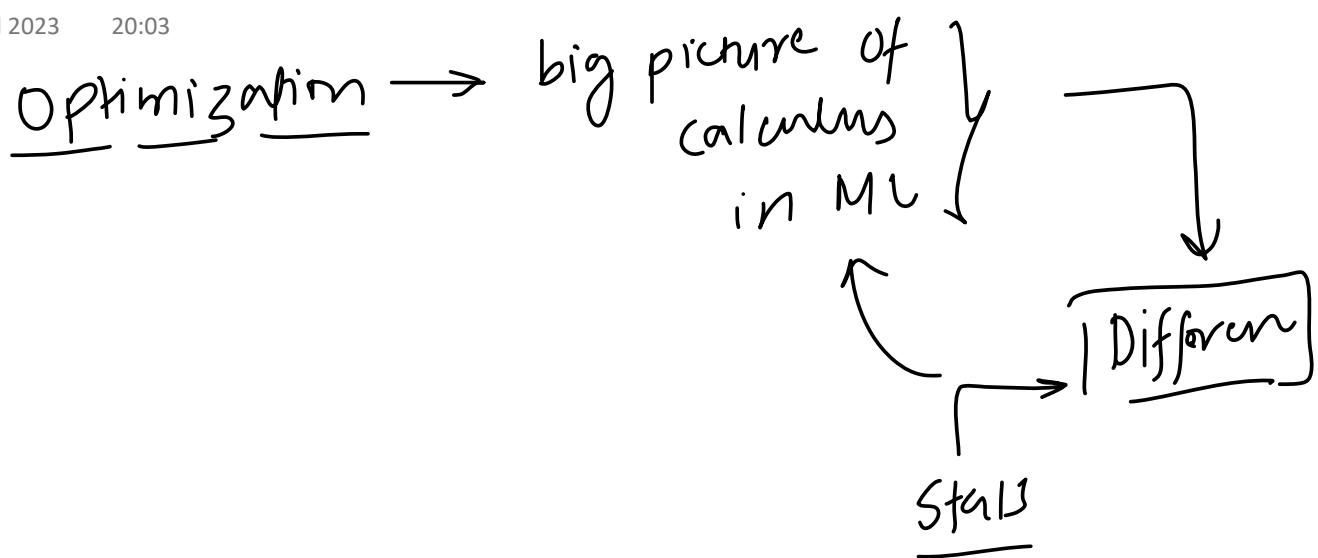
$(\underline{y} - \bar{y})^2$

given $\boxed{\bar{y} = 2}$

constraint $\frac{\text{SVM}}{\text{PCA}}$

constraint $\frac{\text{SVM}}{\text{PCA}}$

The diagram illustrates a constrained optimization problem. It starts with a general statement about gradients and constraints. Below this, it shows the objective function as the sum of squared differences between observed values y and predicted values \bar{y} . A specific constraint is provided: $\bar{y} = 2$, which is enclosed in a box. Finally, it applies this to the context of SVM and PCA.



What is differentiation

22 April 2023 08:29

diff → slope

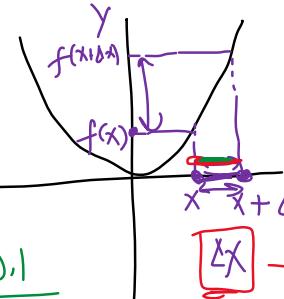
- What is differentiation?
- Why instantaneous?
- Relation with slope
- Maxima and Minima
- How to calculate derivative
- Intuition
- Derivative in ML

$$\frac{dy}{dx} = 0$$

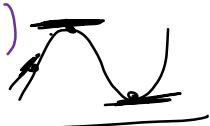
maxima
minima

Differentiation is the process of finding the derivative of a function. The derivative of a function represents the instantaneous rate of change of the function with respect to its variable, typically denoted as 'x'.

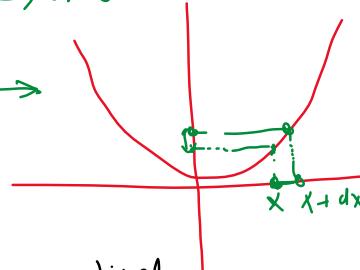
$$y = f(x) = x^2$$



$$\text{rate of change} = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$



$$\frac{\Delta x \rightarrow 0.1}{dx \rightarrow 0.001111111111111111}$$



$$\frac{f(x + \Delta x) - f(x)}{\Delta x} = \frac{df(x)}{dx} \text{ insl}$$

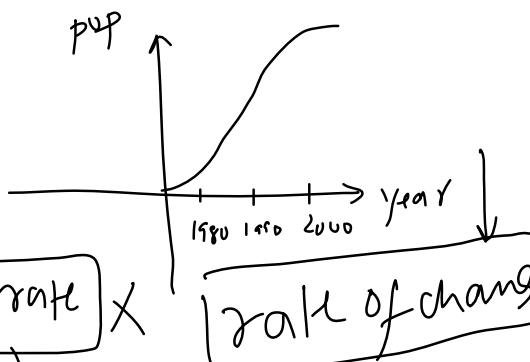
$$\lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

$$y = f(x) = x^3$$

↑ time
pop over time

POP 1990

pop → growth rate



$$\frac{1000 - 800}{10} = 100$$

1990 ← 1980 ←

$$y = f(t) = t^2$$

↑ distance time

$$y = x^2$$

velocity / speed

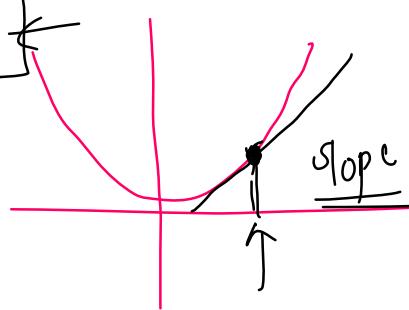
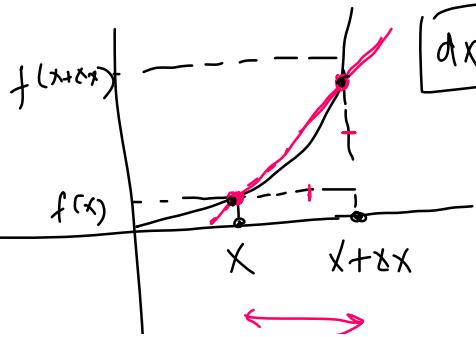
at $t = 4$

$$s = \frac{d}{t}$$

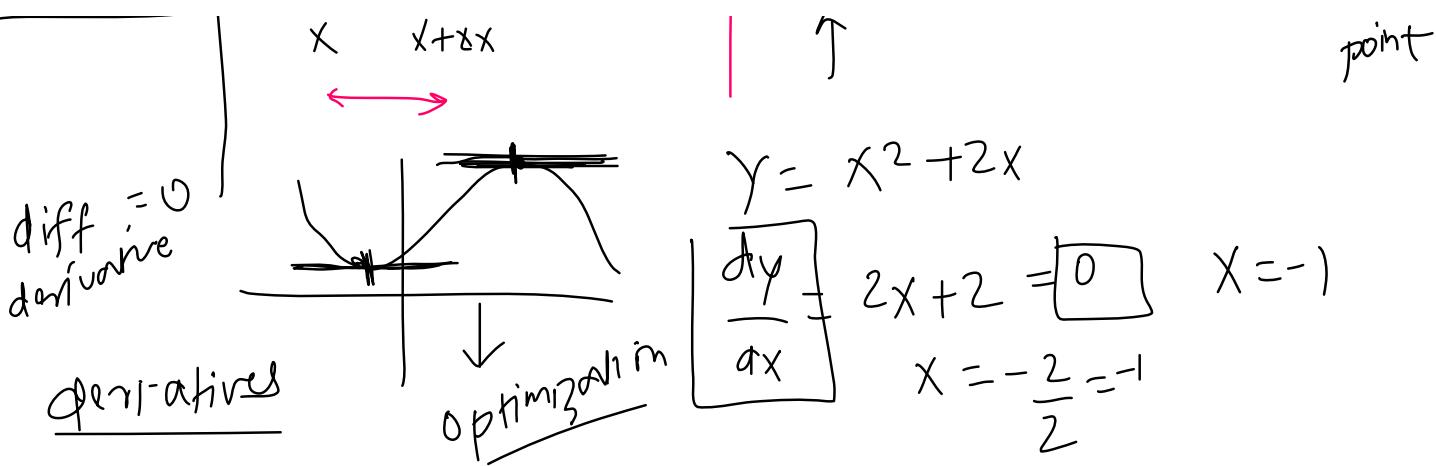
$$0 \rightarrow \text{dist} \rightarrow 0 \\ 4 \rightarrow \text{dist} \rightarrow 16$$

$$\frac{16 - 0}{4 - 0} = 4 \text{ m/s}$$

$$\frac{16 - 0}{4 - 0}$$



$$\frac{df}{dx} \rightarrow \text{slope at that point}$$



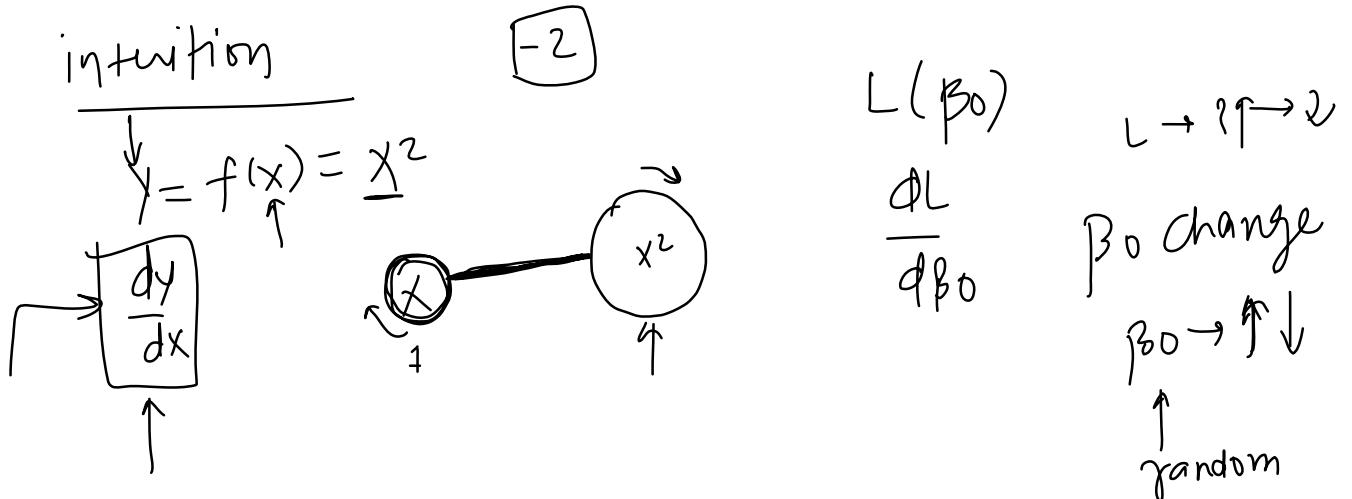
$y = f(x) = x^2$

$$\frac{df}{dx} = \frac{f(x+dx) - f(x)}{dx}$$

$$f(x+dx) = (x+dx)^2 = \frac{(x+dx)^2 - x^2}{dx}$$

$$f(x) = x^2 \quad - \cancel{x^2 + (dx)^2 + 2xdx - x^2}$$

$$\frac{(dx)^2 + 2xdx}{dx} = \underset{0}{\cancel{dx}} + 2x = 2x$$

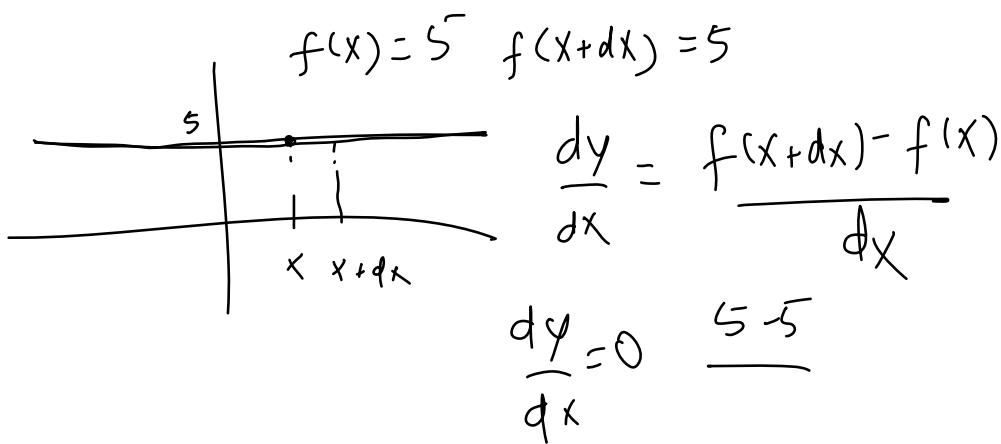


Derivative of a constant

22 April 2023 13:57

$$\frac{d}{dx}(c) = 0 \quad 0 \quad y = 5$$

derivative of
a const = 0



$$y = \boxed{5}$$

$$\frac{dy}{dx} = 0$$

Cheatsheet

22 April 2023 11:01

COMMON DERIVATIVES	
$\left\{ \begin{array}{l} \frac{d}{dx}(x) = 1 \\ \frac{d}{dx}(\sin x) = \cos x \\ \frac{d}{dx}(\cos x) = -\sin x \\ \frac{d}{dx}(\tan x) = \sec^2 x \\ \frac{d}{dx}(\sec x) = \sec x \tan x \\ \frac{d}{dx}(\csc x) = -\csc x \cot x \\ \frac{d}{dx}(\cot x) = -\csc^2 x \\ \frac{d}{dx}(\sin^{-1} x) = \frac{1}{\sqrt{1-x^2}} \\ \frac{d}{dx}(\cos^{-1} x) = -\frac{1}{\sqrt{1-x^2}} \\ \frac{d}{dx}(\tan^{-1} x) = \frac{1}{1+x^2} \\ \frac{d}{dx}(a^x) = a^x \ln(a) \\ \frac{d}{dx}(e^x) = e^x \\ \frac{d}{dx}(\ln(x)) = \frac{1}{x}, x > 0 \\ \frac{d}{dx}(\ln x) = \frac{1}{x} \\ \frac{d}{dx}(\log_a(x)) = \frac{1}{x \ln(a)} \end{array} \right.$	

Power Rule

22 April 2023 10:59

$$\frac{d}{dx}(x^n) = nx^{n-1}$$

$$\rightarrow \frac{dy}{dx} = \frac{f(x+dx) - f(x)}{dx}$$

$$f(x+dx) = x^2 + x dx + x dx + (dx)^2$$

$$= x^2 + 2x dx$$

$$\frac{x^2 + 2x dx - x^2}{dx} = \frac{2x}{dx} = 2x$$

$$y = f(x) = x^3$$

$$\frac{dy}{dx} = \frac{f(x+dx) - f(x)}{dx}$$

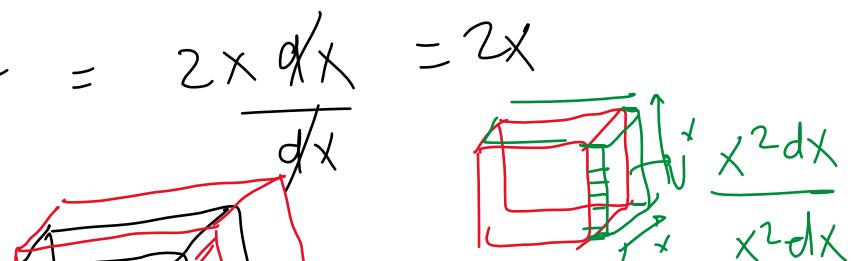
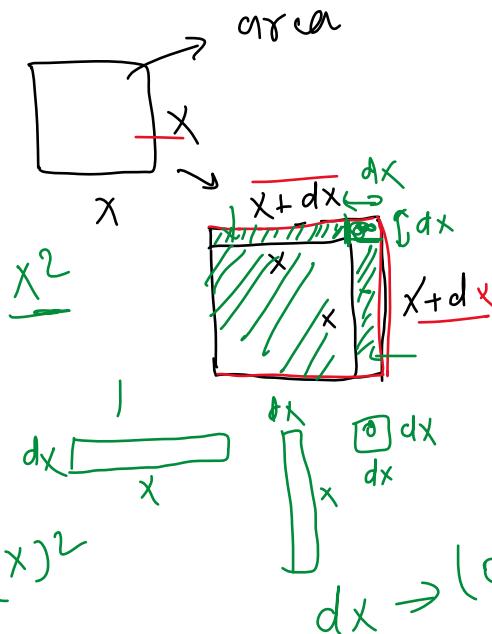
$$\frac{dy}{dx} = 3x^2$$

$$x^4 \rightarrow 4x^3$$

$$x^5 \rightarrow 5x^4$$

$$x^6 \rightarrow 6x^5$$

$$x^n \rightarrow nx^{n-1}$$



$$x^3 + x^2 dx + x^2 dx + x^2 dx - x^3$$

$$\frac{3x^2 dx}{dx} = 3x^2$$

power rule

Sum Rule

22 April 2023 11:01

$$(f(x) + g(x))' =$$



$$(f(x) \pm g(x))' = f'(x) \pm g'(x)$$

$$\frac{d(x^2 + \log x)}{dx} = \frac{d}{dx} x^2 + \frac{d}{dx} \log x$$

$$y = x \quad h(x) = f(x) + g(x)$$

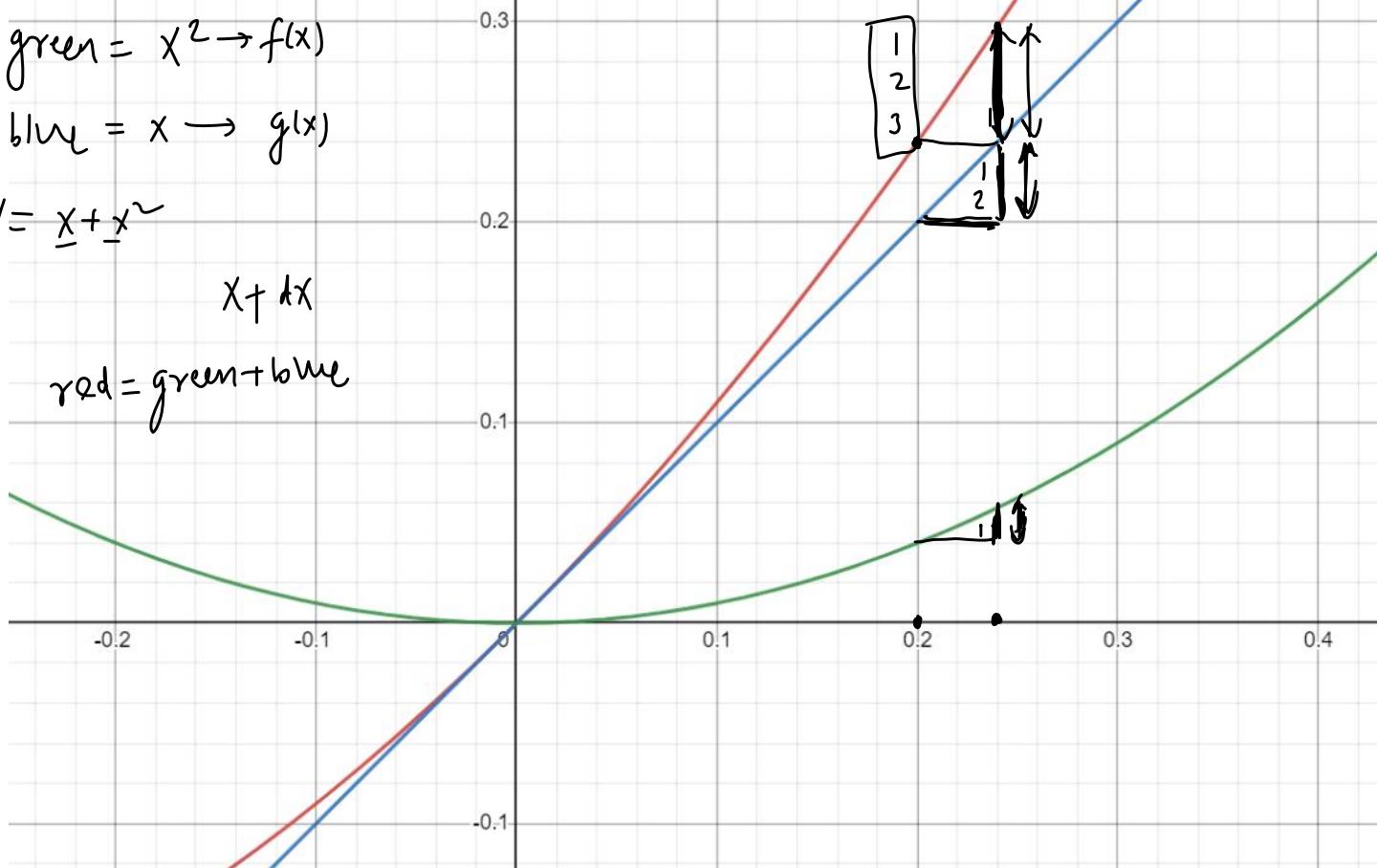
$$\text{green} = x^2 \rightarrow f(x)$$

$$\text{blue} = x \rightarrow g(x)$$

$$y = x + x^2$$

$$x + dx$$

$$\text{red} = \text{green} + \text{blue}$$



Product Rule

22 April 2023 11:01

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

$y = x^2 \sin x$

$$\frac{dy}{dx} = x^2 \sin x$$

$$\frac{x^3 \log x}{x}$$

$$\frac{d}{dx} x^2 \sin x + x^2 \frac{d}{dx} \sin x$$

$$\frac{dy}{dx} = [f(x+dx)] - [f(x)]$$

$y = x^2$

$$\frac{(x+dx)^2 \text{ area - area}}{(x+dx)}$$

$$+ \frac{a}{b} + \frac{c}{d} \rightarrow (x^2 + (dx)^2 + 2x \cdot dx) (x+dx)$$

$$y = x^2 + x(dx)^2 + 2x \cdot dx + x^2 dx + (dx)^3 + 2x(dx)^2 - x^2 \cdot dx$$

$$x^2 + x$$

$$[2x^2 + x^2]$$

$$\frac{x \cdot dx}{2x^2 + x^2 + 3x \cdot dx}$$

$$[dx = 0]$$

$$x^2 \rightarrow x^3$$

$$[3x^2]$$

$$\rightarrow [2x^2 + x^2] = [3x^2]$$

$$y = [5x^2] = 5 \frac{d(x^2)}{dx} = [5 \cdot 2x] = 10x$$

$$x^2$$

$$5x^2$$

$$\frac{(x+dx)^2 \cdot 5 - 5x^2}{dx}$$

$$\frac{(x^2 + (dx)^2 + 2x \cdot dx) \cdot 5 - 5x^2}{dx}$$

$$\frac{5x^2 + 5(dx)^2 + 5(2x)dx - 5x^2}{dx}$$

$$\frac{5(dx)^2 + 5(2x)dx}{dx} = \frac{5dx + 5(2x)}{5(2x)}$$

Quotient Rule

22 April 2023 13:53

$$\frac{d}{dx} \left(\frac{f(x)}{g(x)} \right) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}$$

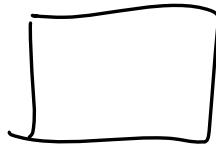
$$\frac{d}{dx} \frac{x^2}{\sin x} =$$

$$x^2 (\sin x)^{-1}$$

a b



$$\frac{\frac{d}{dx} x^2 \sin x - x^2 \frac{d}{dx} \sin x}{(\sin x)^2}$$



$$\boxed{\frac{dy}{dx} = \frac{2x \sin x - x^2 \cos x}{(\sin x)^2}}$$

$$\frac{d}{dx}(f(g(x))) = \underline{f'(g(x))} \underline{g'(x)}$$

$$y = f(g(x))$$

$$\frac{dy}{dx} = \boxed{\frac{df}{dg} \frac{dg}{dx}}$$

$$y = \sin x^2$$

$$\frac{dy}{dx}$$

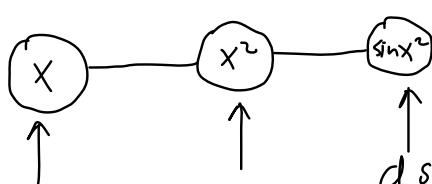
$$f(x) \circ g(x)$$

$$y = f(g(x)) = \sin(x^2)$$

↑
deep learning

$$g(x) = x^2$$

$$f(g(x)) = \sin(x^2)$$



$$\frac{d \sin(x^2)}{dx^2} \times \frac{dx^2}{dx}$$

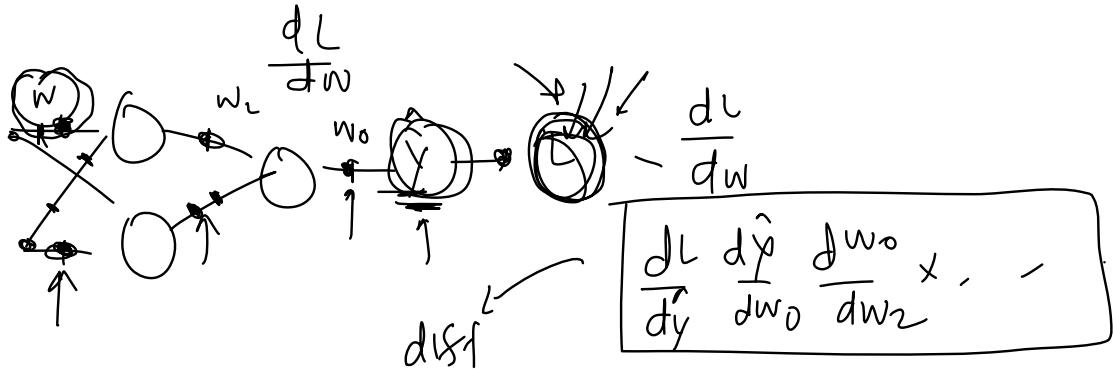
$$\cos x^2 \cdot 2x = \boxed{2x \cos x^2}$$

$$\sin(\log(x^2))$$

$$x \rightarrow x^2 \rightarrow \underline{\log x^2} \rightarrow \sin(\log x^2)$$

$$\frac{d \sin(\log x^2)}{d \log x^2} \frac{d \log x^2}{dx^2} \frac{dx^2}{dx}$$

$$\cos(\log x^2) \cdot \frac{1}{x^2} \cdot 2x = \boxed{\frac{1}{x} \cos(\log x^2)}$$

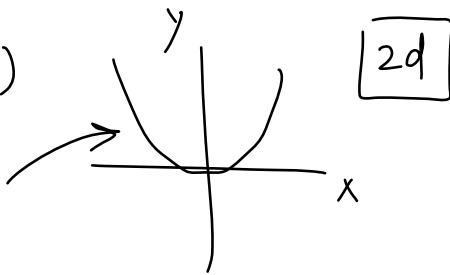


Partial Differentiation

22 April 2023 13:50

$$y = f(x) \quad \frac{dy}{dx} = f'(x)$$

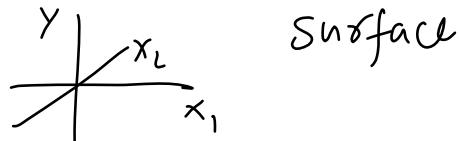
single variable
function



2d

$$y = f(\underline{x_1}, \underline{x_2})$$

$$f(x_1, x_2, \dots, x_n)$$



Surface

$$z = f(x, y) = x^2 + y^2$$

$$\frac{\partial z}{\partial x}$$

$$z = x^2 + \boxed{y^2}$$

$$\boxed{\frac{\partial z}{\partial x}} = \frac{2x}{2(1)} = 2x$$

$$\boxed{=2}$$

$$\boxed{\frac{\partial z}{\partial y}} = 2y$$

$$= \boxed{4}$$

complete

$$\boxed{\frac{\partial z}{\partial x} \frac{\partial z}{\partial y}}$$

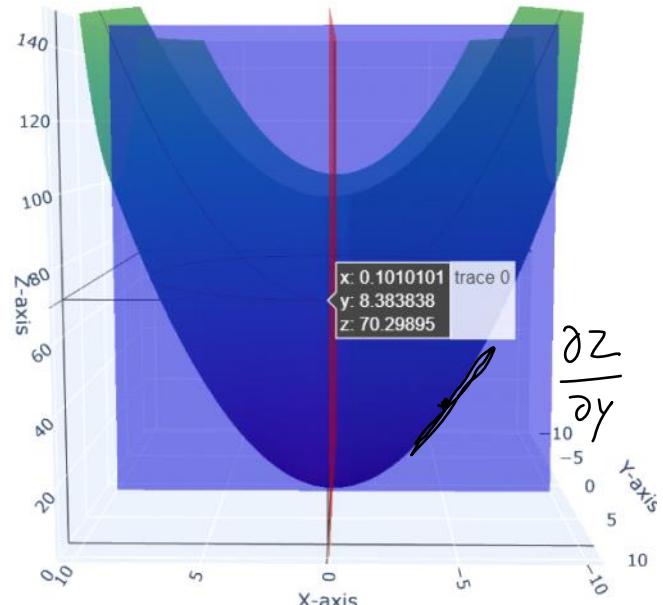
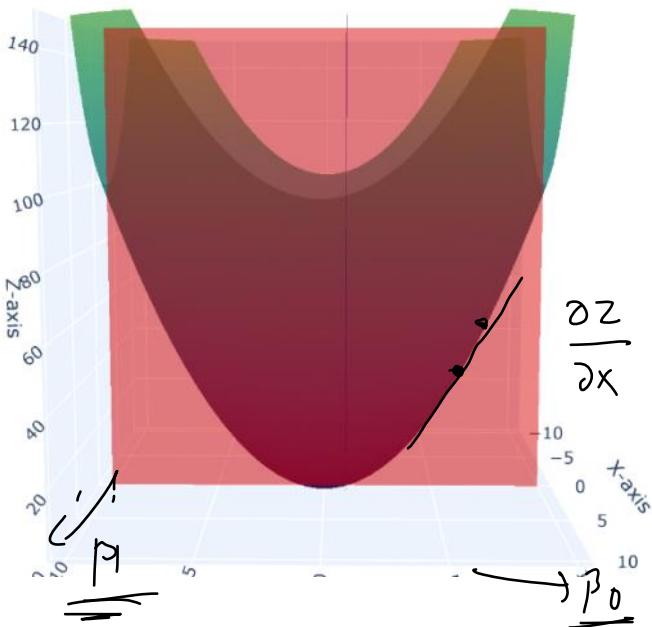
↑
↑
partial derivative

3d parabola

$$(1, 2)$$

$$x=1 \quad y=2$$

$$f_0 \beta_1$$



Higher Order Derivatives

22 April 2023 14:01

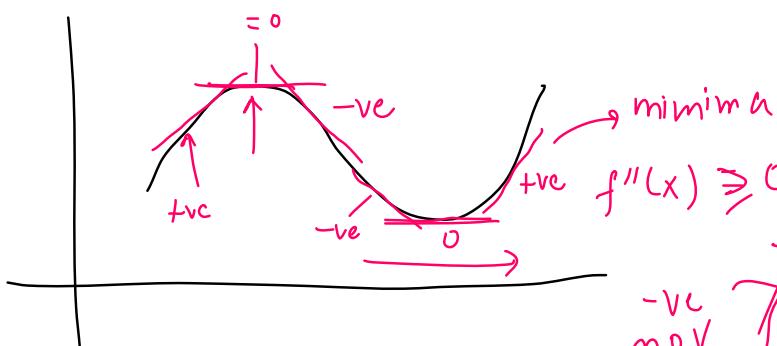
$$y = x^3 \quad \frac{dy}{dx} = 3x^2 \quad S \rightarrow V \rightarrow a$$

$$\frac{d^2y}{dx^2} \leftarrow \frac{d}{dx} \frac{dy}{dx} = 6x \quad \frac{d^3y}{dx^3}$$

instant rate of change of slope

slope = 0 $\frac{dy}{dx} \rightarrow$ maxima minima

maxima $\frac{d^2y}{dx^2}$ maxima



$\frac{d^2y}{dx^2}$ → rate of change of slope
derivative

$$\frac{dy}{dx}$$

$$\frac{d^2y}{dx^2}$$

rate of change of slope

$$\frac{d^2L}{dx^2}$$

Newton-Horner method

$$f''(x)$$

$$f''(x) >$$

$$f''(x)$$

Matrix Differentiation

22 April 2023 13:53

matrix Differentiation

April 2023 13:53

$\frac{d}{dx} \boxed{Ax}$ ←

$\frac{d}{dx} \underline{\underline{Cx}} = \underline{\underline{C}}$

Constant $\frac{d}{dx} 5x = 5$

$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\frac{d}{dx} \underline{\underline{Ax}}$

$\frac{d}{dx} Ax \rightarrow \boxed{A}$

$\frac{d}{dx} \underline{\underline{f_1(x, y)}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} \end{bmatrix}$

$\underline{\underline{f_1(x_1, x_2)}} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix}$

$\frac{d}{dx} \underline{\underline{f_2(x, y)}} = \begin{bmatrix} \frac{\partial f_2}{\partial x_1} \\ \frac{\partial f_2}{\partial x_2} \end{bmatrix}$

$\rightarrow \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \rightarrow \textcircled{A}$

$\frac{d}{dx} Ax = A$

$$y = \underbrace{x^T A x}_{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} \quad A^T = A \quad \begin{bmatrix} a_{11} & a_x \\ a_x & \underline{a_{22}} \end{bmatrix}^T$$

$$\frac{\partial y}{\partial x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_x \\ a_x & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \begin{bmatrix} a_{11} & a_x \\ a_x & a_{22} \end{bmatrix} \leftarrow$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11}x_1 + a_x x_2 \\ a_x x_1 + a_{22} x_2 \end{bmatrix} \quad f(x_1, x_2)$$

$$\underbrace{a_{11}x_1^2 + \cancel{a_x x_2^2} + a_x x_1^2 + \cancel{a_{11} x_2^2}}_{\begin{bmatrix} \frac{df}{dx_1} \\ \frac{df}{dx_2} \end{bmatrix}}$$

$$\begin{bmatrix} 2a_{11}x_1 + 2a_x x_1 \\ 2a_x x_2 + 2a_{22} x_2 \end{bmatrix} = 2 \begin{bmatrix} a_{11}x_1 + a_x x_1 \\ a_x x_2 + a_{22} x_2 \end{bmatrix} \leftarrow$$

$$\left\{ \begin{array}{l} \text{matrix} \\ \text{columns} \end{array} \right\}$$

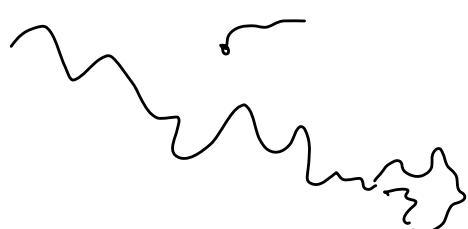
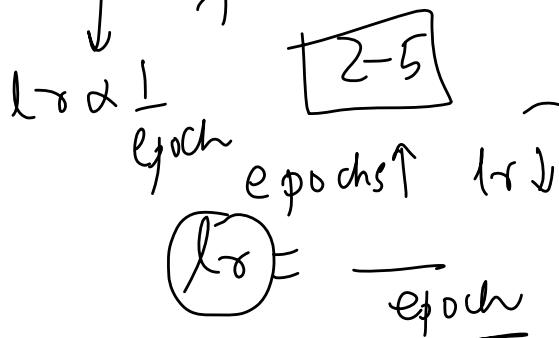
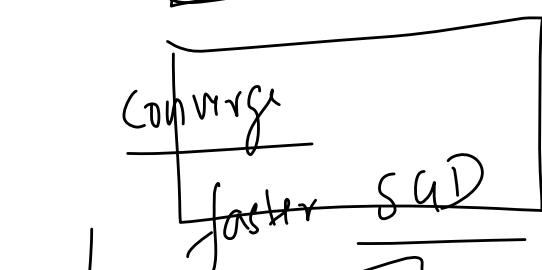
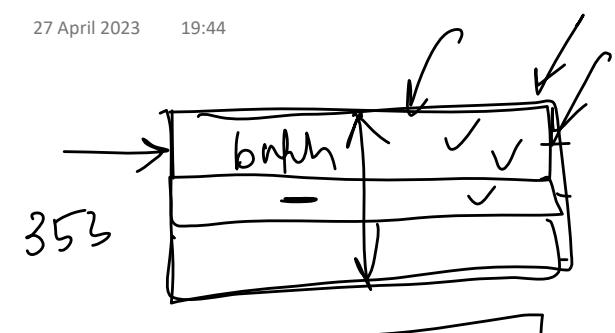
$$2 \begin{bmatrix} a_{11} & a_x \\ a_x & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\rightarrow 2x^2 \quad \rightarrow (2A)^T \quad 2x^T \underline{A^T} =$$

$$\boxed{4x} \leftarrow$$

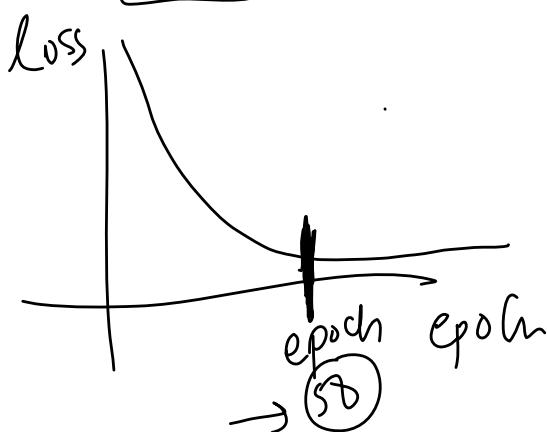
Doubt Clearance

27 April 2023 19:44

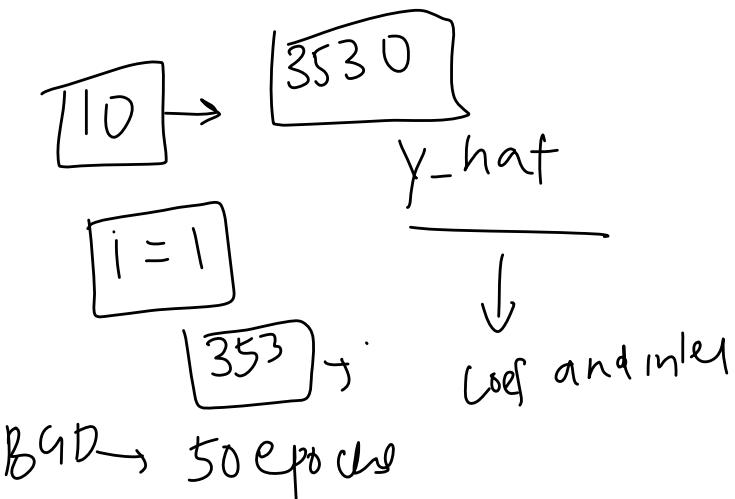


$$\overline{lr} \rightarrow \overline{0.01}$$

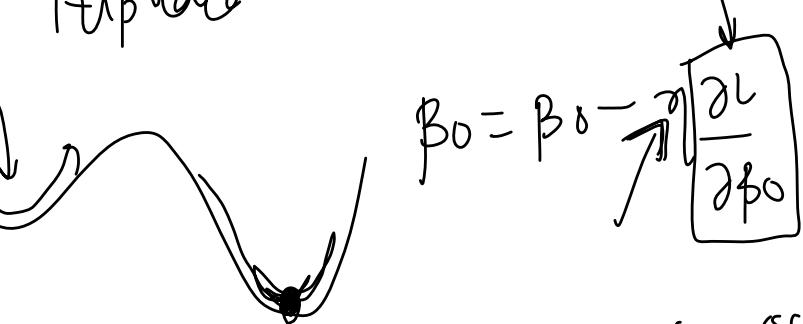
$$10 \rightarrow 50 \rightarrow \underline{10^0}$$



cg1g1g1g1g1



It up date



high dimen → lots of cols
 inverse $\frac{1}{(n^3)}$
 GP → regularization → ridge → least squares

Regression Analysis

$$\rightarrow [B_0 \ B_1 \ B_2 \dots B_n]$$

R^2

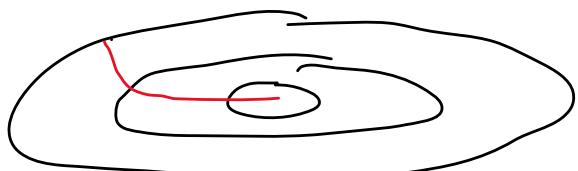
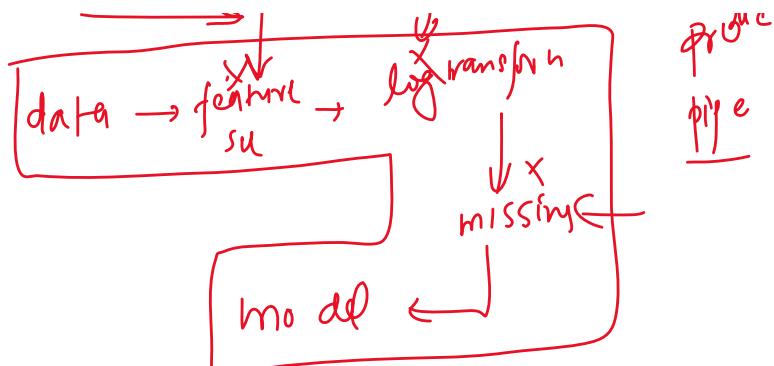
→ $T_{1,1}, \dots, T_{n,n} \rightarrow$ log transform

predictive

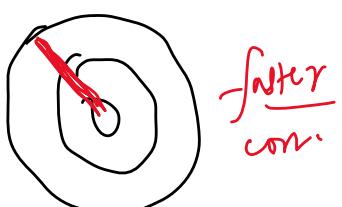
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Diagram showing a linear regression model structure:

- y is the output variable.
- β_0 is the intercept.
- $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients for features x_1, x_2, \dots, x_n .
- A circled β_0 is labeled "fitter const".
- A circled β_1 is labeled "fitter corr".



data → model → prediction
data → pipe



$$\begin{aligned} X &\rightarrow n \times m \\ & \quad \left[\begin{array}{c|c|c} 1 & 1 & 1 \\ \hline 1 & 1 & 1 \end{array} \right] \\ & \quad \text{fitter corr} \end{aligned}$$

Diagram illustrating matrix multiplication:

$$X \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} y \\ \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m \end{bmatrix}$$

$$\begin{aligned} & \quad \left[\begin{array}{c|c|c} 1 & x_1 & x_2 \\ \hline 1 & 1 & 1 \end{array} \right] \\ & \quad \left[\begin{array}{c|c|c} 1 & x_1 & x_2 \\ \hline 1 & 1 & 1 \end{array} \right] \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} y \\ \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m \end{bmatrix} \end{aligned}$$

python anywhere
streamlit
huggingface → gradio

y ← x_{input}

$$y = \beta_0 + \beta_1 x$$
$$\boxed{\beta_0 + \beta_1 x + \beta_2 x^2}$$

Till now

28 April 2023 06:49

Linear Reg $\rightarrow X, Y$

find the
coeff of
linear reg

OLS ✓
slow

high dim
data

GD ✓

linear ref - $\hat{Y} = mX + b$ $\xrightarrow{2d}$ $\boxed{m, b}$ \downarrow
 $\boxed{m, b}$ $\uparrow \uparrow \uparrow$ X, Y

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

stats \rightarrow Regression Analysis

Regression Analysis

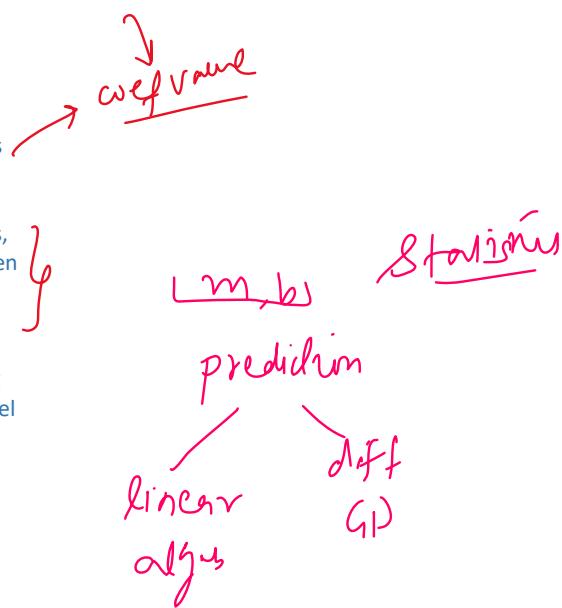
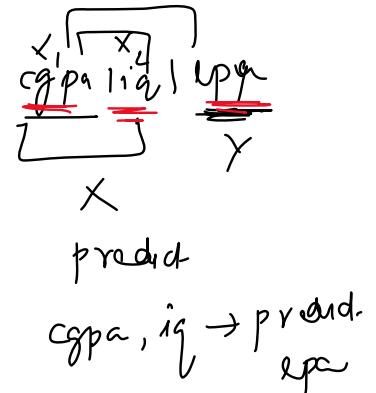
29 April 2023 02:21

$$\textcircled{Y} \times (x_1, x_2, \dots, x_n)$$

Regression analysis is a statistical method used to examine the relationship between one dependent variable and one or more independent variables. The goal of regression analysis is to understand how the dependent variable changes when one or more independent variables are altered, and to create a model that can predict the value of the dependent variable based on the values of the independent variables.

Flow → LR

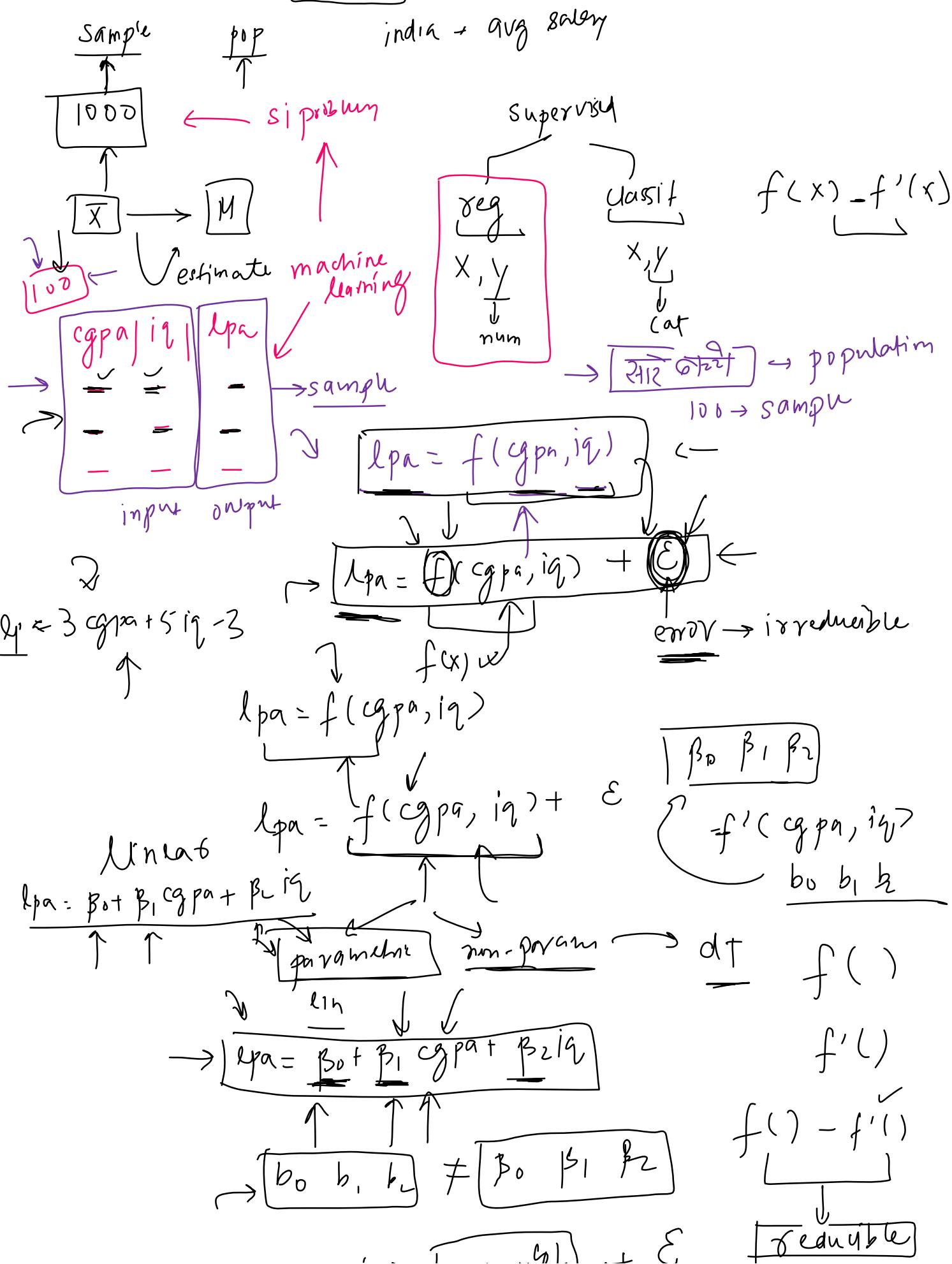
1. Define the research question: Identify the dependent variable (the variable you want to predict or explain) and the independent variable(s) (the variables that you think influence the dependent variable).
2. Collect and prepare data: Gather data for the dependent and independent variables. The data should be organized in a tabular format, with each row representing an observation and each column representing a variable. It's essential to clean and pre-process the data to handle missing values, outliers, and other potential issues that may affect the analysis.
3. Visualize the data: Before fitting a linear regression model, it's helpful to create scatter plots to visualize the relationship between the dependent variable and each independent variable. This can help you identify trends, outliers, and any potential issues with the data.
4. Check assumptions: Linear regression has some underlying assumptions, including linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. You can use diagnostic plots and statistical tests to check whether these assumptions hold for your data.
5. Fit the linear regression model: Use statistical software (e.g., R, Python, or Excel) to fit a linear regression model to your data. The model will estimate the regression coefficients (intercept and slope) that minimize the sum of squared residuals (i.e., the differences between the observed and predicted values of the dependent variable).
6. Interpret the model: Analyse the estimated regression coefficients, their standard errors, t-values, and p-values to determine the statistical significance of the relationship between the dependent and independent variables. The R-squared value and adjusted R-squared value can provide insights into the goodness-of-fit of the model and the proportion of variation in the dependent variable explained by the independent variables.
7. Validate the model: If you have a sufficiently large dataset, you can split it into a training and testing set. Fit the linear regression model to the training set, and then use the model to predict the dependent variable in the testing set. Calculate the mean squared error, root mean squared error, or another performance metric to assess the predictive accuracy of the model.
8. Report results: Summarize the findings of the linear regression analysis in a clear and concise manner, including the estimated coefficients, their interpretation, and any limitations or assumptions that may impact the results.



1. What's the statistics connection? ✓
2. Why is Regression Analysis required? →

Why ML problems are a Statistical Inference Problems? [with Example]

28 April 2023 06:56



$$lpa = f'(cgpa, iq) + \boxed{\text{randomness}} + \epsilon$$

↳ reducible

estimate y un
of x any
based on given data $f'() \approx f()$
Type of x, y for p.pw

$$y = \boxed{2x - 5} + \boxed{\text{some randomness}}$$

$$\underline{f(x)} = \underbrace{2x - 5}_{\uparrow \uparrow}$$

$\beta_0 = -5$
$\beta_1 = 2$

pop
parameters

$$\begin{bmatrix} b_0 & b_1 \end{bmatrix}$$

→ current set of 50 points

Inference Vs Prediction [Why regression analysis is required?]

28 April 2023 06:51

$$\text{cgpa} / \text{iq} / \text{lpa}$$

linear

$$\beta_0 \beta_1 \beta_2$$

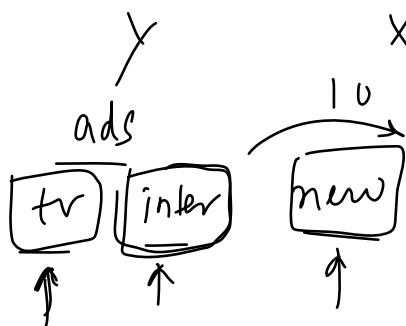
$$(8, 80)$$

$$\underline{\text{lpa}} = \underline{\beta_0 + \beta_1 \text{cgpa} + \beta_2 \text{iq}} = 10$$

Tnp inference (relationship study)

$$\text{lpa} \rightarrow \text{cgpa}, \text{iq}$$

$$\text{lpa} \rightarrow \boxed{\text{cgpa}}$$



$$\text{lpa} \rightarrow \underline{\text{iq}}$$

reg

m, b

next

Black box model

$$\boxed{X \rightarrow Y}$$

explain

inform

product

regression
analysis

ChatGPT

linear

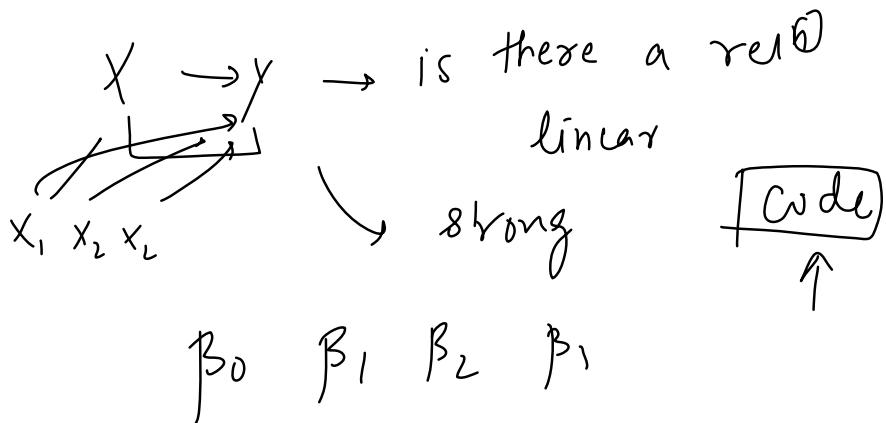
inference \rightarrow predict

ban \rightarrow achieve

trade off

Statsmodel Linear Regression

28 April 2023 06:59



OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Sat, 29 Apr 2023	Prob (F-statistic):	1.58e-96			
Time:	07:32:56	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const (β_0)	2.9389	0.312	9.422	0.000	2.324	3.554
TV (β_1)	0.0458	0.001	32.809	0.000	0.043	0.049
Radio (β_2)	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper (β_3)	-0.0010	0.006	-0.177	0.860	-0.013	0.011
Omnibus:	60.414	Durbin-Watson:	2.084			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241			
Skew:	-1.327	Prob(JB):	1.44e-33			
Kurtosis:	6.332	Cond. No.	454.			

↓ α_{sum}

$\rightarrow \underline{\text{Hypo}} \quad \underline{\text{Ho}}$ → $f\text{-test}$ for overall significance (ANOVA)

$X \rightarrow Y$

↓ TV | radio | newspaper | Sales

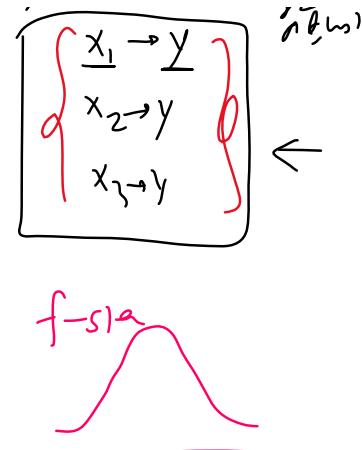
goodness of fit f-test $\rightarrow p\text{-val } 0.05$

OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Sat, 29 Apr 2023	Prob (F-statistic):	1.58e-96			
Time:	07:32:56	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					

- 1) LR → $f\text{-test}$
- 2) $\boxed{X \rightarrow Y}$ Strong m
- 3) $X (x_1, x_2, x_3) \rightarrow$
 $\boxed{\begin{array}{l} x_1 \rightarrow Y \\ \vdots \\ x_n \rightarrow Y \end{array}}$

↑
IS

	coef	std err	t	P> t	[0.025 0.975]
const	2.9389	0.312	9.422	0.000	2.324 3.554
TV	0.0458	0.001	32.809	0.000	0.043 0.049
Radio	0.1885	0.009	21.893	0.000	0.172 0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013 0.011
Omnibus:	60.414	Durbin-Watson:	2.084		
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241		
Skew:	-1.327	Prob(JB):	1.44e-33		
Kurtosis:	6.332	Cond. No.	454.		



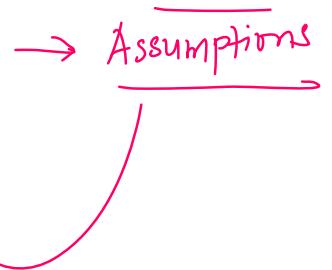
[SE b1]

$$0.0458 \pm 3.18 \times 0.01$$

b1

↓
T-test → formula → Initial derivation

Assumptions



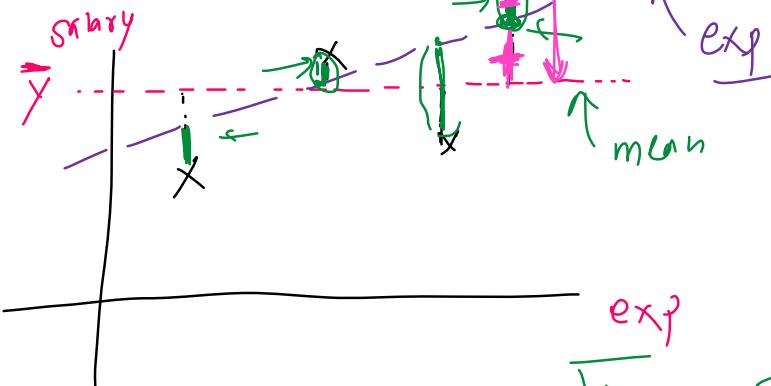
TSS, RSS and ESS

29 April 2023 04:16

$TSS \rightarrow$ Total sum of squares

$RSS \rightarrow$ Residual sum of squares

$ESS \rightarrow$ Explained sum of squares



$$Y = 5X + 2$$

$TSS - RSS$

$ESS \rightarrow$ Explained Variation

$$TSS = \sum (y_i - \bar{y})^2$$

$$RSS = \sum (y_i - \hat{y}_i)^2$$

overall Variance in data

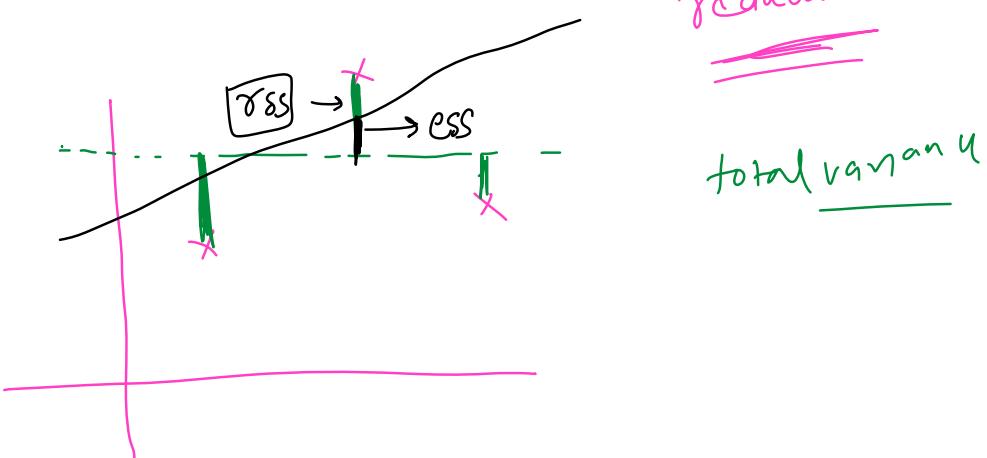
$$TSS \rightarrow ESS + RSS$$

Explained

unexplained

reducible

Irreducible



total variance

Degree of Freedom
28 April 2023 07:00

$$df_{\text{total}} = n - 1 \quad (n-1) \rightarrow df$$

$n \rightarrow \# \text{ of rows}$

In linear regression, the total degrees of freedom (df_{total}) represent the total number of data points minus 1. It represents the overall variability in the dataset that can be attributed to both the model and the residuals.

For a linear regression with n data points (observations), the total degrees of freedom can be calculated as:

$$df_{\text{total}} = n - 1$$

where: n is the number of data points (observations) in the dataset

The total degrees of freedom in linear regression is divided into two components:

1. Degrees of freedom for the model (df_{model}): This is equal to the number of independent variables in the model (k).

2. Degrees of freedom for the residuals ($df_{\text{residuals}}$):

The degrees of freedom for the residuals indicate the number of independent pieces of information that are available for estimating the variability in the residuals (errors) after fitting the regression model.

This is equal to the number of data points (n) minus the number of estimated parameters, including the intercept ($k+1$).

The sum of the degrees of freedom for the model and the degrees of freedom for the residuals is equal to the total degrees of freedom:

$$df_{\text{total}} = df_{\text{model}} + df_{\text{residuals}}$$

$$K \rightarrow \# \text{ of input cols} \quad n - K - 1 + K = [n-1] + K = n - (K+1)$$

$n \rightarrow \# \text{ of rows}$

$$df_{\text{+}} = \frac{df_{\text{-m}}}{K} + \frac{df_{\text{-r}}}{n - K - 1}$$

$n - 1$

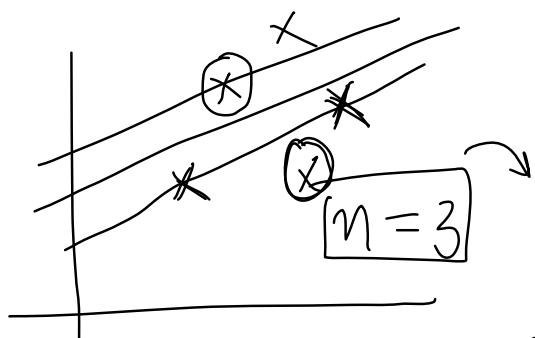
$$\frac{k}{n} \rightarrow \# \text{ of input cols}$$

$n - 1 \rightarrow \# \text{ of rows}$

simple regression

$$x | y$$

$200 - k - 1$
 $200 - 1 - 1 = 198$



$y_{\text{reg}} \rightarrow \text{line}$
min data points

$$df = 1 \quad df = 2$$

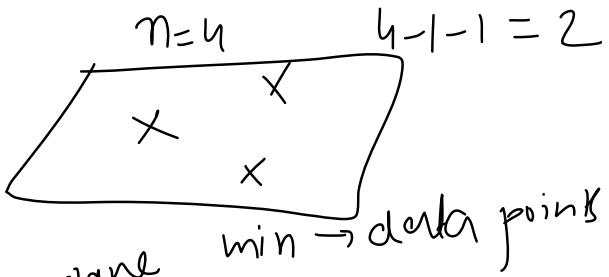
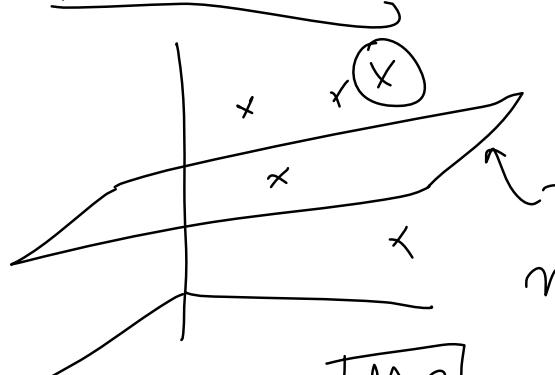
$$n = 3 \quad k = 1$$

$$n - k - 1 \quad 3 - 1 - 1 = 1$$

capitalia | M9

$$n = 4 \quad k = 1 \quad 4 - 1 - 1 = 2$$

cgpa | age



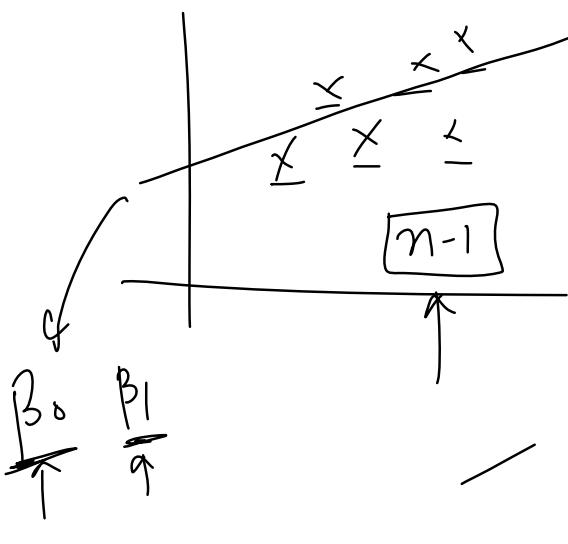
$$n=4 \quad k=2$$

$$n-k-1$$

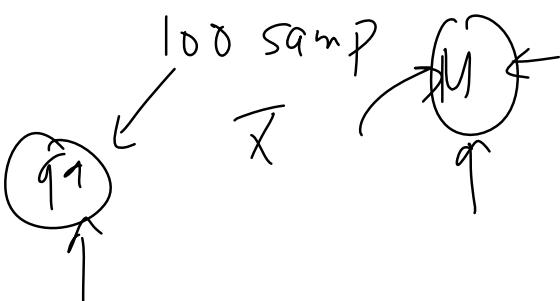
$$4-2-1 = 1$$

df: 2

$\leftarrow n-1$



last nth point



F-statistic & Prob(F-statistic)

28 April 2023 07:01

$$\begin{array}{c} X_1 \quad X_2 \quad X_3 \\ \beta_1 \quad \beta_2 \quad \beta_3 \end{array} \rightarrow \beta_1 = \beta_2 = \beta_3 = 0$$

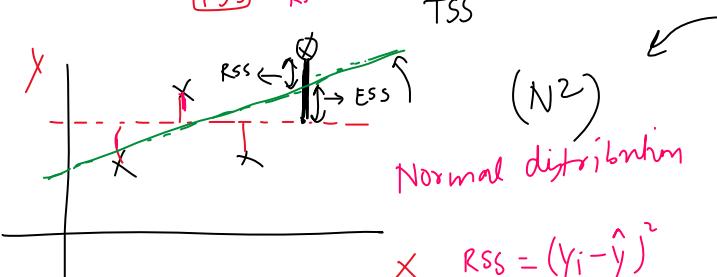
The F-test for overall significance is a statistical test used to determine whether a linear regression model is statistically significant, meaning it provides a better fit to the data than just using the mean of the dependent variable.

Here are the steps involved in conducting an F-test for overall significance:

- State the null and alternative hypotheses:
 - Null hypothesis (H_0): All regression coefficients (except the intercept) are equal to zero ($\beta_1 = \beta_2 = \dots = \beta_k = 0$), meaning that none of the independent variables contribute significantly to the explanation of the dependent variable's variation.
 - Alternative hypothesis (H_1): At least one regression coefficient is not equal to zero, indicating that at least one independent variable contributes significantly to the explanation of the dependent variable's variation.
- Fit the linear regression model to the data, estimating the regression coefficients (intercept and slopes).
- Calculate the Sum of Squares (SS) values:
 - Total Sum of Squares (TSS): The sum of squared differences between each observed value of the dependent variable and its mean.
 - Regression Sum of Squares (ESS): The sum of squared differences between the predicted values of the dependent variable and its mean.
 - Residual Sum of Squares (RSS): The sum of squared differences between the observed values and the predicted values of the dependent variable.
- Compute the Mean Squares (MS) values:
 - Mean Square Regression (MSR): ESS divided by the degrees of freedom for the model (df_{model}), which is the number of independent variables (k). This could also be called Average Explained Variance per independent feature.
 - Mean Square Error (MSE): RSS divided by the degrees of freedom for the residuals ($df_{residuals}$), which is the number of data points (n) minus the number of estimated parameters, including the intercept ($k+1$). This could also be called as average unexplained variance per degree of freedom.
- Calculate the F-statistic: $F\text{-statistic} = MSR / MSE$
- Determine the p-value:
 - Compute the p-value associated with the calculated F-statistic using the F-distribution or a statistical software package.
- Compare the calculated F-statistic to the p-value to the chosen significance level (α):
 - If the p-value $< \alpha$, reject the null hypothesis. This indicates that at least one independent variable contributes significantly to the prediction of the dependent variable, and the overall regression model is statistically significant.
 - If the p-value $\geq \alpha$, fail to reject the null hypothesis. This suggests that none of the independent variables in the model contribute significantly to the prediction of the dependent variable, and the overall regression model is not statistically significant.

Following these steps, you can perform an F-test for overall significance in a linear regression analysis and determine whether the regression model is statistically significant.

TSS RSS



$$TSS = (y_i - \bar{y})^2$$

$$F\text{-stat} = \frac{\frac{ESS}{K}}{\frac{RSS}{n-K-1}}$$

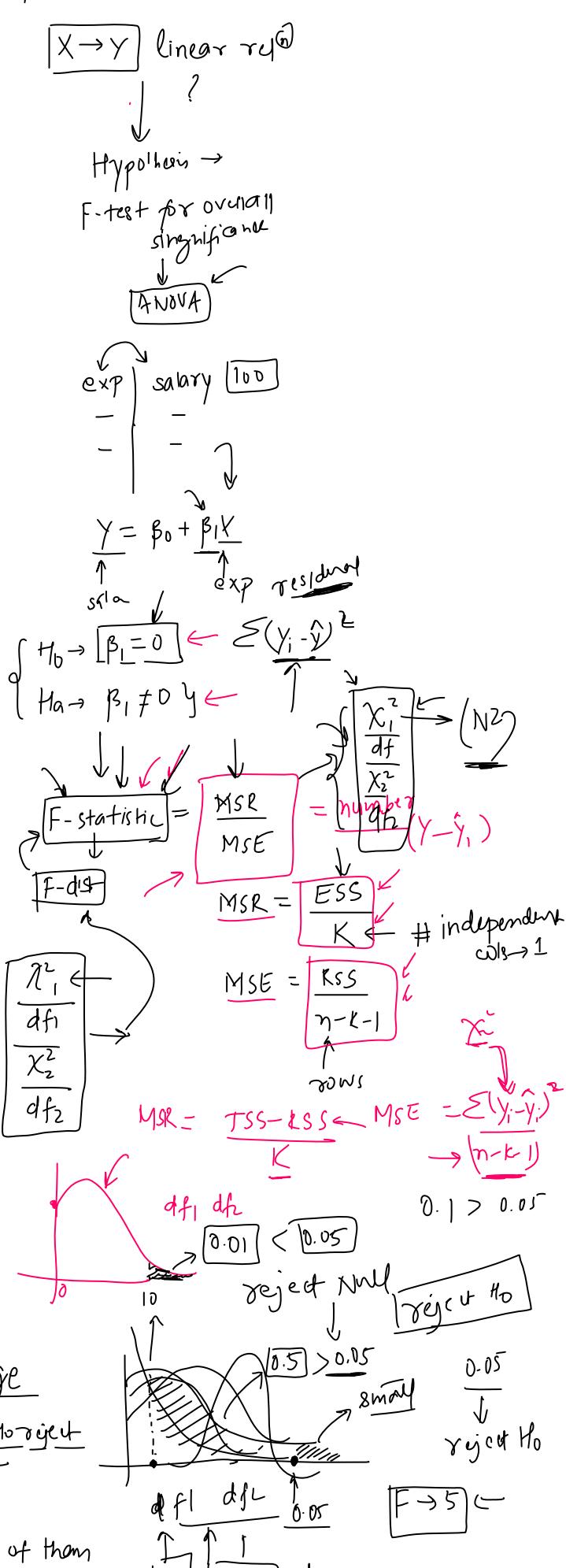
avg explained var per df
avg unexplained var per df

F very small if at least one of $\beta_1, \beta_2, \dots, \beta_k$ is not zero

f-statistic

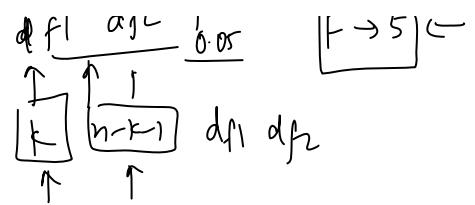
very large if at least one of $\beta_1, \beta_2, \dots, \beta_k$ is not zero

$F \uparrow$ if at least one of $\beta_1, \beta_2, \dots, \beta_k$ is not zero



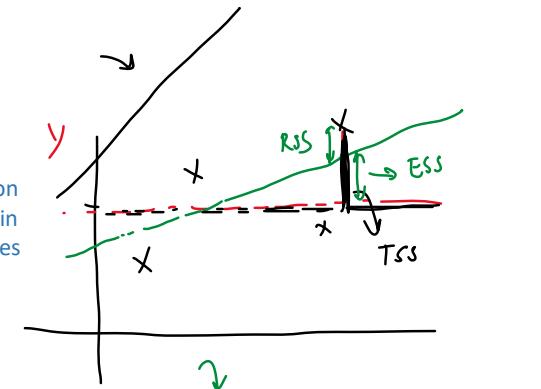
$F \sim \chi^2$

$\overbrace{\quad \quad \quad}^{\text{at least one of them}} \text{ is not } 0$



$R^2 \rightarrow \text{goodness of fit}$

R-squared (R^2), also known as the coefficient of determination, is a measure used in regression analysis to assess the goodness-of-fit of a model. It quantifies the proportion of the variance in the dependent variable (response variable) that can be explained by the independent variables (predictor variables) in the regression model. R-squared is a value between 0 and 1, with higher values indicating a better fit of the model to the observed data.



In the context of a simple linear regression, R^2 is calculated as the square of the correlation coefficient (r) between the observed and predicted values. In multiple regression, R^2 is obtained from the ratio of the explained sum of squares (ESS) to the total sum of squares (TSS):

$$R^2 = ESS / TSS$$

where:

- ESS (Explained Sum of Squares) is the sum of squared differences between the predicted values and the mean of the observed values. It represents the variation in the response variable that can be explained by the predictor variables in the model.
- TSS (Total Sum of Squares) is the sum of squared differences between the observed values and the mean of the observed values. It represents the total variation in the response variable.

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS}$$

$R^2 = 1 - \frac{RSS}{TSS}$

ESS → TSS
RSS > TSS
0.81 → [0.61]

An R-squared value of 0 indicates that the model does not explain any of the variance in the response variable, while an R-squared value of 1 indicates that the model explains all of the variance. However, R-squared can be misleading in some cases, especially when the number of predictor variables is large or when the predictor variables are not relevant to the response variable.

Disadvantage of R^2

Adjusted R^2 score

$$\begin{array}{c|c} x_1 & x_2 & x_3 & x_4 & \dots & x_n \\ \hline & & & & & | y & x_1 \rightarrow 0.4 \end{array}$$

input vars

$$\begin{array}{c|c} x_1 & x_2 \\ \hline | y \end{array}$$

$$\begin{array}{c|c} x_2 & x_3 \\ \hline x_2 \rightarrow y & x_3 \end{array}$$

$$x_2 \neq y$$

③ irr.

$$\begin{array}{c|c} 6 \text{ input} & x_1 \rightarrow x_6 \\ \hline 0.4 & \text{misleading} \end{array}$$

$$\begin{array}{c|c} \text{cgpa} & \text{salary} \\ \downarrow & \uparrow \\ \boxed{1.9} & \boxed{\text{Temp}} x \end{array}$$

$$\begin{array}{c|c} x_2 \text{-score} \rightarrow 0.4 & \\ \hline x_2 \text{-score} \rightarrow 0.6 & \boxed{0.4} \end{array}$$

Adjusted R-squared

28 April 2023 07:01

Adjusted R-squared is a modified version of R-squared (R^2) that adjusts for the number of predictor variables in a multiple regression model. It provides a more accurate measure of the goodness-of-fit of a model by considering the model's complexity.

In a multiple regression model, R-squared (R^2) measures the proportion of variance in the response variable that is explained by the predictor variables. However, R-squared always increases or stays the same with the addition of new predictor variables, regardless of whether those variables contribute valuable information to the model. This can lead to overfitting, where a model becomes too complex and starts capturing noise in the data instead of the underlying relationships.

Adjusted R-squared accounts for the number of predictor variables in the model and the sample size, penalizing the model for adding unnecessary complexity. Adjusted R-squared can decrease when an irrelevant predictor variable is added to the model, making it a better metric for comparing models with different numbers of predictor variables.

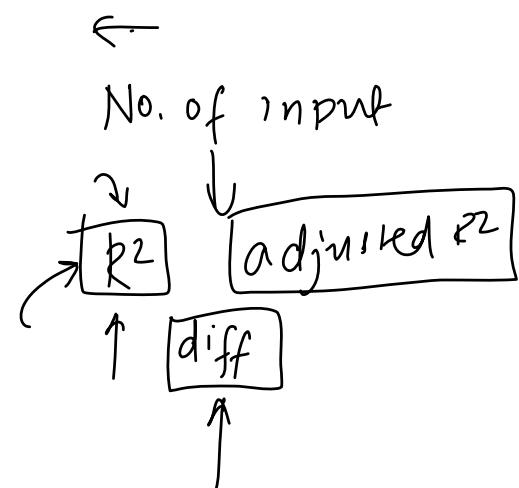
The formula for adjusted R-squared is:

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1-R^2) * (n-1)}{(n-k-1)} \right]$$

where:

- R^2 is the R-squared of the model
- n is the number of observations in the dataset
- k is the number of predictor variables in the model

By using adjusted R-squared, you can more accurately assess the goodness-of-fit of a model and choose the optimal set of predictor variables for your analysis.

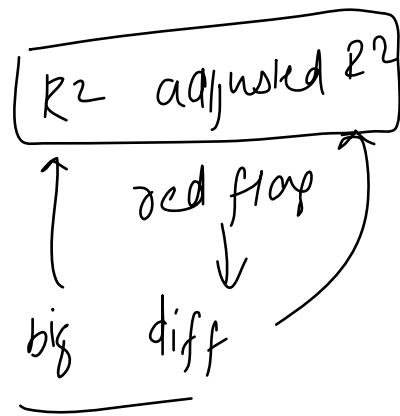


Which one should be used?

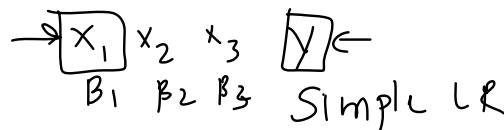
02 May 2023 13:43

The choice between using R-squared and adjusted R-squared depends on the context and the goals of your analysis. Here are some guidelines to help you decide which one to use:

1. **Model comparison:** If you're comparing models with different numbers of predictor variables, it's better to use adjusted R-squared. This is because adjusted R-squared takes into account the complexity of the model, penalizing models that include irrelevant predictor variables. R-squared, on the other hand, can be misleading in this context, as it tends to increase with the addition of more predictor variables, even if they don't contribute valuable information to the model.
2. **Model interpretation:** If you're interested in understanding the proportion of variance in the response variable that can be explained by the predictor variables in the model, R-squared can be a useful metric. However, keep in mind that R-squared does not provide information about the significance or relevance of individual predictor variables. It's also important to remember that a high R-squared value does not necessarily imply causation or a good predictive model.
3. **Model selection and overfitting:** When building a model and selecting predictor variables, it's important to guard against overfitting. In this context, adjusted R-squared can be a helpful metric, as it accounts for the number of predictor variables and penalizes the model for unnecessary complexity. By using adjusted R-squared, you can avoid including irrelevant predictor variables that might lead to overfitting.



In summary, adjusted R-squared is generally more suitable when comparing models with different numbers of predictor variables or when you're concerned about overfitting. R-squared can be useful for understanding the overall explanatory power of the model, but it should be interpreted with caution, especially in cases with many predictor variables or potential multicollinearity.



Performing a t-test for a simple linear regression, including the intercept term and using the p-value approach, involves the following steps:

- State the null and alternative hypotheses for the slope and intercept coefficients:

For the slope coefficient (β_1):

- Null hypothesis (H_0): $\beta_1 = 0$ (no relationship between the predictor variable (X) and the response variable (y))
- Alternative hypothesis (H_1): $\beta_1 \neq 0$ (a relationship exists between the predictor variable and the response variable)

For the intercept coefficient (β_0):

- Null hypothesis (H_0): $\beta_0 = 0$ (the regression line passes through the origin)
- Alternative hypothesis (H_1): $\beta_0 \neq 0$ (the regression line does not pass through the origin)

- Estimate the slope and intercept coefficients (b_0 and b_1): Using the sample data, calculate the slope (b_1) and intercept (b_0) coefficients for the regression model.

- Calculate the standard errors for the slope and intercept coefficients ($SE(b_0)$ and $SE(b_1)$): Compute the standard errors of the slope and intercept coefficients using the following formulas:

$$SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2) \sum (x_i - \bar{x})^2}}$$

$$SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2) \sum (x_i - \bar{x})^2} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]}$$

$$t_{\text{stat}} = \frac{b_1 - 0}{SE(b_1)}$$

$$t_{\text{stat}} = \frac{b_0 - 0}{SE(b_0)}$$

- Compute the t-statistics for the slope and intercept coefficients:
Calculate the t-statistics for the slope and intercept coefficients using the following formulas:

$$t_{\text{value } b_0} = \frac{b_0 - 0}{SE(b_0)}$$

p-value

$$e_i \sim N(0, \sigma^2)$$

$$\frac{\bar{x} - \mu}{SE}$$

- Calculate the p-values for the slope and intercept coefficients: Using the t-statistics and the degrees of freedom, look up the corresponding p-values from the t-distribution table or use a statistical calculator.

- Compare the p-values to the chosen significance level (α): A common choice for α is 0.05, which corresponds to a 95% confidence level.
Compare the calculated p-values to α :

- If the p-value is less than or equal to α , reject the null hypothesis.
- If the p-value is greater than α , fail to reject the null hypothesis.

Confidence Intervals for Coefficients

29 April 2023 02:35

1. Estimate the slope and intercept coefficients (b_0 and b_1): Using the sample data, calculate the slope (b_1) and intercept (b_0) coefficients for the regression model.
2. Calculate the standard errors for the slope and intercept coefficients ($SE(b_0)$ and $SE(b_1)$):

$$SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2) \sum (x_i - \bar{x})^2}}$$

$$SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]}$$

3. Determine the degrees of freedom: In a simple linear regression, the degrees of freedom (df) is equal to the number of observations (n) minus the number of estimated parameters (2: the intercept and the slope coefficient).
 $df = n - 2$

4. Find the critical t-value: Look up the critical t-value from the t-distribution table or use a statistical calculator based on the chosen confidence level (e.g., 95%) and the degrees of freedom calculated in step 3.
5. Calculate the confidence intervals for the slope and intercept coefficients: Compute the confidence intervals for the slope (b_1) and intercept (b_0) coefficients using the following formulas:

$$CI_{b_0} = b_0 \pm t\text{-value} * SE(b_0)$$

$$CI_{b_1} = b_1 \pm t\text{-value} * SE(b_1)$$

These confidence intervals represent the range within which the true population regression coefficients are likely to fall with a specified level of confidence (e.g., 95%)

t-dist

significance ↓

0.05

→ 95% prob

$$b_1 \pm 3.18 \times SE(b_1)$$

lower + upper
b₁

Others

28 April 2023 07:02

OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Sat, 29 Apr 2023	Prob (F-statistic):	1.58e-96			
Time:	07:32:56	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
Omnibus:	60.414	Durbin-Watson:	2.084			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241			
Skew:	-1.327	Prob(JB):	1.44e-33			
Kurtosis:	6.332	Cond. No.	454.			

Recap

05 May 2023 19:29



OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Sat, 29 Apr 2023	Prob (F-statistic):	1.58e-96			
Time:	07:32:56	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					
<i>t-test</i> WLS						
	coef	std err	t	P> t	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011
Omnibus:	60.414	Durbin-Watson:	2.084			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241			
Skew:	-1.327	Prob(JB):	1.44e-33			
Kurtosis:	6.332	Cond. No.	454.			

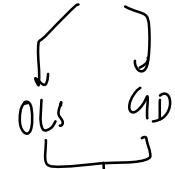
Assumption

test of normality
of residuals

Linear



Coefficient

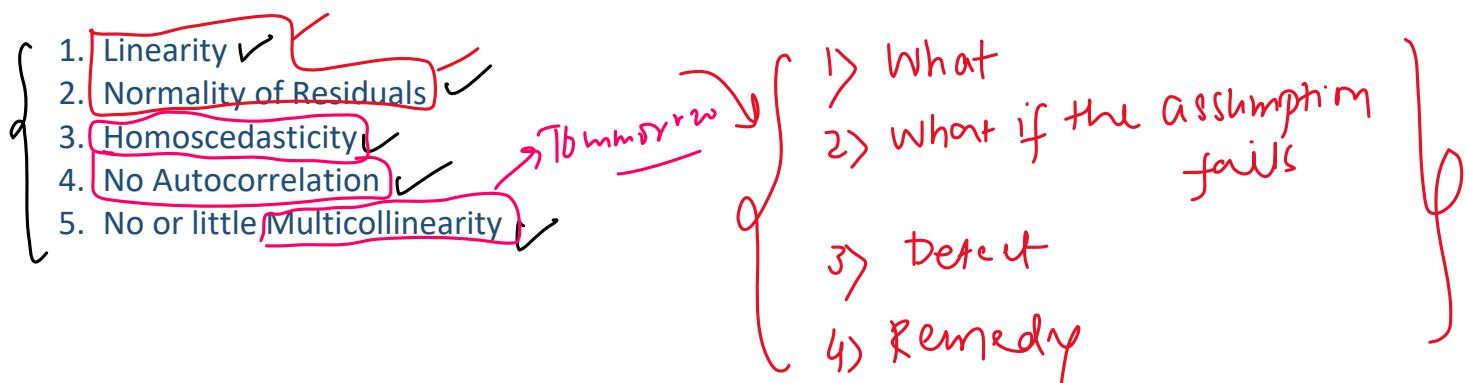


half Regression Analysis
→ Assumption of LR
→ Multicollinearity

Assumptions of Linear Regression

03 May 2023 14:17

Linear regression relies on several assumptions to ensure the validity and reliability of the estimates and inferences. The key assumptions of linear regression are:



1. Linearity

03 May 2023 14:21

The Assumption

{ There is a linear relationship between the independent variables and the dependent variable.
The model assumes that changes in the independent variables lead to proportional changes in the dependent variable.

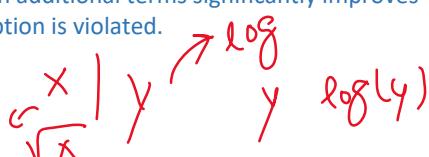
What happens when this assumption is violated?

1. Bias in parameter estimates: When the true relationship is not linear, the estimated regression coefficients can be biased, leading to incorrect inferences about the relationship between the independent and dependent variables.
2. Reduced predictive accuracy: A mis-specified linear model may not accurately capture the underlying relationship, which can result in poor predictive performance. The model might underfit the data, missing important patterns and trends.
3. Invalid hypothesis tests and confidence intervals: The violation of the linearity assumption can affect the validity of hypothesis tests and confidence intervals, leading to incorrect inferences about the significance of the independent variables and the effect sizes.

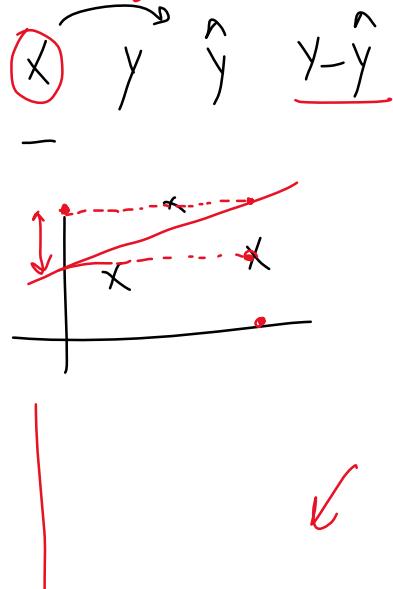
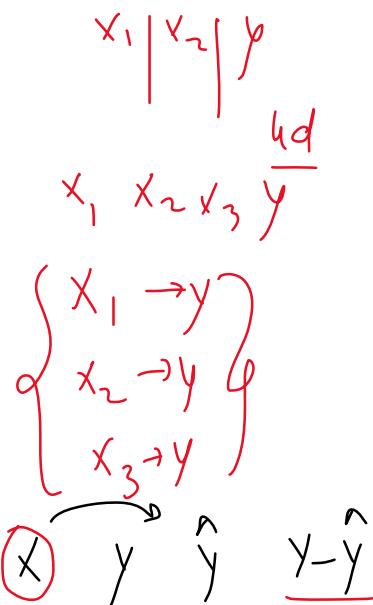
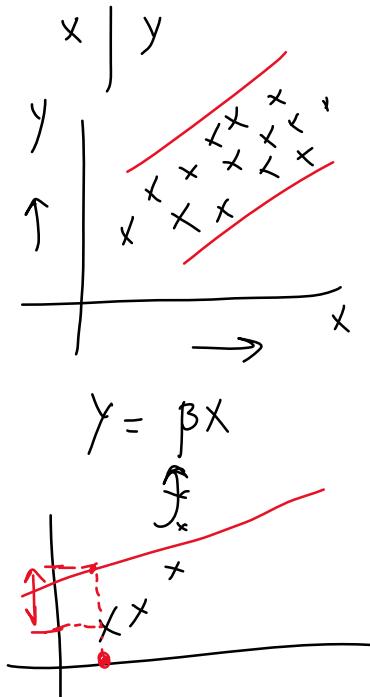
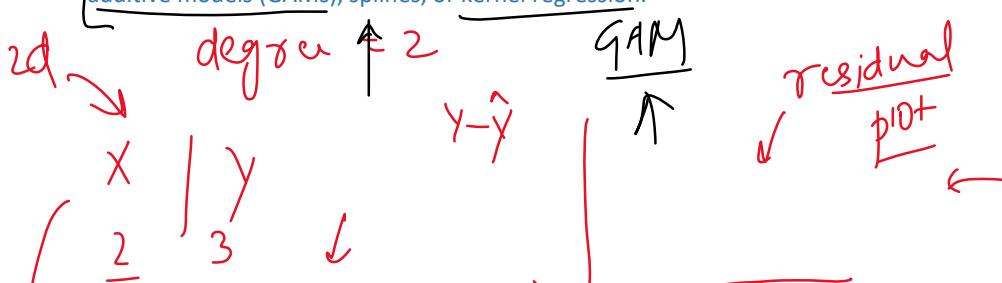
How to check this assumption

1. Scatter plots: Create scatter plots of the dependent variable against each independent variable. If the relationship appears to be linear, the linearity assumption is likely satisfied. Nonlinear patterns or other trends may indicate that the assumption is violated.
2. Residual plots: Plot the residuals (the differences between the observed and predicted values) against the predicted values or against each independent variable. If the linearity assumption holds, the residuals should be randomly scattered around zero, with no discernible pattern. Any trends, curvature, or heteroscedasticity in the residual plots suggest that the linearity assumption may be violated.
3. Polynomial terms: Add polynomial terms to your model and compare the model fit with the original linear model. If the new model with additional terms significantly improves the fit, it may suggest that the linearity assumption is violated.
polynomial & log

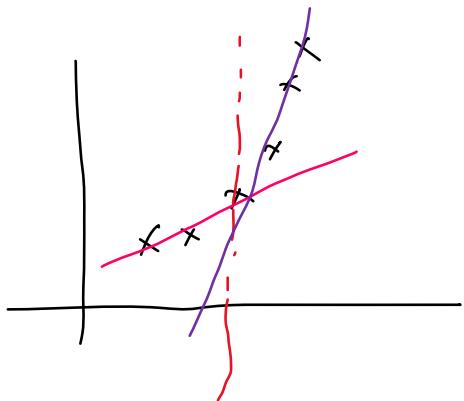
What to do when the assumption fails?



1. Transformations: Apply transformations to the dependent and/or independent variables to make their relationship more linear. Common transformations include logarithmic, square root, and inverse transformations.
2. Polynomial regression: Add polynomial terms of the independent variables to the model to capture non-linear relationships.
3. Piecewise regression: Divide the range of the independent variable into segments and fit separate linear models to each segment.
4. Non-parametric or semi-parametric methods: Consider using non-parametric or semi-parametric methods that do not rely on the linearity assumption, such as generalized additive models (GAMs), splines, or kernel regression.



$$\begin{array}{c}
 \text{2} \quad 1 \quad 3' \quad \downarrow \\
 Y = \beta_0 + \beta_1 X \\
 X^0 \quad X^1 \quad X^2 \\
 | \quad 2 \quad 4 \quad 3
 \end{array}
 \quad
 \begin{array}{c}
 \checkmark \quad \hat{y} \\
 \text{4d} \\
 Y
 \end{array}
 \quad
 \boxed{Y = \beta_0 + \beta_1 X^0 + \beta_2 X^1 + \beta_3 X^2}$$



2. Normality of Residual

03 May 2023 16:36

The Assumption

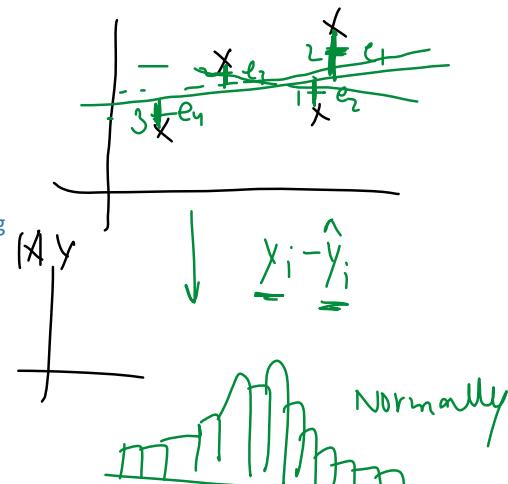
The error terms (residuals) are assumed to follow a normal distribution with a mean of zero and a constant variance.

$$\text{residuals} \sim f(x)$$



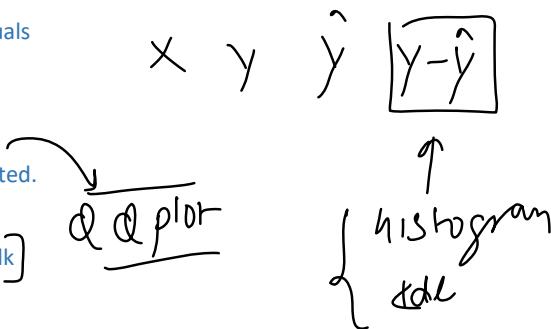
What happens when this assumption is violated?

1. Inaccurate hypothesis tests: The t-tests and F-tests used to assess the significance of the regression coefficients and the overall model rely on the normality assumption. If the residuals are not normally distributed, these tests may produce inaccurate results, leading to incorrect inferences about the significance of the independent variables.
2. Invalid confidence intervals: The confidence intervals for the regression coefficients are based on the assumption of normally distributed residuals. If the normality assumption is violated, the confidence intervals may not be accurate, affecting the interpretation of the effect sizes and the precision of the estimates.
3. Model performance: The violation of the normality assumption may indicate that the chosen model is not the best fit for the data, potentially leading to reduced predictive accuracy.



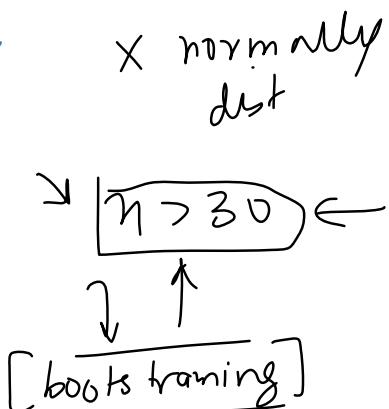
How to check this assumption

1. Histogram of residuals: Plot a histogram of the residuals to visually assess their distribution. If the histogram resembles a bell-shaped curve, it suggests that the residuals are normally distributed.
2. Q-Q plot: A Q-Q (quantile-quantile) plot compares the quantiles of the residuals to the quantiles of a standard normal distribution. If the points in the Q-Q plot fall approximately along a straight line, it indicates that the residuals are normally distributed. Deviations from the straight line suggest deviations from normality.
3. Statistical tests: Statistical tests like Omnibus test, Jarque-Bera test or even Shapiro wilk test can test this assumption.



What to do when the assumption fails?

1. Model selection techniques: Employ model selection techniques like cross-validation, AIC, or BIC to choose the best model among different candidate models that can handle non-normal residuals.
2. Robust regression: Use robust regression techniques that are less sensitive to the distribution of the residuals, such as M-estimation, Least Median of Squares (LMS), or Least Trimmed Squares (LTS). Transformation may also help.
3. Non-parametric or semi-parametric methods: Consider using non-parametric or semi-parametric methods that do not rely on the normality assumption, such as generalized additive models (GAMs), splines, or kernel regression.
4. Use bootstrapping: Bootstrap-based inference methods do not rely on the normality of residuals and can provide more accurate confidence intervals and hypothesis tests.



Remember that the normality of residuals assumption is not always critical for linear regression, especially when the sample size is large, due to the Central Limit Theorem.

Omnibus Test

04 May 2023 10:03

The Omnibus test is a statistical test used to check if the residuals from a linear regression model follow a normal distribution. The test is based on the skewness and kurtosis of the residuals. Here's a step-by-step guide on how to conduct the Omnibus test:

- Decide the Null and Alternate Hypothesis: The Null hypothesis states that the residuals are normally distributed and the Alternate Hypothesis says that the residuals are not normally distributed.

→ 2. Fit the linear regression model: Fit the linear regression model to your data to obtain the predicted values.

→ 3. Calculate the residuals: Compute the residuals (error terms) by subtracting the predicted values from the observed values of the dependent variable.

4. Calculate the skewness: Calculate the skewness of the residuals. Skewness measures the asymmetry of the distribution. For a normal distribution, skewness is expected to be close to zero.

5. Calculate the kurtosis: Calculate the kurtosis of the residuals. Kurtosis measures the "tailedness" of the distribution. For a normal distribution, kurtosis is expected to be close to zero (in excess kurtosis terms).

6. Calculate the Omnibus test statistic: Compute the Omnibus test statistic (K^2) using the skewness and kurtosis values. The formula for the Omnibus test statistic is:

$$K^2 = n \left[\frac{(\text{skewness})^2}{6} + \frac{(\text{kurtosis})^2}{24} \right]$$

$n \rightarrow \text{number of observations}$

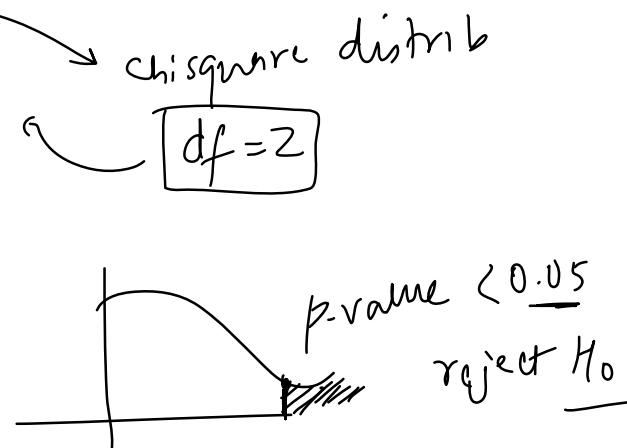
chi square distrib
 $df = 2$

6. Determine the p-value: The Omnibus test statistic follows a chi-square distribution with 2 degrees of freedom. Use this distribution to calculate the p-value corresponding to the test statistic.

7. Compare the p-value to the significance level: Compare the p-value obtained in step 6 to your chosen significance level (e.g., 0.05). If the p-value is greater than the significance level, you can accept the null hypothesis that the residuals are normally distributed. If the p-value is smaller than the significance level, you reject the null hypothesis, suggesting that the residuals may not follow a normal distribution.

$$\text{Coef } \hat{y} \rightarrow \hat{y} - \hat{y}$$

$$\hat{y} - \hat{y}$$



$< 0.05 \rightarrow$
reject H_0

3. Homoscedasticity

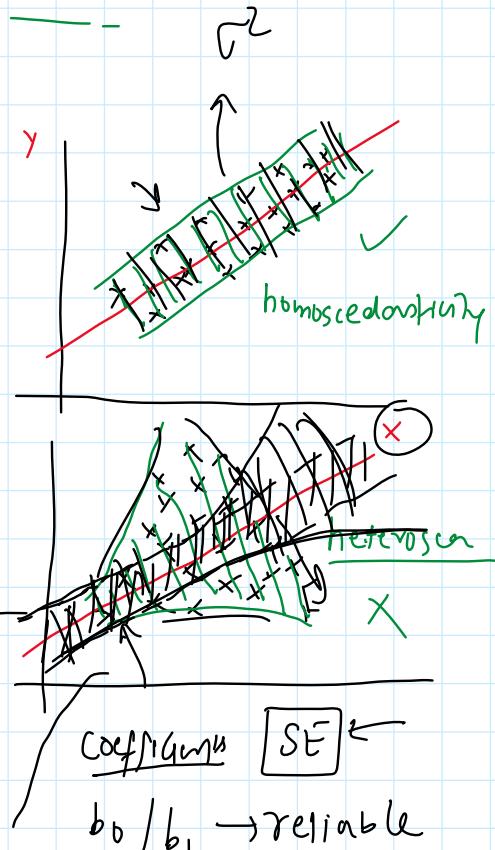
03 May 2023 16:49

The Assumption

The spread of the error terms (residuals) should be constant across all levels of the independent variables. If the spread of the residuals changes systematically, it leads to heteroscedasticity, which can affect the efficiency of the estimates.

What happens when this assumption is violated?

What is the problem



How to check this assumption

t-stats
coefficients
SE

1. Residual plot: Create a scatter plot of the residuals (the differences between the observed and predicted values) against the predicted values or against each independent variable. If the plot shows a random scattering of points around zero with no discernible pattern, it suggests homoscedasticity. If there is a systematic pattern, such as a funnel shape or a curve, it indicates heteroscedasticity.
2. Breusch-Pagan test: This is a formal statistical test for heteroscedasticity. The null hypothesis is that the error variances are constant (homoscedastic). If the resulting p-value is less than a chosen significance level (e.g., 0.05), the null hypothesis is rejected, indicating heteroscedasticity.

What to do when the assumption fails?

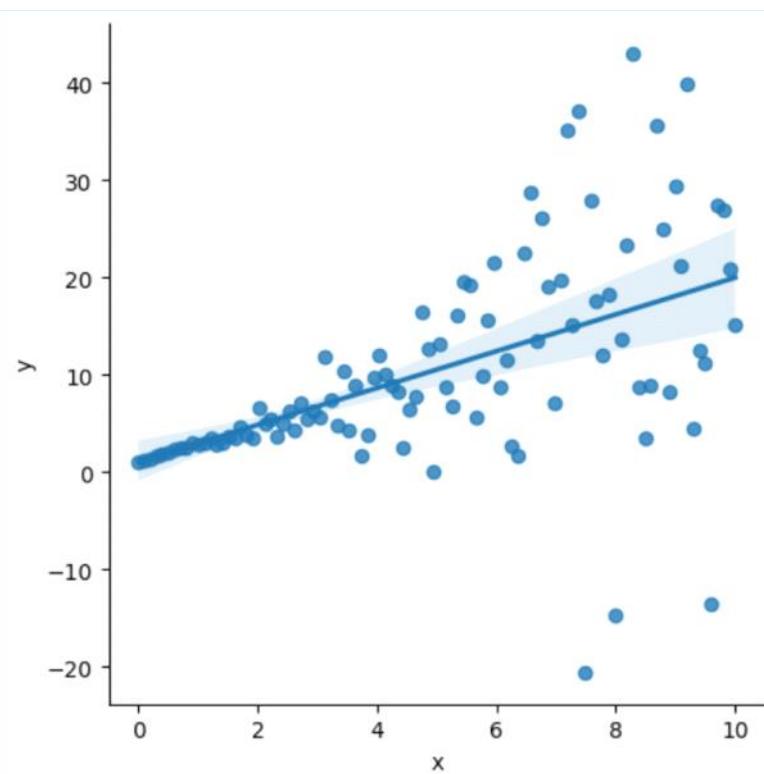
1. Transformations: Apply transformations to the dependent and/or independent variables to stabilize the variance of the residuals. Common transformations include logarithmic, square root, and inverse transformations.
2. Weighted Least Squares (WLS): Use a weighted least squares approach, which assigns different weights to the observations based on the magnitude of their residuals. This method can help account for heteroscedasticity by giving more importance to observations with smaller residuals and less importance to those with larger residuals.
3. Robust standard errors: Calculate robust (or heteroscedasticity-consistent) standard errors for the regression coefficients. These standard errors are more reliable under heteroscedasticity and can be used to perform more accurate hypothesis tests and construct valid confidence intervals.

$$\text{error} = \epsilon_i$$
$$\downarrow$$
$$\text{residual}$$
$$\boxed{\text{Var}(\epsilon_i) = f(x)}$$

homosced.

$$\text{var}(\epsilon_i) = \boxed{f(x)}$$

heterosced.



Standard Error

05 May 2023 21:29

$$\text{DB} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{30000}} = \mu$$

\bar{x}_1

$\mu = \frac{\sum x_i}{n}$ std dev

india
IT
employee
salary

avg salary
IT employee

pop mean

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

avg dist from
the mean

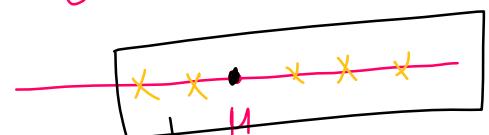
xxxx
x x + x

CLT

$$100 \rightarrow \bar{x}_1$$

$$100 \rightarrow \bar{x}_2$$

$$100 \rightarrow \bar{x}_3$$



how

$$SE = \frac{\sigma}{\sqrt{n}}$$

pop std dev
 $n \rightarrow \text{obs}$

std → standard error

$$N(\mu, \sigma^2/n)$$

$$SE(b_1)$$

$$SE(b_0)$$



$$b_{1(a)}, b_{1(b)}, b_{1(c)}, \dots, b_{1(l)}$$

const

SE

t-test
z-stat

$$\frac{\bar{x} - \mu}{SE}$$

$$SE = \sqrt{s}$$

$$\frac{b_1 - 0}{SE}$$

$$\frac{1000 - 0}{t-\text{stat}^2}$$

x unit SE

T2 SP

p-value

$$\exp | \text{salary} | 100$$

$$0.880001$$

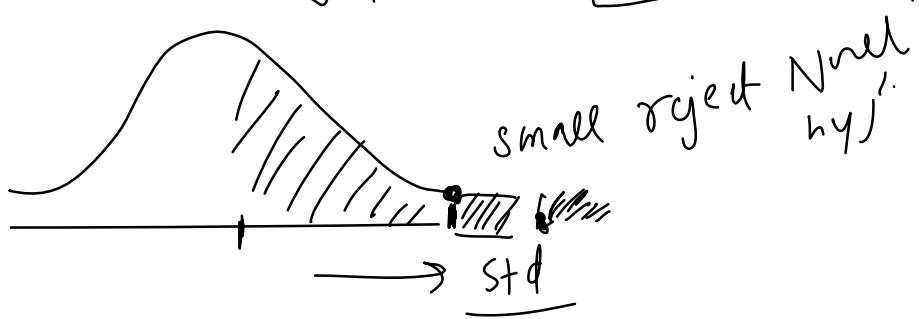
$$B_1$$

$$\frac{\text{std dev}}{\text{pred}}$$

p-value

4, 5

$$SE = \frac{s}{\sqrt{n}} + SE = \frac{s}{\sqrt{n}} \approx \boxed{3SE} \quad \underline{4,5}$$



Breusch-Pagan Test

05 May 2023 08:15

The Breusch-Pagan test, also known as the Cook-Weisberg test, is a statistical test used to detect heteroscedasticity in a linear regression model. The test is based on the assumption that the variance of the errors is a function of one or more independent variables. Here are the steps to perform the Breusch-Pagan test:

1. Estimate the linear regression model: Fit a linear regression model to the data using the ordinary least squares (OLS) method. Obtain the residuals (errors) from this model.
2. Calculate the squared residuals: Square each residual obtained in step 1.
3. Regress squared residuals on the independent variables: Perform another linear regression, this time with the squared residuals as the dependent variable and the same set of independent variables used in the original model. Obtain the R-squared value from this regression.
4. Calculate the test statistic: The Breusch-Pagan test statistic, known as the Lagrange Multiplier (LM) statistic, is calculated as follows:
$$LM = n * R^2$$
where n is the number of observations and R^2 is the R-squared value obtained in step 3.
5. Determine the p-value: The LM statistic follows a chi-squared distribution with k degrees of freedom, where k is the number of independent variables (excluding the constant term). Calculate the p-value for the LM statistic using the chi-squared distribution.
6. Make a decision based on the p-value: Compare the calculated p-value to a chosen significance level (usually $\alpha = 0.05$). If the p-value is less than or equal to α , reject the null hypothesis and conclude that there is evidence of heteroscedasticity in the data. If the p-value is greater than α , do not reject the null hypothesis and assume that the data exhibits homoscedasticity (constant variance of the residuals).

Note that the Breusch-Pagan test assumes a linear relationship between the independent variables and the variance of the errors. If the relationship is not linear, the test may not be appropriate, and other tests for heteroscedasticity should be considered.

4. No Autocorrelation

03 May 2023 16:49

The Assumption

There should be no apparent correlation or pattern in the residuals, as this would suggest that the error terms are not independent.

What happens when this assumption is violated?

1. **Inefficient estimates:** The parameter estimates (coefficients) remain unbiased, but they are no longer the best linear unbiased estimators (BLUE). The inefficiency of the estimates implies that the standard errors may be larger than they should be, which may reduce the statistical power of hypothesis tests.
2. **Inaccurate hypothesis tests:** The t-tests and F-tests used to assess the significance of the regression coefficients and the overall model rely on the assumption of no autocorrelation. If the error terms exhibit autocorrelation, these tests may produce misleading results, leading to incorrect inferences about the significance of the independent variables.
3. **Invalid confidence intervals:** The confidence intervals for the regression coefficients are based on the assumption of no autocorrelation. If this assumption is violated, the confidence intervals may not be accurate, affecting the interpretation of the effect sizes and the precision of the estimates.

How to check this assumption

1. **Durbin-Watson test:** This is a formal statistical test for autocorrelation, specifically first-order autocorrelation. The Durbin-Watson test statistic ranges from 0 to 4, with a value of 2 indicating no autocorrelation. Values less than 2 suggest positive autocorrelation, while values greater than 2 indicate negative autocorrelation. It is important to note that the Durbin-Watson test is only applicable for first-order autocorrelation and may not detect higher-order autocorrelation.

What to do when the assumption fails?

1. **Lagged variables:** Include lagged values of the dependent variable or the

independent variables as predictors in the model to account for the autocorrelation.

2. **Differencing:** Apply differencing to the dependent and/or independent variables, which can help remove the autocorrelation by focusing on the changes between consecutive observations rather than the absolute values.
3. **Generalized least squares (GLS):** Use a generalized least squares approach that accounts for the autocorrelation structure in the error terms, leading to more efficient and reliable estimates.
4. **Time series models:** Consider using specialized time series models, such as autoregressive (AR), moving average (MA), autoregressive integrated moving average (ARIMA), or seasonal decomposition of time series (STL), which are designed to handle autocorrelation.
5. **Robust standard errors:** Calculate robust standard errors that are more reliable under autocorrelation, such as Newey-West standard errors or HAC (heteroscedasticity and autocorrelation consistent) standard errors.

5. No Multicollinearity

03 May 2023 16:49

Extra

04 May 2023 14:19

What is Multicollinearity

06 May 2023 08:23

Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a multiple regression model are highly correlated. In other words, these variables exhibit a strong linear relationship, making it difficult to isolate the individual effects of each variable on the dependent variable.

0.9 \rightarrow 0.8

$$\begin{array}{c} \boxed{100} \\ \xrightarrow{\quad} \begin{array}{c|c|c} \text{cgpa} & \text{iq} & \text{lpa} \\ \hline 8 & 80 & 8 \end{array} \end{array}$$

Corr \rightarrow linear relation

$$\begin{array}{c|c|c} \text{iq} & \text{backlog} & \text{lpa} \\ \hline \downarrow & & \end{array}$$

iq \uparrow backlog \downarrow multicoll

$$\begin{array}{c} \xrightarrow{\quad} \begin{array}{c|c|c} \text{cgpa} & \text{doub} & \text{lpa} \sim 3 \\ \hline T & \uparrow & \end{array} \\ \beta_0 \beta_1 \beta_2 \rightarrow \text{unreliable} \end{array}$$

$$lpa = \beta_0 + \beta_1 \text{cgpa} + \beta_2 iq$$

interpret

When is Multicollinearity bad?

06 May 2023 14:33

1. Inference:

- Inference focuses on understanding the relationships between the variables in a model.
- It aims to draw conclusions about the underlying population or process that generated the data.
- Inference often involves hypothesis testing, confidence intervals, and determining the significance of predictor variables.
- The primary goal is to provide insights about the structure of the data and the relationships between variables.
- Interpretability is a key concern when performing inference, as the objective is to understand the underlying mechanisms driving the data.
- Examples of inferential techniques include linear regression, logistic regression, and ANOVA.

2. Prediction:

Linear ref → predict → mult

- Prediction focuses on using a model to make accurate forecasts or estimates for new, unseen data.
- It aims to generalize the model to new instances, based on the patterns observed in the training data.
- Prediction often involves minimizing an error metric, such as mean squared error or cross-entropy loss, to assess the accuracy of the model.
- The primary goal is to create an accurate and reliable model for predicting outcomes, rather than understanding the relationships between variables.
- Interpretability may be less important in predictive modelling, as the main objective is to create accurate forecasts rather than understanding the underlying structure of the data.
- Examples of predictive techniques include decision trees, support vector machines, neural networks, and ensemble methods like random forests and gradient boosting machines.

In summary, inference focuses on understanding the relationships between variables and interpreting the underlying structure of the data, while prediction focuses on creating accurate forecasts for new, unseen data based on the patterns observed in the training data.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$\uparrow \quad \uparrow \quad \uparrow$

$\beta_0 \quad \beta_1 \quad \beta_2$

$y \rightarrow x_1 \quad y \rightarrow x_2$

$$\boxed{100} \rightarrow \boxed{cgpa} \parallel \boxed{iq} \parallel \boxed{lpa}$$

$\uparrow \quad \uparrow \quad \uparrow$

Finance → banking
model

Cust user data → loan

User activity → bank/allow

$$\boxed{X_1 \mid X_2 \mid Y}$$

$$\boxed{X_1 = \alpha_0 + \alpha_1 X_2 + \eta}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

stochastic error

$$Y = \beta_0 + \beta_1 (\alpha_0 + \alpha_1 X_2 + \eta) + \epsilon$$

$$Y = \beta_0 + \beta_1 \alpha_0 + \beta_1 \alpha_1 X_2 + \beta_1 \eta + \epsilon$$

$$\rightarrow Y = (\beta_0 + \beta_1 \alpha_0) + \beta_1 \alpha_1 X_2 + (\beta_1 \eta + \epsilon)$$

$$\boxed{Y \rightarrow X_2}$$

$$\xrightarrow{\text{op hm}} \boxed{Y = c_0 + c_1 X_2 + \gamma}$$

What exactly happens in Multicollinearity(Mathematically?)

06 May 2023 08:29

When multicollinearity is present in a model, it can lead to several issues, including:

$$\hat{y}_{pa} = \beta_0 + \beta_1 \underline{cgpa} + \beta_2 \underline{iq}$$

$SE(\beta) \rightarrow \text{high value}$

1. **Difficulty in identifying the most important predictors:** Due to the high correlation between independent variables, it becomes challenging to determine which variable has the most significant impact on the dependent variable.
2. **Inflated standard errors:** Multicollinearity can lead to larger standard errors for the regression coefficients, which decreases the statistical power and can make it challenging to determine the true relationship between the independent and dependent variables.
3. **Unstable and unreliable estimates:** The regression coefficients become sensitive to small changes in the data, making it difficult to interpret the results accurately.

statsmodel → summary

OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Sat, 29 Apr 2023	Prob (F-statistic):	1.58e-96			
Time:	07:32:56	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	p > t	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV → y_1	0.0458	0.001	32.809	0.000	0.043	0.049
Radio → y_2	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper → y_3	-0.0010	0.006	-0.177	0.860	-0.013	0.011
Omnibus:	60.414	Durbin-Watson:	2.084			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241			
Skew:	-1.327	Prob(JB):	1.44e-33			
Kurtosis:	6.332	Cond. No.	454.			

$TV | \underline{\text{radio}} | \underline{\text{newspaper}} | \underline{\text{sales}}$

$x_1 \quad x_2 \quad x_3 \quad y$

$SE \rightarrow \underline{\text{precise}}$

← σ_{cgf} analy
impact

multicollinearity

multicollinearity
(30) >

→ unstable coefficients

→ high SE

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$| \underline{cgpa} | \underline{iq} | \underline{\hat{y}_{pa}} \longleftrightarrow \underline{\text{sample (100)}}$

$$\text{Var}(\beta) = \text{SE}(\beta)$$

$$\hat{y}_{pa} = \beta_0 + \beta_1 \underline{cgpa} + \beta_2 \underline{iq}$$

$\uparrow \quad \uparrow \quad \uparrow$
 $\beta_0 \quad \beta_1 \quad \beta_2$

$$\underline{\beta_0} \quad \underline{\beta_1} \quad \underline{\beta_2}$$

$$\underline{\beta_0} \quad \underline{\beta_1} \quad \underline{\beta_2}$$

$$\underline{\text{SE}(\beta_0)} \quad \underline{\text{SE}(\beta_1)} \quad \underline{\text{SE}(\beta_2)}$$

$$\uparrow \quad \uparrow \quad \uparrow$$

$$\text{Var}(\beta) = \text{SE}$$

linearity

$$\sqrt{\text{Var}(\beta)} = (\text{SE}) \quad \text{uncertainty}$$

normally dist $\rightarrow N \sim (0, \sigma^2)$

$$\beta \rightarrow (3)$$

Variance unc

$$\text{Var}(\beta) = \sigma^2 (X^T X)^{-1}$$

σ^2 $\boxed{3 \times 3}$

$$\text{SE } a \rightarrow \sqrt{\text{Var}(\beta_0)}$$

$$\text{SE } b \rightarrow \sqrt{\text{Var}(\beta_1)}$$

$$\text{SE } c \rightarrow \sqrt{\text{Var}(\beta_2)}$$

$$\text{SE}(\beta) = \sqrt{\text{diag}(\sigma^2 (X^T X)^{-1})}$$

perfect multicollin' $\rightarrow \det(X^T X) = 0$

$$\beta = (X^T X)^{-1} X^T y$$

sensitive to input $(X^T X)^{-1}$

\uparrow
strong multi

$\det(X^T X)$ very small

inv \det very small \uparrow \rightarrow high
 \uparrow inflate

$$\frac{1}{(\beta) \beta_0 \beta_1 \beta_2} (X^T X)^{-1} \frac{C\& P}{8.1} \rightarrow 8.2$$

\uparrow sensitive to data

$$(\beta) \rightarrow (X^T X)^{-1}$$

Perfect Multicollinearity

06 May 2023 08:35

Perfect multicollinearity occurs when one independent variable in a multiple regression model is an exact linear combination of one or more other independent variables. In other words, there is an exact linear relationship between the independent variables, making it impossible to uniquely estimate the individual effects of each variable on the dependent variable.

corr linear \downarrow

corr $x_1 = a_1 x_2 + q_0 + \text{error}$

$\hookrightarrow x_1 = a_1 x_2 + q_0$ \leftarrow perfect multicollinear

$$\begin{array}{c} \downarrow \\ \text{cgpa | percent | lpa} \\ \hline 8.5 & 85 & 7 \\ 9.12 & 91.2 & 6 \end{array}$$

percent = $10 \times \text{cgpa} + 0$

\uparrow

$a_1 = 10 \quad q_0 = 0$

cgpa | percent | lpa

8.5	83	?
9.12	95	

$10 \times \text{cgpa} + 0 + \underline{\text{lwa}}$

$\left[\begin{array}{c|c|c} \text{cgpa} & \text{percent} & \text{lpa} \\ \hline 8 & 80 & 3 \\ 6 & 60 & 4 \end{array} \right] \quad \beta \rightarrow$

$\text{lpa} = \beta_0 + \beta_1 \text{cgpa} + \beta_2 \text{percent} + \text{error}$

OLS / GD

$\beta = (X^T X)^{-1} X^T y$

$\beta_0 \quad \beta_1 \quad \beta_2$

X
↑
design
matrix

$$\begin{bmatrix} 1 & 8 & 80 \\ 1 & 6 & 60 \end{bmatrix}$$

inverse X

singular

matrix

X^T

$$\begin{bmatrix} 1 & 1 \\ 8 & 6 \\ 80 & 60 \end{bmatrix} \quad \begin{bmatrix} 1 & 8 & 80 \\ 1 & 6 & 60 \end{bmatrix} = \begin{bmatrix} 2 & 14 & 140 \\ 14 & 84 & 8400 \\ 140 & 8400 & 840000 \end{bmatrix}$$

$\boxed{\text{Det}} \rightarrow 2(0) - 14(0) + 140(0) = \boxed{0} \rightarrow$

Types of Multicollinearity

06 May 2023 08:30

1 [Structural multicollinearity]: Structural multicollinearity arises due to the way in which the variables are defined or the model is constructed. It occurs when one independent variable is created as a linear combination of other independent variables or when the model includes interaction terms or higher-order terms (such as polynomial terms) without proper scaling or centering.

2 [Data-driven multicollinearity]: Data-driven multicollinearity occurs when the independent variables in the dataset are highly correlated due to the specific data being analysed. In this case, the high correlation between the variables is not a result of the way the variables are defined or the model is constructed but rather due to the observed data patterns.



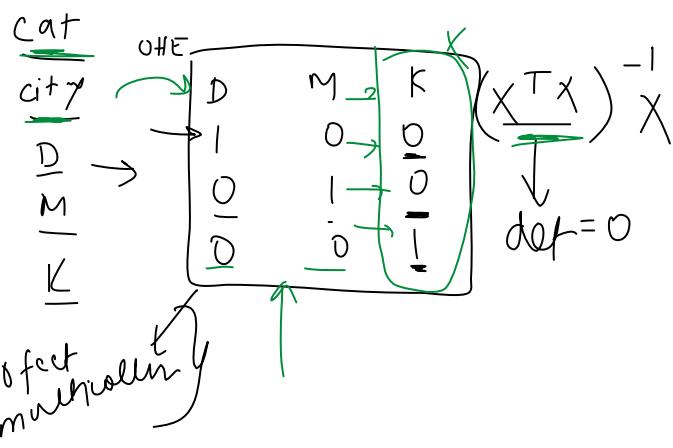
$$\text{OHE} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$1 - 0 - 1 = 0$$

$$\boxed{\beta} \text{ cal } X$$

$$1 - 0 - 0 = \boxed{1}$$

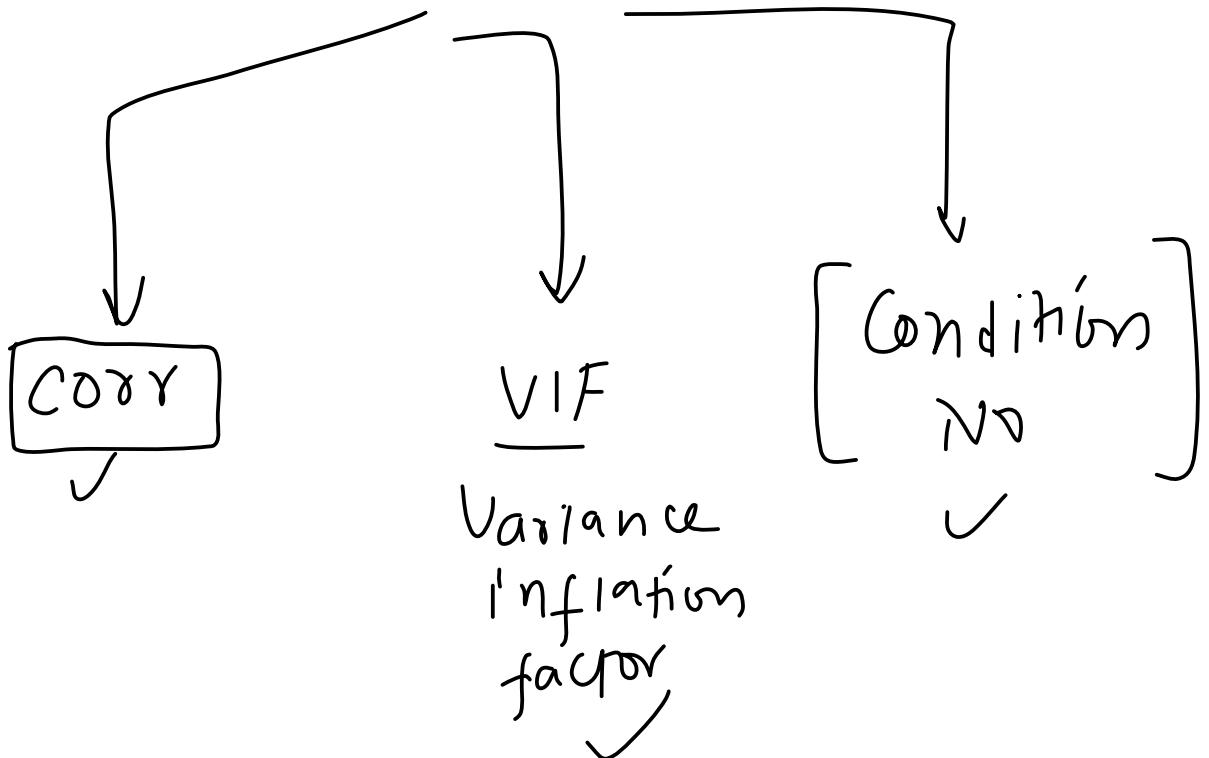
One hot encoding



$$\begin{matrix} & x & | & y \\ & \underline{x^0} & \underline{x^1} & \underline{x^2} & x \end{matrix}$$

How to Detect Multicollinearity

06 May 2023 08:30



Correlation

06 May 2023 14:23

Correlation is a measure of the linear relationship between two variables, and it is commonly used to identify multicollinearity in multiple linear regression models. Multicollinearity occurs when two or more predictor variables in the model are highly correlated, making it difficult to determine their individual contributions to the output variable.

To detect multicollinearity using correlation, you can calculate the correlation matrix of the predictor variables. The correlation matrix is a square matrix that shows the pairwise correlations between each pair of predictor variables. The diagonal elements of the matrix are always equal to 1, as they represent the correlation of a variable with itself. The off-diagonal elements represent the correlation between different pairs of variables.

In the context of multicollinearity, you should look for off-diagonal elements with high absolute values (e.g., greater than 0.8 or 0.9, depending on the specific application and the level of concern about multicollinearity). High correlation values indicate that the corresponding predictor variables are highly correlated and may be causing multicollinearity issues in the regression model.

It's important to note that while correlation can be a useful tool for detecting multicollinearity, it doesn't provide a complete picture of the severity of the issue or its impact on the regression model. Other diagnostic measures, such as Variance Inflation Factor (VIF) and condition number, can also be used to assess the presence and severity of multicollinearity in a regression model.



$$\rightarrow \underline{X_1} = \frac{a_1 X_2 + \text{error}}{r}$$

corr()

Variance Inflation Factor

06 May 2023 08:30

Variance Inflation Factor (VIF) is a metric used to quantify the severity of multicollinearity in a multiple linear regression model. It measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity.

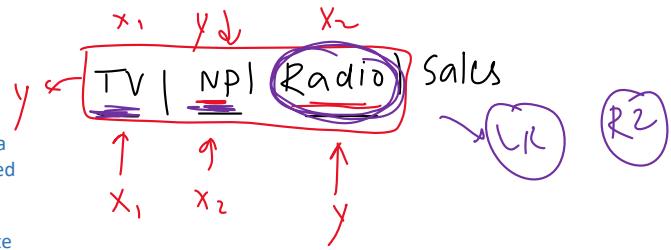
For each predictor variable in the regression model, VIF is calculated by performing a separate linear regression using that predictor as the response variable and the remaining predictor variables as the independent variables. The VIF for the predictor variable is then calculated as the reciprocal of the variance explained by the other predictors, which is equal to $1 / (1 - R^2)$. Here, R^2 is the coefficient of determination for the linear regression using the predictor variable as the response variable.

The VIF calculation can be summarized in the following steps:

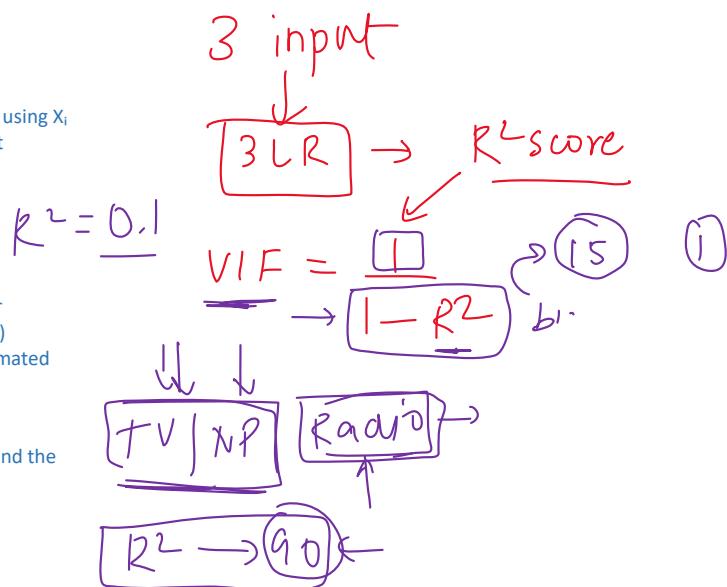
1. For each predictor variable X_i in the regression model, perform a linear regression using X_i as the response variable and the remaining predictor variables as the independent variables.
2. Calculate the R^2 value for each of these linear regressions.
3. Compute the VIF for each predictor variable X_i as $VIF_i = 1 / (1 - R^2_i)$.

A VIF value close to 1 indicates that there is very little multicollinearity for the predictor variable, whereas a high VIF value (e.g., greater than 5 or 10, depending on the context) suggests that multicollinearity may be a problem for the predictor variable, and its estimated coefficient might be less reliable.

Keep in mind that VIF only provides an indication of the presence and severity of multicollinearity and does not directly address the issue. Depending on the VIF values and the goals of the analysis, you might consider using techniques like variable selection, regularization, or dimensionality reduction methods to address multicollinearity.



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$



Condition No

16 May 2023 14:23

→ Eigen value
Eigen vectors

In the context of multicollinearity, the condition number is a diagnostic measure used to assess the stability and potential numerical issues in a multiple linear regression model. It provides an indication of the severity of multicollinearity by examining the sensitivity of the linear regression to small changes in the input data.

The condition number is calculated as the ratio of the largest eigenvalue to the smallest eigenvalue of the matrix $X^T X$, where X is the design matrix of the regression model (each row representing an observation and each column representing a predictor variable). A high condition number suggests that the matrix $X^T X$ is ill-conditioned and can lead to numerical instability when solving the normal equations for the regression coefficients.

In the presence of multicollinearity, the design matrix X has highly correlated columns, which can cause the eigenvalues of $X^T X$ to be very different in magnitude (one or more very large eigenvalues and one or more very small eigenvalues). As a result, the condition number becomes large, indicating that the regression model may be sensitive to small changes in the input data, leading to unstable coefficient estimates.

Typically, a condition number larger than 30 (or sometimes even larger than 10 or 20) is considered a warning sign of potential multicollinearity issues. However, the threshold for the condition number depends on the specific application and the level of concern about multicollinearity.

It's important to note that a high condition number alone is not definitive proof of multicollinearity. It is an indication that multicollinearity might be a problem, and further investigation (e.g., using VIF, correlation matrix, or tolerance values) may be required to confirm the presence and severity of multicollinearity.

$[X^T X] \rightarrow$ linear
trans

cond number

↓
ill cond
of
the matrix
 $(\det(A) \approx 0)$

> 30

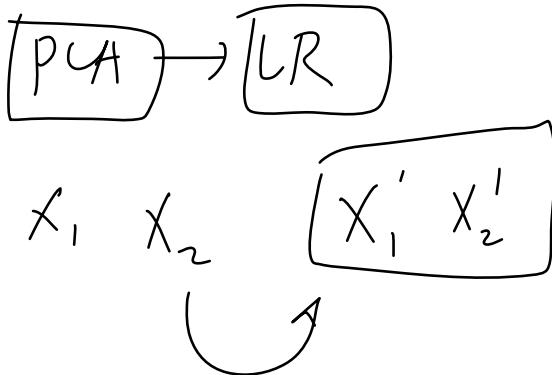
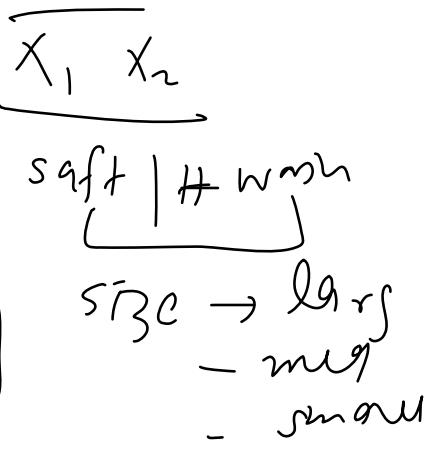
↓
ill cond
multicoll

How to remove multicollinearity

06 May 2023 08:31

$$VIF = 3 \quad 4 \quad 18$$

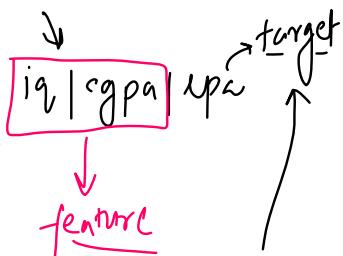
1. **Collect more data:** In some cases, multicollinearity might be a result of a limited sample size. Collecting more data, if possible, can help reduce multicollinearity and improve the stability of the model.
2. **Remove one of the highly correlated variables:** If two or more independent variables are highly correlated, consider removing one of them from the model. This step can help eliminate redundancy in the model and reduce multicollinearity. Choose the variable to remove based on domain knowledge, variable importance, or the one with the highest VIF.
3. **Combine correlated variables:** If correlated independent variables represent similar information, consider combining them into a single variable. This combination can be done by averaging, summing, or using other mathematical operations, depending on the context and the nature of the variables.
4. **Use partial least squares regression (PLS):** PLS is a technique that combines features of both principal component analysis and multiple regression. It identifies linear combinations of the predictor variables (called latent variables) that have the highest covariance with the response variable, reducing multicollinearity while retaining most of the predictive power.



What is Feature Selection

10 May 2023 14:46

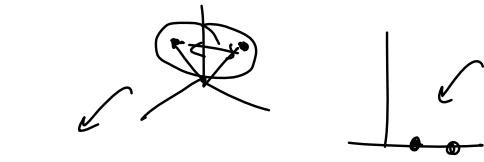
feature \rightarrow input col



$10 \rightarrow 5 \rightarrow$ important

$$\boxed{P > S}$$

dimension \rightarrow feature



dimension

2d 2 point \rightarrow 1d
100d 2 points \rightarrow 1d

sparsity

1) Curse of dimensionality \rightarrow

Certain features \rightarrow optimum results

Sqft | #beds | locn | price

\hookrightarrow 10 features \rightarrow 11 \rightarrow 12

ml algo \rightarrow

2) Computation complexity \rightarrow $\frac{500 \text{ cols}}{50 \text{ cols}} \frac{5000 \text{ rows}}{5000 \text{ rows}}$

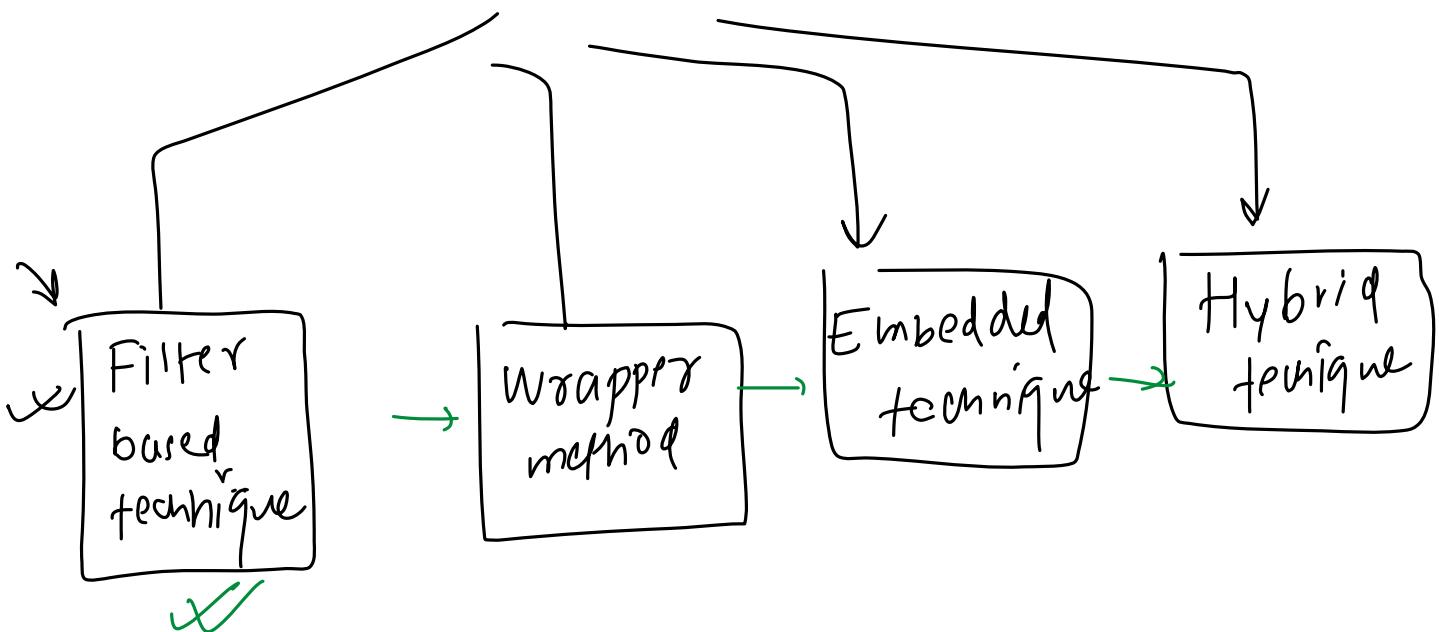
\downarrow
 $\frac{50 \text{ cols}}{500 \text{ cols}} \frac{5000 \text{ rows}}{5000 \text{ rows}}$

3) Interpretability \rightarrow $\frac{\text{predsum}}{\text{inference}} \rightarrow \frac{500 \text{ cols}}{500 \text{ factors}} \rightarrow$
 $\frac{500 \text{ cols}}{50 \text{ factors}} \rightarrow$ loan reject

5 factors

Types of Feature Selection

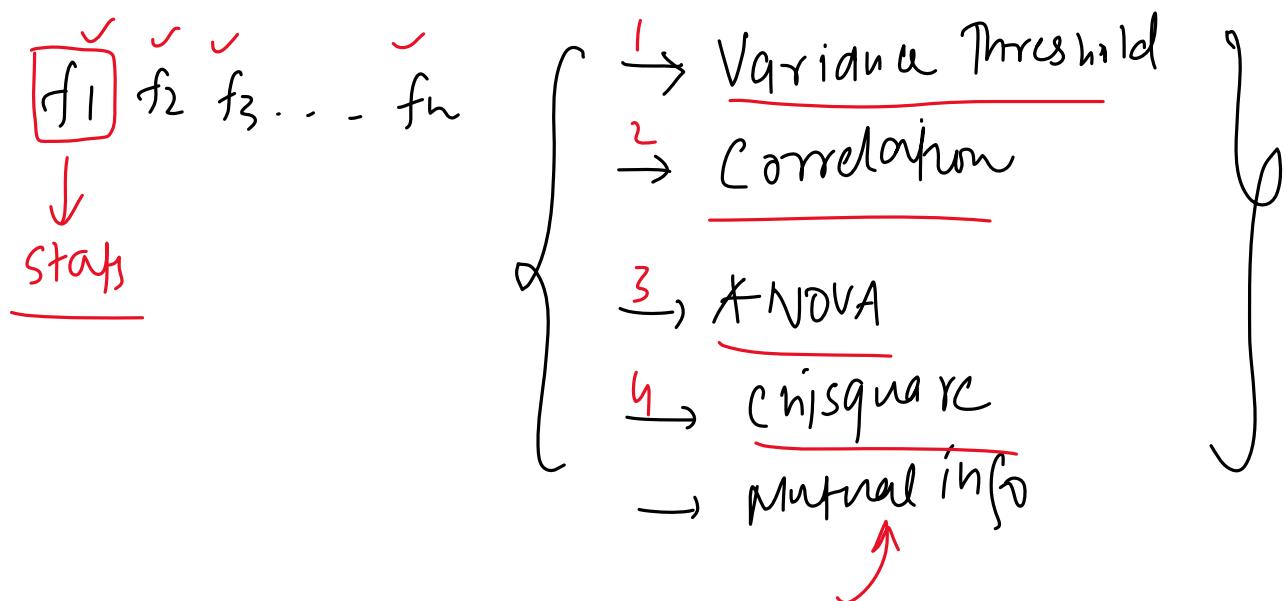
10 May 2023 14:47



Filter Based Feature Selection ←

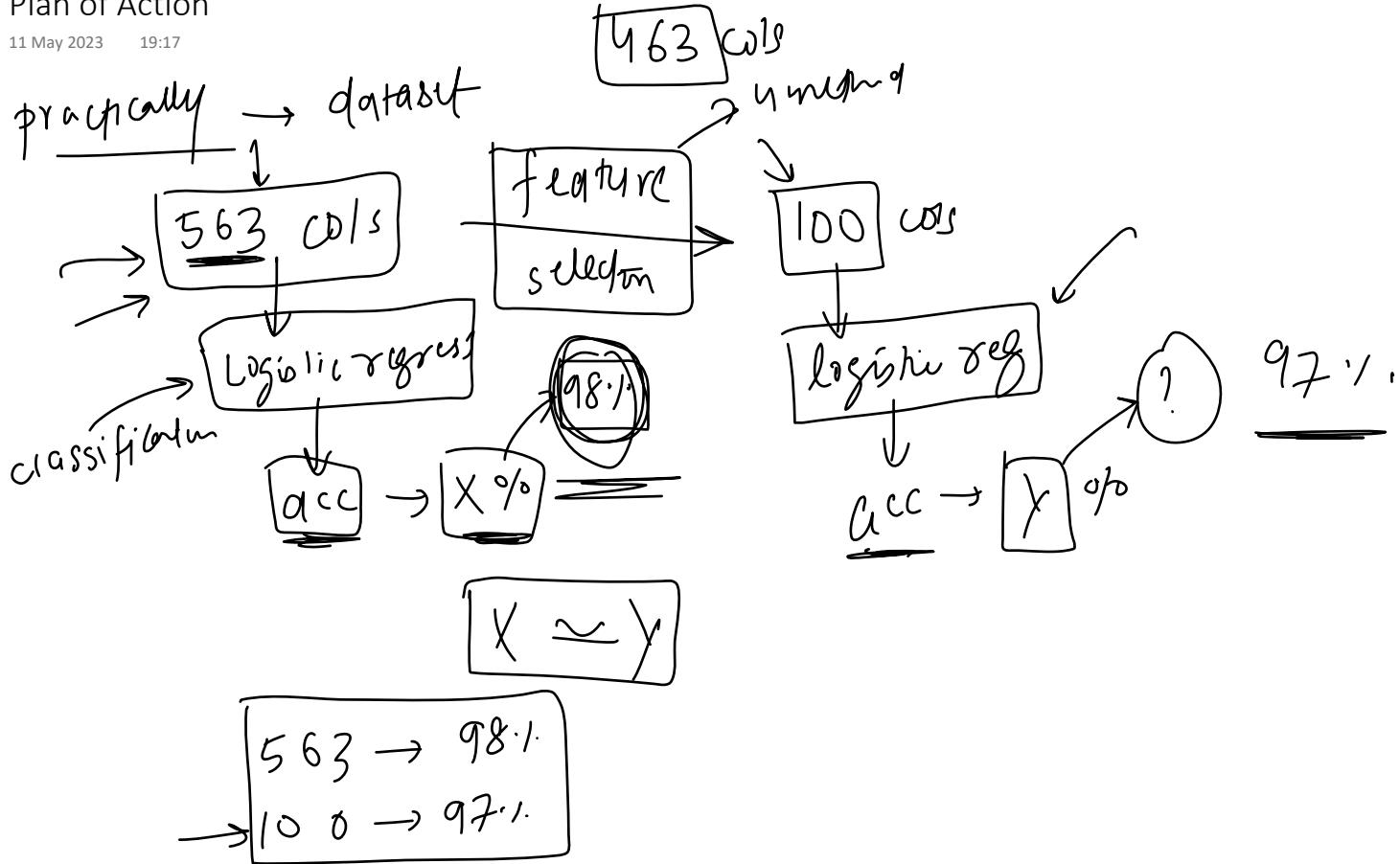
10 May 2023 14:47

{ Filter-based feature selection techniques are methods that use statistical measures to score each feature independently, and then select a subset of features based on these scores. These methods are called "filter" methods because they essentially filter out the features that do not meet some criterion.



Plan of Action

11 May 2023 19:17



1. Duplicate Features

10 May 2023 14:47

f_1	f_2	f_3	f_4	f_5	Output
1	2	1	2	3	X
2	1	2	2	3	N
3	3	3	2	3	N

same

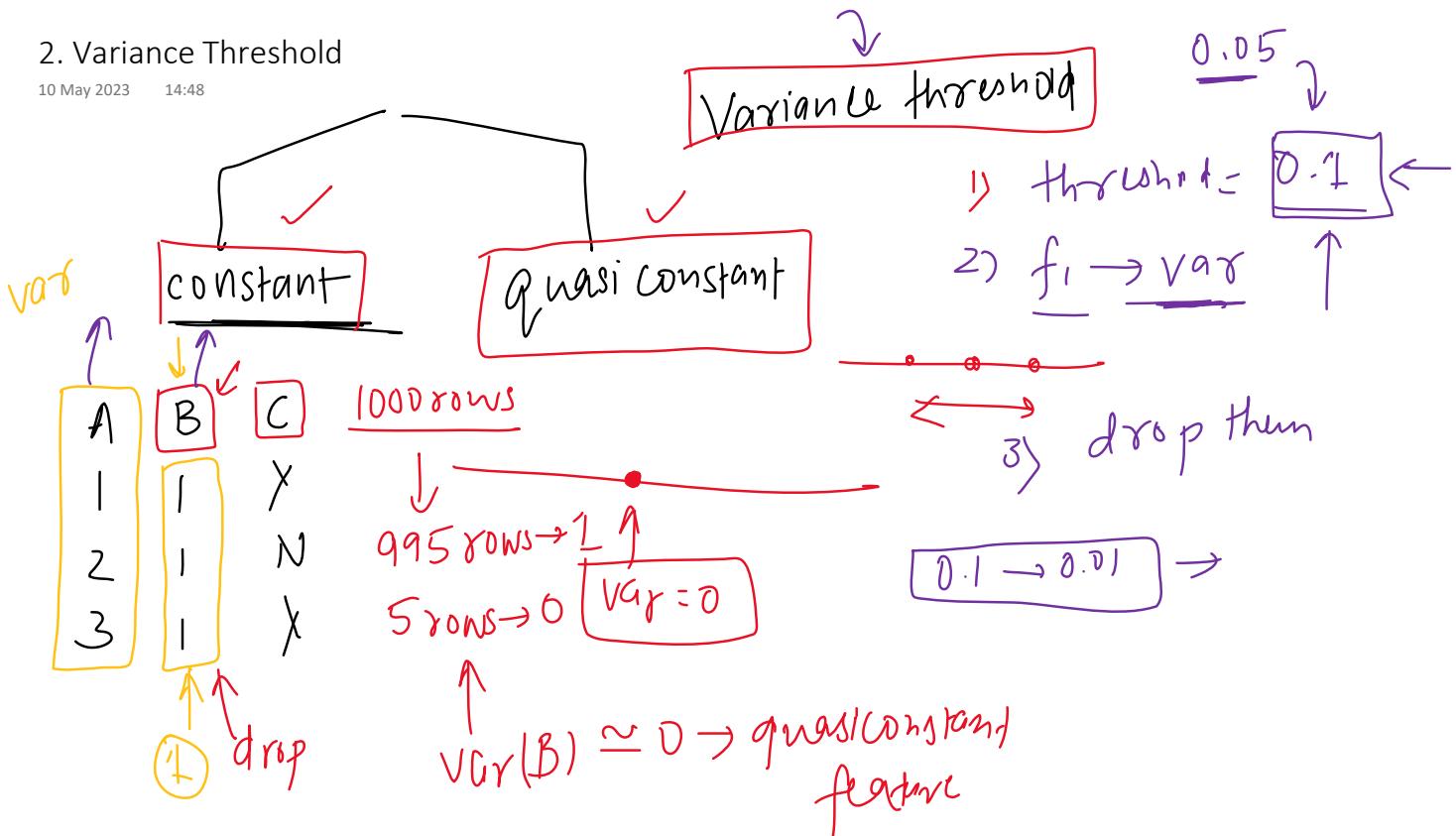
561 → 540

21 duplicate

dict → keep keys → dict → values
key → original cols
values → duplicates

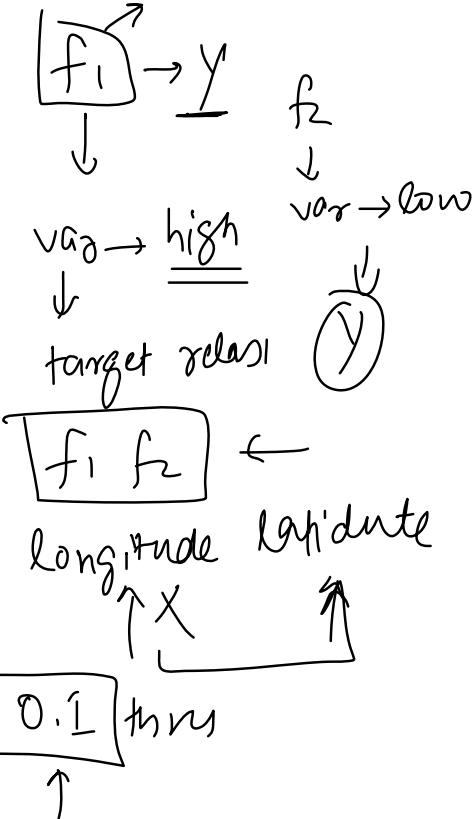
2. Variance Threshold

10 May 2023 14:48



Points to Consider

- 1. Ignores Target Variable: Variance Threshold is a univariate method, meaning it evaluates each feature independently and doesn't consider the relationship between each feature and the target variable. This means it may keep irrelevant features that have a high variance but no relationship with the target, or discard potentially useful features that have a low variance but a strong relationship with the target.
- 2. Ignores Feature Interactions: Variance Threshold doesn't account for interactions between features. A feature with a low variance may become very informative when combined with another feature.
- 3. Sensitive to Data Scaling: Variance Threshold is sensitive to the scale of the data. If features are not on the same scale, the variance will naturally be higher for features with larger values. Therefore, it is important to standardize the features before applying Variance Threshold.
- 4. Arbitrary Threshold Value: It's up to the user to define what constitutes a "low" variance. The threshold is not always easy to define and the optimal value can vary between datasets.



$0.1 \quad 0.01 \quad 10 \quad 50$

$100 \quad 0.00 \quad 0.2$

Avg

0.1 thru

3. Correlation

10 May 2023 14:48

pearson corr coeff
↓

$f_1 \rightarrow y$ $\rightarrow -1 \leftrightarrow +1$
strong inverse linear relationship

$X \rightarrow y$
0.9

-0.9 0

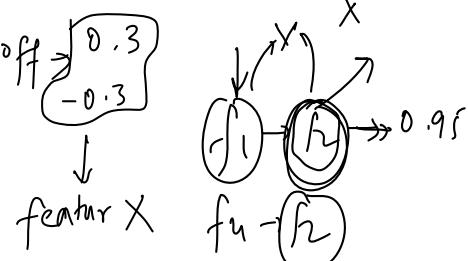
$X \rightarrow y$ $X \leftarrow y$

349 →

brute force
corr 0.95

$f_1 f_2 f_3 \dots f_n \rightarrow y$

$f_1 \rightarrow y \rightarrow$
cut off 0.3
-0.3



$f_1 \rightarrow f_2 \rightarrow f_1 - f_2 \rightarrow 0.9$
multicollinearity
groups $f_1 \rightarrow f_3 \rightarrow 0.85$
ols →

$f_1 \rightarrow y \rightarrow$
 $f_2 \rightarrow y \rightarrow$
 \vdots
 $f_n \rightarrow y \rightarrow$

561 → 541 → 341 → 152 → 100
dupl rm var corr ANOVA

Disadvantages

- Linearity Assumption:** Correlation measures the linear relationship between two variables. It does not capture non-linear relationships well. If a relationship is nonlinear, the correlation coefficient can be misleading.
- Doesn't Capture Complex Relationships:** Correlation only measures the relationship between two variables at a time. It may not capture complex relationships involving more than two variables.
- Threshold Determination:** Just like variance threshold, defining what level of correlation is considered "high" can be subjective and may vary depending on the specific problem or dataset.
- Sensitive to Outliers:** Correlation is sensitive to outliers. A few extreme values can significantly skew the correlation coefficient.

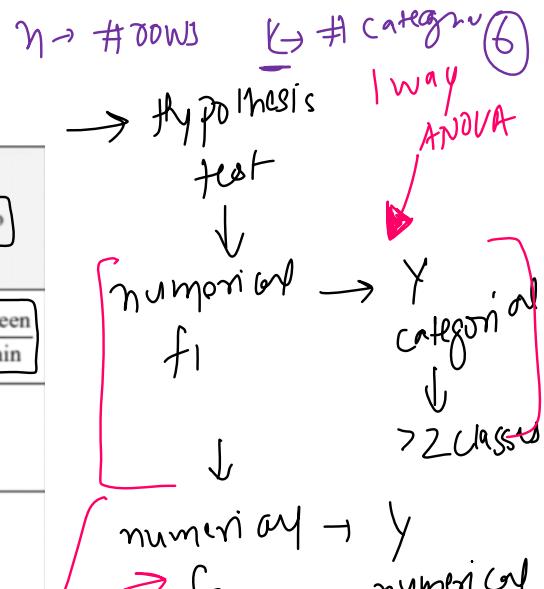
$f_1 \rightarrow f_2 \rightarrow f_1 f_2 \rightarrow y$

0.95 → 0.9 → 0.8

4. ANOVA

10 May 2023 14:49

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS) (This is SS Divided by d.f.) and is an Estimation of Variance to be Used in F-ratio	F-ratio
Between samples or categories	$n_1(\bar{X}_1 - \bar{X})^2 + \dots + n_k(\bar{X}_k - \bar{X})^2$	(k-1)	MS between SS between $\rightarrow (k-1)$	$\frac{MS \text{ between}}{MS \text{ within}}$
Within samples or categories	$\sum_{i=1, 2, 3, \dots} (X_{it} - \bar{X}_i)^2$	(n-k)	MS within SS within $\rightarrow (n-k)$	
Total	$\sum_{i, j=1, 2, 3, \dots} (X_{ij} - \bar{X})^2$	(n-1)		



f-statistic → p-value → f → y

$f_1 \ f_2 \dots f_{152}$

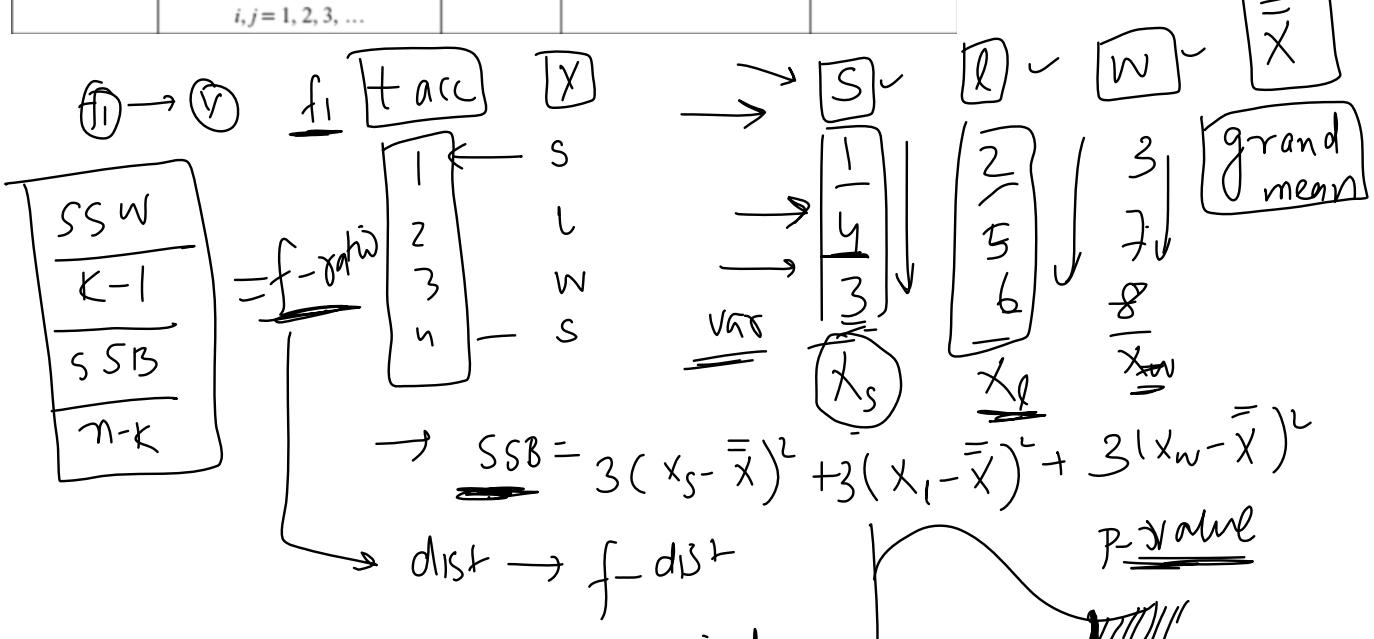
ANOVA

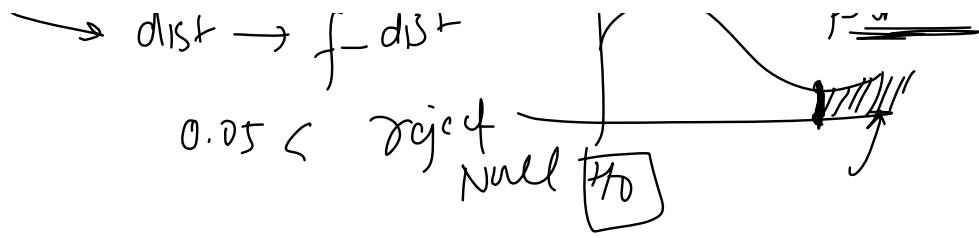
→ f → y → no rel → skimm

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS) (This is SS Divided by d.f.) and is an Estimation of Variance to be Used in F-ratio	F-ratio
Between samples or categories	$n_1(\bar{X}_1 - \bar{X})^2 + \dots + n_k(\bar{X}_k - \bar{X})^2$	(k-1)	SS between (k-1)	$\frac{MS \text{ between}}{MS \text{ within}}$
Within samples or categories	$\sum_{i=1, 2, 3, \dots} (X_{it} - \bar{X}_i)^2$	(n-k)	SS within (n-k)	
Total	$\sum_{i, j=1, 2, 3, \dots} (X_{ij} - \bar{X})^2$	(n-1)		

$$SSW = 150$$

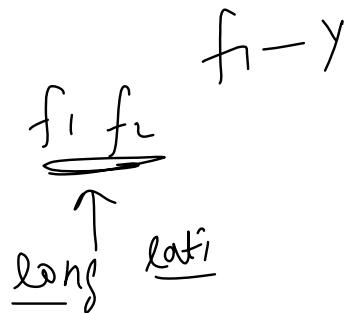
$$(1 - \bar{x}_S)^2 + (4 - \bar{x}_S)^2 + (3 - \bar{x}_S)^2 \\ + (2 - \bar{x}_L) + (5 - \bar{x}_L)^2 + (6 - \bar{x}_L)^2 \\ + (3 - \bar{x}_W)^2 + (7 - \bar{x}_W)^2 + (8 - \bar{x}_W)^2$$





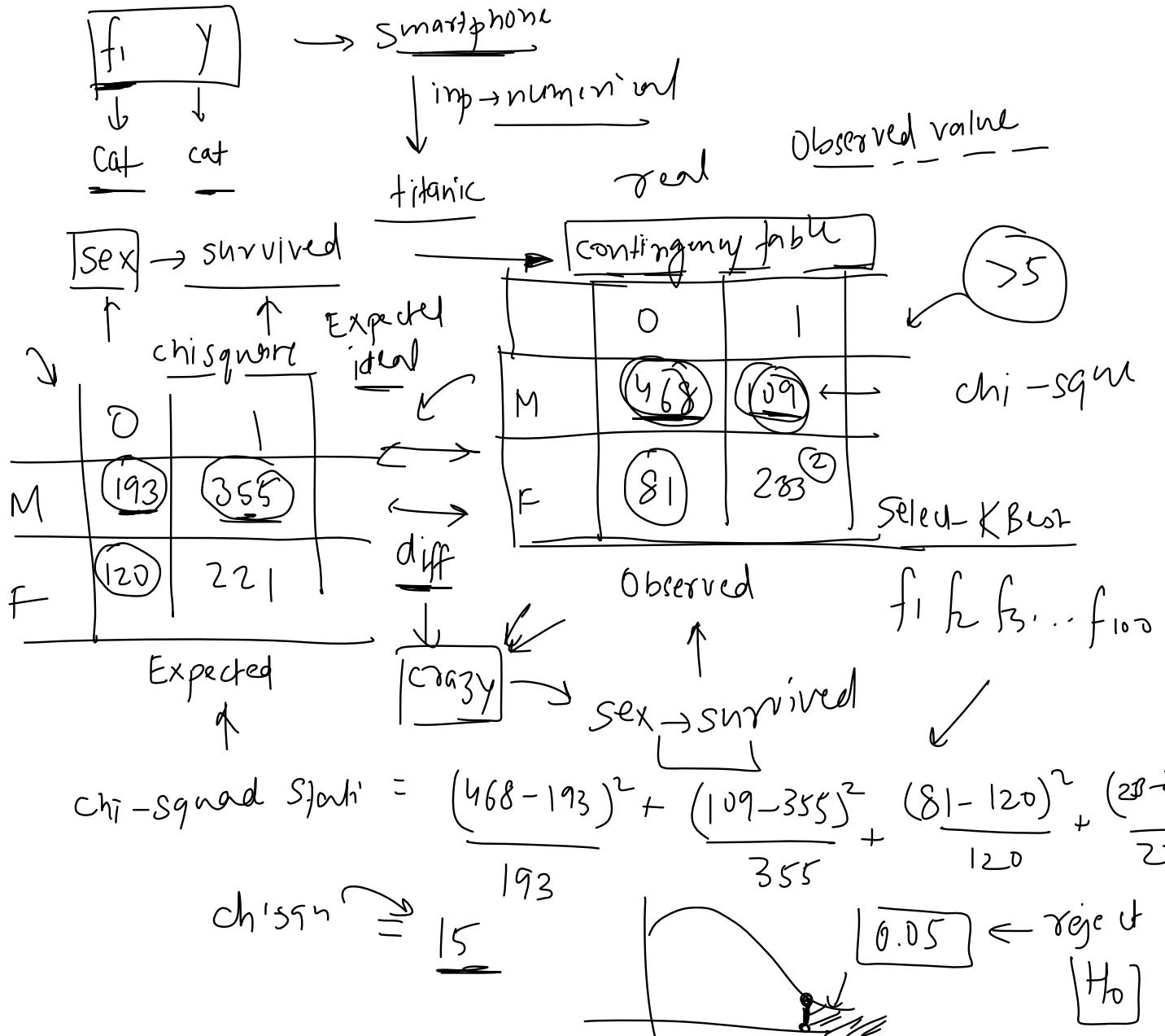
Disadvantages

1. Assumption of Normality: ANOVA assumes that the data for each group follow a normal distribution. This assumption may not hold true for all datasets, especially those with skewed distributions.
2. Assumption of Homogeneity of Variance: ANOVA assumes that the variances of the different groups are equal. This is the assumption of homogeneity of variance (also known as homoscedasticity). If this assumption is violated, it may lead to incorrect results.
3. Independence of Observations: ANOVA assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).
4. Effect of Outliers: ANOVA is sensitive to outliers. A single outlier can significantly affect the F-statistic leading to a potentially erroneous conclusion.
5. Doesn't Account for Interactions: Just like other univariate feature selection methods, ANOVA does not consider interactions between features.



5. Chi-Square

10 May 2023 14:48



Disadvantages

- Categorical Data Only:** The chi-square test can only be used with categorical variables. It is not suitable for continuous variables unless they have been discretized into categories, which can lead to loss of information.

Age

0 - 10
10 - 20
20 - 30

- Independence of Observations:** The chi-square test assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).

- Sufficient Sample Size:** Chi-square test requires a sufficiently large sample size. The results may not be reliable if the sample size is too small or if the frequency count in any category is too low (typically less than 5).

- No Variable Interactions:** Chi-square test, like other univariate feature selection methods, does not consider interactions between features. It might miss out on identifying important features that are significant in combination with other features.

$f_1 f_2$

$f_1 \rightarrow y$

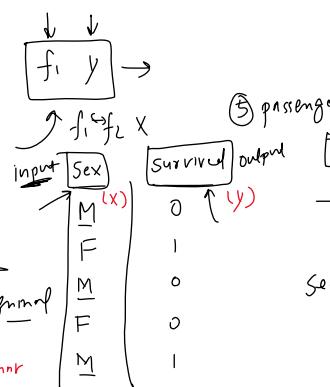
6. Mutual Information

15 May 2023 14:50

Mutual Information (MI) is a measure of the dependency between two variables. It quantifies the amount of information obtained about one random variable through observing the other random variable. It is a fundamental quantity in information theory.

$$MI = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

where
 $p(x,y) \rightarrow$ Joint prob of X and Y
 $p(x) \rightarrow$ marginal prob of X
 $p(y) \rightarrow$ marginal prob of Y



1. Joint Probability: This is the probability of two (or more) simultaneous events. For example, if we have two random variables, X and Y, the joint probability of X and Y is denoted as $P(X, Y)$, and it represents the probability that X takes on a specific value and Y takes on a specific value at the same time. In other words, it represents the probability of both events happening at the same time.

2. Marginal Probability: This is the probability of an event occurring regardless of the outcome of another event. If we have two random variables, X and Y, the marginal probability of X is simply denoted as $P(X)$, and it represents the probability that X takes on a specific value irrespective of the values of Y. The term "marginal" refers to the process of summing or integrating over the distribution of the other variable(s) to obtain the distribution of the variable of interest.

$$MI = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Mutual Information has several properties that make it useful for feature selection:

1. **It is non-negative:** MI is always zero or positive, with zero indicating that the variables are independent (i.e., no information about one variable can be obtained by observing the other variable).
2. **It is symmetric:** $MI(X, Y) = MI(Y, X)$. The mutual information from X to Y is the same as from Y to X.
3. **It can capture any kind of statistical dependency:** Unlike correlation, which only captures linear relationships, mutual information can capture any kind of relationship, including nonlinear ones.

How to deal with numerical variables

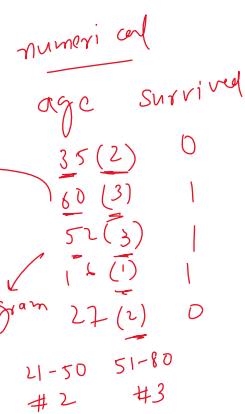
Disadvantages

1. **Estimation Difficulty:** Estimating MI from data can be challenging, especially when the dimensionality of the data is high or the number of samples is low. This is because MI estimation often relies on techniques like binning or density estimation, which can be sensitive to the chosen parameters or assumptions.
2. **Assumes Large Sample Sizes:** MI works best with large sample sizes. With smaller sample sizes, the estimates of MI can be noisy and less reliable, which might lead to incorrect conclusions about the dependencies between variables.
3. **Computationally Intensive:** Calculating MI for many features can be computationally expensive, especially for continuous variables. This might be problematic for large datasets or for applications where computational resources or time are limited.
4. **Difficulty with Continuous Variables:** While MI theoretically applies to continuous variables, in practice it's often difficult to estimate MI between continuous variables due to the need for accurate density estimation, which is a challenging problem in its own right.
5. **No Direct Indication of the Nature of Relationships:** Although MI can identify the existence of a relationship between variables, it does not provide direct information about the nature of this relationship (e.g., linear, quadratic, etc.). This contrasts with methods such as correlation, which directly indicate the strength and direction of a linear relationship.
6. **Doesn't Account for Redundancy:** Mutual Information measures the relevance of individual features to the target variable, but it doesn't take into account the redundancy among features. Two features might individually have high MI with the target, but if they are highly correlated, they might not provide much unique information. This can lead to the selection of redundant features.

$$M \rightarrow 0 \quad \frac{2}{5} \quad P(X=m, Y=0)$$

$$P(X=m \text{ AND } Y=1) \quad \downarrow \quad \text{sex}$$

		Survived	
		0	1
Sex	0	2/5	3/5
	1	1/5	2/5
Age	3/5	2/5	
	1/5	1/5	



Sex \rightarrow Survived

survived

linear \rightarrow chi square

age

survived

sex surv

↓ ↓

15 15 \rightarrow 125

-	-
-	-

→ binning

MI ↑ \rightarrow r^2

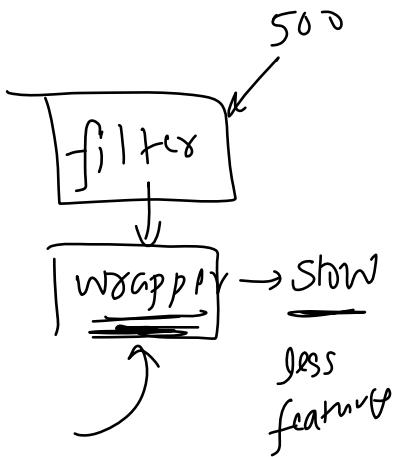
$f_1 f_2 \rightarrow$ multi

Advantages and Disadvantages

11 May 2023 16:07

Advantages

1. Simplicity: Filter methods are generally straightforward and easy to understand. They involve calculating a statistic that measures the relevance of each feature, and selecting the top features based on this statistic.
2. Speed: These methods are usually computationally efficient. Because they evaluate each feature independently, they can be much faster than wrapper methods or embedded methods, which need to train a model to evaluate feature importance.
3. Scalability: Filter methods can handle a large number of features effectively because they don't involve any learning methods. This makes them suitable for high-dimensional datasets.
4. Pre-processing Step: They can serve as a pre-processing step for other feature selection methods. For instance, you could use a filter method to remove irrelevant features before applying a more computationally expensive method, such as a wrapper method.



Disadvantages

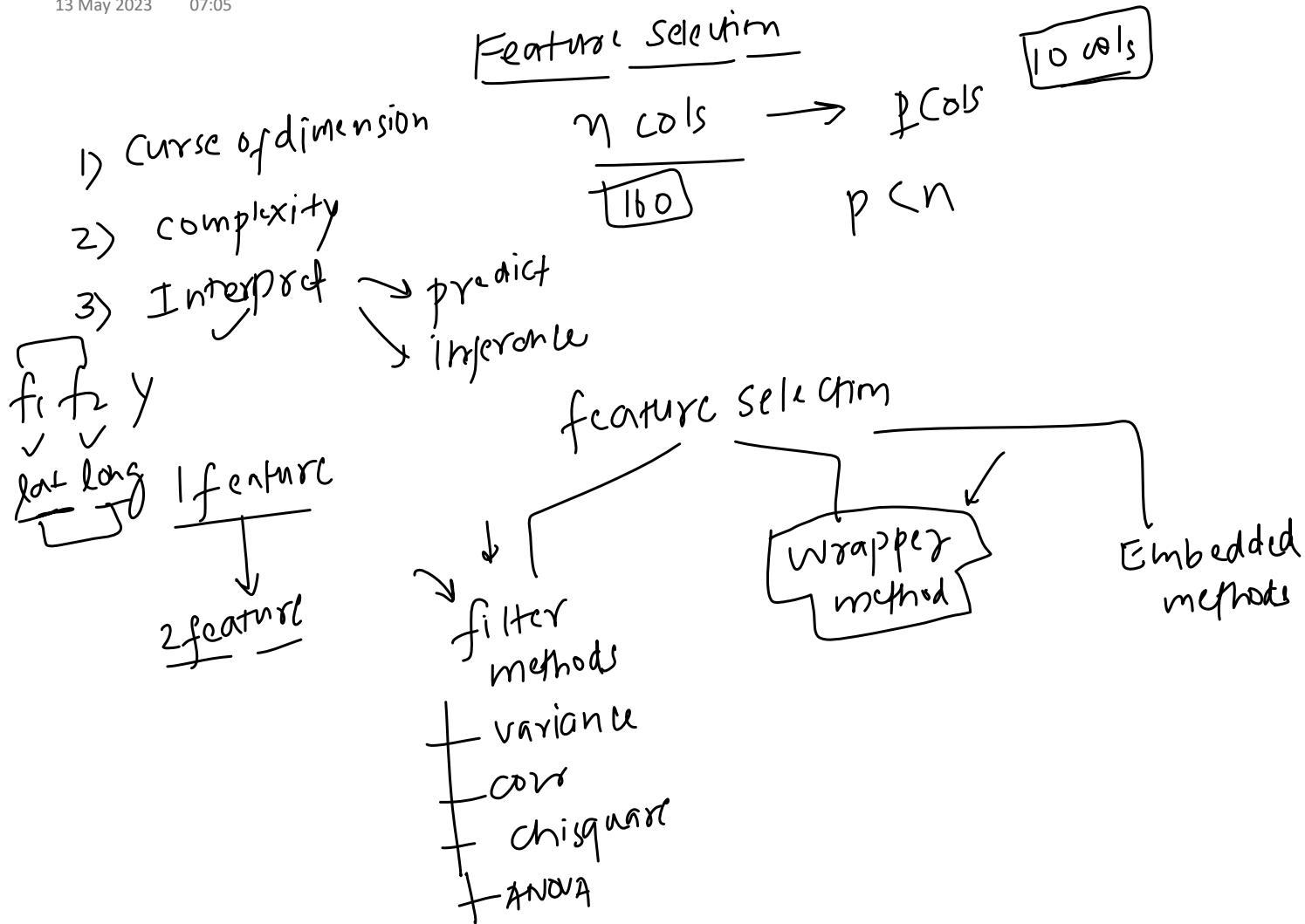
Fi-h

1. Lack of Feature Interaction: Filter methods treat each feature individually and hence do not consider the interactions between features. They might miss out on identifying important features that don't appear significant individually but are significant in combination with other features.
2. Model Agnostic: Filter methods are agnostic to the machine learning model that will be used for the prediction. This means that the selected features might not necessarily contribute to the accuracy of the specific model you want to use.
3. Statistical Measures Limitation: The statistical measures used in these methods have their own limitations. For example, correlation is a measure of linear relationship and might not capture non-linear relationships effectively. Similarly, variance-based methods might keep features with high variance but low predictive power.
4. Threshold Determination: For some methods, determining the threshold to select features can be a bit subjective. For example, what constitutes "low" variance or "high" correlation might differ depending on the context or the specific dataset.

Var
ANOVA
CHISQ

Recap

13 May 2023 07:05



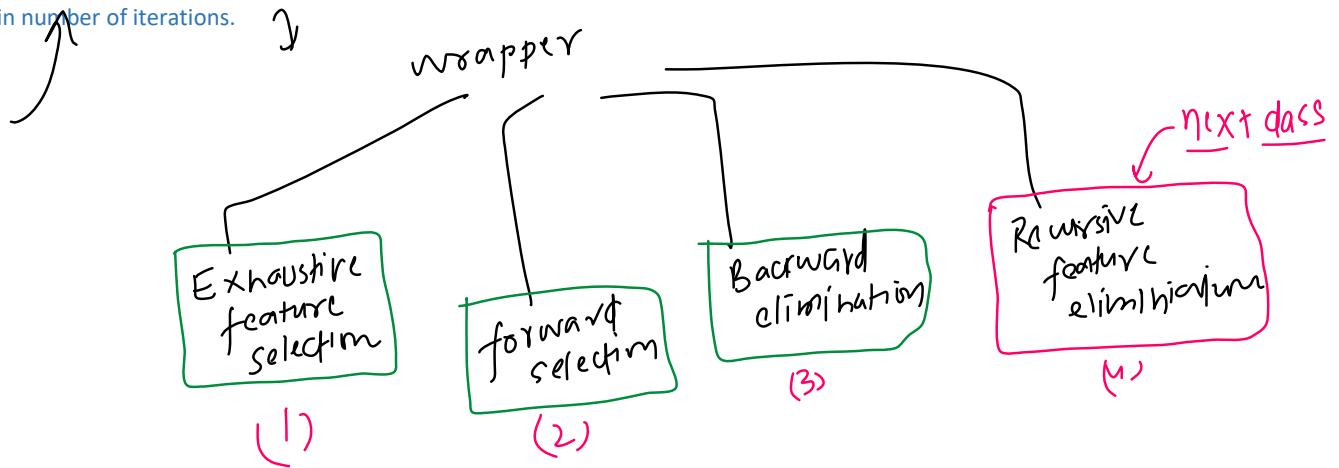
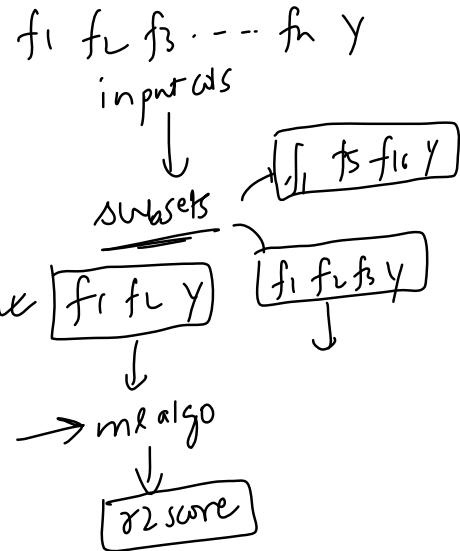
Wrapper Methods

13 May 2023 07:05

Wrapper methods for feature selection are a type of feature selection methods that involve using a predictive model to score the combination of features. They are called "wrapper" methods because they "wrap" this type of model-based evaluation around the feature selection process.

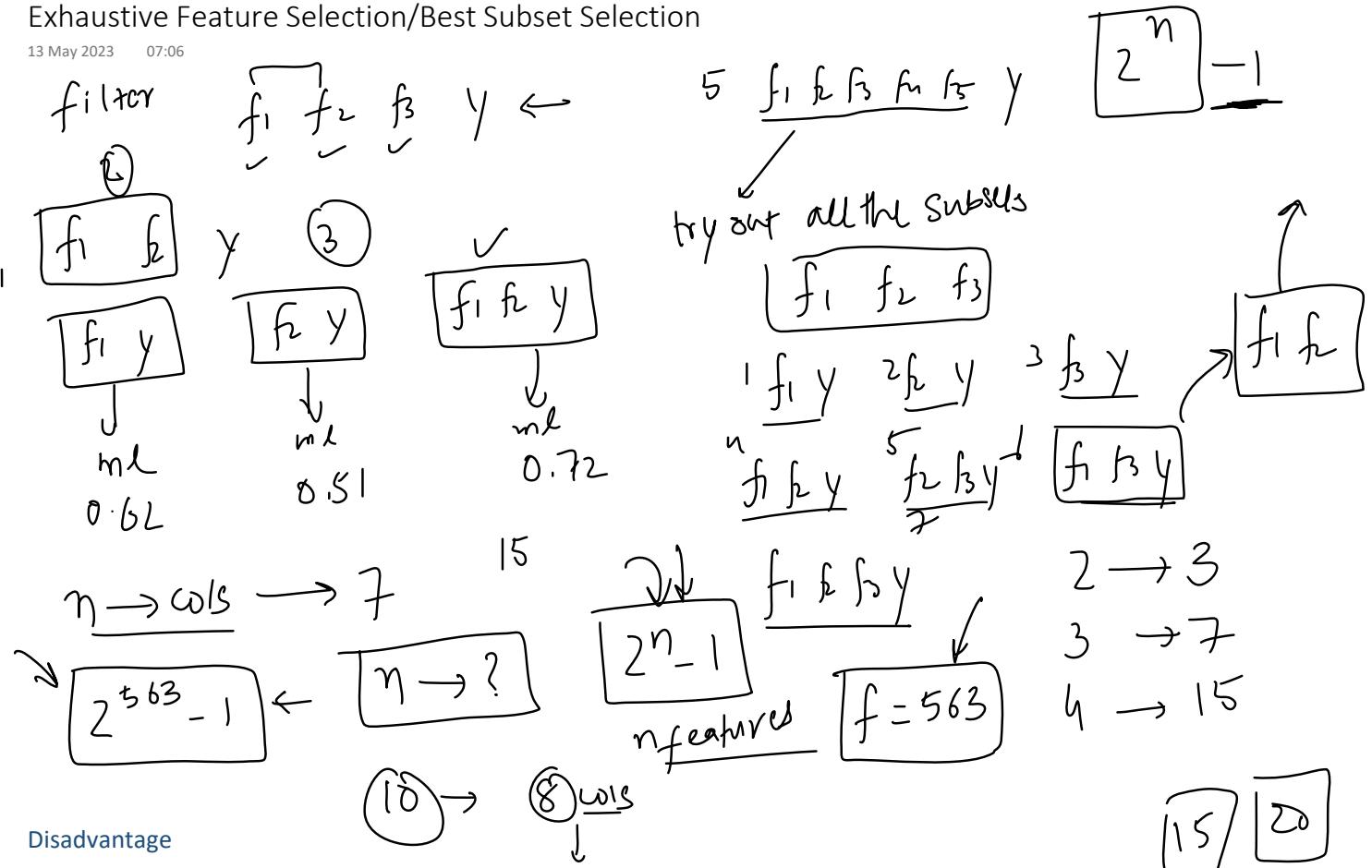
Here's how wrapper methods work in general:

1. **Subset Generation:** First, a subset of features is generated. This can be done in a variety of ways. For example, you might start with one feature and gradually add more, or start with all features and gradually remove them, or generate subsets of features randomly. The subset generation method depends on the specific type of wrapper method being used.
2. **Subset Evaluation:** After a subset of features has been generated, a model is trained on this subset of features, and the model's performance is evaluated, usually through cross-validation. The performance of the model gives an estimate of the quality of the features in the subset.
3. **Stopping Criterion:** This process is repeated, generating and evaluating different subsets of features, until some stopping criterion is met. This could be a certain number of subsets evaluated, a certain amount of time elapsed, or no improvement in model performance after a certain number of iterations.



Exhaustive Feature Selection/Best Subset Selection

13 May 2023 07:06



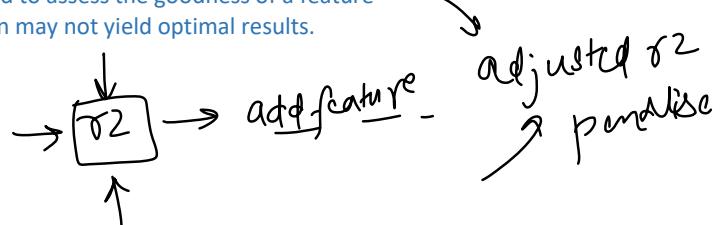
1. [Computational Complexity]: The biggest drawback is its computational cost. If you have n features, the number of combinations to check is 2^n . So, as the number of features grows, the number of combinations grows exponentially, making this method computationally expensive and time-consuming. For datasets with a large number of features, it may not be practical.

→ 10 cols

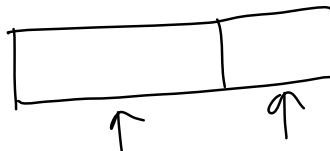
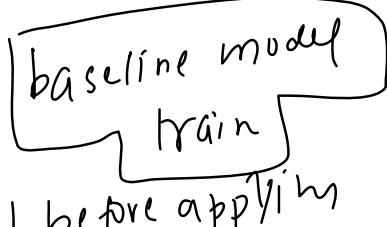
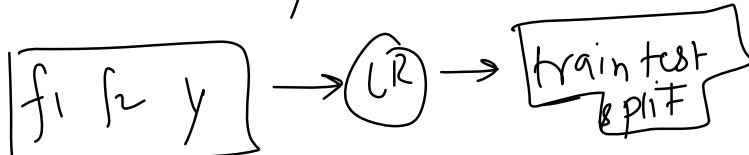
→ overfitting

2. Risk of Overfitting: By checking all possible combinations of features, there's a risk of overfitting the model to the training data. The feature combination that performs best on the training data may not necessarily perform well on unseen data.

3. Requires a Good Evaluation Metric: The effectiveness of exhaustive feature selection depends on the quality of the evaluation metric used to assess the goodness of a feature subset. If a poor metric is used, the feature selection may not yield optimal results.



$f_1 f_2 f_3 f_4 y$



r^2 score with high cols
↓
adjusted r^2

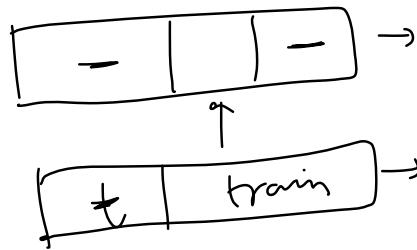


avg score
(13) cols

before applying any filter

LR

$$\gamma_2 \rightarrow 0.65$$



(13) wls

To wls

best
subset

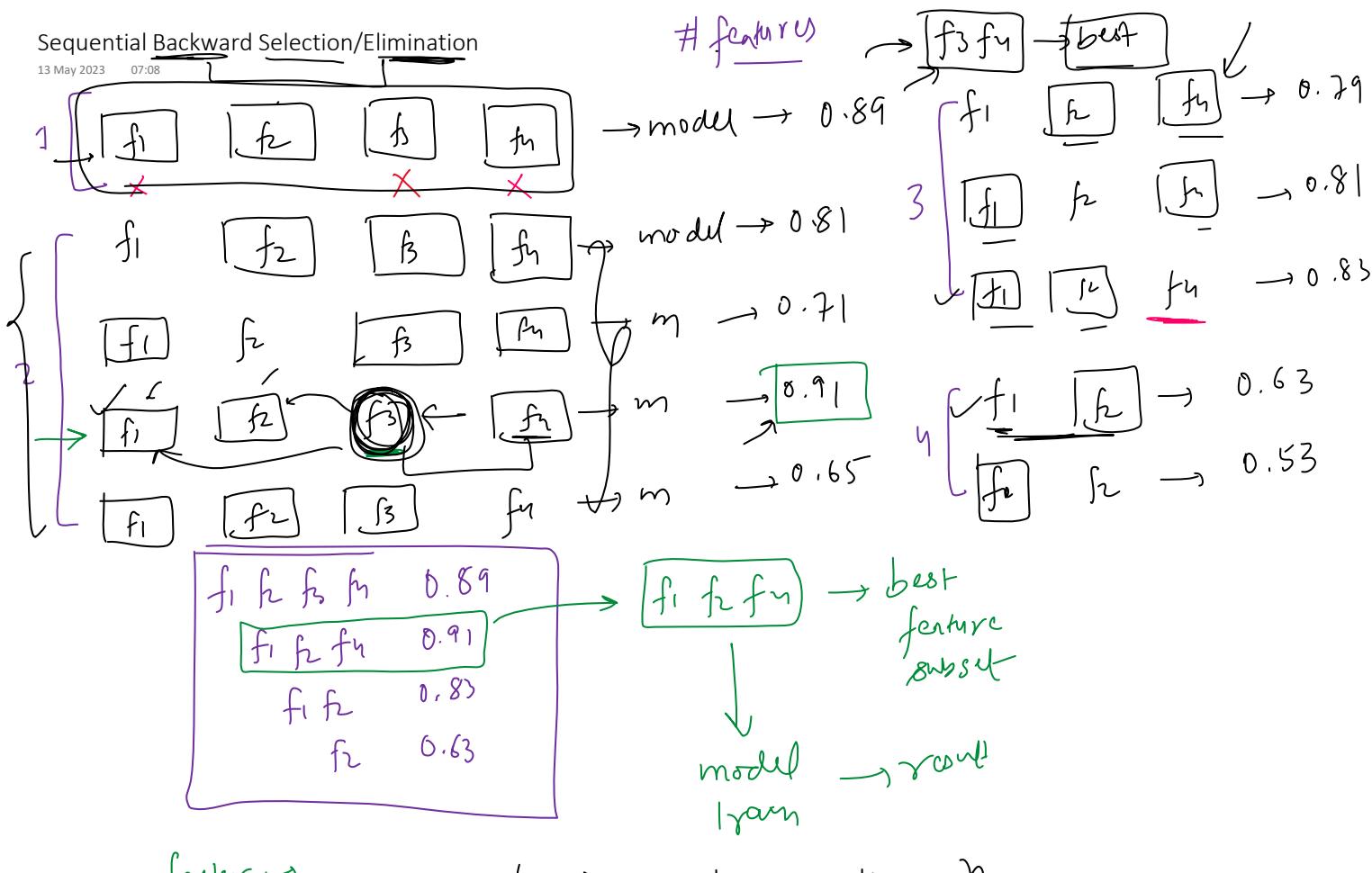
EFS $\rightarrow f_1 f_2 f_3$

$\checkmark [f_1 f_2] \rightarrow 97\%$

$f_1 f_2 f_3 \rightarrow 97$

Sequential Backward Selection/Elimination

13 May 2023 07:08



faster →

$$\frac{n(n+1)}{2} \rightarrow 1 + 2 + 3 + \dots + n$$

Ex → 2^{n-1}

$$\frac{13 \times 14}{2} = \frac{13 \times 7}{81}$$

$$\begin{cases} n \rightarrow n \\ 3 \rightarrow 3 \\ 2 \rightarrow 2 \\ 1 \rightarrow 1 \end{cases} \quad \underbrace{n \rightarrow n}_{\substack{n-1 \\ n-2 \\ \vdots \\ 1}} \quad \text{(13) cols}$$

$$1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}$$

swsy for itera

$$\begin{array}{c} [100] \text{ cols} \\ \downarrow \\ Z^{100} \rightarrow \end{array} \quad \begin{array}{c} n^2 \rightarrow \\ \hline [100 \times 100] = [50 \times 100] \end{array} \quad \begin{array}{c} \downarrow \\ [50 \times 50] \end{array}$$

disadvantage

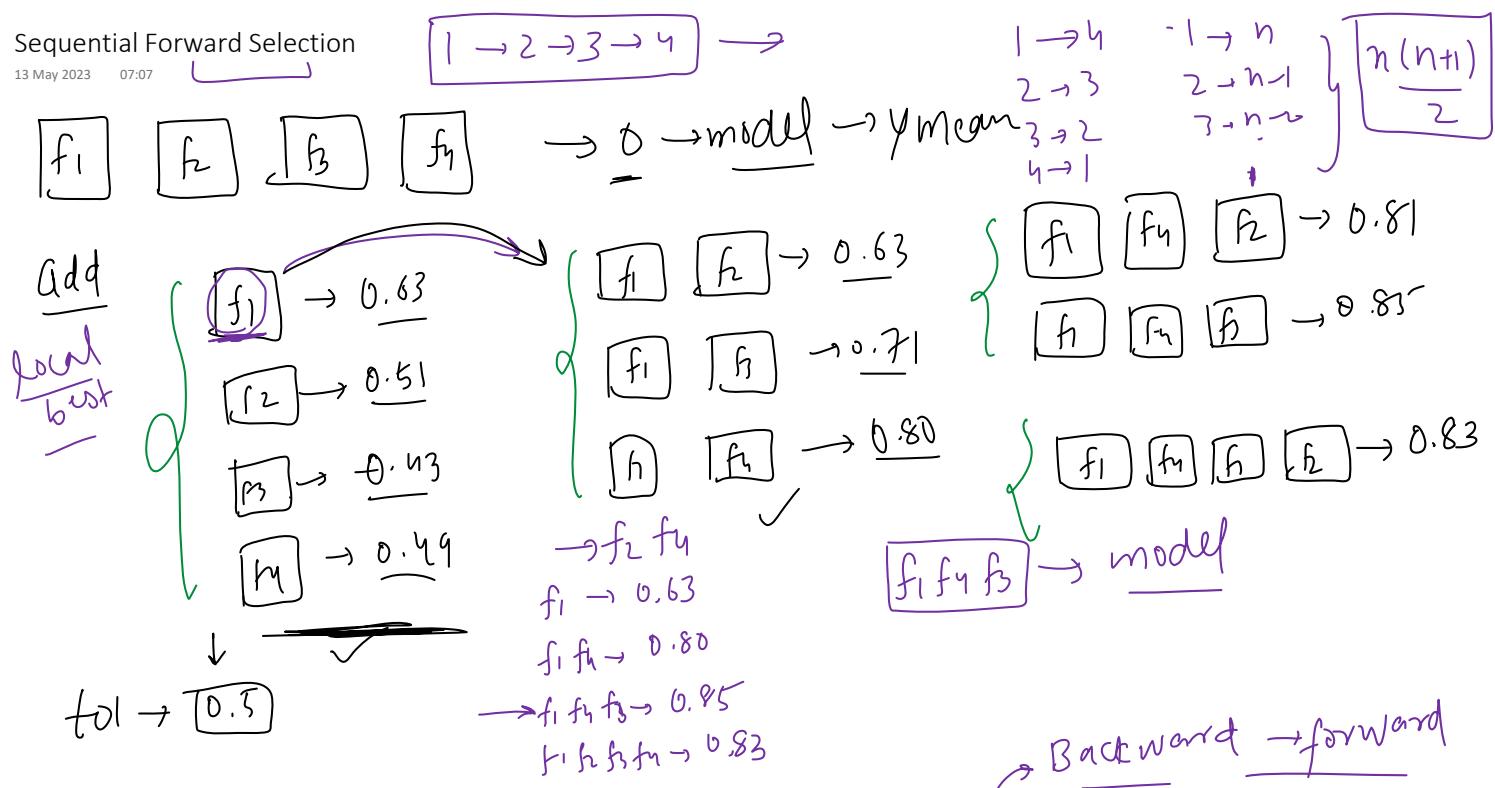
Exhaust Time

iteration → best → local selection

miss the best

Sequential Forward Selection

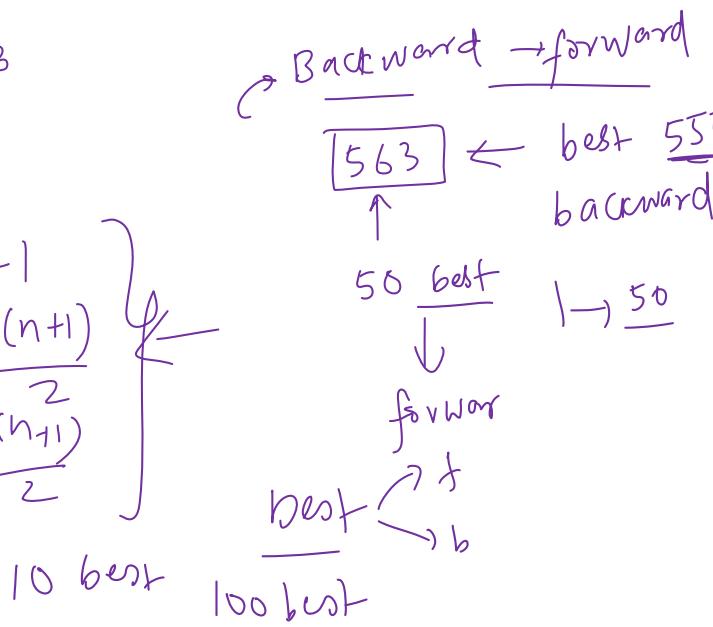
13 May 2023 07:07



$n - \text{features}$

$$\left. \begin{array}{l} \text{Exhaustive} \rightarrow 2^{n-1} \\ \text{Backward sel} \rightarrow \frac{n(n+1)}{2} \\ \text{Forward sel} \rightarrow \frac{n(n+1)}{2} \end{array} \right\}$$

100 feature
go best \rightarrow back
 $\underline{100 \text{ best}} \rightarrow \text{form}$



Advantages and Disadvantages

13 May 2023 09:06

Advantages

filter

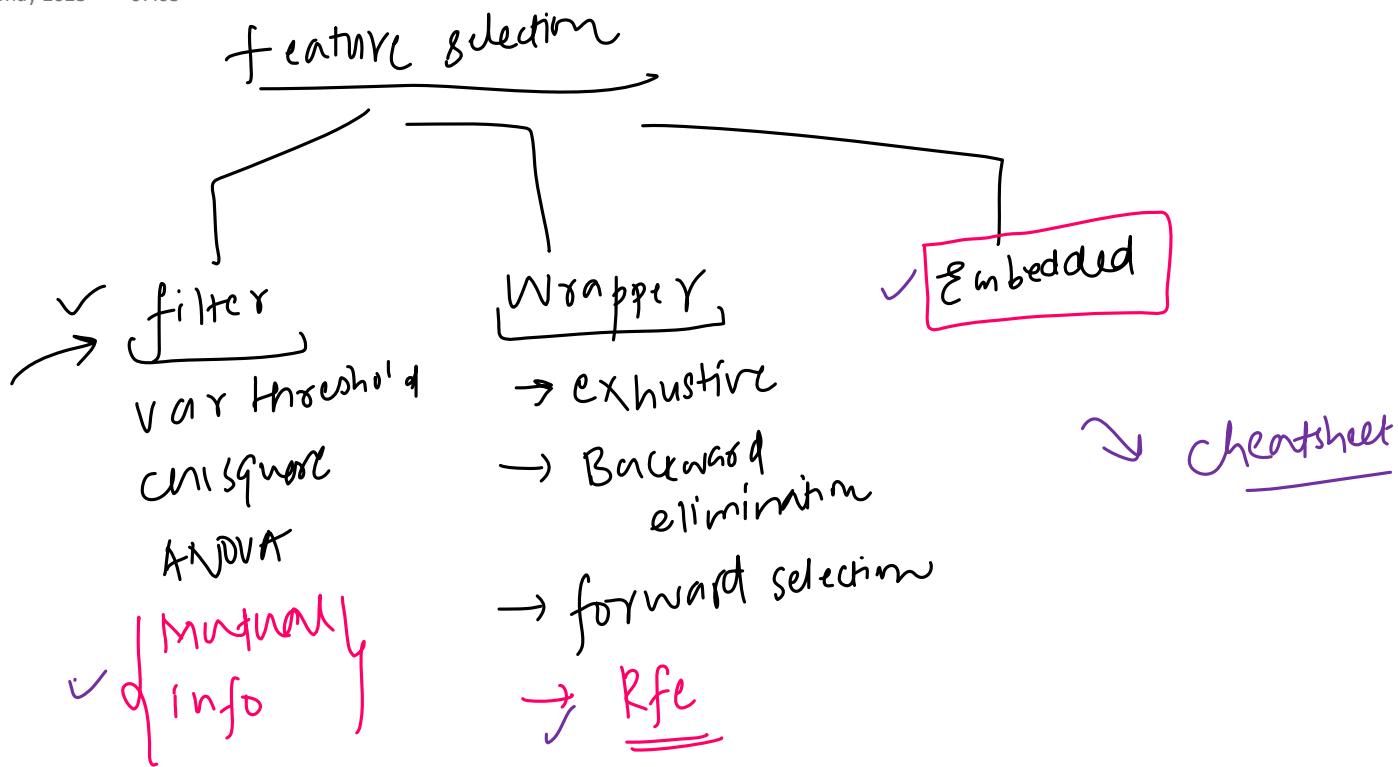
1. **Accuracy:** Wrapper methods usually provide the best performing feature subset for a given machine learning algorithm because they use the predictive power of the algorithm itself for feature selection.
2. **Interaction of Features:** They consider the interaction of features. While filter methods consider each feature independently, wrapper methods evaluate subsets of features together. This means that they can find groups of features that together improve the performance of the model, even if individually these features are not strong predictors.

Disadvantages

1. **Computational Complexity:** The main downside of wrapper methods is their computational cost. As they work by generating and evaluating many different subsets of features, they can be very time-consuming, especially for datasets with a large number of features.
2. **Risk of Overfitting:** Because wrapper methods optimize the feature subset to maximize the performance of a specific machine learning model, they might select a feature subset that performs well on the training data but not as well on unseen data, leading to overfitting.
3. **Model Specific:** The selected feature subset is tailored to maximize the performance of the specific model used in the feature selection process. Therefore, this subset might not perform as well with a different type of model.

Recap

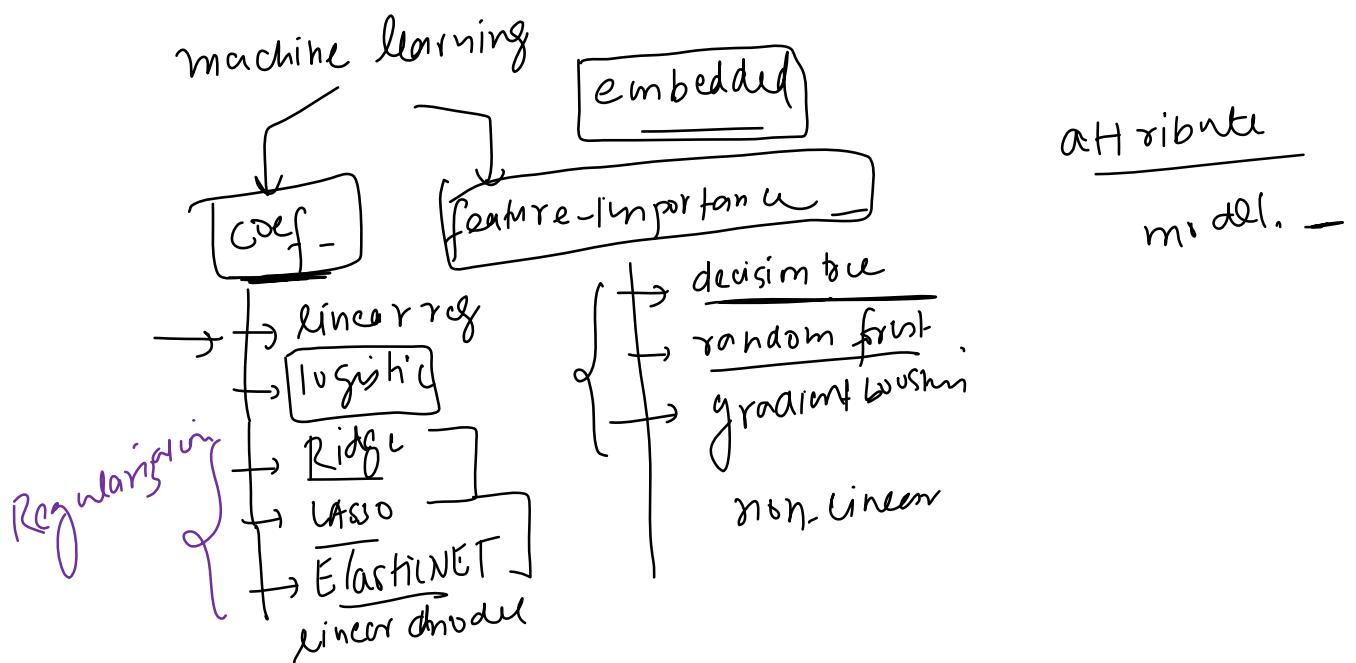
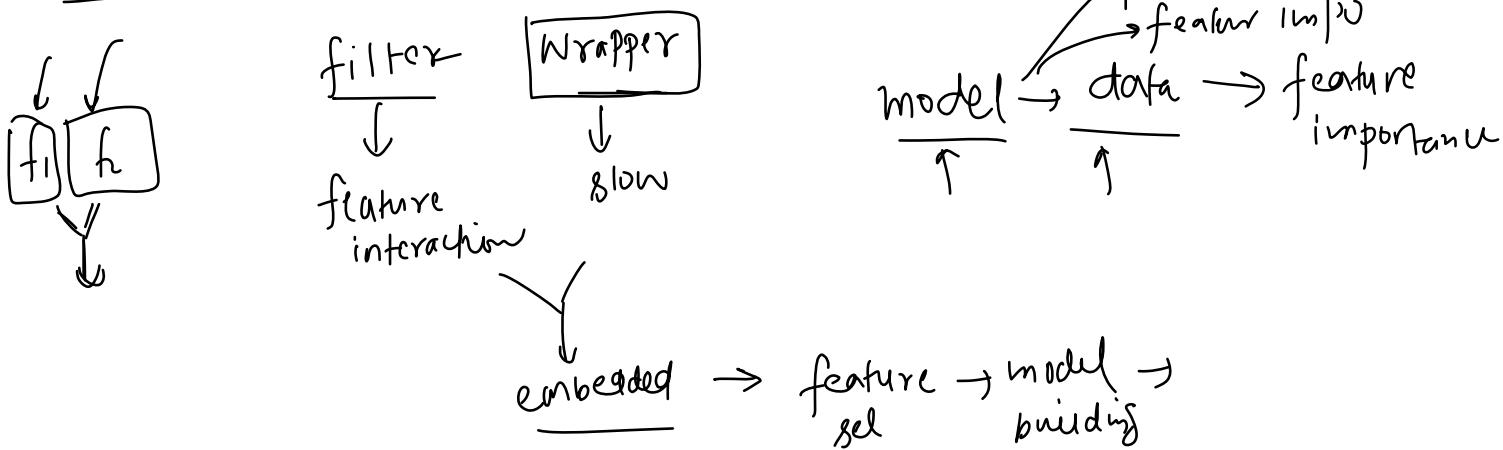
15 May 2023 07:05



Embedded Methods

15 May 2023 07:05

{ Embedded methods are feature selection techniques which perform feature selection as part of the model construction process. They are called embedded methods because feature selection is embedded within the construction of the machine learning model. These methods aim to solve the limitations of filter and wrapper methods by including the interactions of the features while also being more computationally efficient.



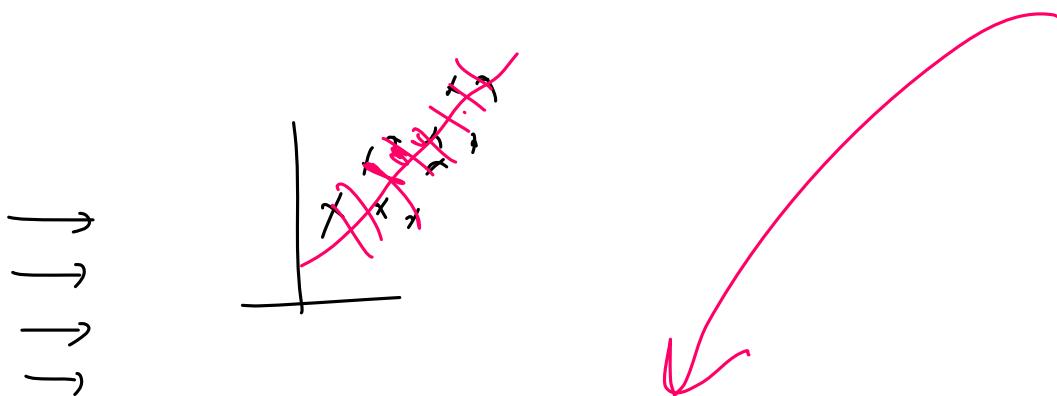
Linear Regression

15 May 2023 07:05

$$cgpa = \beta_0 + \beta_1 \text{cgpa} + \beta_2 \text{iq}$$

Diagram illustrating the linear regression model:

- The dependent variable lpa is influenced by two independent variables: $cgpa$ and iq .
- The coefficient β_1 is labeled "importance" below it.
- The coefficient β_2 is labeled "impn" below it.
- A bracket labeled "feature importance" covers the coefficients β_1 and β_2 .
- A bracket labeled "feature" covers the entire equation $lpa = \beta_0 + \beta_1 \text{cgpa} + \beta_2 \text{iq}$.

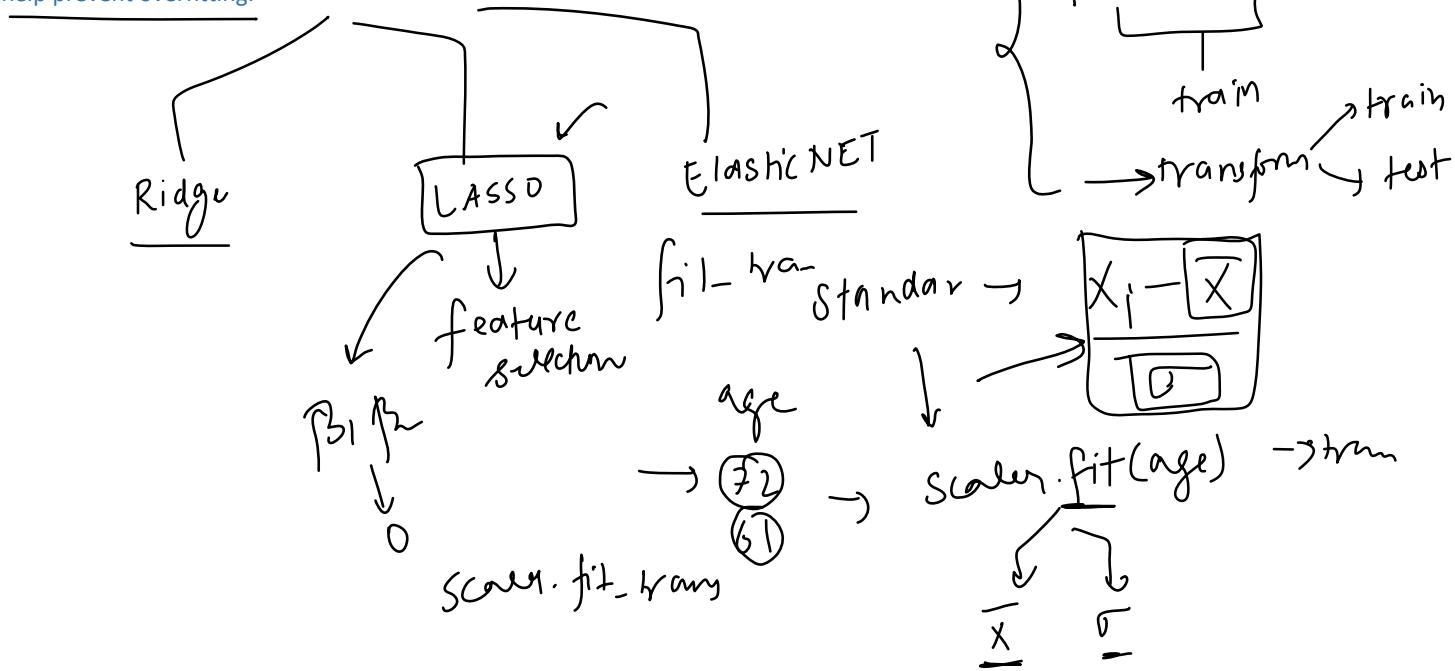


1. **Linearity:** The relationship between the independent and dependent variables is linear. This also means the change in the dependent variable for a unit change in the independent variable(s) is constant.
2. **Independence:** The observations are independent of each other. This implies that the residuals (the differences between the observed and predicted values) are independent.
3. **Homoscedasticity:** The variance of the residuals is constant across all levels of the independent variables.
4. **Normality:** The residuals are normally distributed.
5. **No Multicollinearity:** The independent variables are not highly correlated with each other. This assumption is really important when you want to interpret the regression coefficients.

Regularized Models

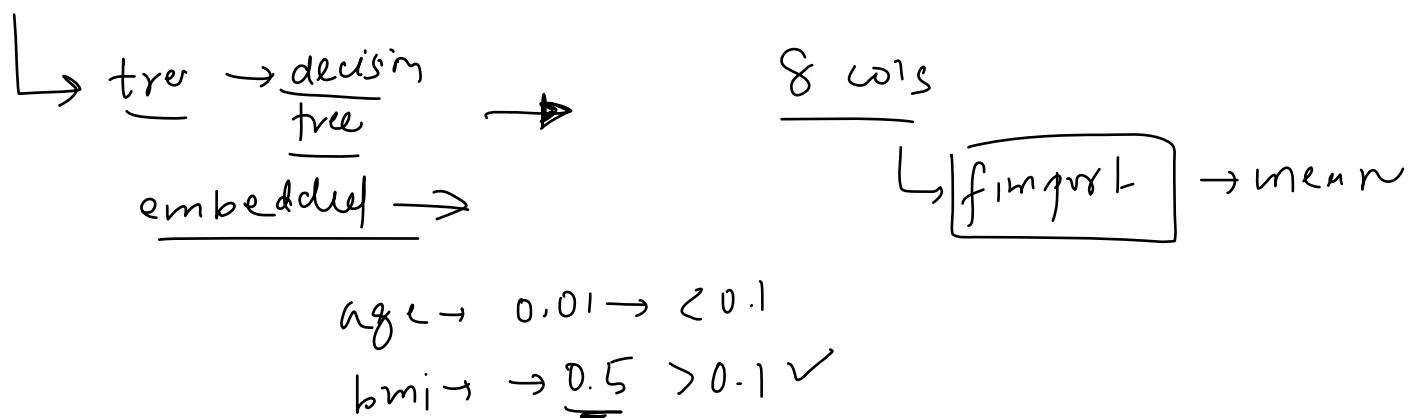
15 May 2023 07:55

Regularized linear models are linear models that include a penalty term in the loss function during training. The penalty term discourages the learning of a too complex model, which can help prevent overfitting.



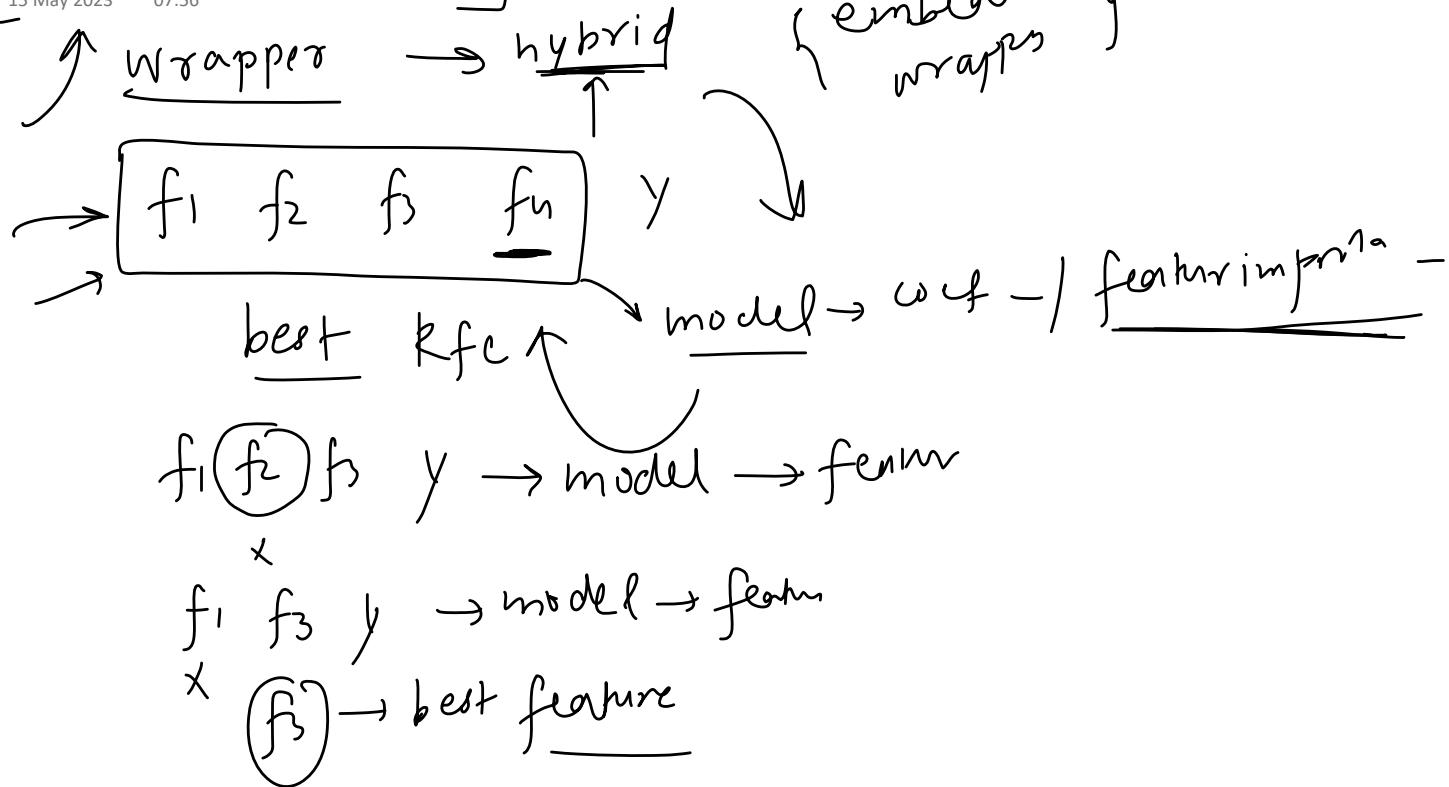
Tree Based Models

15 May 2023 07:56



Recursive Feature Elimination

15 May 2023 07:56



Advantages and Disadvantages

15 May 2023 14:45

Advantages:

1. Performance: They are generally more accurate than filter methods since they take the interactions between features into account.
2. Efficiency: They are more computationally efficient than wrapper methods since they fit the model only once.
3. Less Prone to Overfitting: They introduce some form of regularization, which helps to avoid overfitting. For example, Lasso and Ridge regression add a penalty to the loss function, shrinking some coefficients to zero.

Lasso
Ridge

Disadvantages:

1. Model Specific: Since they are tied to a specific machine learning model, the selected features are not necessarily optimal for other models.
2. Complexity: They can be more complex and harder to interpret than filter methods. For example, understanding why Lasso shrinks some coefficients to zero and not others can be non-trivial.
3. Tuning Required: They often have hyperparameters that need to be tuned, like the regularization strength in Lasso and Ridge regression. $\alpha = 0.01 \quad 0.1$
4. Stability: Depending on the model and the data, small changes in the data can result in different sets of selected features. This is especially true for models that can fit complex decision boundaries, like decision trees.

Cheatsheet

15 May 2023 17:38

1. Filter Methods:

- **Variance Threshold:** Removes all features whose variance doesn't meet a certain threshold. Use this when you have many features and you want to remove those that are constants or near constants.
- **Correlation Coefficient:** Finds the correlation between each pair of features. Highly correlated features can be removed since they contain similar information. Use this when you suspect that some features are highly correlated.
- **Chi-Square Test:** This statistical test is used to determine if there's a significant association between two variables. It's commonly used for categorical variables. Use this when you have categorical features and you want to find their dependency with the target variable.
- **Mutual Information:** Measures the dependency between two variables. It's a more general form of the correlation coefficient and can capture non-linear dependencies. Use this when you want to measure both linear and non-linear dependencies between features and the target variable.
- **ANOVA (Analysis of Variance):** ANOVA is a statistical test that stands for "Analysis of Variance". ANOVA tests the impact of one or more factors by comparing the means of different samples. Use this when you have one or more categorical independent variables and a continuous dependent variable.

2. Wrapper Methods:

- **Recursive Feature Elimination (RFE):** Recursively removes features, builds a model using the remaining attributes, and calculates model accuracy. It uses model accuracy to identify which attributes contribute the most. Use this when you want to leverage the model to identify the best features.
- **Sequential Feature Selection (SFS):** Adds or removes one feature at the time based on the classifier performance until a feature subset of the desired size k is reached. Use this when computational cost is not an issue and you want to find the optimal feature subset.
- **Exhaustive Feature Selection:** This is a brute-force evaluation of each feature subset. This method, as the name suggests, tries out all possible combinations of variables and returns the best subset. Use this when the

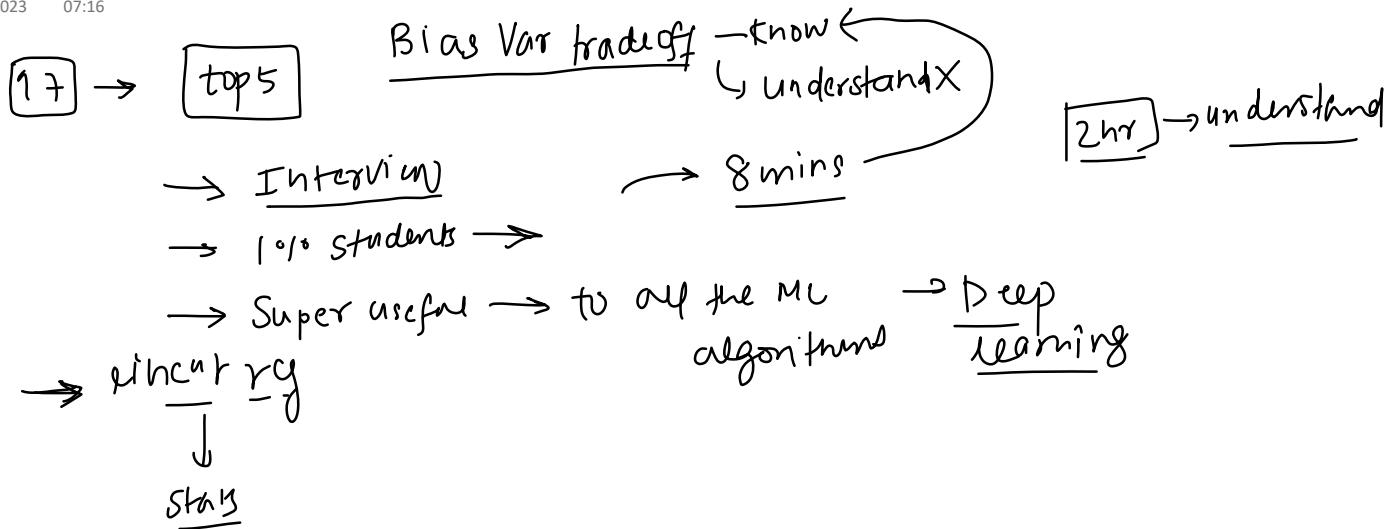
number of features is small, as it can be computationally expensive.

3. Embedded Methods:

- **Lasso Regression:** Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization. Use this when you want to create a simple and interpretable model.
- **Ridge Regression:** Ridge regression is a method used to analyze multiple regression data that suffer from multicollinearity. Unlike Lasso, it doesn't lead to feature selection but rather minimizes the complexity of the model.
- **Elastic Net:** This method is a combination of Lasso and Ridge. It incorporates penalties from both methods and is particularly useful when there are multiple correlated features.
- **Random Forest Importance:** Random forests provide a straightforward method for feature selection, namely mean decrease impurity (MDI). Use this when you want to leverage the power of random forests for feature selection.

Why this lecture is important?

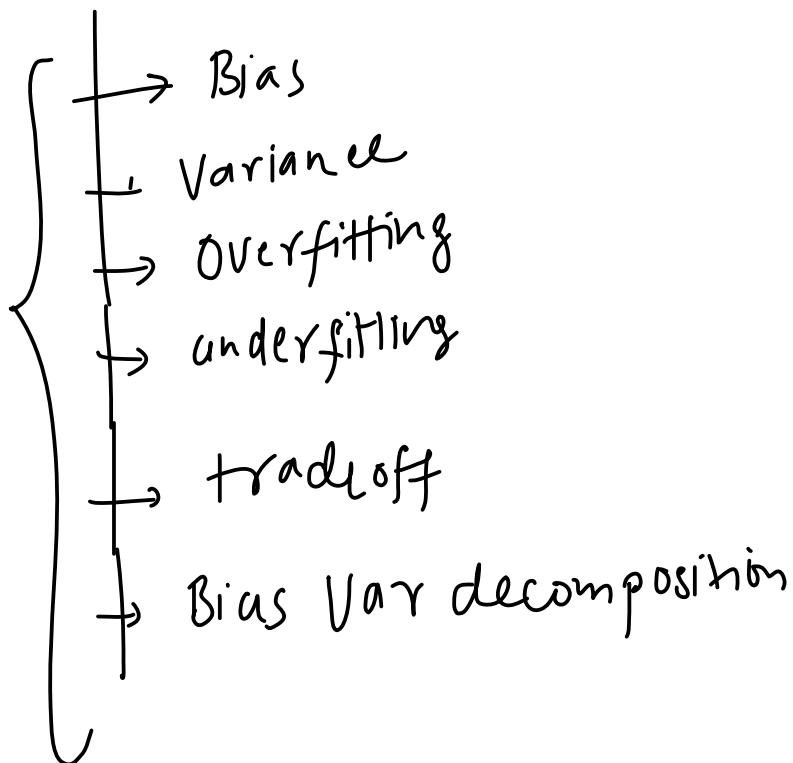
17 May 2023 07:16



What are we going to study?

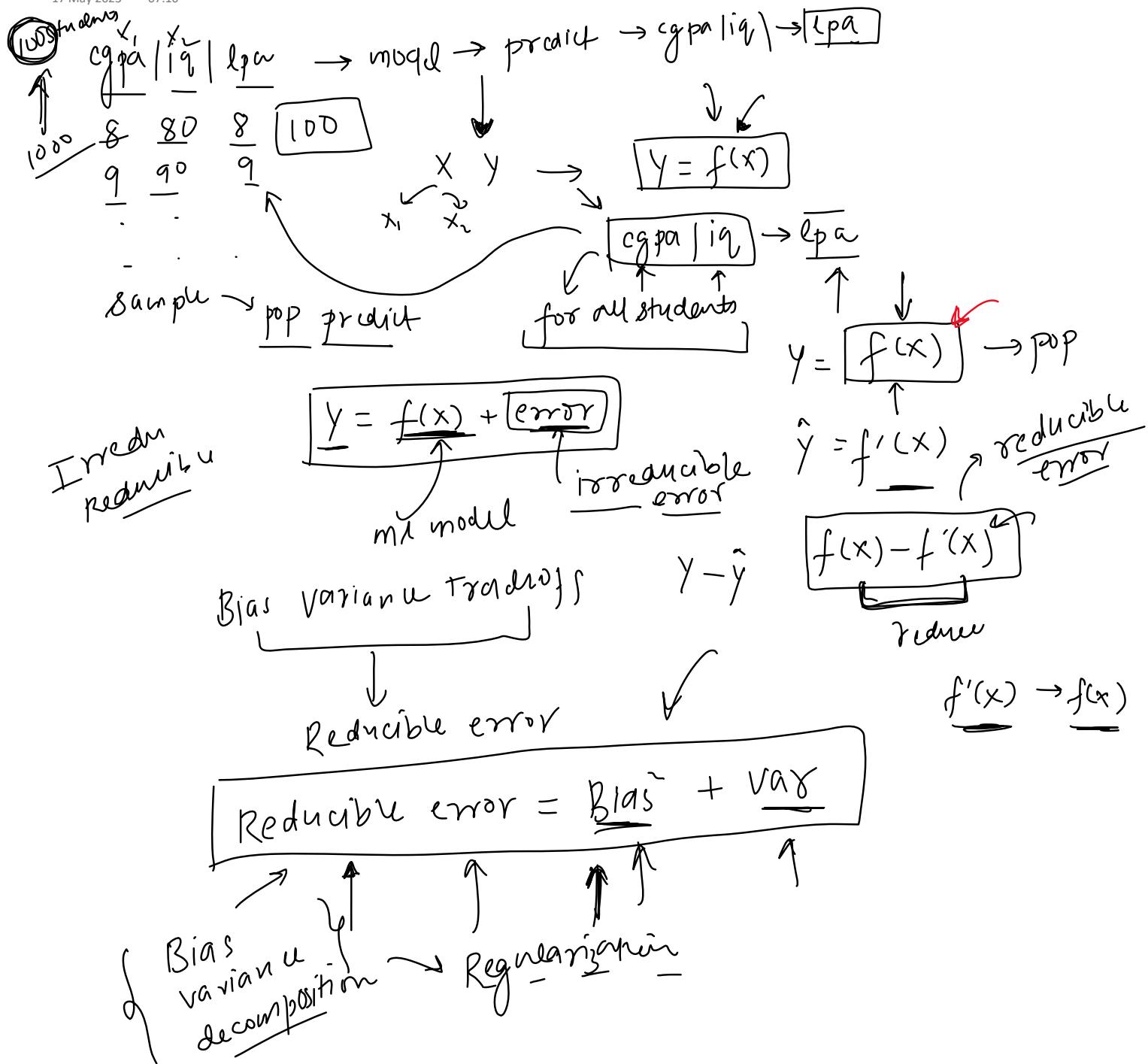
17 May 2023 17:38

Bias Var tradeoff



The Hidden Truth

17 May 2023 07:16



Bias Variance Tradeoff

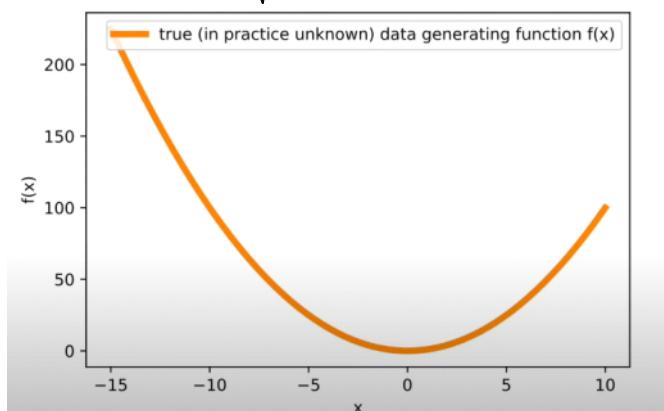
17 May 2023 07:16

$$y = f(x) = x^2 \quad [-15, 10]$$

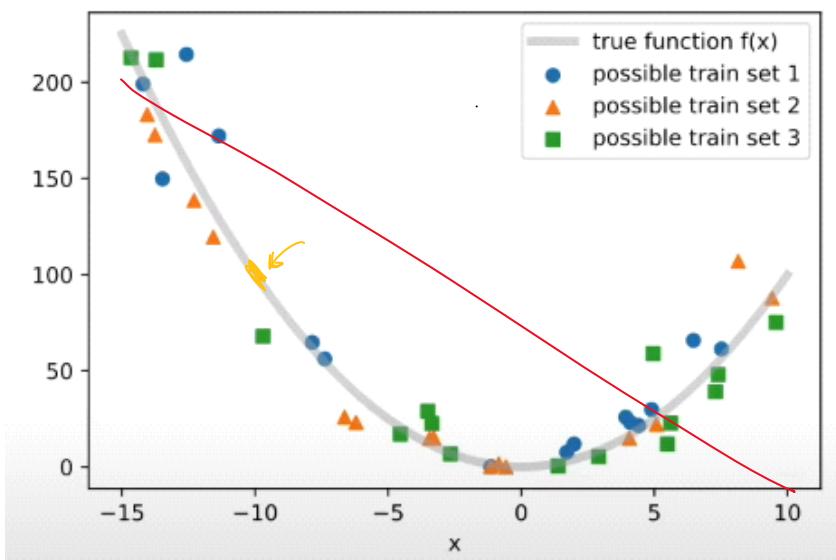
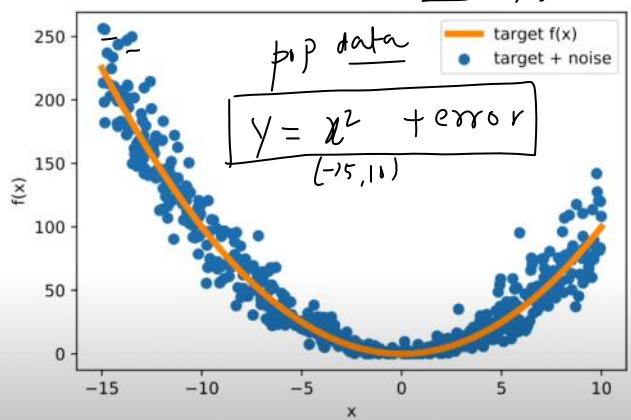
$$y = x^2$$

$$y = x^2 + \text{error}$$

1000 points
pop ↑



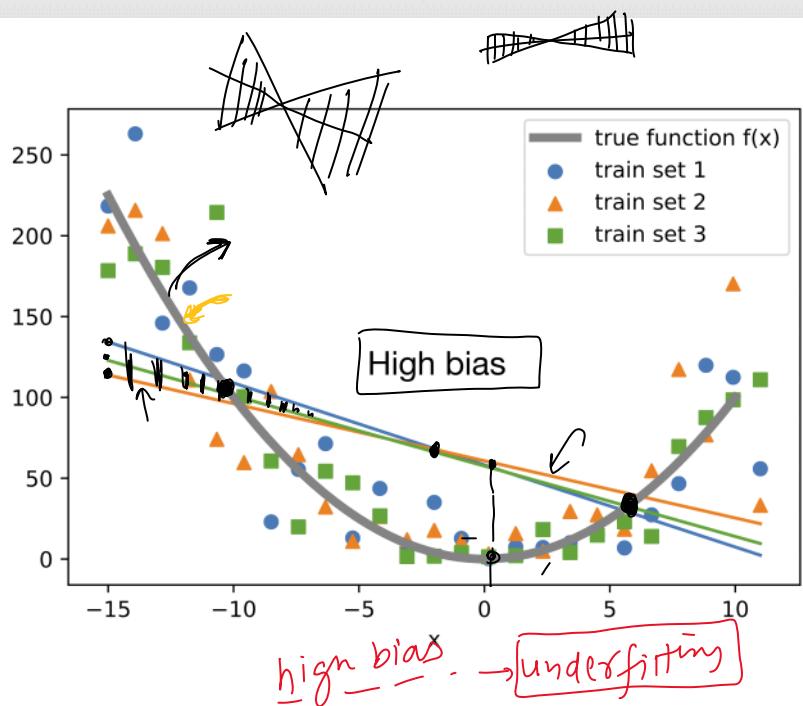
3 samples
random



- → 1
- ▲ → 2
- → 3

3 student

linear model
y



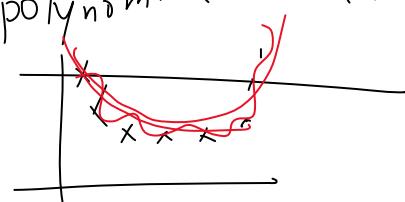
Bias → the inability of a ML model to fit the training data
Variance

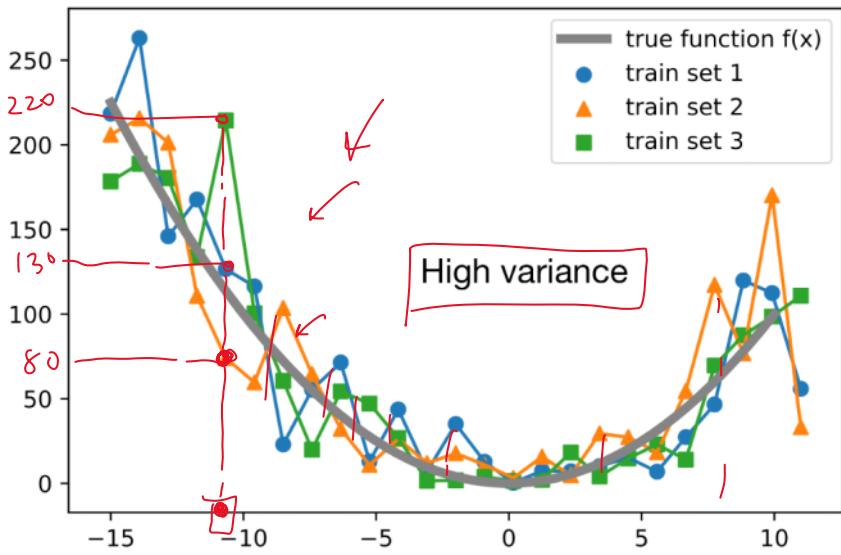
high bias ✓
low bias

Low variance

ml model predict
When the training
data has
degree = 3

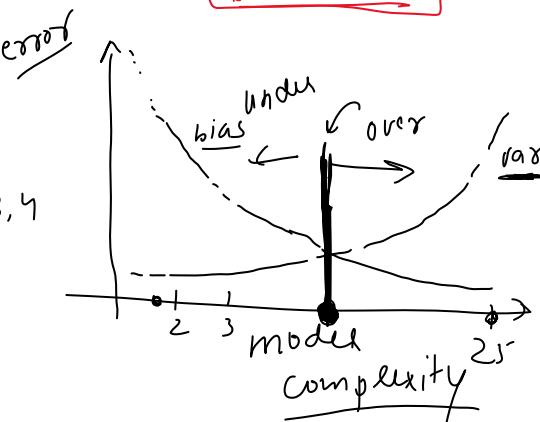
polynomial





The "trade-off" in bias-variance trade-off refers to the fact that minimizing bias will usually increase variance and vice versa.

$\{ \begin{matrix} \text{bias} \\ \text{variance} \end{matrix} \}$
 $\text{poly} \rightarrow \text{degru } 2, 3, 4$



polynomial
high degree
Overfitting

high bias
low bias ✓

high var
low bias high variance
high var overfitting

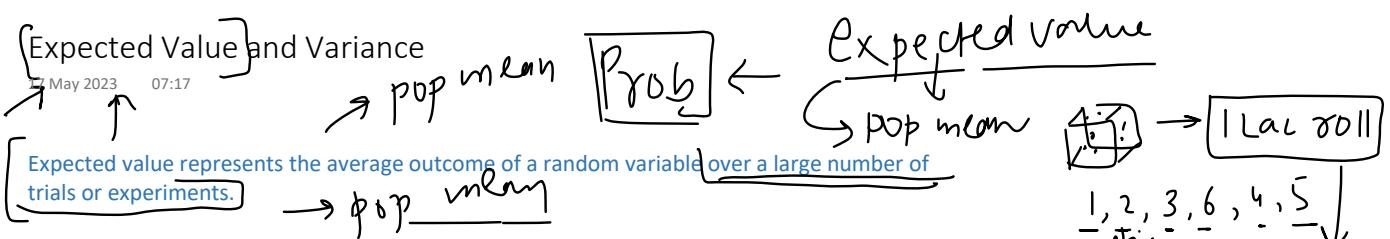
$\boxed{\begin{matrix} \text{low bias} \\ \text{low variance} \end{matrix}} \rightarrow$ bias ↓ var ↑
var ↓ bias ↑

Some questions

17 May 2023 07:16

1. How would you define bias and variance mathematically?
2. How is bias and variance related to overfitting and underfitting mathematically?
3. Why is there a tradeoff between bias and variance mathematically?

Expected Value and Variance
5 May 2023 07:17



In a simple sense, the expected value of a random variable is the long-term average value of repetitions of the experiment it represents. For example, the expected value of rolling a six-sided die is 3.5 because, over many rolls, we would expect to average about 3.5.

$X = \text{rolling a die}$

$$\underline{E[X]} = \underline{x_1 p(x_1)} + \underline{x_2 p(x_2)} + \dots + \underline{x_n p(x_n)}$$

$$= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

$$= \frac{1+2+3+4+5+6}{6}$$

$$= \frac{21}{6} = 3.5$$

overall avg

mean

$$[4.1]$$

$$5 + 3 + 4 + 5 + 3 + 5 = \frac{25}{6}$$

$$3(5) + 2(3) + 1(4)$$

$$= \frac{3}{6}(5) + \frac{2}{6}(3) + \frac{1}{6}(4)$$

$$= x_1 p(x_1) + x_2 p(x_2) + x_3 p(x_3)$$

$$\underline{\underline{E[X]}}$$

mean

expected value

discrete random var (X)

$$E[X] = \sum_{i=1}^n x_i p(x_i)$$

→ expected value
→ pop mean

$$\underline{\underline{E[X]}}$$

continuous random var

$$\rightarrow E[X] = \int x_i f(x_i) dx$$

$\boxed{\text{Var}(X)}$ → var of pop pop var of X

$$\rightarrow \boxed{\text{Var}(X) = E[X^2] - (E[X])^2}$$

↑ sample

$$\begin{aligned}
 \text{Var}(x) &= \frac{\sum (x_i - \bar{x})^2}{n} = E[(x - E[x])^2] = \text{Var}(x) \\
 &\quad \text{avg} - \text{num} \quad \text{constant} \quad E[\frac{1}{n}] \\
 &= E[(x - E[x])^2] \\
 &= E[x^2 + (E[x])^2 - 2xE[x]] \\
 &= E[x^2] + E[(E[x])^2] - E[2xE[x]] \\
 &= E[x^2] + E[(E[x])^2] - E[2] E[x] E[E[x]] E[E[x]] \\
 &= E[x^2] + E[(E[x])^2] - 2 E[x] E[x] \\
 &= E[x^2] + \underline{E[(E[x])^2]} - 2(E[x]) \\
 &= E[x^2] + \underline{(E[x])^2} - 2(E[x])^2 \\
 &= E[x^2] - (E[x])^2
 \end{aligned}$$

$E[X]$
 $E[X+Y] = E[X] + E[Y]$
 $E[XY] = E[X]E[Y]$
 given X and Y are independent

$\text{Var}(x) = E[x^2] - (E[x])^2$

$$\begin{aligned}
 \text{Var}(x) &= E[x^2] - (E[x])^2 \\
 &= E[(x - E[x])^2] \\
 &\quad \text{mean} \quad \text{pop} \quad \text{var} \\
 &\quad E[x] \quad \text{mean} \quad E[(x - E[x])^2] \\
 &\quad \text{discrete} \quad \text{continuous} \\
 &\quad \boxed{2^2} \\
 &\quad \text{feature} \\
 &\quad [x]
 \end{aligned}$$

$\text{Var} = \frac{\sum (x_i - \bar{x})^2}{n}$

$\text{Bias} \quad \dots \text{anh.} \quad E[x]$

Bias ?
Var ?

→ mathemath. μ
 $E[x]$

What exactly are Bias and Variance Mathematically?

17 May 2023 07:17

Bias

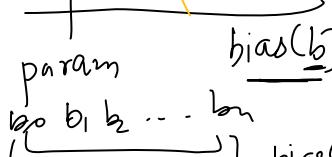
In the context of machine learning and statistics, bias refers to the systematic error that a model introduces because it cannot capture the true relationship in the data. It represents the difference between the expected prediction of our model and the correct value which we are trying to predict. More bias leads to underfitting, where the model does not fit the training data well.

Variance

In the context of machine learning and statistics, variance refers to the amount by which the prediction of our model will change if we used a different training data set. In other words, it measures how much the predictions for a given point vary between different realizations of the model.

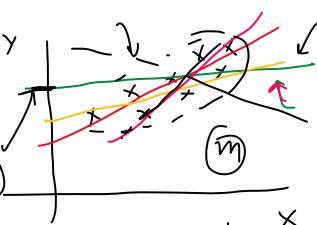
$$m - \bar{m} = 0$$

$$\text{bias} = 0$$



$$E[m] - \bar{m}$$

bias



true function

$$f(x)$$

$$f'(x)$$

$$f''(x)$$

$$E[m]$$

100 samples

LR $\rightarrow \bar{m}$

$$\text{Bias} = E[f'(x)] - f(x)$$

$$f(x) \uparrow \uparrow 0$$

$$f'(x) \downarrow \downarrow 0$$

$$y = \underline{f(x) + \text{error}} \quad \overbrace{\text{mse}}^{\text{Var} -} = \frac{1}{n} \hat{\theta} - \theta$$

$$\hat{y} = \underline{f'(x)} \quad \left\{ \begin{array}{l} \text{Bias}(f'(x)) = E[f'(x)] - f(x) \\ \text{Var}(f'(x)) = E[(f'(x) - E[f'(x)])^2] \end{array} \right.$$

Bias Variance Decomposition

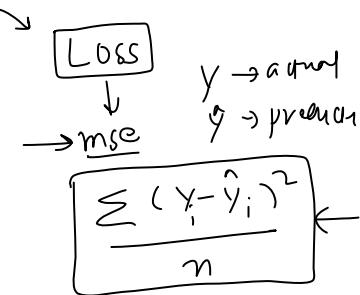
$$\text{Bias} = E[\hat{\theta}] - \theta$$

Bias Variance Decomposition

17 May 2023 07:17

Bias-variance decomposition is a way of analysing a learning algorithm's expected generalization error with respect to a particular problem by expressing it as the sum of three very different quantities: bias, variance, and irreducible error.

1. **Bias:** This is the error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
2. **Variance:** This is the error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).
3. **Irreducible Error:** This is the noise term. This part of the error is due to the inherent noise in the problem itself, and can't be reduced by any model.



$$\begin{aligned} \text{Loss} &= \text{bias} + \text{variance} + \text{irreducible} \\ \text{Loss} &= [\text{bias}^2 + \text{variance}] + [\text{var}(\epsilon)] \\ &\quad \text{reducible} \qquad \text{noise} \\ &\quad \text{(random)} \end{aligned}$$

derive

gpa | int | lpa | pred
 $\bar{g} \quad \bar{q} \quad \bar{l} \quad \bar{p}$

$$\begin{array}{r} \bar{g} \\ \bar{q} \\ \bar{l} \\ \bar{p} \end{array} \rightarrow \text{reducible}$$

$$\begin{array}{r} \bar{g} \\ \bar{q} \\ \bar{l} \\ \bar{p} \end{array} \rightarrow \text{irreducible}$$

$$\begin{array}{r} \bar{g} \\ \bar{q} \\ \bar{l} \\ \bar{p} \end{array} \rightarrow \text{irreducible}$$

$$\begin{array}{r} \bar{g} \\ \bar{q} \\ \bar{l} \\ \bar{p} \end{array} \rightarrow \text{irreducible}$$

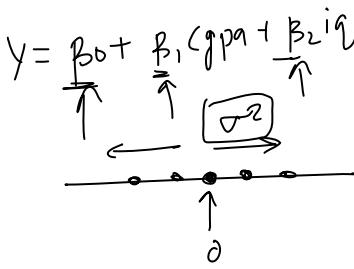
$$\begin{array}{r} \bar{g} \\ \bar{q} \\ \bar{l} \\ \bar{p} \end{array} \rightarrow \text{irreducible}$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \leftarrow \text{minimise mse}$$

$$y = \beta_0 + \beta_1 gpa + \beta_2 lpa$$

$$\text{mean} = \bar{y}$$

$$\text{Var} = \sigma^2$$



Derivation

$$\text{mse} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} = E[(y - \hat{y})^2] = (\theta - \hat{\theta})^2 + \epsilon^2$$

$$y = f(x) + \epsilon = \theta + \epsilon$$

$$\hat{y} = f'(x) = \hat{\theta} \quad \theta - \hat{\theta} + \epsilon \leftarrow \text{epsilon irreduc}$$

$$\begin{aligned} E[x+y] &= E[x] + E[y] \\ &= E[(\theta + \hat{\theta})] + E[\epsilon^2] + E[2\epsilon(\theta - \hat{\theta})] \end{aligned}$$

$$\begin{aligned} E[xy] &= E[x]E[y] \\ &= E[(\theta + \hat{\theta})^2] + E[\epsilon^2] + E[2\epsilon E[\epsilon] E[\theta - \hat{\theta}]] \end{aligned}$$

$$\begin{aligned} E[x]E[y] &= E[(\theta + \hat{\theta})^2] + E[\epsilon^2] \\ &\leftarrow E[\epsilon^2] = 0 \end{aligned}$$

$$E[X]E[Y]$$

mse

$$\text{mse} = E[(\theta - \hat{\theta})^2] + \boxed{E[\epsilon^2]} \leftarrow$$

$$E[\epsilon^2] \rightarrow \underline{\text{var}(\epsilon)} = \sigma^2 = E[(\epsilon - \overbrace{E[\epsilon]}^0)^2]$$

$$\uparrow E[\epsilon^2] = \text{Var}(\epsilon) \quad = E[(\epsilon - 0)^2] = E[\epsilon^2]$$

$$\text{mse} = E[(\theta - \hat{\theta})^2] + \boxed{\text{var}(\epsilon)}$$

irreducible

$$E[(\theta - \hat{\theta})^2] = E[(\theta - E[\hat{\theta}] + E[\hat{\theta}] - \hat{\theta})^2] \leftarrow$$

$$E[(\theta - E[\hat{\theta}])^2 + (E[\hat{\theta}] - \hat{\theta})^2 + 2(\theta - E[\hat{\theta}])(E[\hat{\theta}] - \hat{\theta})]$$

$$E[(\theta - E[\hat{\theta}])^2] + E[(E[\hat{\theta}] - \hat{\theta})^2]$$

$$E[2(\theta - E[\hat{\theta}])(E[\hat{\theta}] - \hat{\theta})] = 0$$

$$\downarrow$$

$$E[\epsilon] E[(\theta - E[\hat{\theta}])] \quad E[(E[\hat{\theta}] - \hat{\theta})]$$

$$2(\theta - E[\hat{\theta}]) \quad E[E[\hat{\theta}]] - E[\hat{\theta}]$$

$$\theta - E[\hat{\theta}] \quad E[\hat{\theta}] - E[\hat{\theta}]$$

$$E[(\theta - E[\hat{\theta}])^2] + E[(E[\hat{\theta}] - \hat{\theta})^2]$$

var

$$(\theta - E[\hat{\theta}])^2$$

$$\rightarrow (\theta - E[\hat{\theta}])^2$$

$\gamma = \theta + \text{var}$

$$(\text{Bias})^2 + \text{var}$$

$$\text{mse} = (\text{bias})^2 + \boxed{\text{variance}} + \text{noise}$$

irreducible
 $\text{var}(\epsilon)$

↓

reducible error

Bias - variance decomposition

$$\text{mse} = \text{reducible} + \text{variance}$$

$$\rightarrow \boxed{\text{bias} + \text{var}}$$

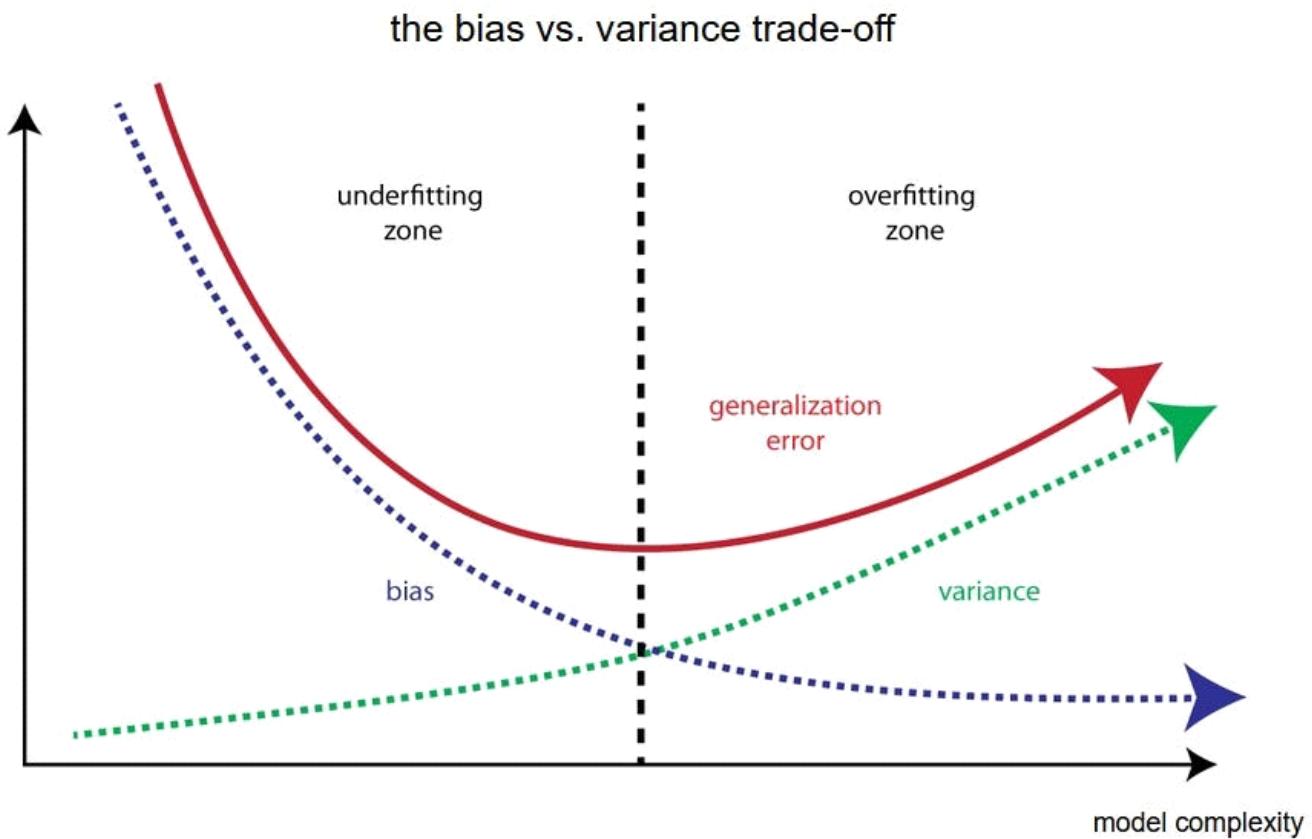
variance

$$\begin{array}{ccccccc} & \text{copper} & \text{nickel} & \text{lead} & & & \\ & 8 & 80 & 8 & - & 8.1 & 0.05 \\ \rightarrow & 7 & 70 & 7 & - & 6.9 & 0.1 \\ & & & & & & 0.04 \end{array}$$

$$\overbrace{bias + variance}^{\text{Total Error}} \rightarrow \text{Bias} + \frac{\text{Variance}}{n}$$

Diagram

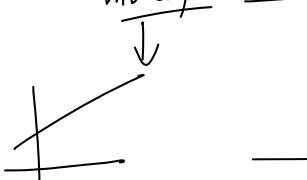
17 May 2023 17:52



irreducible

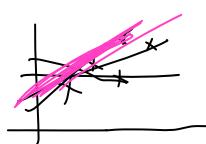
$$\text{mse} = (\text{bias})^2 + \text{variance} + \boxed{\text{varianu}(\epsilon) = 0} \quad y = f(x)$$

multiple shots



bias

$$E[\hat{\theta}] - \theta = \text{bias}$$

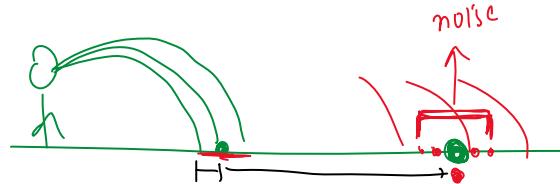
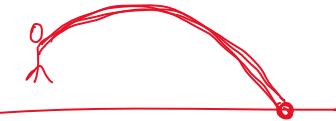
var on avg
mean

variance

$$\text{Var}(\hat{\theta})$$

$$E[\hat{\theta}]$$

$$\hat{\theta} = f(x)$$



$$\text{var}(\epsilon) = 0$$

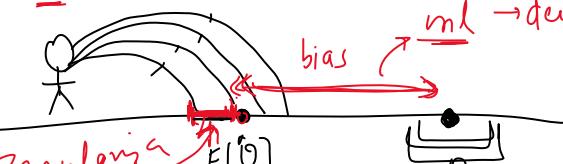
$$\text{loss} = 0$$

$$\text{bias} + \frac{\text{var}}{\text{std}}$$

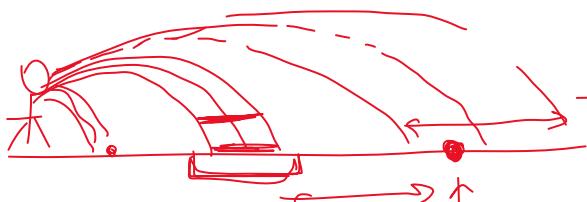
$$\rightarrow (y - \hat{y})^2 = (\text{bias})^2 + \frac{\text{var}}{\text{std}}$$

$$\text{Regulation} \rightarrow E[\hat{\theta}]$$

$$\frac{\text{Bias Variance}}{2}$$

polynomial
degree

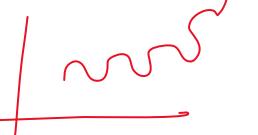
$$\begin{array}{c} \text{loss} \\ \swarrow \quad \searrow \\ \text{bias} \quad \text{var} \\ \downarrow \quad \downarrow \\ \text{target} \end{array}$$



Refuse

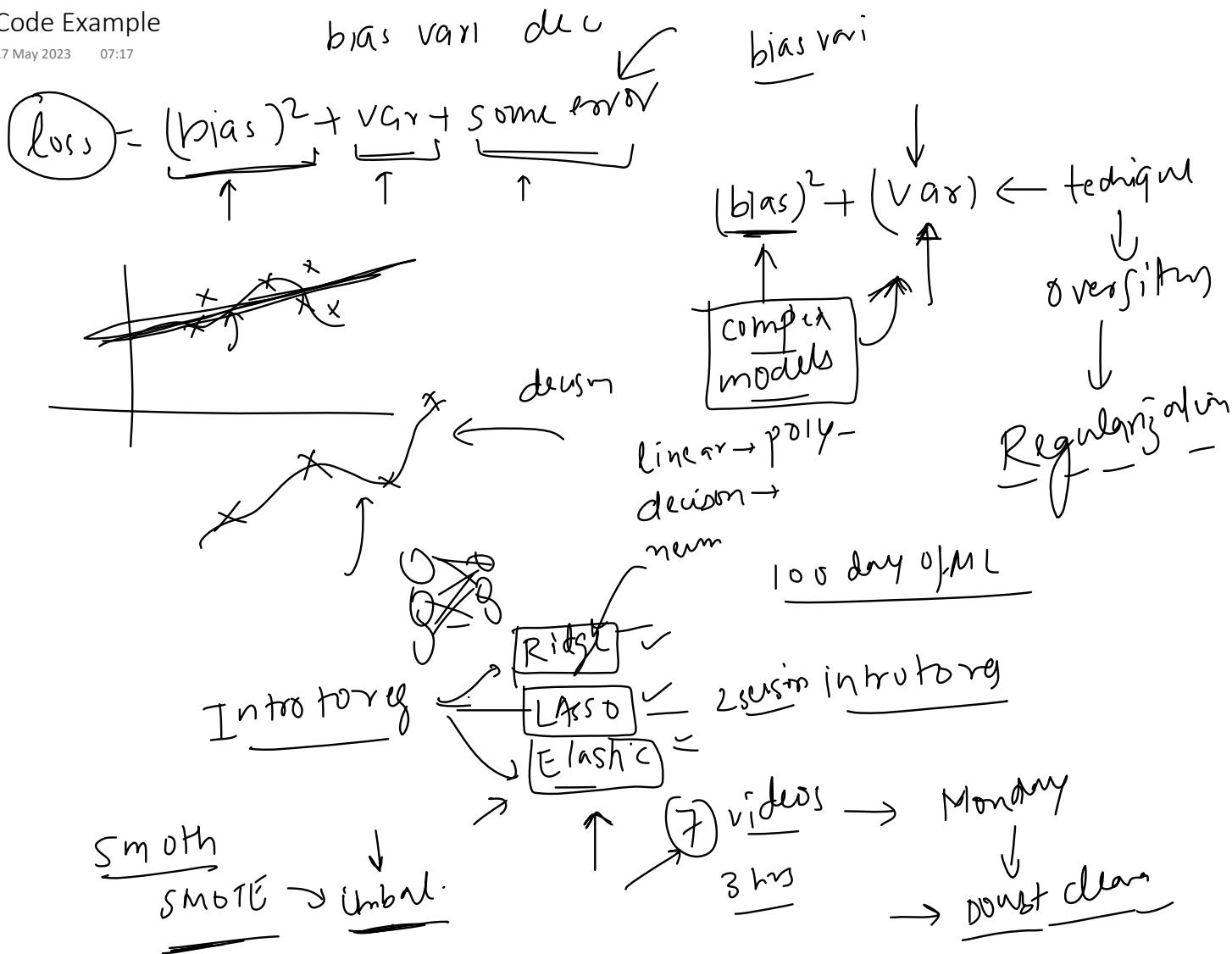
reduce variance

reduce overfitting



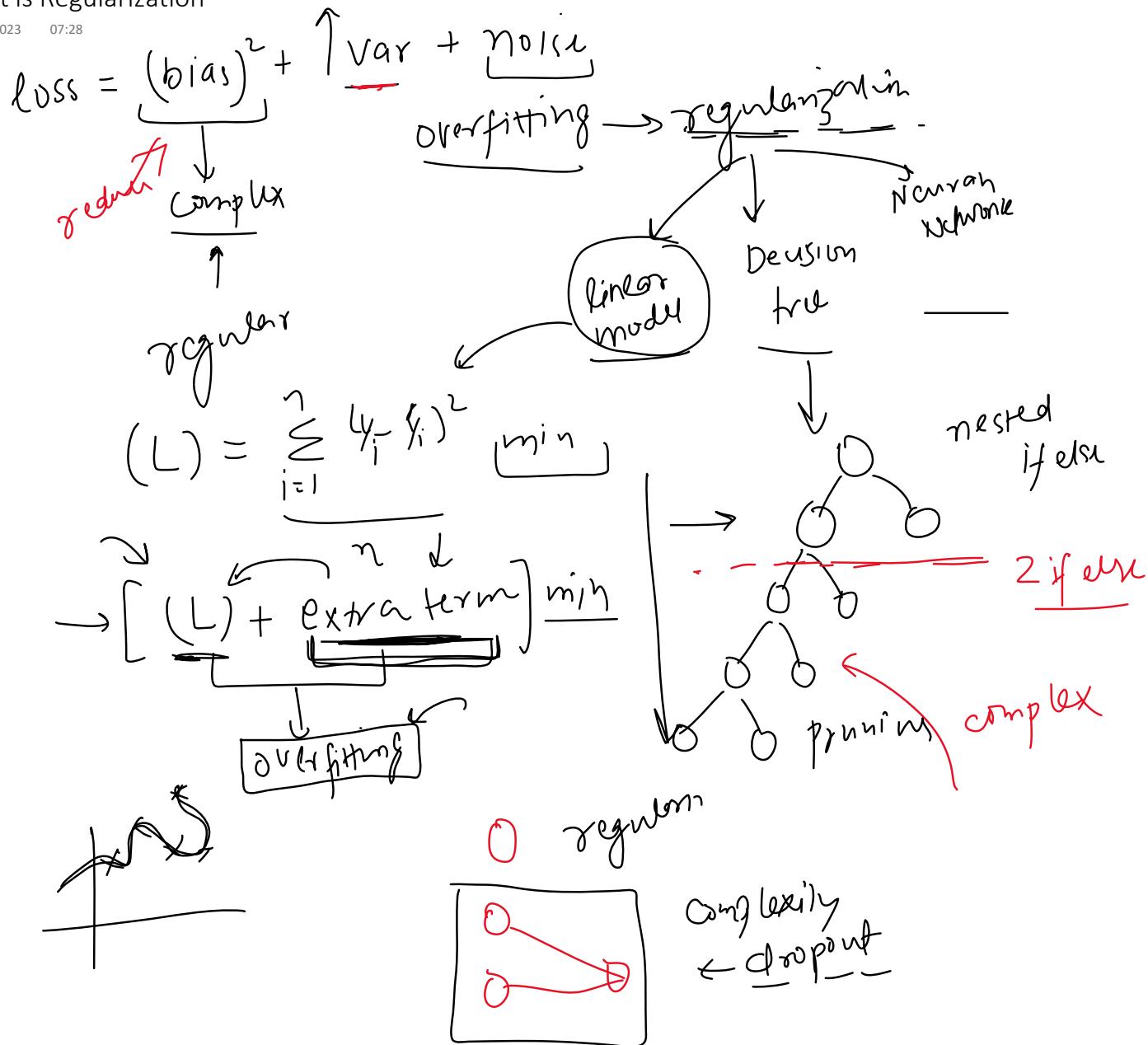
Code Example

17 May 2023 07:17



What is Regularization

19 May 2023 07:28



When to use Regularization?

19 May 2023 07:31

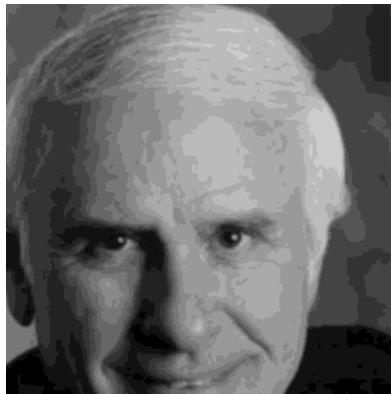
$$\rightarrow \underset{\text{X}}{\underbrace{x_1, x_2, \dots, x_n}} \underset{\text{Y}}{\rightarrow}$$

$f_1, f_2, f_3, \dots, f_{100}$
Overfit

1. Preventing Overfitting: Regularization is most commonly used as a tool to prevent overfitting. If your model performs well on the training data but poorly on the validation or test data, it might be overfitting, and regularization could help.
2. High Dimensionality: Regularization is particularly useful when you have a high number of features compared to the number of data points. In such scenarios, models tend to overfit easily, and regularization can help by effectively reducing the complexity of the model.
3. Multicollinearity: When features are highly correlated (multicollinearity), it can destabilize your model and make the model's estimates sensitive to minor changes in the model. L2 regularization (Ridge regression) can help in such cases by distributing the coefficient estimates among correlated features.
4. Feature Selection: If you have a lot of features and you believe many of them might be irrelevant, L1 regularization (Lasso) can help. It tends to produce sparse solutions, driving the coefficients of irrelevant features to zero, thus performing feature selection.
5. Interpretability: If model interpretability is important and you want a simpler model, regularization can help achieve this by constraining the model's complexity.
6. Model Performance: Even if you're not particularly worried about overfitting, regularization might still improve your model's out-of-sample prediction performance.

\rightarrow regularization \rightarrow feature select
 \rightarrow embedded-LASSO
forwards
selection

$$\overset{\curvearrowleft}{\mathcal{F}} \overset{\curvearrowright}{\mathcal{S}}$$



\rightarrow Knn \rightarrow simple elegant

100 student

$(\text{cgpa} | \text{iq})$ placement

	8	80	1
2D	7	70	0
.	.	.	.
.	.	.	.
.	.	.	.

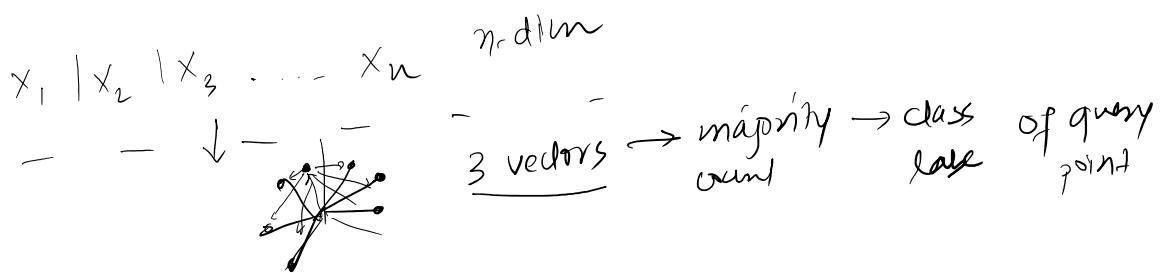
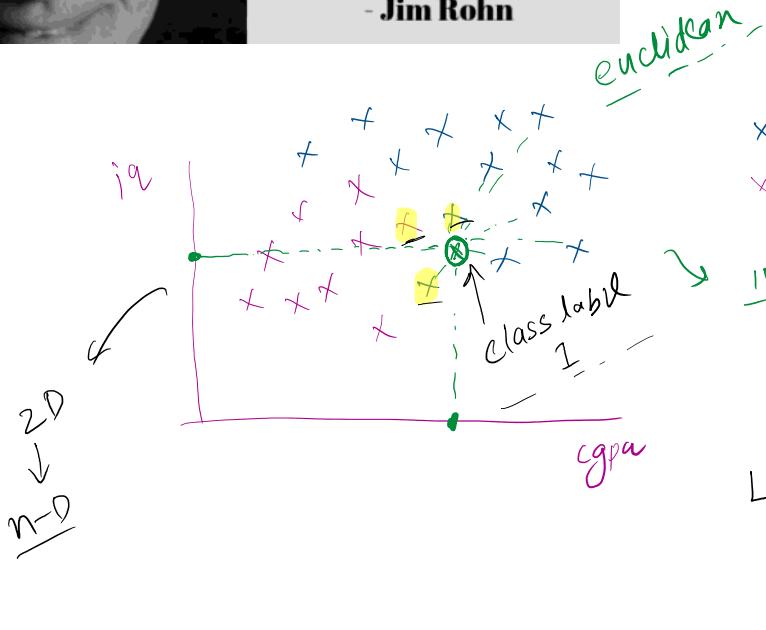
ml model

\downarrow

$\{\text{cgpa} \rightarrow \text{placement}$

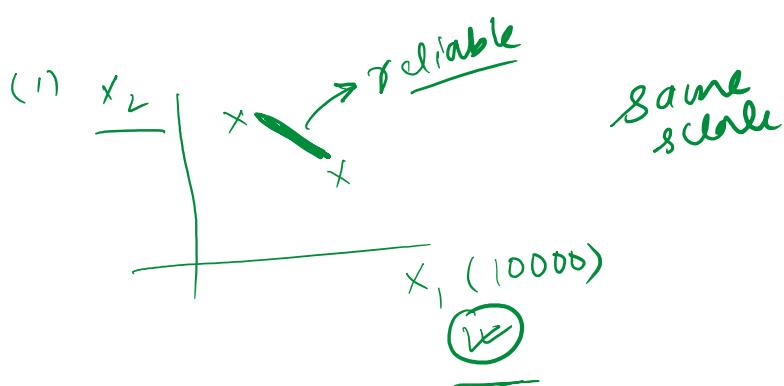
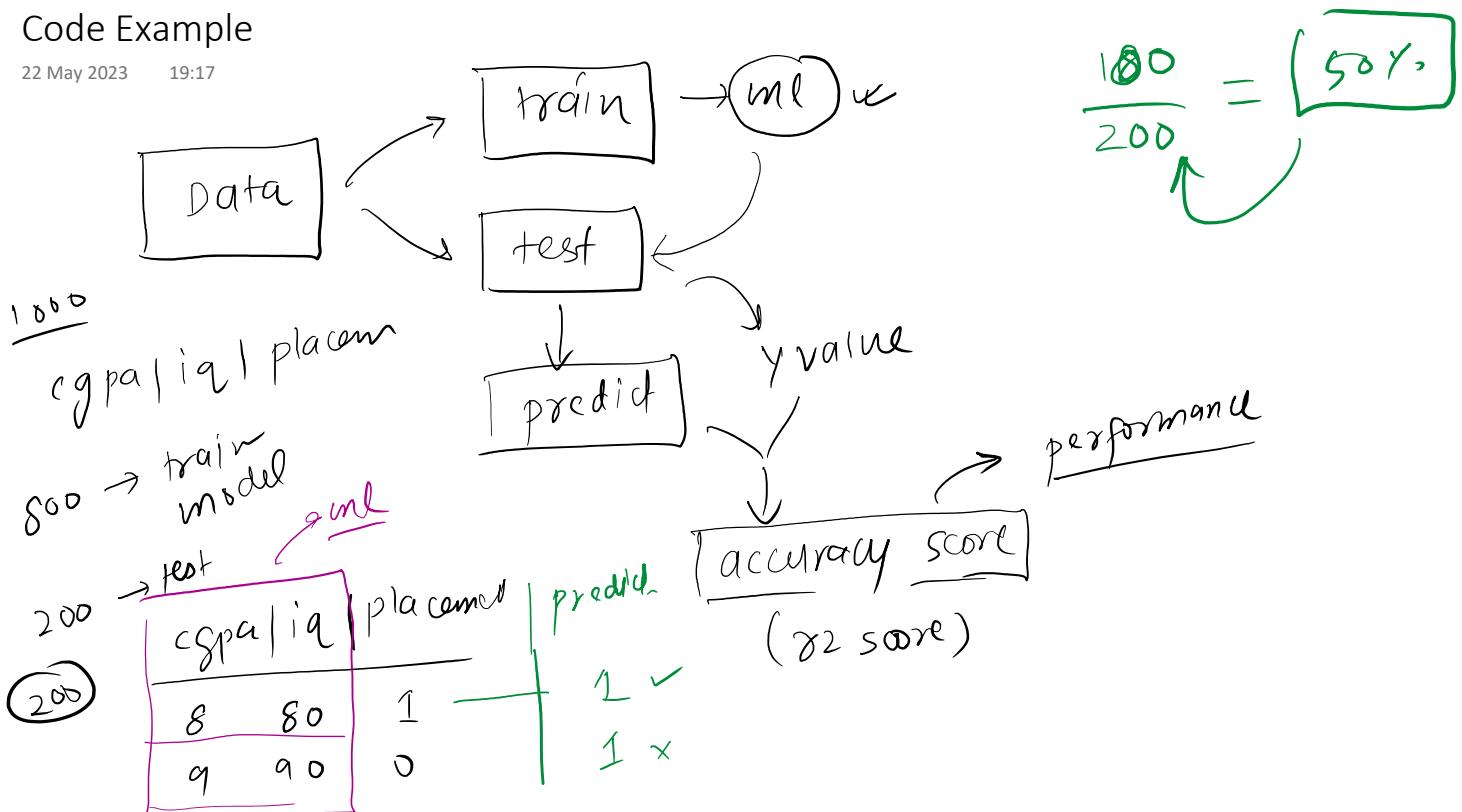
$\{\text{iq} \rightarrow \text{placement}$

Knn



Code Example

22 May 2023 19:17



age

$$\frac{72 - \mu}{\sigma}$$

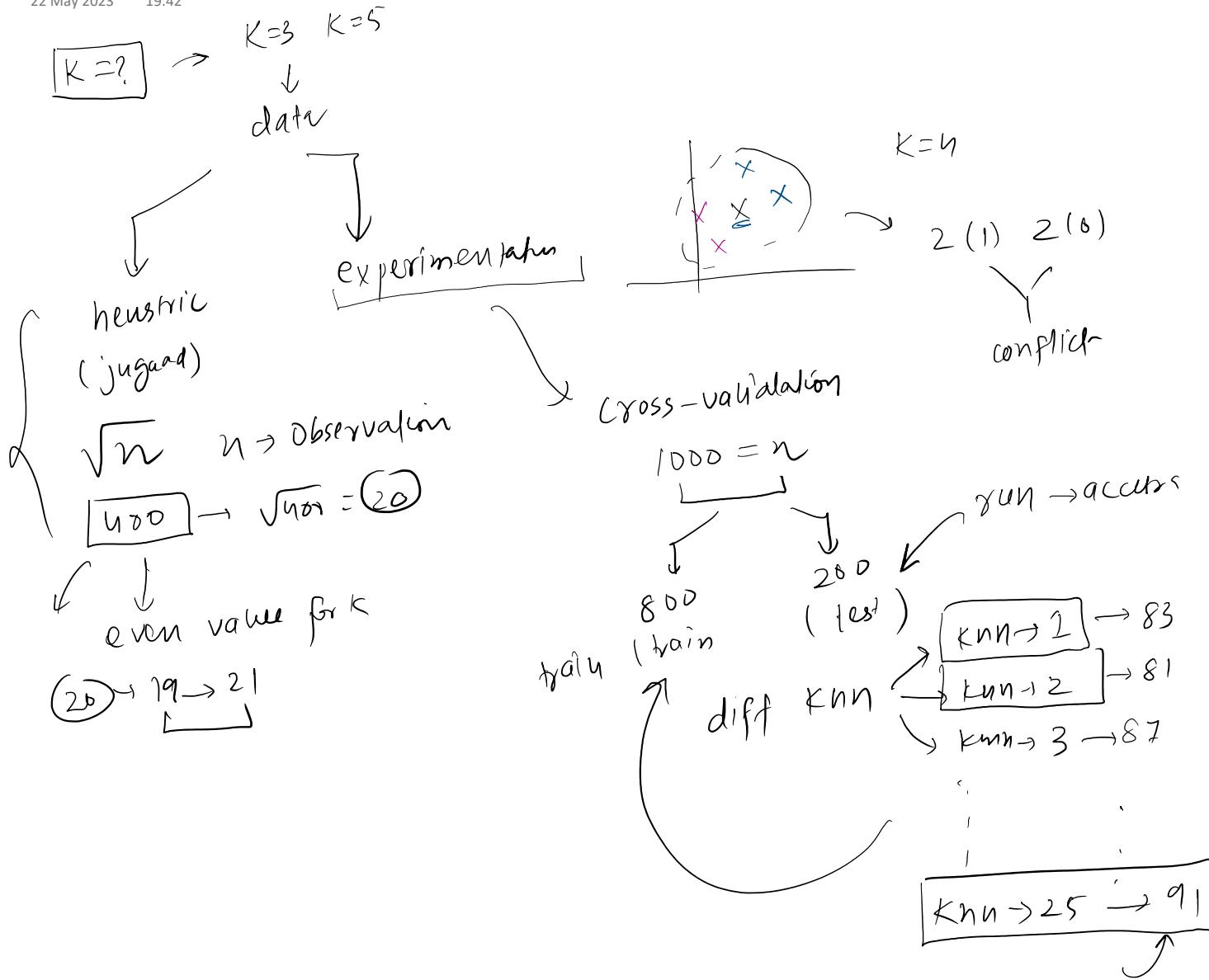
$$\frac{73 - \mu}{\sigma}$$

$$\rightarrow 61$$

$$\rightarrow 37$$

How to select K?

22 May 2023 19:42



[Decision Surface] → tool → classification → kNN → SVM

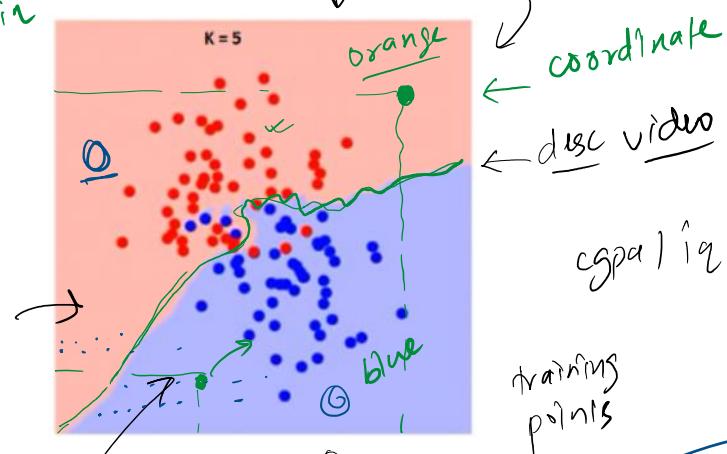
22 May 2023 19:15

↓ ↘ 1D 2D 3D

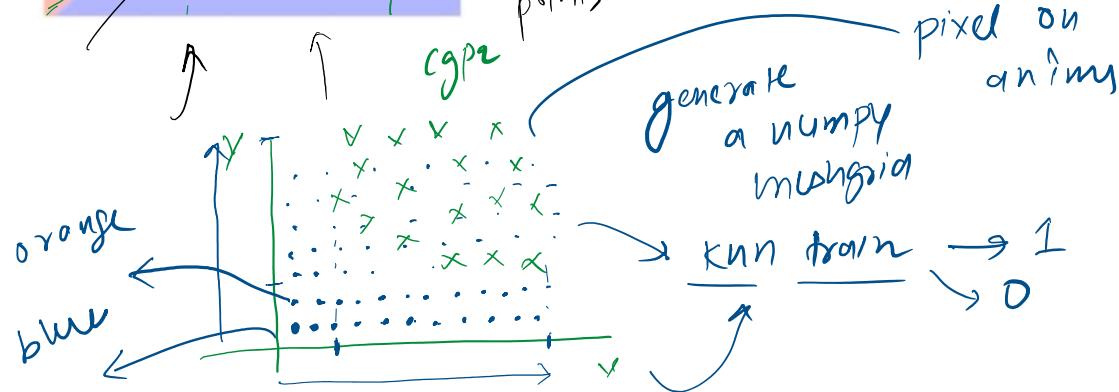
lr
dt
nn

upcoming → extend

plot decision surface

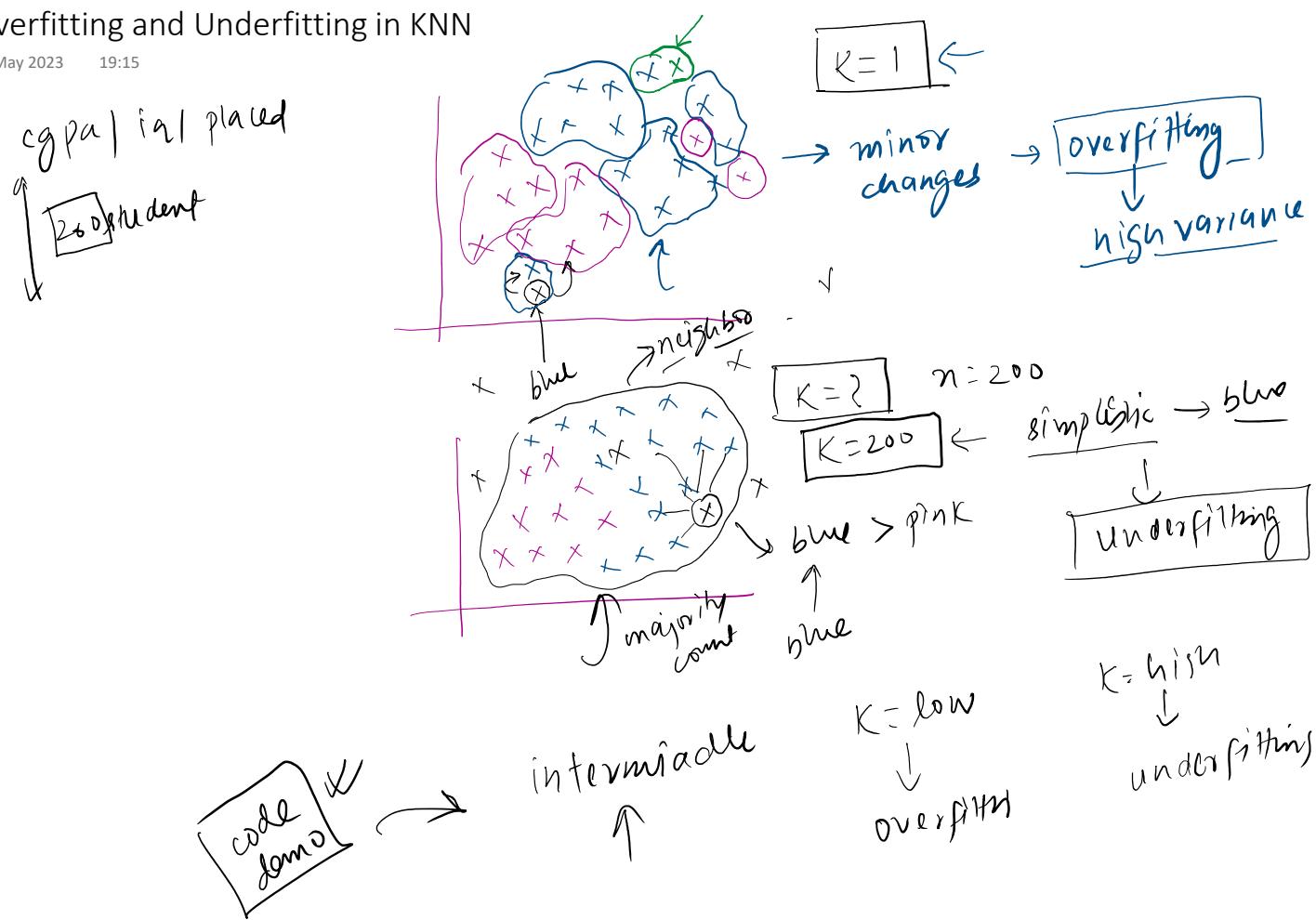


cgpal iq | place binary [0,1] ↑ ↓



Overfitting and Underfitting in KNN

22 May 2023 19:15



Limitations of KNN

22 May 2023 19:32

failure case

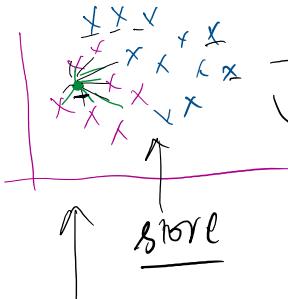
1) large datasets $\rightarrow n = 5L$, $f = 100$

\hookrightarrow Knn is a lazy learning techni

$(500000^0, 10^0)$

prediction

slow \rightarrow dataset

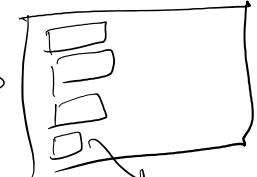


query point \rightarrow prediction

training \rightarrow nothing

5L distance \rightarrow sort

majority



low latency

[3 sec] \rightarrow slows

2) High dim data

$f = 500$

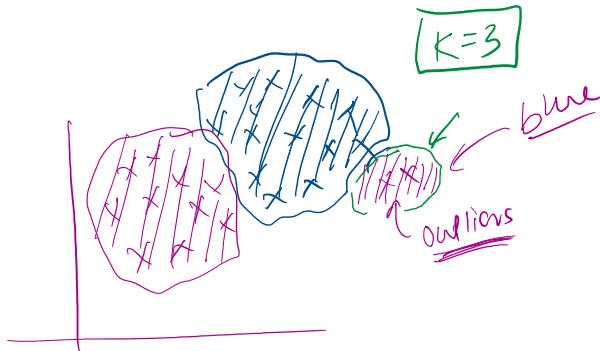
curse of dimension

distance concept \rightarrow redundant

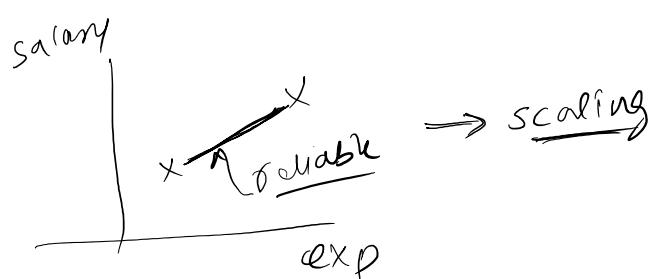
\uparrow

KNN

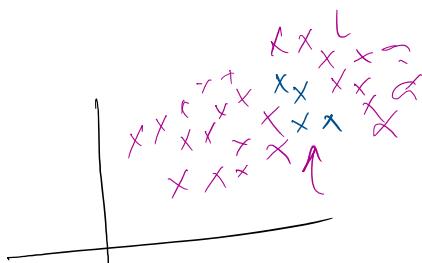
3) Outliers



4) Non-homogeneous scales
 \times dominate
 exp | salary | fire
 0-25 20K-10K



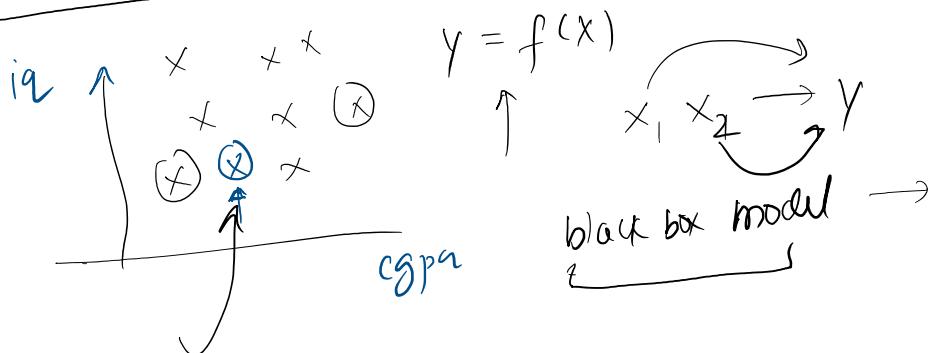
5) Imbalanced dataset
 \hookrightarrow Yes $\rightarrow 98\%$



5) Imbalance

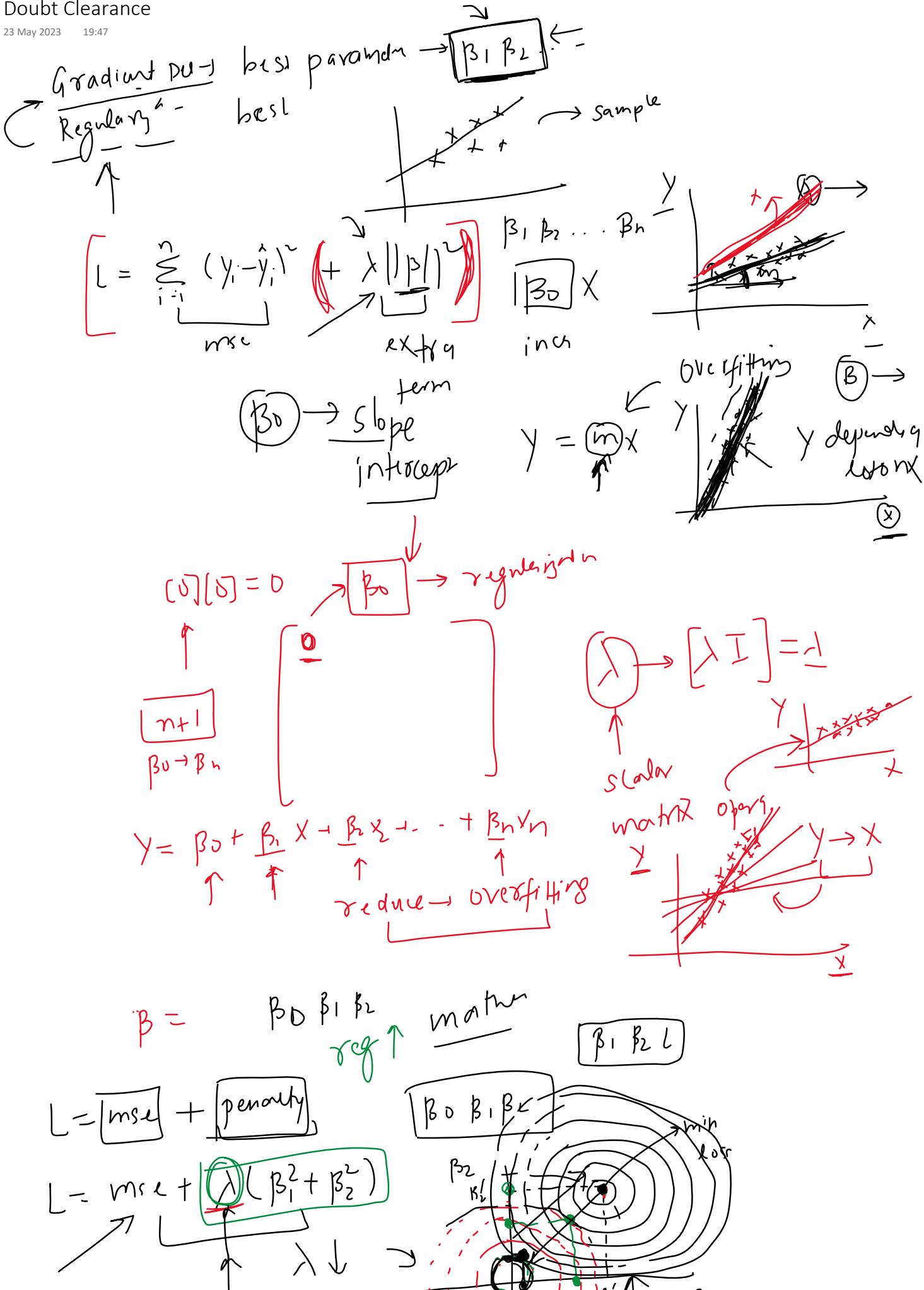
↳ Yes \rightarrow 98%
No \rightarrow 2%
↳ biased \rightarrow precise

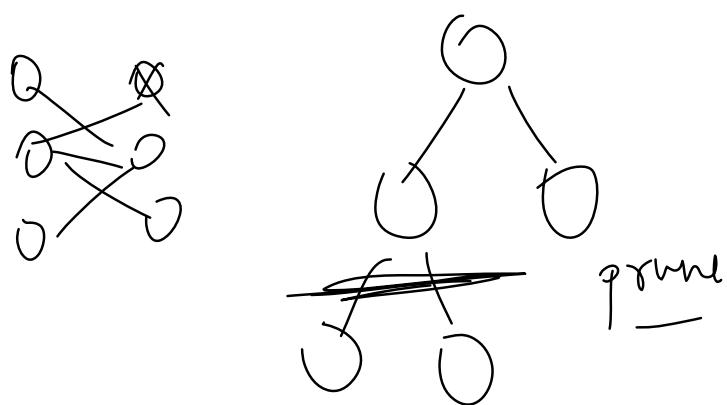
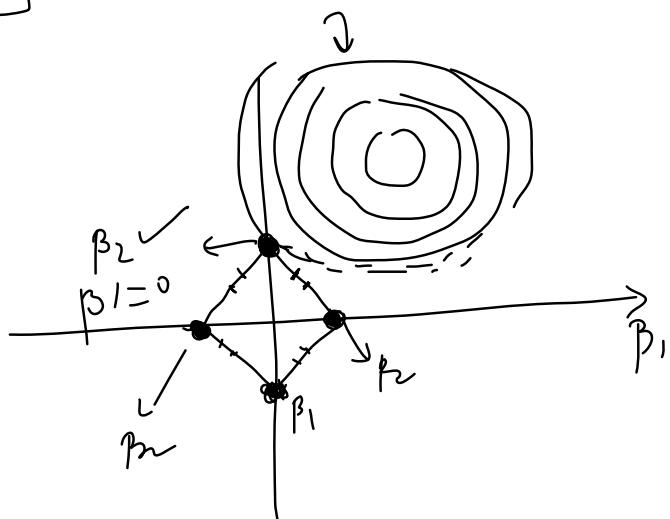
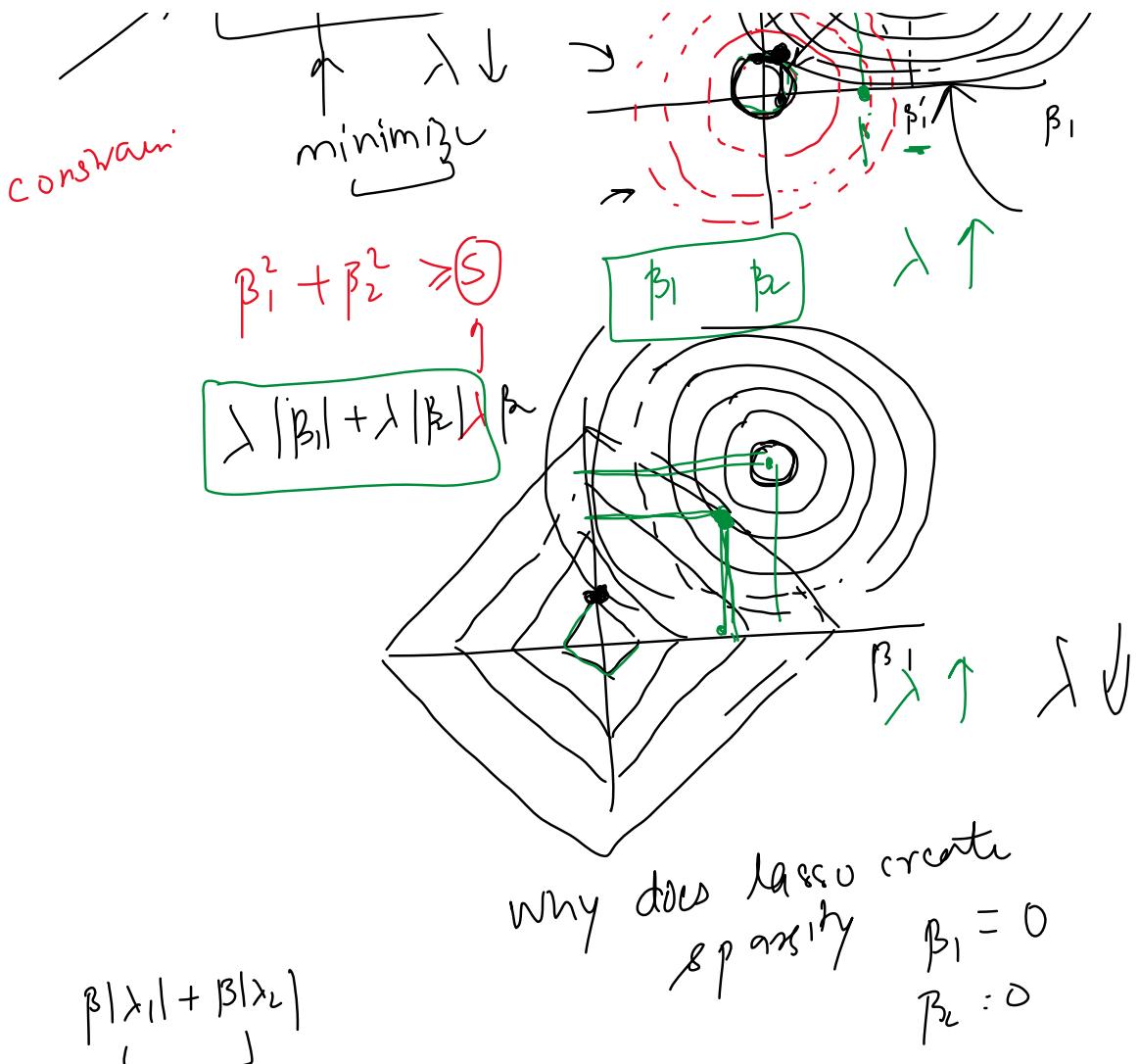
6) Inference and not for prediction



Doubt Clearance

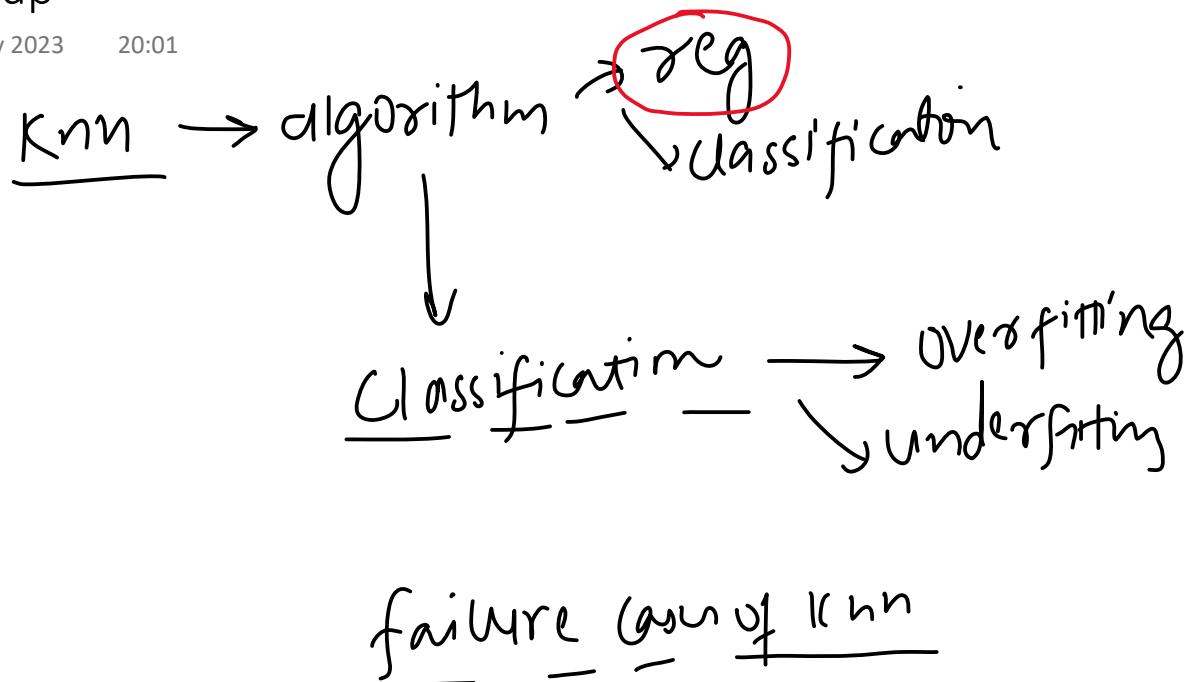
23 May 2023 19:47





Recap

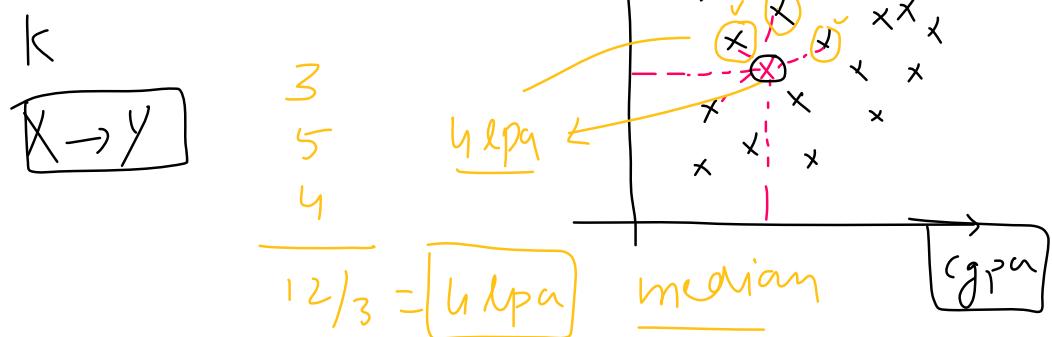
24 May 2023 20:01



KNN Regressor

24 May 2023 14:44

\downarrow
KNN \rightarrow regression



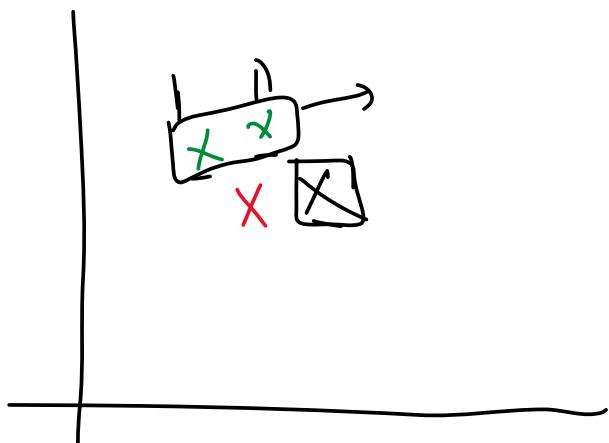
↳ find distance
 x_g and all the
 x_{-train}

K high \rightarrow underfit
 K low \rightarrow overfitting

Hyperparameters

24 May 2023 14:44

$K=3$

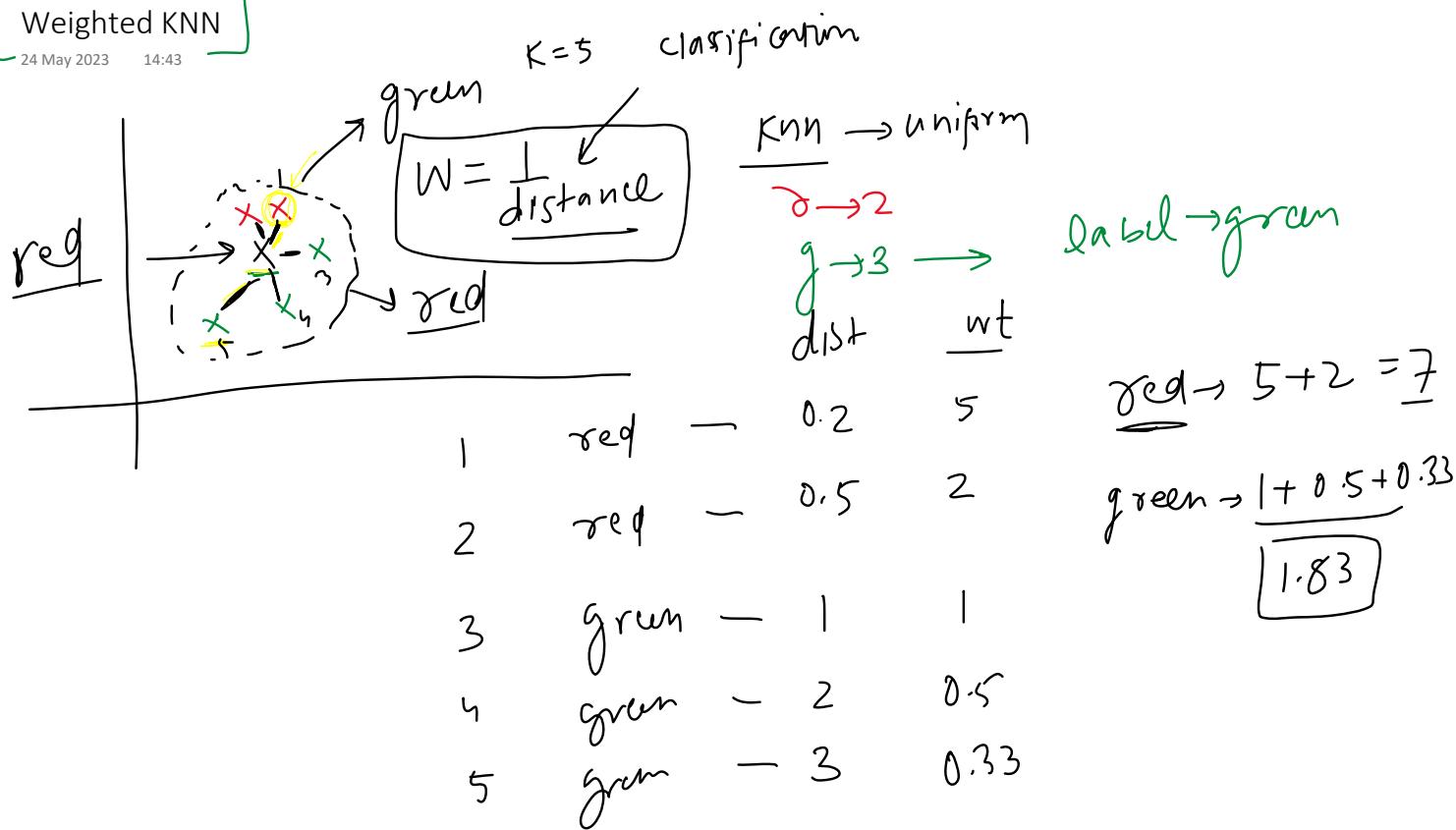


uniform

weighted sum → KNN

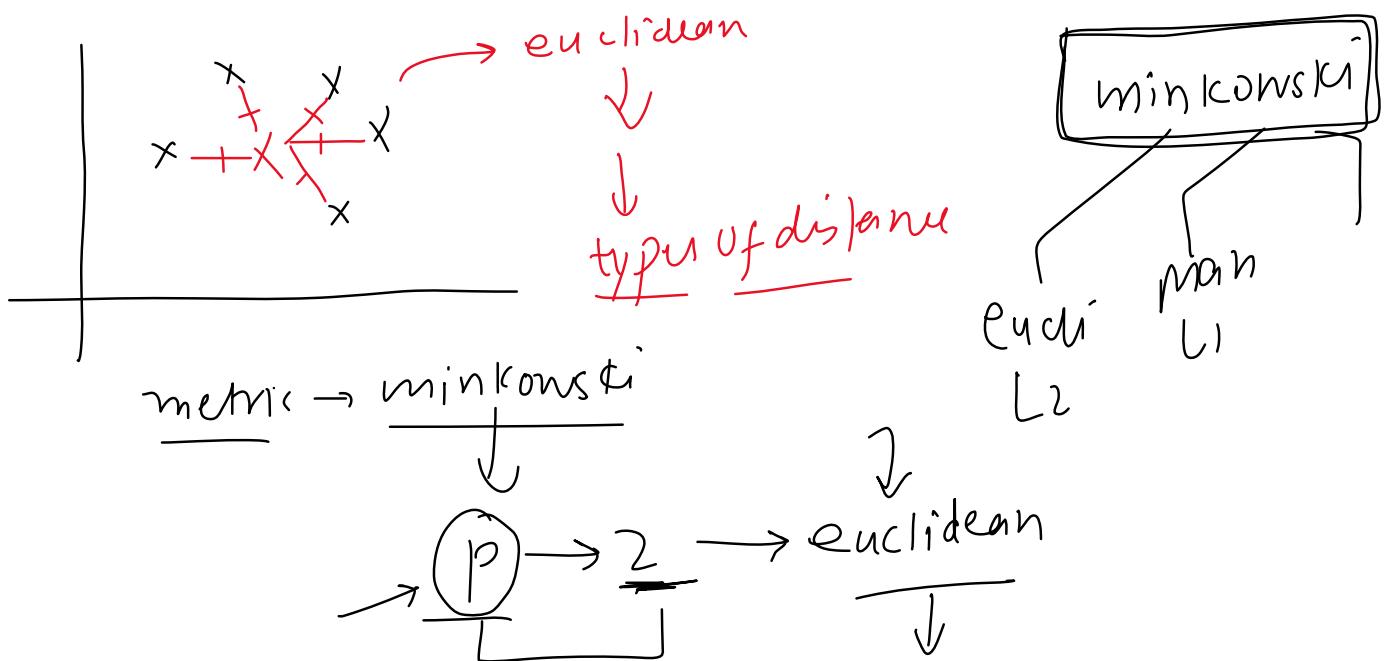
Weighted KNN

24 May 2023 14:43

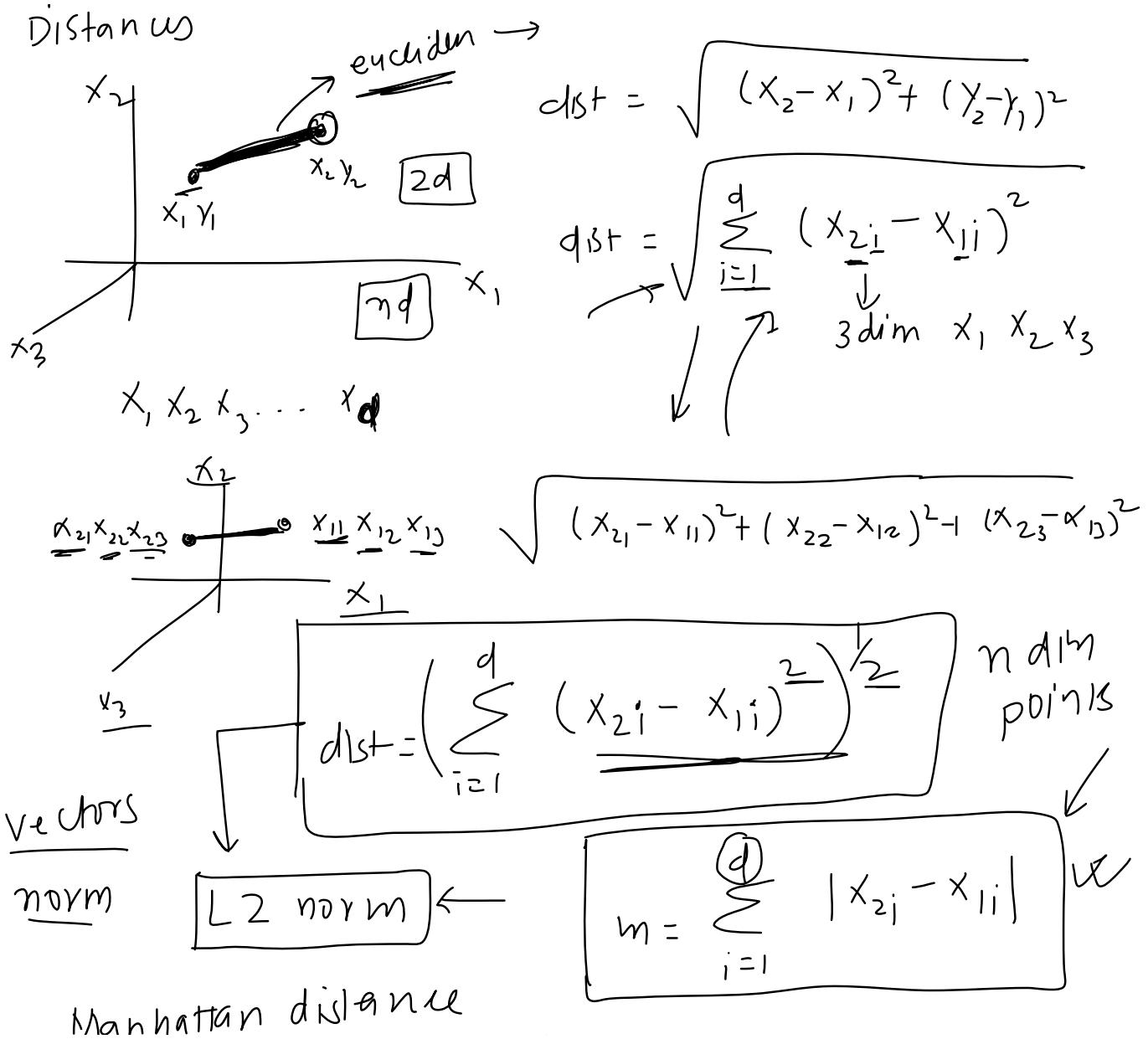


Types of Distances

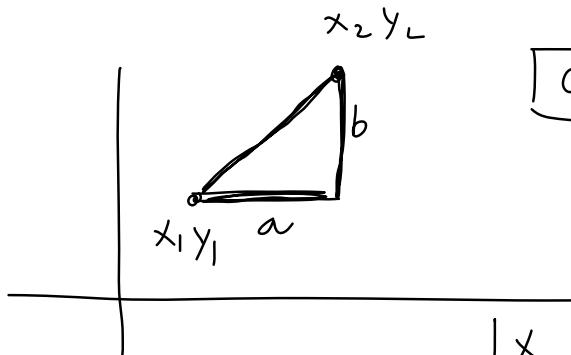
24 May 2023 14:44



Distances



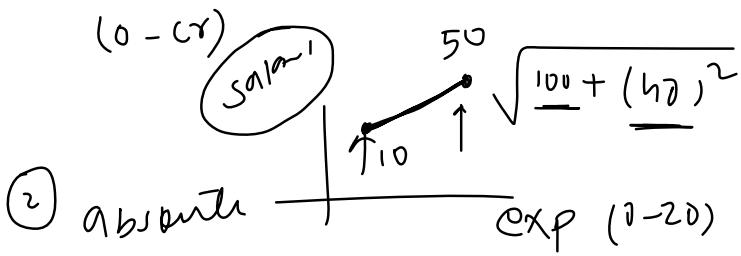
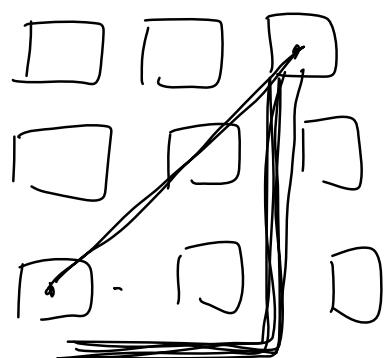
Manhattan distance



$$a+b \rightarrow \text{manhattan distance}$$

taxi cab distance

$$|x_2 - x_1| + |y_2 - y_1| \rightarrow \text{mann}$$



problems \rightarrow euclidean

$$d = 100$$

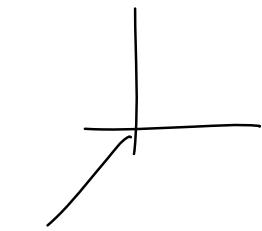
1) same scale \leftarrow

euclidean

2) wise of dimension

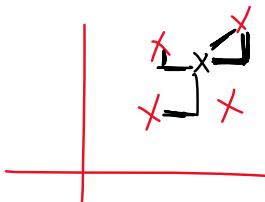
\uparrow manhattan

reliable



p=2 \rightarrow euclidean

p=1 manhattan



minkowski

$$\left(\sum_{i=1}^d (x_{2i} - x_{1i})^2 \right)^{\frac{1}{2}}$$

$p=2$ L2

eucm \rightarrow

$$\left(\sum_{i=1}^d |x_{2i} - x_{1i}|^1 \right)^{\frac{1}{1}}$$

$p=1$ L1

man

man

$$\left(\sum_{i=1}^d |x_{2i} - x_{1i}| \right)$$

↓

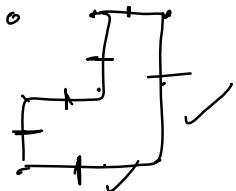
general

$\theta \rightarrow$ higher

$$\left(\sum_{i=1}^d (|x_{2i} - x_{1i}|)^p \right) \quad \begin{matrix} p \\ \leftarrow 2 \text{ (each)} \\ \leftarrow 1 \text{ (man)} \\ \underline{\underline{3}} \text{ L3} \\ \underline{\underline{1}} \text{ L4} \end{matrix}$$

$p > 0$

$2.5 \quad \underline{L2.5}$



Space and Time Complexity

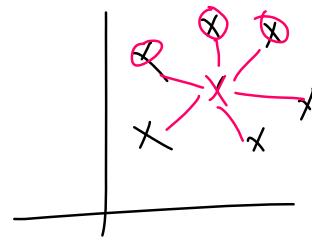
24 May 2023 14:46

Knn → slow algorithm

↓
product
↑

→ production

→ train



Sort
↓
K nearest
↓
major

← result t

time complexity

outlining $f(n)$ down
no of rows

Time complexity

$\rightarrow \mathcal{O}(nd)$ $n \rightarrow \# \text{ of rows in training data}$
 $d \rightarrow \# \text{ features}$

$|x_1 \ x_2 \ x_3 \dots \ x_d | y$
 $(10L, 100D)$

$\text{eucl} \downarrow$

$$\sqrt{\sum_{i=1}^d (x_{2i} - x_{1i})^2}$$

$\uparrow d \text{ times (2)}$ → 1 ur

$2 \times 10^8 \rightarrow \text{distance}$

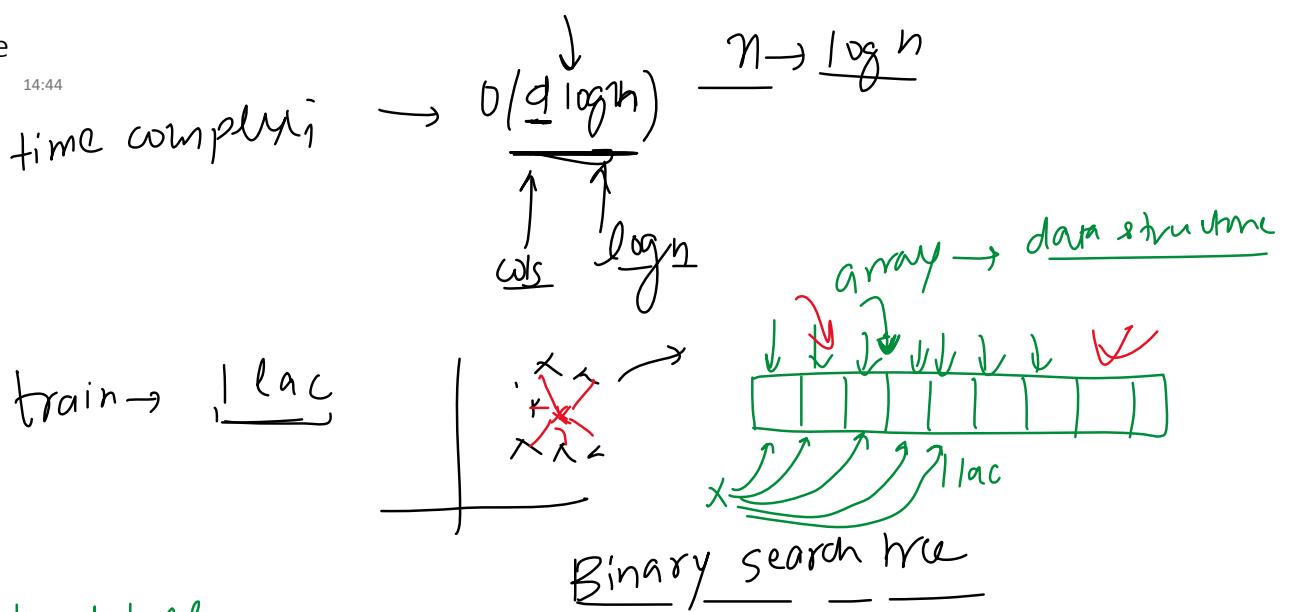
space complexity

$x \ x \ x \ x$ → $\mathcal{O}(nd)$

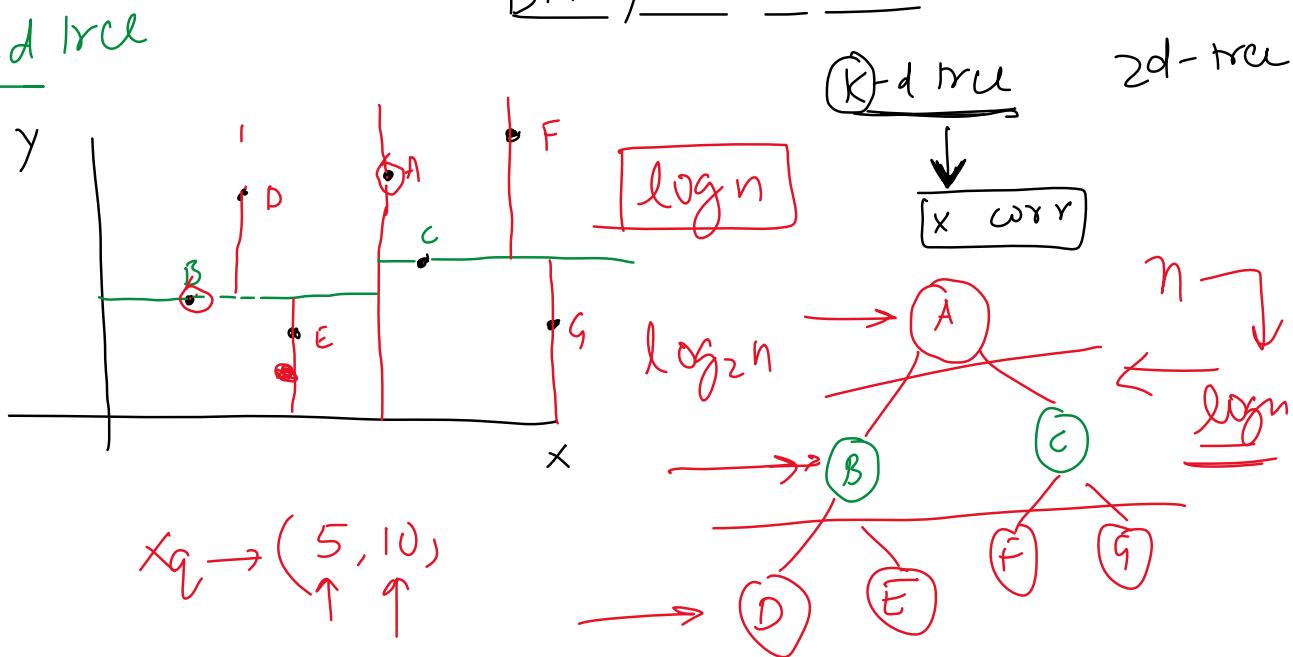
1GB → 2am

KD-Tree

24 May 2023 14:44



K-d tree



What are Matrices?

30 May 2023 18:33

Linear transform

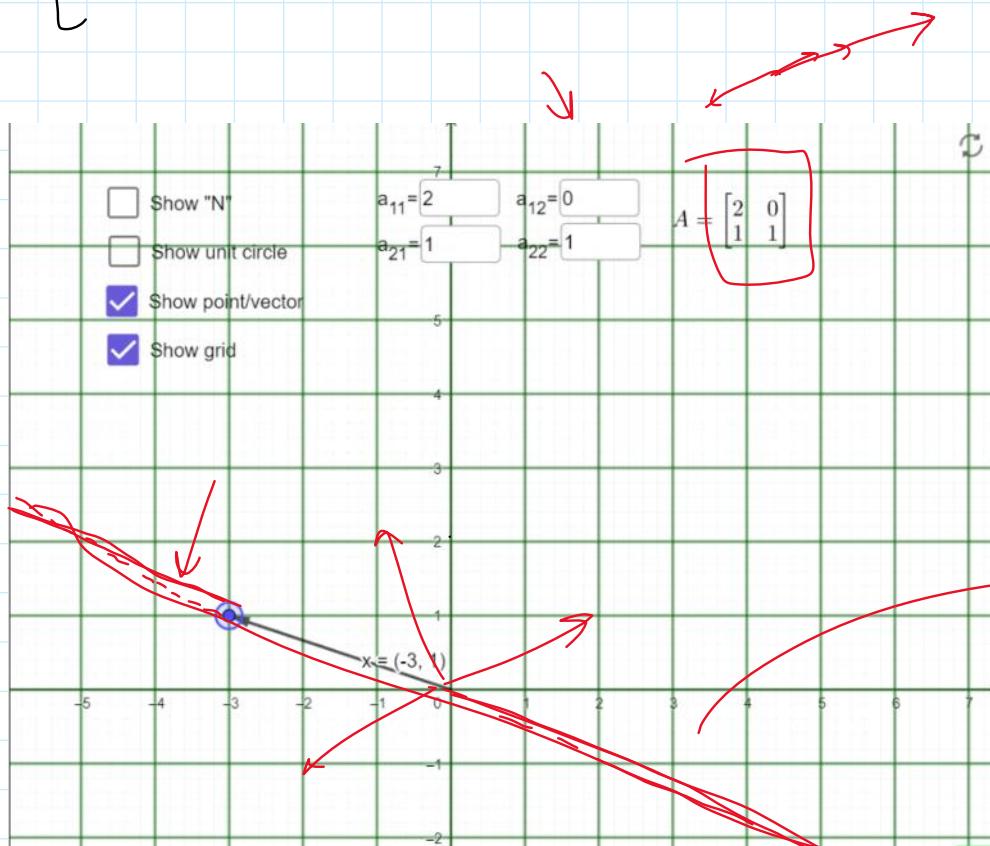
$$\begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix} \rightarrow \text{Linear transform}$$
$$\begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

2-dim

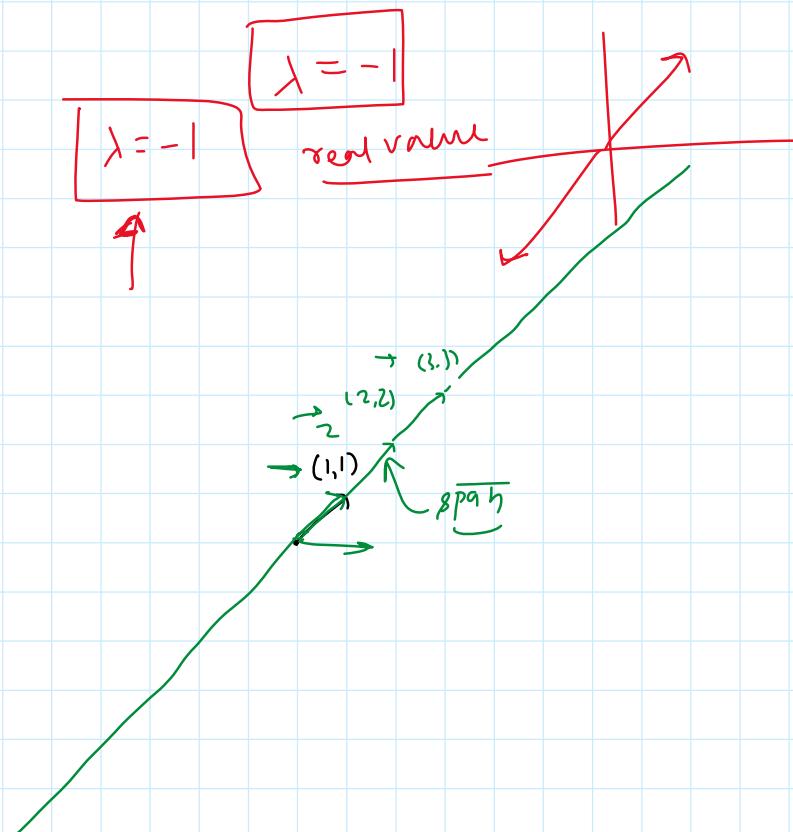
What are Eigen Vectors and Eigen Values

30 May 2023 10:34

$$\begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}$$



Eigen vectors
Eigen value
 $\vec{v} = \lambda \vec{V} \rightarrow (2, 0)$
 \vec{v} is scalar
 $\lambda = 2$
Span of vector
 $\begin{bmatrix} 3 \\ 6 \end{bmatrix}$
 $\begin{bmatrix} 6 \\ 12 \end{bmatrix}$
 $\begin{bmatrix} 1.5 \\ 3 \end{bmatrix} \times 0.5$



Intuition - axis of rotation

30 May 2023 18:45

How to calculate Eigen Vectors and Eigen Values

30 May 2023 18:34

$$A = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} \rightarrow A\vec{x} = \lambda\vec{x}$$

\uparrow
 $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

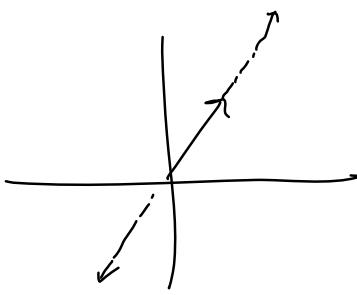
\uparrow
 $A \rightarrow \text{matrix}$
 $\vec{x} \rightarrow \text{vector}$
 $\lambda \rightarrow \text{scalar}$

$$\frac{1}{2}A\vec{x} = \frac{\lambda}{2}\vec{x}$$

$$A\vec{x} - \lambda I\vec{x} = 0$$

$$(A - \lambda I)\vec{x} = 0$$

\downarrow
 $\text{matrix } \vec{x} \text{ vector}$
 non-invertible
 $\hookrightarrow \det(\text{matrix}) = 0$



$$(A - \lambda I)$$

\downarrow
 $(2\lambda) \rightarrow \text{id}$

$C_{1,N}$

non invertible $\rightarrow \det(A - \lambda I) = 0$

$\det(A) = 0$

$$\det \left(\begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = 0$$

$$\det \left(\begin{bmatrix} 2-\lambda & 3 \\ 1 & 1-\lambda \end{bmatrix} \right) = 0$$

$$(2-\lambda)(1-\lambda) + 3 = 0$$

$$2 - 2\lambda - \lambda + \lambda^2 + 3 = 0$$

$$\boxed{\lambda^2 - 3\lambda + 5 = 0} \quad \leftarrow \underline{\text{eigen values}}$$

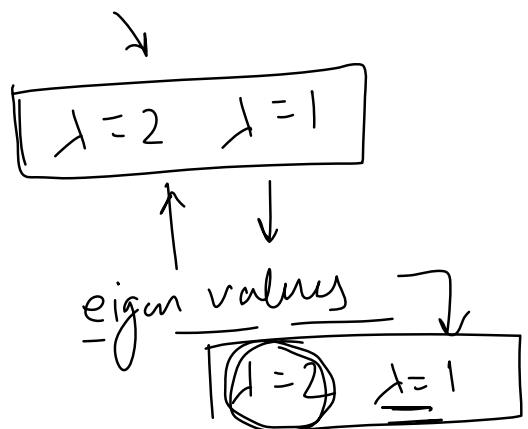
$$\det \left(\begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) = 0$$

$$\det \begin{pmatrix} 2-\lambda & 3 \\ 0 & 1-\lambda \end{pmatrix} = 0$$

$$(2-\lambda)(1-\lambda) = 0$$

$$2 - 2\lambda - \lambda + \lambda^2 = 0$$

$$\lambda^2 - 3\lambda + 2 = 0$$



$$(A - \lambda I) \vec{v} = 0$$

$$\left(\begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 3 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$0x_1 + 3x_2 = 0$$

$$0x_1 - x_2 = 0$$

$$3x_2 = 0$$

$$x_2 = 0$$

$$[1, 0] \leftarrow [2, 0, 3, 0]$$

$$\begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$x_2 = 0, x_1 = 1$

\downarrow

\nearrow

\swarrow

$2v$

$\boxed{\lambda=2}$

$$(A - \lambda I) \vec{v} = 0$$

— r. r. r. —

$$(A - \lambda I) \vec{v} = 0$$

$$\left(\begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\lambda = 1 \quad x_1 + 3x_2 = 0 \quad (1) \quad \begin{bmatrix} -3, 1 \\ -6, 2 \end{bmatrix}$$

$$x_1 = -3x_2$$

$$\begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ 1 \end{bmatrix} = \begin{bmatrix} -6+3 \\ 0+1 \end{bmatrix} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 \\ 4 & 1 \end{bmatrix} \rightarrow \text{eigen vec}$$

$$x=1$$

$$A \vec{v} = 0$$

$$A \vec{v} = \lambda \vec{v}$$

$$n=1 \rightarrow$$

$$n=2 \rightarrow$$

$$784 \rightarrow \begin{matrix} n \\ \downarrow \\ 784 \end{matrix}$$

$$\rightarrow \begin{matrix} n=2 \\ \leftarrow 782 \end{matrix}$$

$$A \vec{v} = 0$$

$$\frac{3d}{2d} \approx 1.5$$

Properties

30 May 2023 18:34

$$c \begin{bmatrix} 0 & 1 \\ 3 & \lambda \end{bmatrix} \xrightarrow{\text{PCA}} \begin{bmatrix} c\lambda & c \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{trace}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow \boxed{\lambda = 1, \lambda = 1}$$

λ_1, λ_2

$$\begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} \rightarrow 2 \times 1 - 0 = 2$$

$$\lambda = 2, \lambda = 1$$

max

(2)

$$\begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix}$$

$$\lambda^T = \lambda$$

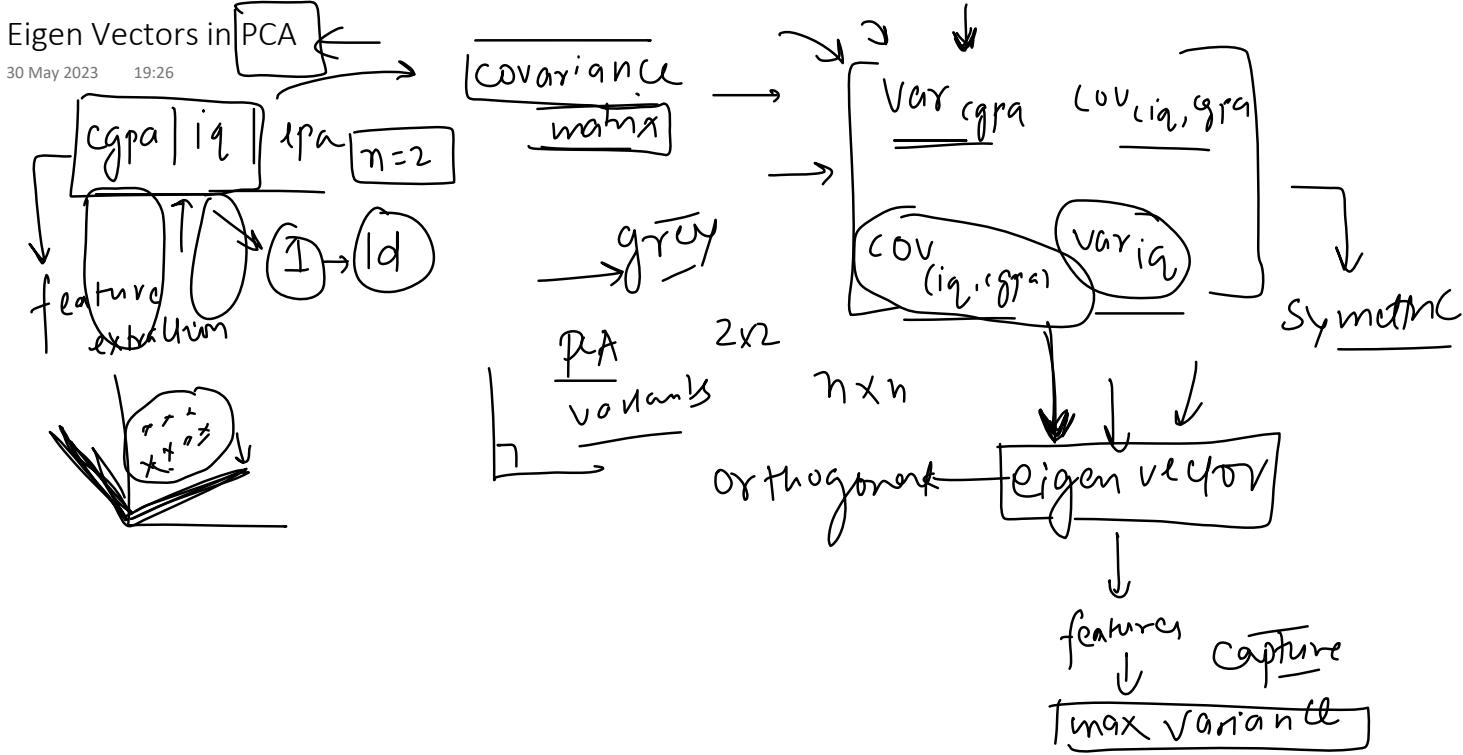
$$\text{eigen vector}$$

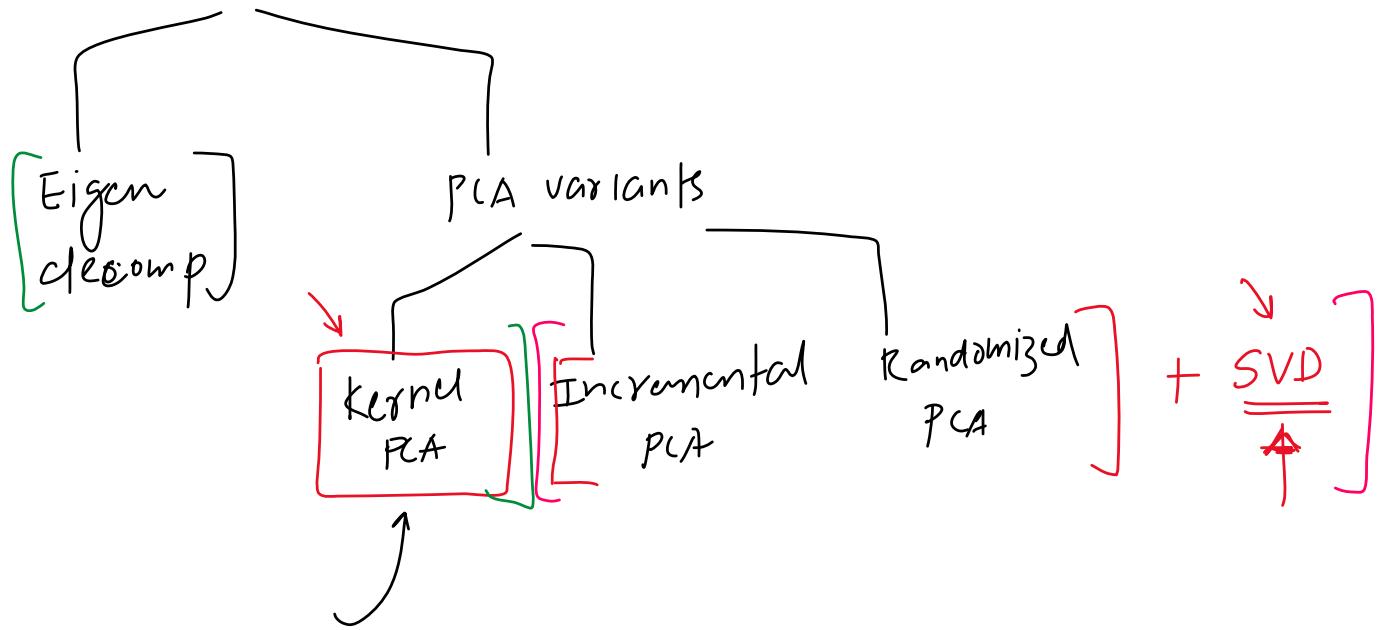
$$\begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix} \xrightarrow{\text{PCA}} \boxed{\lambda = 3, \lambda = 4}$$

max $\rightarrow (n)$

$$(2)d \rightarrow (2) \min \rightarrow (0)$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{independent}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{not correlated}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{CGPA}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{CGPA}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{CGPA}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{CGPA}}$$





Some Special Matrices

01 June 2023 16:32

$$A^{100}$$

$$\begin{bmatrix} a & b \\ 0 & b \end{bmatrix} \quad \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} \leftarrow$$

1. Diagonal Matrix

A diagonal matrix is a type of square matrix where the entries outside the main diagonal are all zero; the main diagonal is from the top left to the bottom right of the square matrix.

- a. Powers: The nth power of a diagonal matrix (where n is a non-negative integer) can be obtained by raising each diagonal element to the power of n.
- b. Eigenvalues: The eigenvalues of a diagonal matrix are just the values on the diagonal. The corresponding eigenvectors are the standard basis vectors.
- c. Multiplication by a Vector: When a diagonal matrix multiplies a vector, it scales each component of the vector by the corresponding element on the diagonal.
- d. Matrix Multiplication: The product of two diagonal matrices is just the diagonal matrix with the corresponding elements on the diagonals multiplied.

2. Orthogonal Matrix

An orthogonal matrix is a square matrix whose columns and rows are orthogonal unit vectors (i.e., orthonormal vectors), meaning that they are all of unit length and are at right angles to each other.

Perfect rotation no scaling or shearing.

$$AT = A^{-1}$$

- a. Inverse Equals Transpose: The transpose of an orthogonal matrix equals its inverse, i.e., $A^T = A^{-1}$. This property makes calculations with orthogonal matrices computationally efficient.

3. Symmetric Matrix

A symmetric matrix is a type of square matrix that is equal to its own transpose. In other words, if you swap its rows with columns, you get the same matrix.

- a. Real Eigenvalues: The eigenvalues of a real symmetric matrix are always real, not complex.
- b. Orthogonal Eigenvectors: For a real symmetric matrix, the eigenvectors corresponding to different eigenvalues are always orthogonal to each other. If the eigenvalues are distinct, you can even choose an orthonormal basis of eigenvectors.

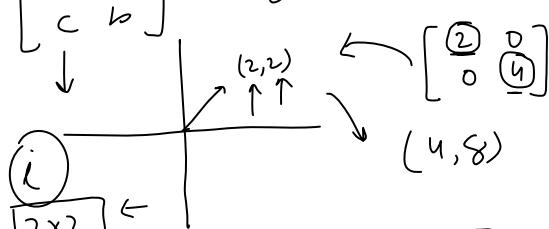
$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \rightarrow A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \leftarrow A^{100}$$

linear transformation

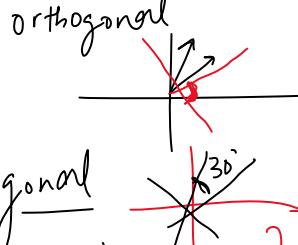
$$A^{100} = A \cdot A \cdot A \cdot \dots \cdot A \quad \text{100 times}$$

$$A = \begin{bmatrix} 5 & 0 \\ 0 & 6 \end{bmatrix} \leftarrow A^{100} = \begin{bmatrix} 5^{100} & 0 \\ 0 & 6^{100} \end{bmatrix}$$

$$A = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \rightarrow \lambda = a, \lambda = b$$



$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \quad \begin{bmatrix} c & 0 \\ 0 & d \end{bmatrix}$$



$$ab + cd = 0$$

$$\sqrt{a^2 + c^2} = 1 \quad \sqrt{b^2 + d^2} = 1$$

$$\text{identity} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

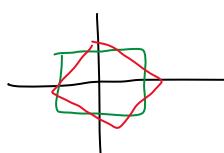
$$\begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \rightarrow \text{orthogonal} \quad \theta \rightarrow 30^\circ$$

$$\begin{bmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{bmatrix}$$

$$\frac{\sqrt{3}}{2} \frac{1}{2} - \frac{1}{2} \frac{\sqrt{3}}{2} = 0$$

$$\sqrt{\left(\frac{\sqrt{3}}{2}\right)^2 + \left(\frac{1}{2}\right)^2} =$$

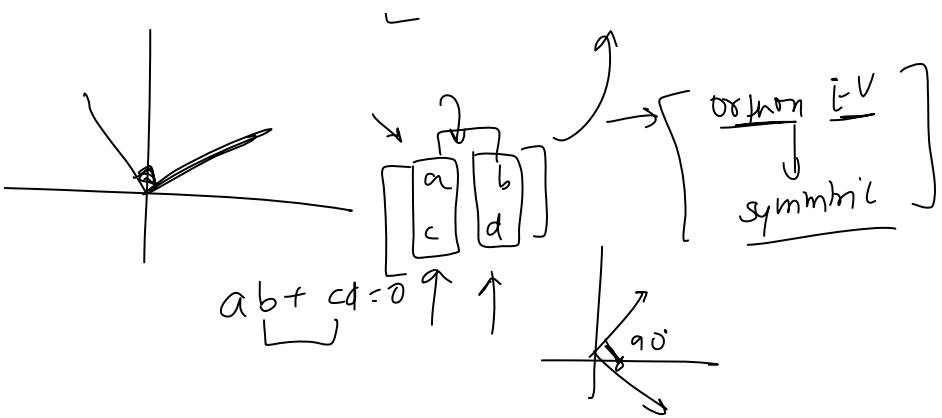
$$\sqrt{\frac{3}{4} + \frac{1}{4}} = \sqrt{\frac{4}{4}} = 1$$



$$\theta = 30^\circ$$

$$\text{cgpq}_{ij} \rightarrow \begin{bmatrix} \text{var}(\text{cgpq}_{ij}) & \text{cov}(\text{cgpq}_{ij}, \text{cgpq}_{ij'}) \\ \text{cov}(\text{cgpq}_{ij'}, \text{cgpq}_{ij}) & \text{var}(\text{cgpq}_{ij'}) \end{bmatrix}$$

$$\text{max } E[V]$$



Matrix Composition

01 June 2023 16:32

Matrix composition

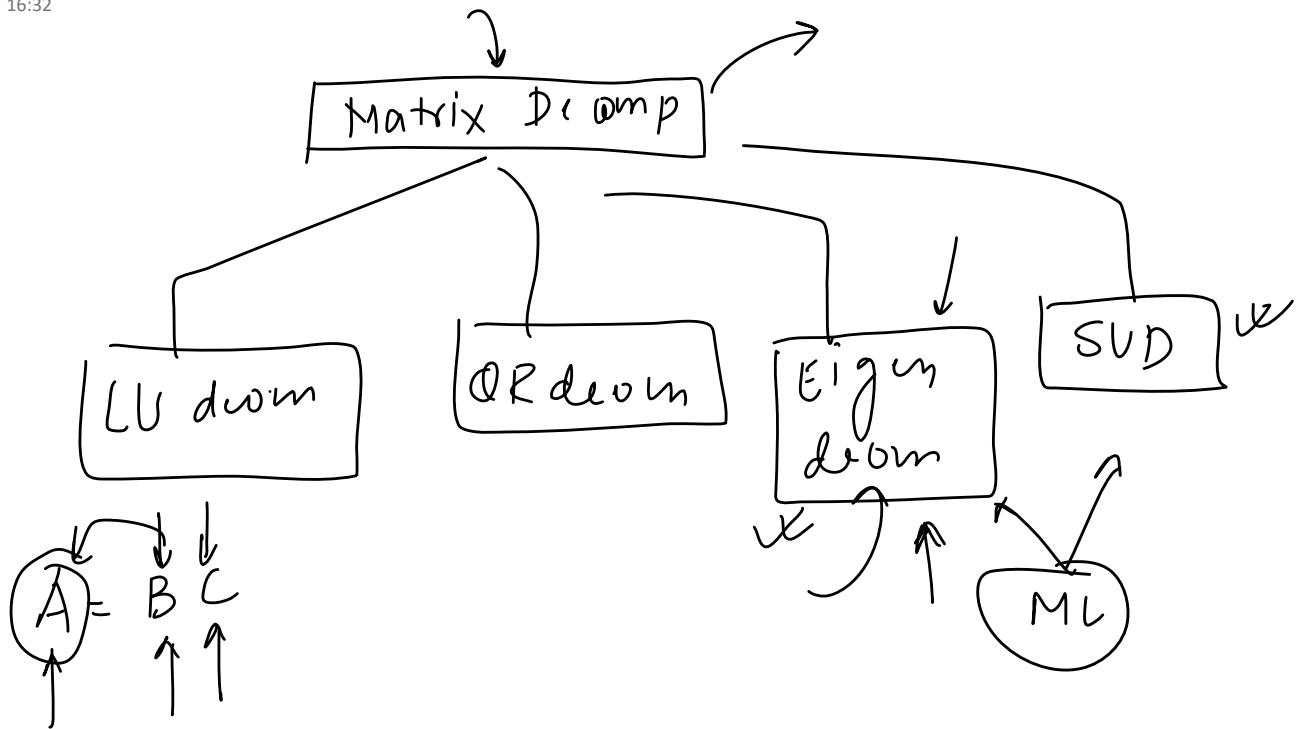
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} \begin{bmatrix} k & l \\ m & n \end{bmatrix} = \begin{bmatrix} \text{comp} \\ ABC \end{bmatrix}$$

$$ABC = D$$

$$D = ABC$$

Matrix Decomposition

01 June 2023 16:32



Eigen Decomposition

01 June 2023 16:32

$$A \rightarrow A = V \Lambda V^{-1}$$

$\uparrow \uparrow$ matrices

The eigen decomposition of a matrix A is given by the equation:

$$A = V \Lambda V^{-1}$$

Where:

$$\boxed{A = V \Lambda V^{-1}}$$

\uparrow Symmetric

- V is a matrix whose columns are the eigenvectors of A
- Λ is a diagonal matrix whose entries are the eigenvalues of A
- V^{-1} is the inverse of V

Assuming

- Square matrix: Eigen decomposition is only defined for square matrices
- Diagonalizability: For a $n \times n$ matrix it should have n linearly independent eigen vectors.

$\begin{matrix} 2 \times 2 \\ \downarrow \\ 2 \text{ eigenvalues} \end{matrix}$

$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$

$$\boxed{2 \times 2}$$

$$\boxed{A \vec{v} = \lambda \vec{v}}$$

↑ eigenvalue ↓ eigenvector

$$\begin{aligned} A \vec{v}_1 &= \lambda_1 \vec{v}_1 \\ A \vec{v}_2 &= \lambda_2 \vec{v}_2 \end{aligned}$$

$$V = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 \end{bmatrix}$$

$$\rightarrow \boxed{A V = V \Lambda}$$

$$\begin{aligned} \vec{v}_1 &= \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \\ v_2 &= \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \end{aligned}$$

$$V = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix}$$

$$\vec{v}_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$A \vec{v}_1 = \lambda_1 \vec{v}_1$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$A \vec{v}_2 = \lambda_2 \vec{v}_2$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \lambda_1 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$\begin{bmatrix} ax_1 + by_1 \\ cx_1 + dy_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 x_1 \\ \lambda_1 y_1 \end{bmatrix} \rightarrow \begin{cases} \frac{ax_1 + by_1}{cx_1 + dy_1} = \lambda_1 \\ ax_2 + by_2 = \lambda_2 x_2 \\ cx_2 + dy_2 = \lambda_2 y_2 \end{cases}$$

$$A V = V \Lambda$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$\begin{bmatrix} ax_1 + by_1 \\ cx_1 + dy_1 \\ ax_2 + by_2 \\ cx_2 + dy_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 x_1 & \lambda_2 x_2 \\ x_1 y_1 & x_2 y_2 \end{bmatrix}$$

$$ax_1 + by_1 = \lambda_1 x_1$$

$$A \vec{v}_1 = \lambda_1 \vec{v}_1$$

$$A \vec{v}_2 = \lambda_2 \vec{v}_2$$

$$A V = V \Lambda$$

$$A = V \Lambda V^{-1}$$

eigen
vectors

Eigen decomposition

diag (eigen value)

Eigen Decomposition of Symmetric Matrix

01 June 2023 16:33

$$A \rightarrow \begin{bmatrix} a & c \\ c & b \end{bmatrix} \quad \underline{\text{symmetric}}$$

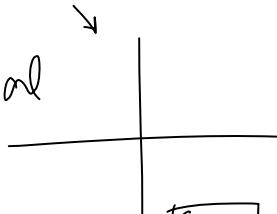
$$\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

diagonal
Orthog
symmetric

$$A = V \Lambda V^{-1} \rightarrow \begin{array}{l} \text{spectral} \\ \text{decomposition} \end{array}$$

eigen vector $V \rightarrow$ orthogonal

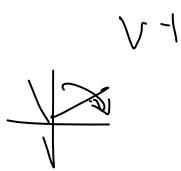
$\Lambda \rightarrow$ diagonal



$[Cov] \rightarrow$ symm

(A)

Symctr \rightarrow Ortho (diag) Orth



$$A = V \Lambda V^{-1}$$

rotates scales \rightarrow rotation

linear from

$A \rightarrow$ symmetric

linear

rotate scale rotate

Advantages of Eigen Decomposition

01 June 2023 16:33

$\rightarrow ml \rightarrow pca$

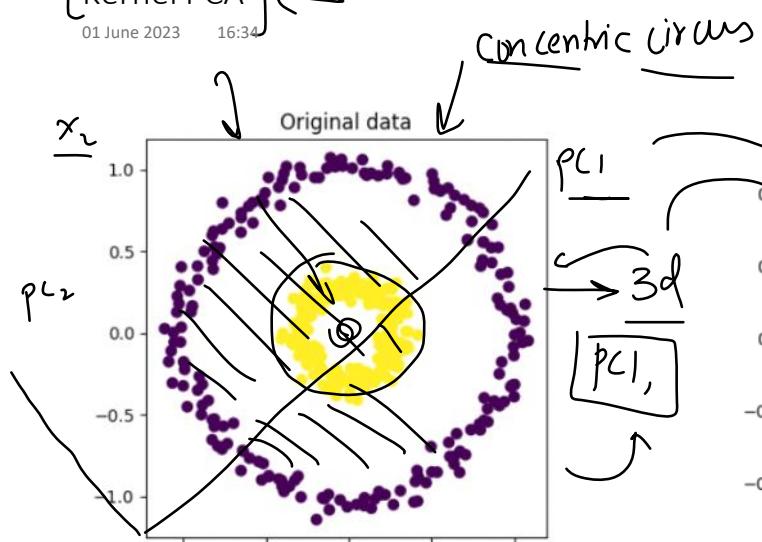
→ Physics

→ singular theory

\rightarrow grammar

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \xrightarrow{\sim} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

[Kernel PCA] ←
01 June 2023 16:34



SVM → Kernel trick

Data after PCA in 1D

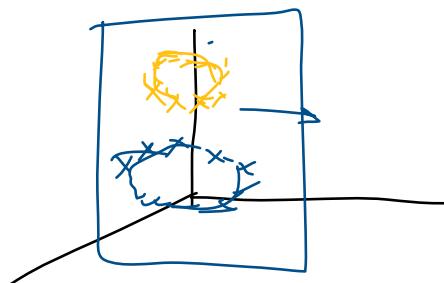
PCA fails

Kernel PCA

mathematical form \rightarrow Kernel $\rightarrow 2d \rightarrow 3d$

$$y = e^{-X^2} \rightarrow \text{soft}$$

PCA assumes that the principal components are a linear combination of the original features. It can't handle complex polynomial relationships between features.

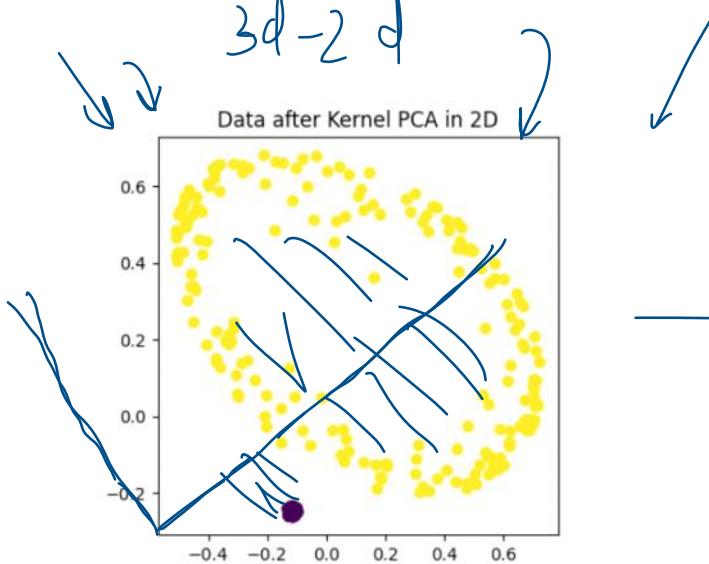


Kernel PCA

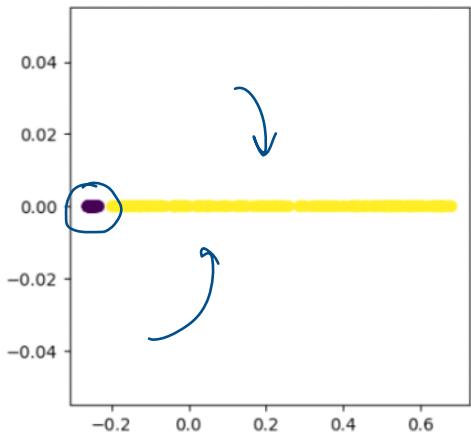
Kernel SVM

3d -> 2d

Data after Kernel PCA in 2D



Data after Kernel PCA in 1D



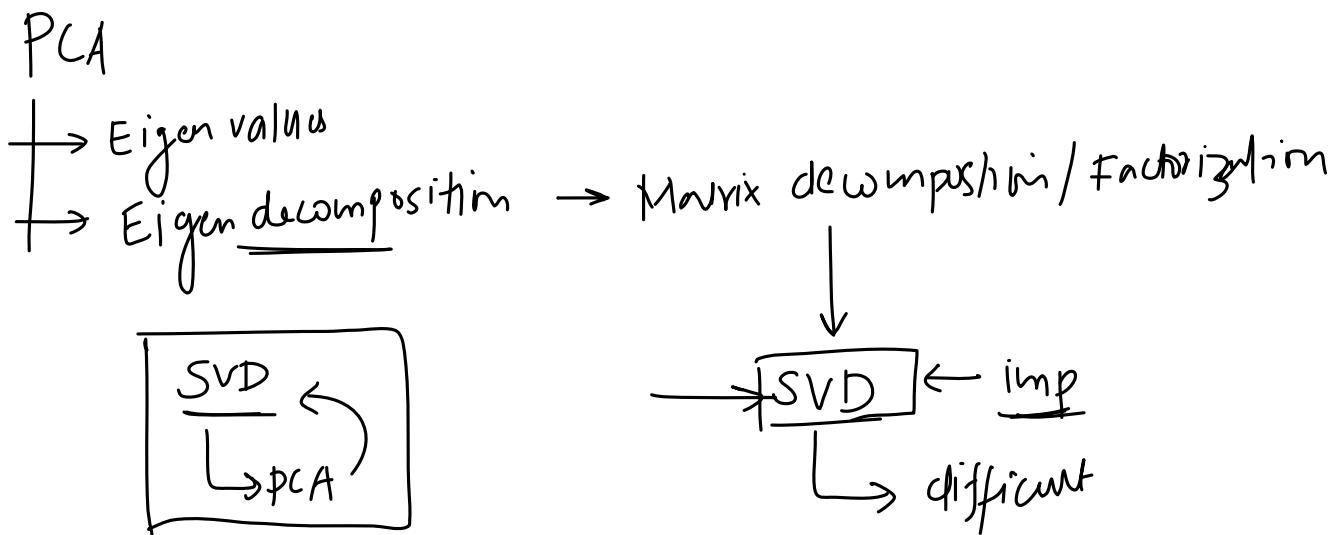
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \stackrel{x^2}{\mapsto} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \stackrel{y}{\mapsto}$$

$$\boxed{X = [100, 2]} \rightarrow$$

$\begin{matrix} 400 \\ \neq \\ 1 \end{matrix}$ | $X = [400, 2] \rightarrow$
 ↓ \downarrow $\boxed{\text{PCI}}$ $\rightarrow \text{best sup}$
 $e^- (\text{gamma} * \underline{\text{distance}}^2)$ $\rightarrow e^- \xrightarrow{\chi^2}$ $\xrightarrow{\text{distance}}$
 $\xrightarrow{\text{gamma}}$ $\xrightarrow{\text{similarity matrix}}$
 $\xrightarrow{(400, 2)}$ $\xrightarrow{\text{high}}$
 $\xrightarrow{2d}$ $\xrightarrow{(400, 400) \rightarrow 400 \text{ d}}$
 $\xrightarrow{\uparrow}$
 $\xrightarrow{\text{down}} \xrightarrow{\text{environ}}$
 $\xrightarrow{\text{down}}$
 $\xrightarrow{\text{down}}$
 $\xrightarrow{\text{down}}$
 $\xrightarrow{\text{down}}$
 $\xrightarrow{400 \times 400}$ $\xrightarrow{\text{400x400}}$
 $\xrightarrow{\text{distance square}}$
 $\xrightarrow{\text{symmetric}}$ $\rightarrow \text{eigen decomposition}$
 $400 \text{ d} \rightarrow \underline{400} \xrightarrow{\text{eigenvalue}}$
 $\xrightarrow{(2) \rightarrow (1) \rightarrow}$

Recap

03 June 2023 10:22



[Non-square Matrix (rectangular)]

03 June 2023 10:23

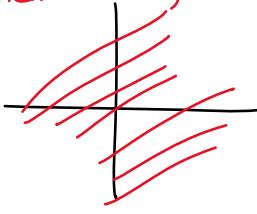
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

2 × 2 2 × 3

→ Square

 Non-square

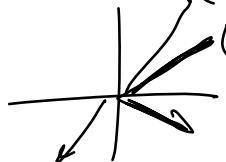
Linear transform



$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

2×2 2×1 $\rightarrow 2 \times 1$

matrix \rightarrow linear transformation



$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix}$$

The diagram illustrates the calculation of element $A_{i,j}$ as the dot product of row i of matrix B and column j of matrix C .

Matrix B (2x2) is shown with elements $B_{1,1} = 1$, $B_{1,2} = 2$, $B_{2,1} = 3$, and $B_{2,2} = 4$. Row i is highlighted in red.

Matrix C (2x2) is shown with elements $C_{1,1} = 1$, $C_{1,2} = 0$, $C_{2,1} = 0$, and $C_{2,2} = 1$. Column j is highlighted in red.

The result of the dot product is $A_{i,j} = 1 \cdot 1 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot 1 = 5$.

$$\begin{matrix} [1 & 3 & 5] \\ [2 & 4 & 6] \end{matrix} \rightarrow \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

2×3

$\uparrow \quad \uparrow$

3×1

output
 $(2d)$ | input $(3d)$

A hand-drawn diagram of a rectangle representing a matrix. The label "m x n" is written inside the rectangle. A bracket below the rectangle spans its width and is labeled "m dim". Another bracket to the right of the rectangle spans its height and is labeled "n dim".

$$\begin{pmatrix} 1 & 0 \\ 2 & b \\ 3 & 1 \end{pmatrix} \quad \begin{pmatrix} 0, 0, 1 \\ 3d \end{pmatrix}$$

$$\begin{array}{c} (1, 2, 3) \\ \hline 3d \text{ cond} \end{array} \quad \begin{array}{c} 3 \times 2 \\ \uparrow \quad \uparrow \end{array} \rightarrow \begin{array}{c} 2d \\ J \\ 3d \end{array}$$

output - input
space space

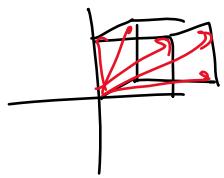
2x1

$$\text{Input}(2d)$$

$m \times n$

$$2 \times 3$$





[Rectangular Diagonal Matrix] ← 2 transformation

05 June 2023 14:05

A matrix that would be diagonal if it were square, but instead is rectangular due to extra rows or columns of zeros

$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

square non-zero items

$$\xrightarrow{\text{rc 4}} \begin{bmatrix} a & 0 & [0] \\ 0 & b & [0] \end{bmatrix}$$

diag
matrix

$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

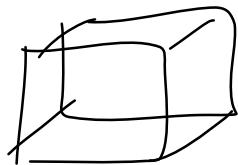
$$\begin{bmatrix} a & 0 \\ 0 & b \\ [0] & [0] \end{bmatrix} \rightarrow \text{diag matrix}$$

2 transformation

$$\begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \end{bmatrix}$$

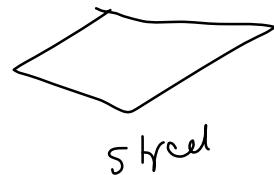
2×3

rect diag → linear



$$\begin{bmatrix} 2 \times 3 \end{bmatrix}$$

$3d \rightarrow 2d$



$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \xrightarrow{\text{2x2}} \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

non-square

(2)

(1)

$\begin{bmatrix} a & 0 \\ 0 & b \\ 0 & 0 \end{bmatrix} \rightarrow \text{linear transformation}$

2×3

3×2

$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

2×2

3×2

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \rightarrow \begin{bmatrix} 4 \times 2 \end{bmatrix}$$

a

$$\textcircled{b} \quad \begin{matrix} 0 & 0 & 1 \\ 3 \times 2 \end{matrix} \quad \begin{matrix} 4 \times 2 \\ a \end{matrix}$$

What is SVD

03 June 2023

10:23

Singular value decomposiⁿ

SVD is a matrix decomposition/factorization method that decomposes a matrix into three other matrices. Given a matrix A, the singular value decomposition of A is usually written as:

$$A = U\Sigma V^T$$

Here:

$$\rightarrow \boxed{A} = \boxed{U} \boxed{\Sigma} \boxed{V^T}$$

- U and V are orthogonal matrices. U is the left singular vectors and V is the right singular vectors.
- Σ is a diagonal matrix containing what we call the singular values.

$$A = U \Sigma V^T$$

↑
Square matrix

Applications of SVD

05 June 2023 05:56

1. Machine Learning and Data Science: SVD is used in Principal Component Analysis (PCA), a technique for dimensionality reduction. This is helpful when dealing with high-dimensional data. It's also used in various recommendation systems, for example in collaborative filtering which is used in Netflix movie recommendation.
2. Natural Language Processing (NLP): SVD is used in Latent Semantic Analysis (LSA), a technique for extracting the underlying meaning (semantic information) from textual data. LSA uses SVD to reduce the dimensionality of a term-document matrix, which helps identify relationships between terms and documents.
3. Computer Vision: In computer vision, SVD is used in image compression. By keeping only the largest singular values and corresponding singular vectors, we can represent an image using less data without losing too much information.
4. Signal Processing: SVD is used to separate useful signals from noise. This is useful in applications like mobile communications and audio signal processing.
5. Numerical Linear Algebra: SVD is used for matrix inversion and solving systems of linear equations. It is often a numerically stable way to solve ill-conditioned systems.
6. Psychometrics: In psychology and education, SVD is used in the construction and scoring of psychological and educational tests, where it is often important to extract underlying latent traits.
7. Bioinformatics: SVD and related techniques are often used to analyze gene expression data, where it is important to identify the underlying patterns of gene activity.
8. Quantum Computing: SVD is also used in quantum state tomography to understand the state of a quantum system.

→ Linear Algebra
↓
SVD

SVD The Equation

03 June 2023 10:23

any matrix

definite

Orthogonal matrix

rectangular

diagonal

or orthogonal

$$A = \underline{U} \Sigma \underline{V}^T$$

$\begin{matrix} A \\ \downarrow \\ (\underline{m \times n}) \end{matrix}$

$\begin{matrix} \Sigma \\ \downarrow \\ (\underline{m \times m}) \end{matrix}$

$\begin{matrix} V^T \\ \downarrow \\ (\underline{n \times n}) \end{matrix}$

$m \times n$

$m \times n$

$\rightarrow \begin{bmatrix} a & c \\ b & d \end{bmatrix}$

$a_c + b_d = 6$

$90^\circ \rightarrow \sqrt{a^2 + b^2} = 1$

$\cup \rightarrow ?$ $\sum \rightarrow ?$ $\vee \rightarrow ?$

Eigen decomposition

$$A = V \lambda V^{-1}$$

$$A = V \Lambda V^{-1}$$

↓

\rightarrow

Symmetric

\uparrow \uparrow

diagonal
orthogonal

Relationship with Eigen Decomposition

03 June 2023 10:39

$$\sqrt{a^2 + b^2}$$

$$\boxed{\bar{A} = U \Sigma V^T}$$

(m x n) non-square square

$$\begin{matrix} a & b \\ \downarrow & \downarrow \end{matrix} \quad \begin{matrix} A = V \Lambda V^{-1} \\ \downarrow \text{symmetric} \\ V^{-1} = V^T \\ \boxed{VV^{-1} = I} \end{matrix}$$

$$\left\{ \begin{array}{l} \boxed{A A^T = U \Sigma V^T (U \Sigma V^T)^T} \\ \boxed{A^T A = U \Sigma V^T V \Sigma^T U^T} \\ = U \Sigma \Sigma^T U^T \end{array} \right.$$

$$V \rightarrow$$

$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \\ 2 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \\ 2 & 0 & 1 \end{bmatrix}$	2×3	$\frac{3 \times 3}{\text{Sqr}}$
$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 1 & 1 \end{bmatrix}$	2×3	$\frac{3 \times 2}{\text{Sqr}}$
$\begin{bmatrix} 6 & 1 \\ 1 & 1 \end{bmatrix}$	$\boxed{2 \times 2}$	Sqr	

symmetric

$$\boxed{A A^T = U X U^T} \quad \text{where } X = \Sigma \Sigma^T$$

$$\begin{aligned} VV^T &= I \\ V^T &= V^{-1} \end{aligned}$$

$$\begin{aligned} A^T A &= (U \Sigma V^T)^T U \Sigma V^T \\ &= V \Sigma^T U^T U \Sigma V^T \\ &= V \Sigma^T \Sigma V^T = \end{aligned}$$

$$\boxed{A^T A = V Y V^T} \quad \text{where } Y = \Sigma^T \Sigma$$

$$\boxed{U^{-1} = U^T}$$

$$\boxed{A A^T = U X U^T} \quad \xrightarrow{\text{eigen values}} \quad X = \Sigma \Sigma^T$$

$$\boxed{A^T A = V Y V^T} \quad \xrightarrow{\text{eigen values}} \quad Y = \Sigma^T \Sigma \quad A = U$$

symmetric Eigen $U \rightarrow$ matrix whose cols contain eigenvectors of $A A^T$ all contain

mv 1

$\overbrace{A}^{\text{left singular vector}} \rightarrow \underline{U}$ $\overbrace{A}^{\text{right singular vectors}} \rightarrow \underline{V}$ $\overbrace{\Sigma}^{\text{eigenvalues of } A^T A}$
 $\overbrace{A}^{\text{left singular vector}} \rightarrow \underline{U}$ $\overbrace{A}^{\text{right singular vectors}} \rightarrow \underline{V}$ $\overbrace{\Sigma}^{\text{matrix whose cols contain eigenvectors of } A^T A}$

$$X = \Sigma \Sigma^T$$

$$Y = \Sigma^T \Sigma$$

$$A = U \Sigma V^T$$

$(2 \times 3) \quad 2 \times 2 \quad 2 \times 3 \quad (3 \times 3)$

$$\Sigma = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \end{bmatrix}$$

$$X = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & b \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} a^2 & 0 & 0 \\ 0 & b^2 & 0 \end{bmatrix} \rightarrow a^2 b^2$$

↓ eigenvalues $A^T A$

$$Y = \begin{bmatrix} a & 0 \\ 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} a^2 & 0 & 0 \\ 0 & b^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow a^2 b^2$$

↓ eigenvalues $A^T A$

$$(A^T A) \xrightarrow{a^2 b^2} \frac{\sqrt{a^2}}{\sqrt{b^2}} \xrightarrow{\downarrow} \frac{\sqrt{a^2}}{\sqrt{b^2}} X, Y \rightarrow \frac{\sqrt{a^2}}{\sqrt{b^2}} \xrightarrow{\downarrow} \frac{a}{b} \rightarrow \underline{\underline{SVD}}$$

$\underline{\underline{a, b}} \rightarrow SVD$

$\xrightarrow{\text{sing. value}} \frac{a}{b} \xrightarrow{\text{sing. value}} \frac{\sqrt{a^2}}{\sqrt{b^2}} \xrightarrow{\text{sing. value}} \frac{\sqrt{a^2}}{\sqrt{b^2}} \xrightarrow{\text{sing. value}} \frac{\sqrt{a^2}}{\sqrt{b^2}}$

$\underline{\underline{\text{sqrt(eigenvalue)}}}$

$$A = U \Sigma V^T$$

$\uparrow \quad \uparrow \quad \uparrow$

left singular right singular $(A^T A)$

$\underline{(a, b)} \rightarrow \frac{\sqrt{a^2}}{\sqrt{b^2}}$

$$\begin{array}{l} \text{singular} \\ \text{vector} \\ \text{vektor} \\ (\mathbf{A}\mathbf{A}^T) \end{array} \quad \begin{array}{c} \text{singular} \\ \text{values} \\ \text{singular} \\ \text{Wertes} \end{array} \quad \sqrt{a^2 + b^2} \quad \begin{array}{l} \text{sqrt(eigenvalues)} \\ \hookrightarrow \mathbf{A}^T \mathbf{A} \quad \mathbf{A} \mathbf{A}^T \end{array}$$

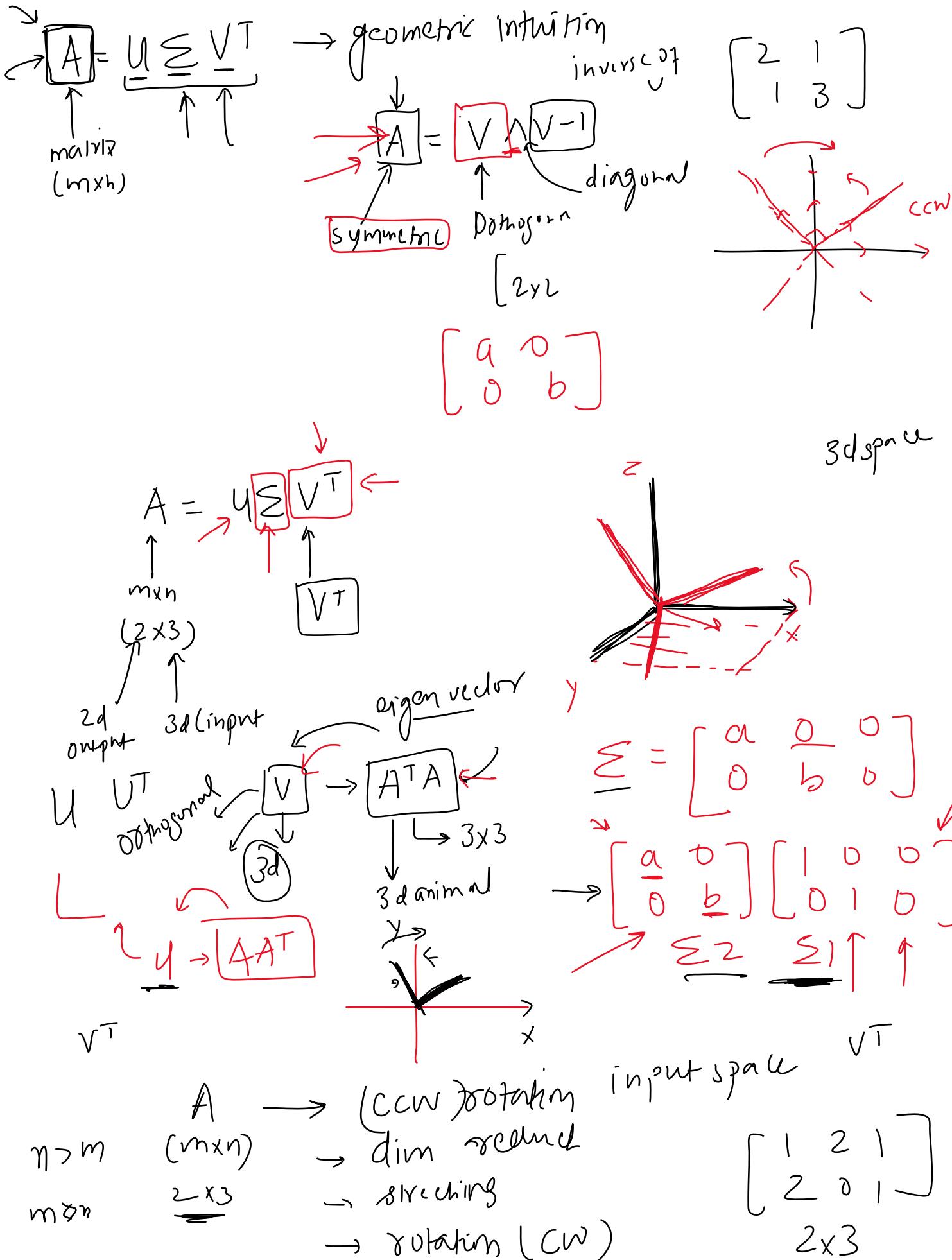
$$\begin{array}{l} \mathbf{A} \mathbf{A}^T \rightarrow a^2 \quad b^2 \\ \mathbf{A}^T \mathbf{A} \rightarrow a^2 \quad b^2 \end{array} \quad \begin{array}{l} \text{eigen} \\ \text{value} \end{array}$$

$$\sum \rightarrow \sqrt{a^2} \quad \sqrt{b^2} \rightarrow \begin{array}{l} \text{eigen} \\ \text{value} \end{array}$$

$$\begin{array}{l} \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \\ \mathbf{U} \rightarrow \mathbf{A} \mathbf{A}^T \\ \mathbf{V} \rightarrow \mathbf{A}^T \mathbf{A} \end{array} \quad \begin{array}{l} \Sigma \rightarrow \sqrt{a^2} \quad \sqrt{b^2} \\ \downarrow \quad \downarrow \\ \frac{a}{\sqrt{a^2}} \quad \frac{b}{\sqrt{b^2}} \\ \text{Singular} \\ \text{value} \end{array}$$

Geometric Intuition

03 June 2023 10:23



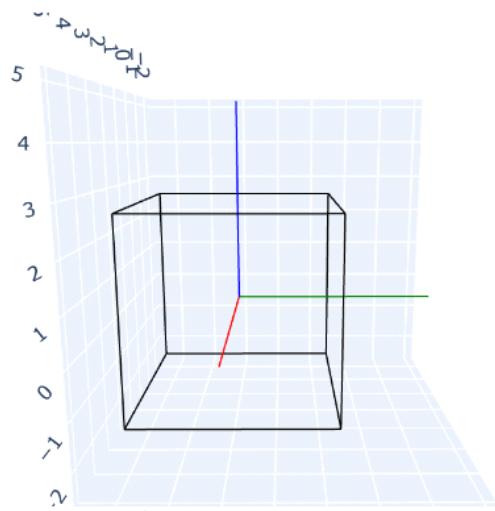
\rightarrow rotation (cw)

2x3

$$A = U \Sigma V^T$$

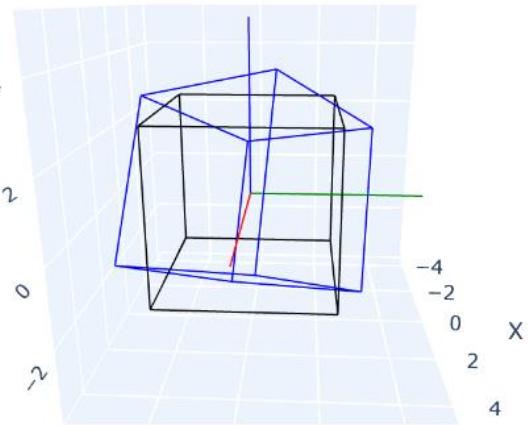
Demo 1 [2x3]

05 June 2023 14:59

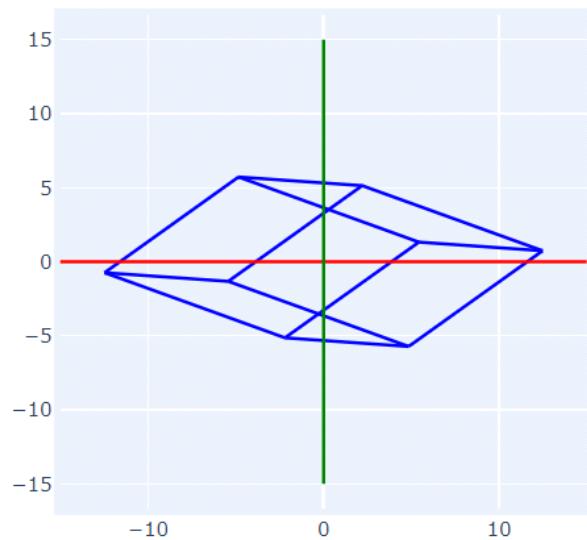


input

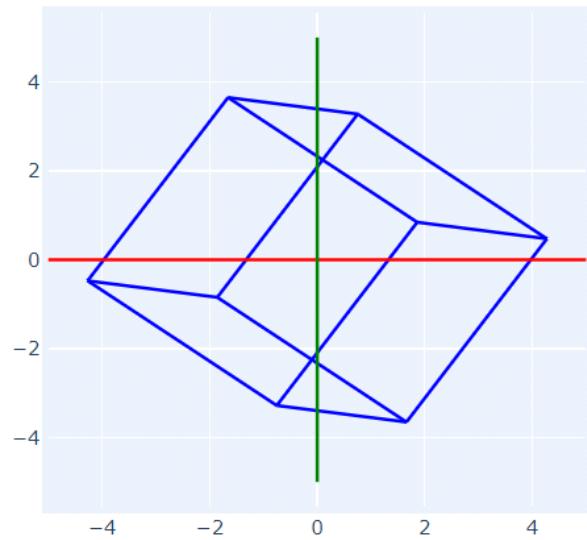
\sqrt{T}



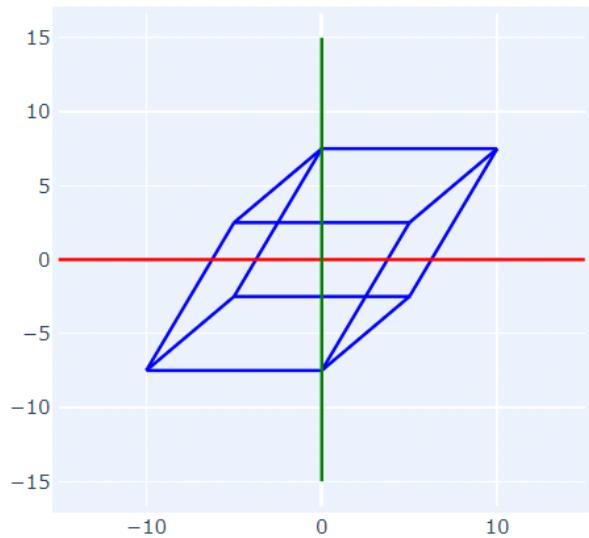
Σ_1



Σ_2

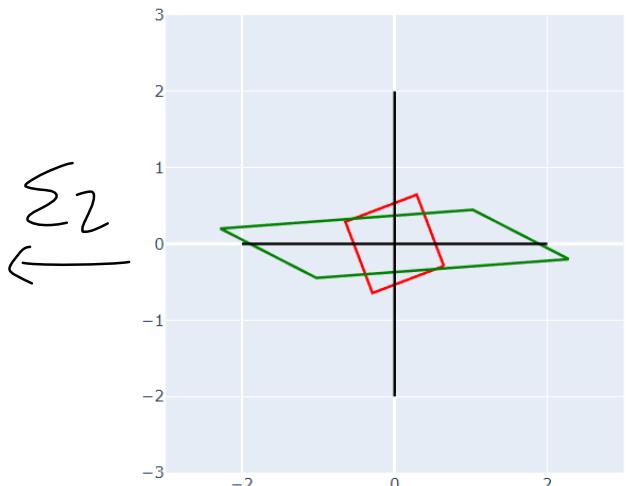
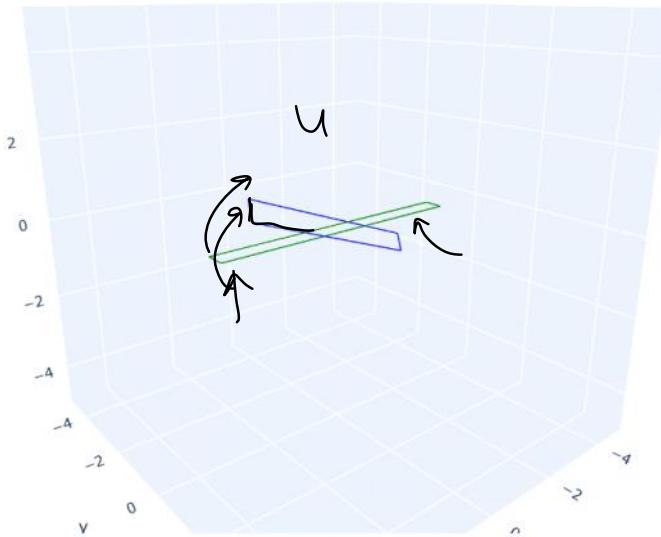
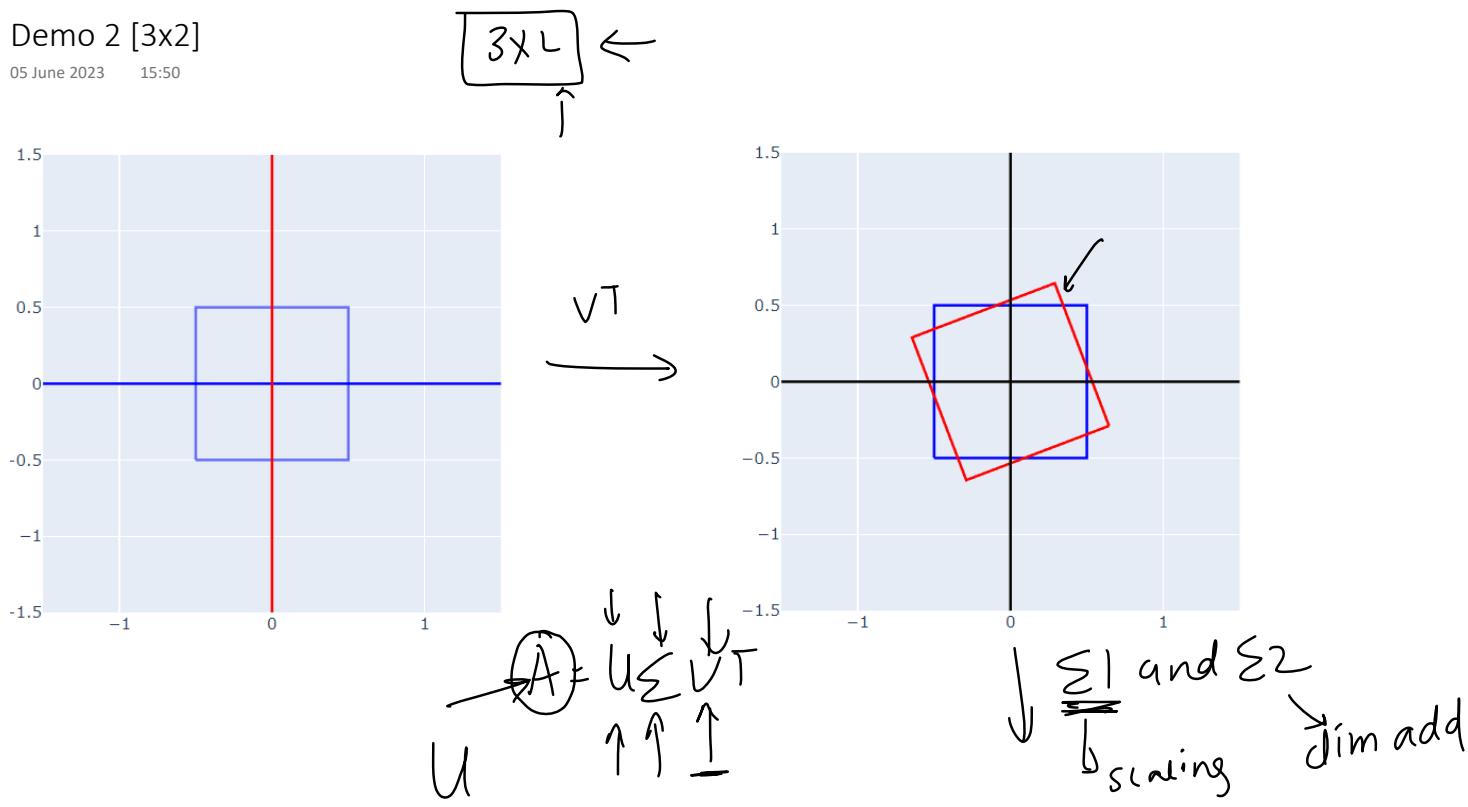


U



Demo 2 [3x2]

05 June 2023 15:50



$$A = U \Sigma V^T \rightarrow \text{Special form of eigen decomposition}$$

$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$ ← scaling

Dimensions: $U: 2 \times 2$, $\Sigma: 2 \times 2$, $V^T: 2 \times 2$

How to Calculate SVD

05 June 2023 08:03

$$\underline{A} = \underline{U} \underline{\Sigma} \underline{V}^T$$

$\boxed{A A^T} = \underline{U} \underline{\Lambda} \underline{U^T}$

$A^T A = \underline{V} \underline{\Lambda} \underline{V^T}$

\downarrow

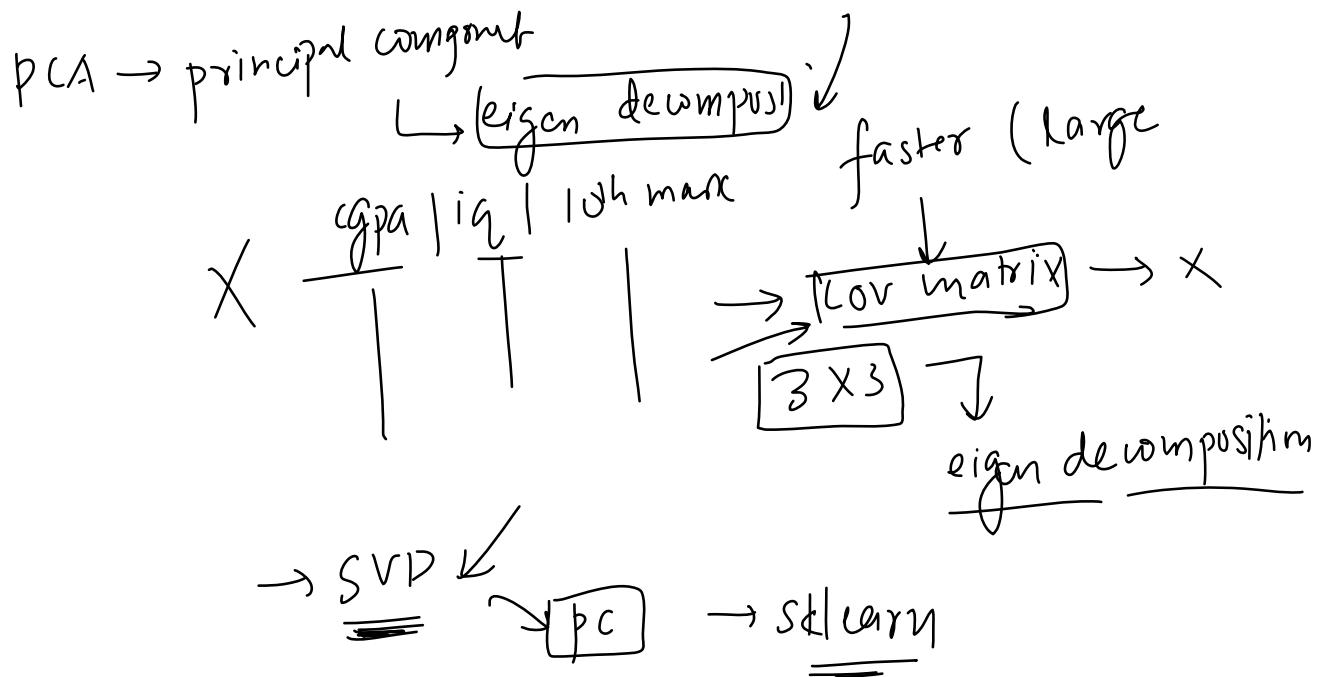
np.linalg.svd(A)

\downarrow

$\boxed{U, S, V^T}$

SVD in PCA

03 June 2023 10:23



iris $X \rightarrow (150, 4)$ $\xrightarrow{\text{wv}} (4, 4)$ $\begin{bmatrix} \dots & \dots & \dots \\ - & - & - \\ - & - & - \\ - & - & - \end{bmatrix}$

$SL | SW | PL | PW$
 $\frac{+}{+} \quad \frac{+}{+} \quad \frac{+}{+} \quad \frac{+}{+}$
 $\text{Cov}(X) \rightarrow$

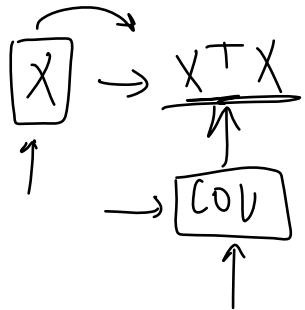
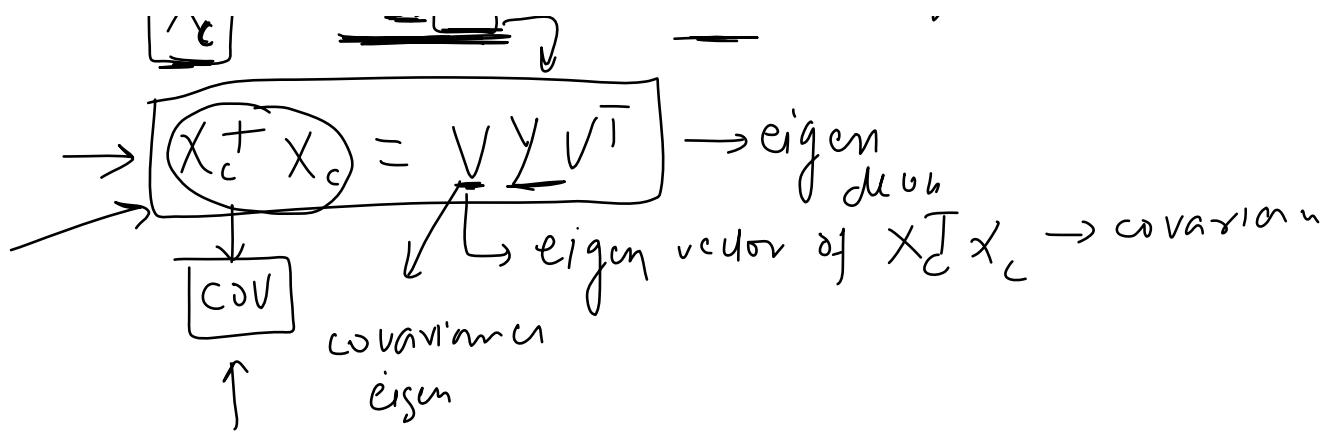
1) mean center all the w's

$\frac{X_C^T X_C}{150-1} \xrightarrow{X \rightarrow X_C (150, 4)} \xrightarrow{2) \frac{X^T X}{n-1} = \text{cov matrix}}$

$$\text{COV}(X) = \frac{X_C^T \cdot X_C}{n-1}$$

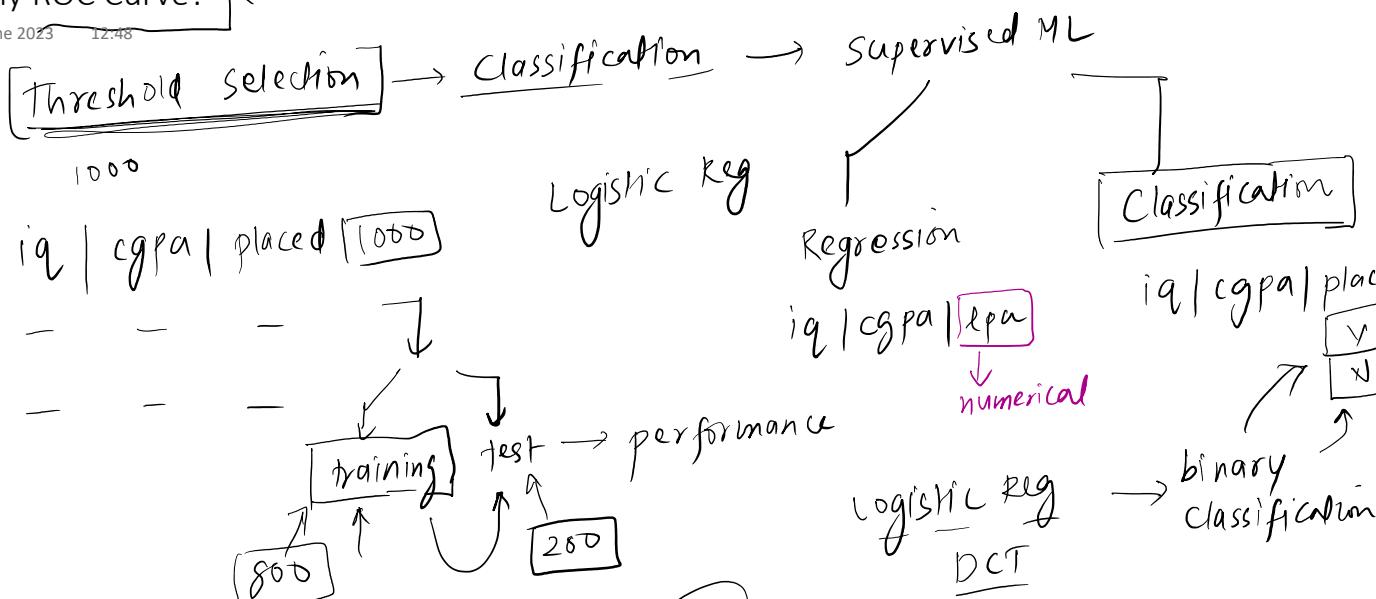
$X_C = U \Sigma V^T$ $\xrightarrow{\text{SVD}}$

eigenvector $\xrightarrow{\Sigma}$ eigenvalues



Why ROC Curve?

06 June 2023 12:48



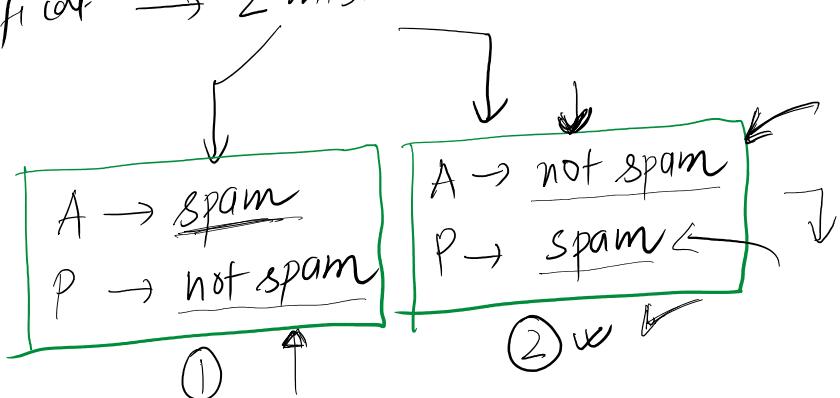
iq cgpa placed			prediction	pred-prob
7	70	1		
8	80	0	0	0.39
9	90	1	1	0.61

email spam classification

→ emails → new email
↓
spam or not spam

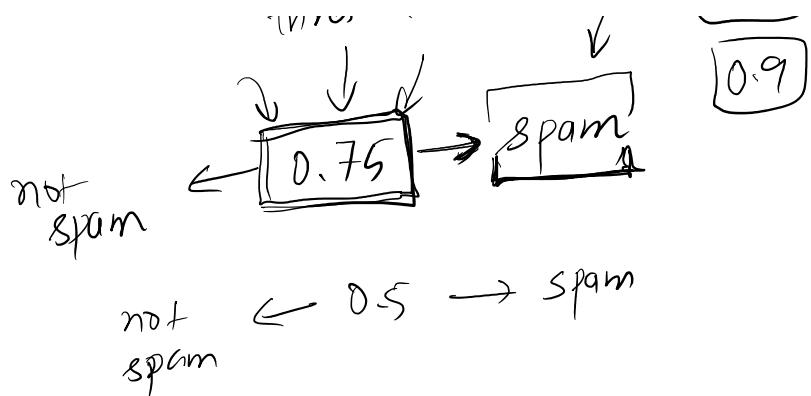
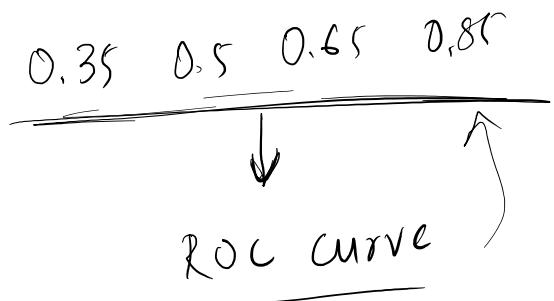
binary classifier → 2 mistakes

placement
A → 1 A → 0
P → 0 P → 1



threshold → 0.5
↓ ↓ ↓
0.85 0.9

$r \rightarrow 0$



Confusion Matrix

06 June 2023 12:48

		grid		Correct		Predicted	
		Predicted		1000	800	1	0
		Actual	1	100	800	1	0
1			True Positive	100	800	1	0
0			False Positive	0	0	1	0
	1		False Negative	0	0	0	1
	0		True Negative	15	80	15	80

Report card → binary classification

True Positive Rate (TPR) \rightarrow Benefit $\leftarrow \underline{100\%}$

06 June 2023 12:48

	1	0
1	TP	FN
0	FP	TN

Prediction

email spam classifier

1 \rightarrow spam
0 \rightarrow not spam

$\leftarrow \underline{100\%}$

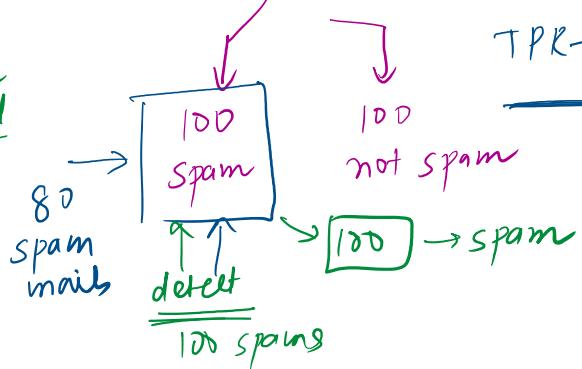
$$\underline{\text{TPR}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Spam		not spam	
Spam	80	20	not spam
Spam	20	80	not spam

$$\text{TPR} = \frac{80}{80+20} = \frac{80}{100}$$

TPR \rightarrow 80%

test \rightarrow 200 emails



churn rate predict

\rightarrow Netflix

\rightarrow 100 customers

leave

\downarrow

detect

80	20
----	----

$$\text{TPR} = 80\%$$

	1	0
1	TP	FN
0	FP	TN

1 \rightarrow leave
0 \rightarrow not leave

False Positive Rate (FPR) \rightarrow Cost model

06 June 2023 12:49

		1	0
Actual	1	TP	<u>FN</u>
	0	<u>FP = 0</u>	TN = 0
Prediction			

email spam \rightarrow

churn predict \rightarrow

$$\underline{FPR} = \frac{FP}{FP + TN}$$

email spam classifier

100 \rightarrow not spam
 \hookrightarrow spam

not spam \rightarrow spam

platforms \rightarrow usex

100 not spam \rightarrow 20 \leftarrow
 \hookrightarrow spam \leftarrow

20
100

100 \rightarrow 10 \rightarrow
 \hookrightarrow discount

$$\underline{FPR} = \frac{FP}{FP + TN}$$

1 \rightarrow 100%
TN \rightarrow D

0 \leftarrow
FP = 0

The best case

$$\begin{cases} TPR = 100\% \text{ or } 1 \\ FPR = 0\% \text{ or } 0 \end{cases}$$

best case

FN = 0
FP = 0

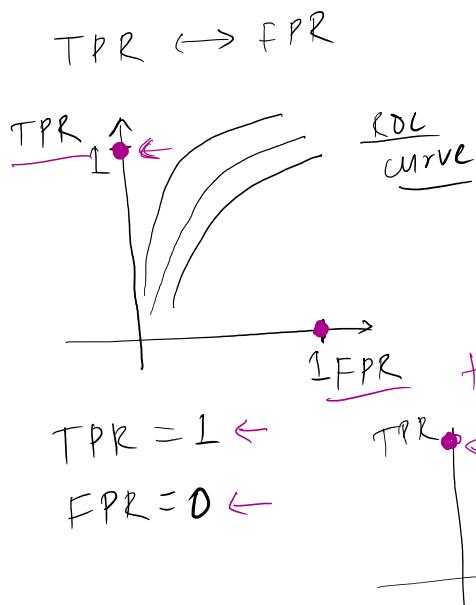
		1	0
Actual	1	TP	<u>FN</u>
	0	<u>FP</u>	TN
Prediction			

best case \rightarrow

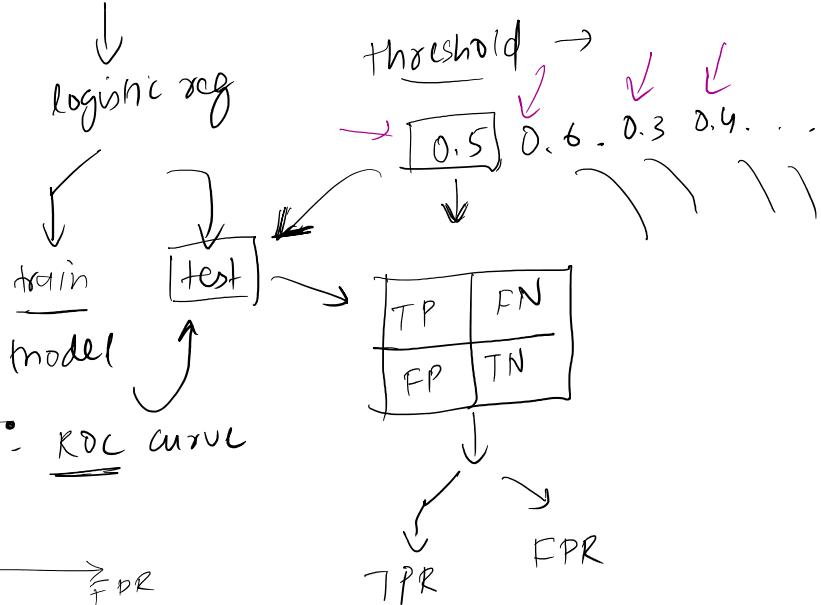
<u>100</u>	<u>0</u>
<u>0</u>	<u>100</u>

ROC Curve

06 June 2023 12:49

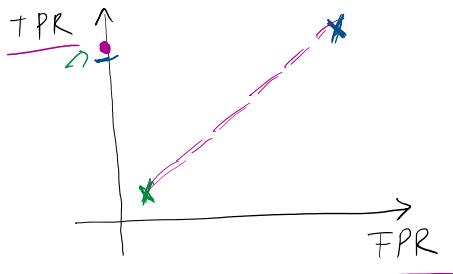


cgpa | iq | placement



Different Cases

06 June 2023 12:49



	1	0
1	TP ↓	FN ↑
0	FP ↓	TN ↑

Predicted

spam $\underline{0.95} >$ spam
email spam

mails \rightarrow spam

$[0.1] \rightarrow$

$[0.2] \rightarrow$ spam

$0.09 \rightarrow$ not spam

$$TPR = \frac{TP}{TP + FN}$$

increase

$$FPR = \frac{FP}{FP + TN}$$

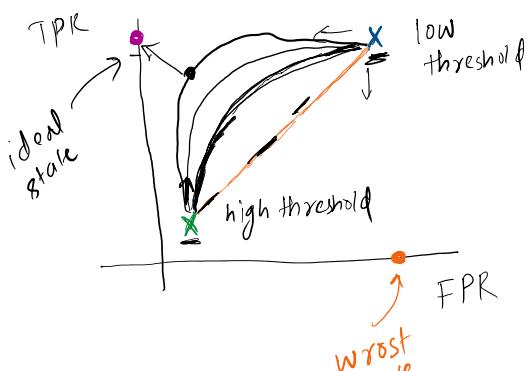
increase

threshold → $\frac{FPR}{TPR}$

$FPR \downarrow$
 $TPR \uparrow$



decrease threshold → threshold
increase threshold



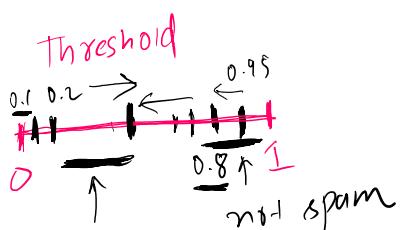
	1	0
1	TP ↑	FN ↓
0	FP ↑	TN

Prediction

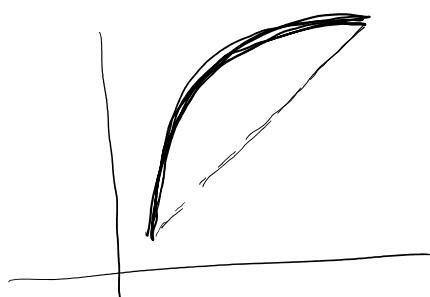
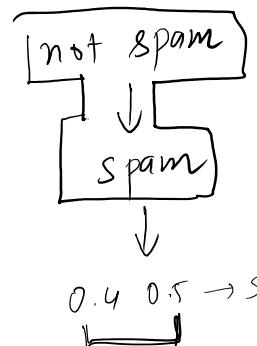
TPR

FPR

$|TPR| = \uparrow$



$TPR \downarrow FPR \downarrow$
threshold → $[0.8] \rightarrow$ relax → spam
 $0.95 \rightarrow$ prob-spam
 $TPR \uparrow FPR \uparrow$
 $0.8 \rightarrow$ spam



Code Example

06 June 2023 12:49

AUC-ROC

06 June 2023 12:49

The AUC-ROC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds.

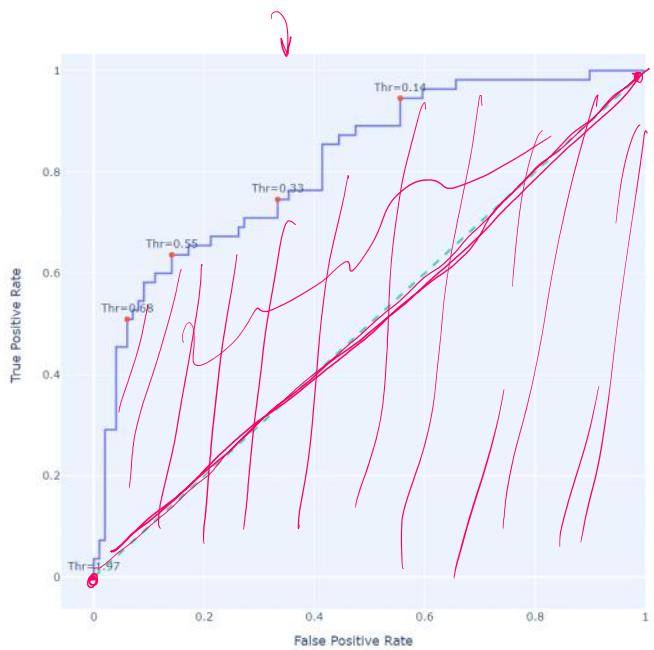
- An AUC of 1 indicates that the model has perfect discrimination: it correctly classifies all positive and negative instances.
- An AUC of 0.5 suggests the model has no discrimination ability: it is as good as random guessing.
- An AUC of 0 indicates that the model is perfectly wrong: it classifies all positive instances as negative and all negative instances as positive.

In practice, AUC values usually fall between 0.5 (random) and 1 (perfect), with higher values indicating better classification performance.

AUC = 1

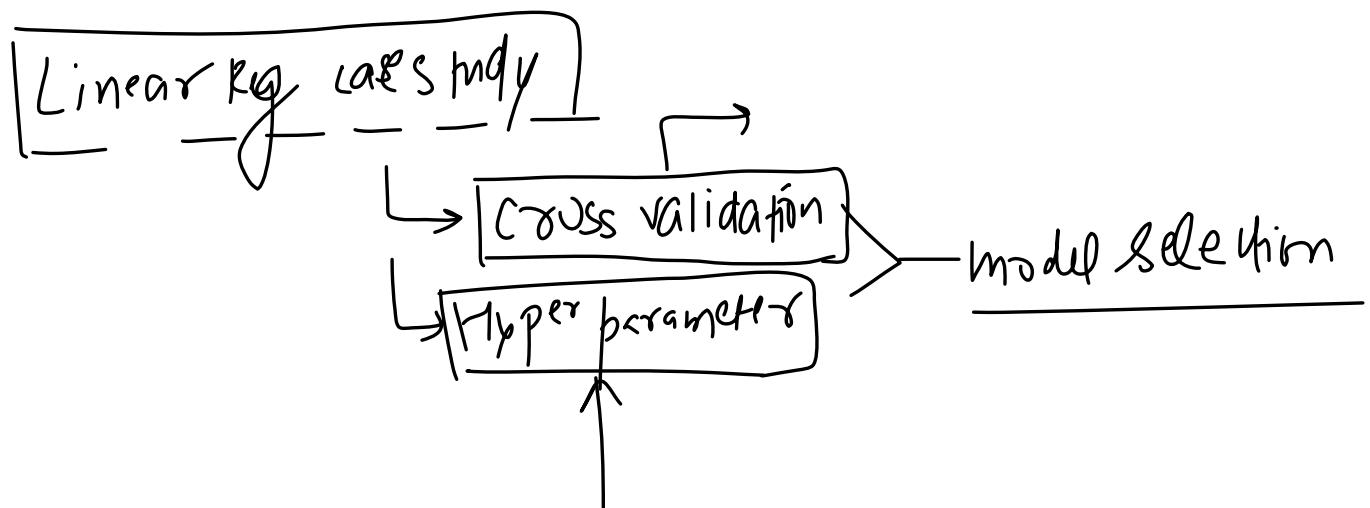
model

AUC → compare



Recap

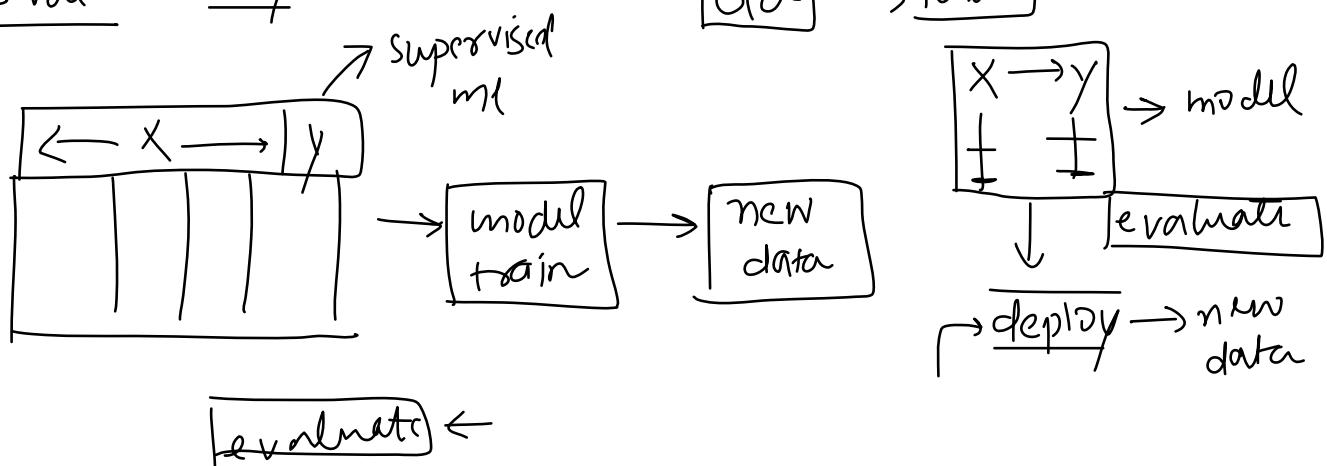
09 June 2023 19:56



The Problem

10 June 2023 12:31

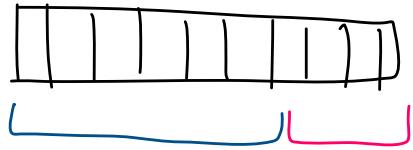
Cross val \rightarrow why



The Hold-out Approach

10 June 2023 12:31

1000 customers \hookrightarrow train-test-split



$\xrightarrow{0.75}$ training $\xrightarrow{0.25}$ test

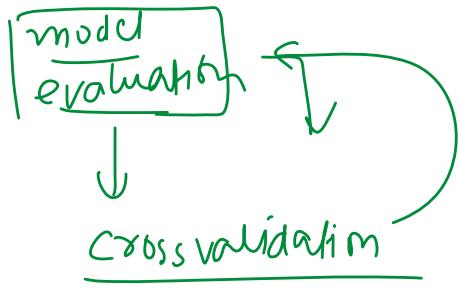
train \rightarrow model

- 1) shuffle
- 2) train, test

y_{pred}

\downarrow

y_{true}

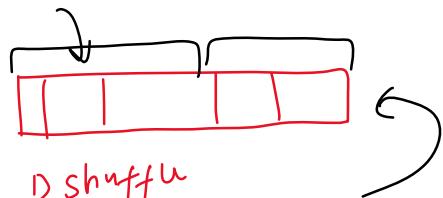


Problem with Hold-out Approach

10 June 2023 12:32

\downarrow $\mu \pm \sigma$

more data better model



1. **Variability**: The performance of the model can be very sensitive to how the data is divided into training and testing sets. If the split is unfortunate, the training set may not be representative of the overall distribution of data, or the test set might contain unusually easy or difficult examples. This leads to high variance in the estimation of the model's performance.

2. **Data inefficiency**: The holdout method only uses a portion of the data for training and a different portion for testing. This means that the model doesn't get to learn from all available data, which can be particularly problematic if the dataset is small.

3. **Bias in performance estimation**: If some classes or patterns are over- or under-represented in the training set or the test set due to the random split, it can lead to a biased performance estimation. \downarrow High bias

data leakage

4. Less reliable for hyperparameter tuning: If the holdout method is used for hyperparameter tuning, there's a risk of overfitting to the test set because information might leak from the test set into the model. This means that the model's performance on the test set might be overly optimistic and not representative of its performance on unseen data.

100%
80%

80% training

20% percent

0.72, 0.76, 0.77

test

model deploy

citus \uparrow
reliable
variance

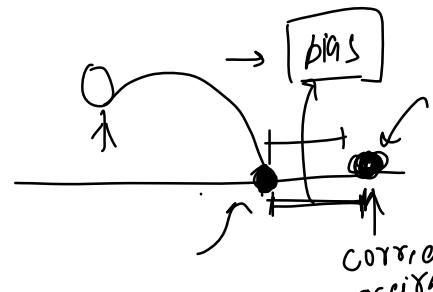
Bias variance trade off

loss

bias

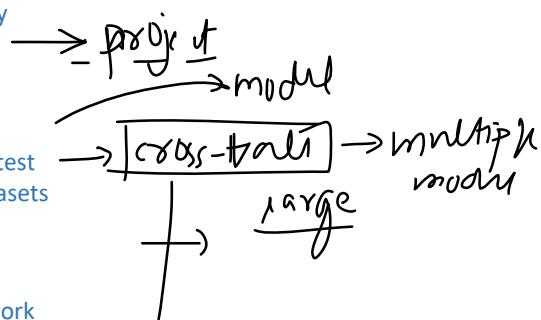
80% \rightarrow bias \uparrow

+ var + noise

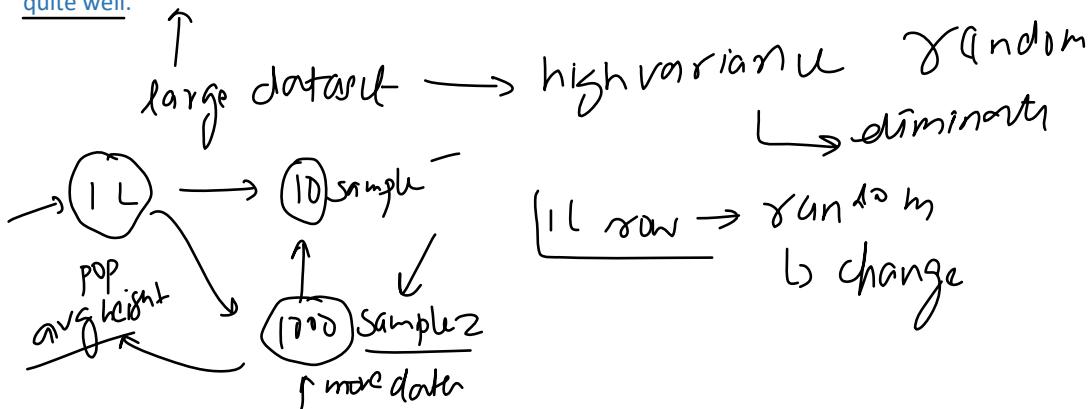


Why is hold-out approach used then?

10 June 2023 12:31



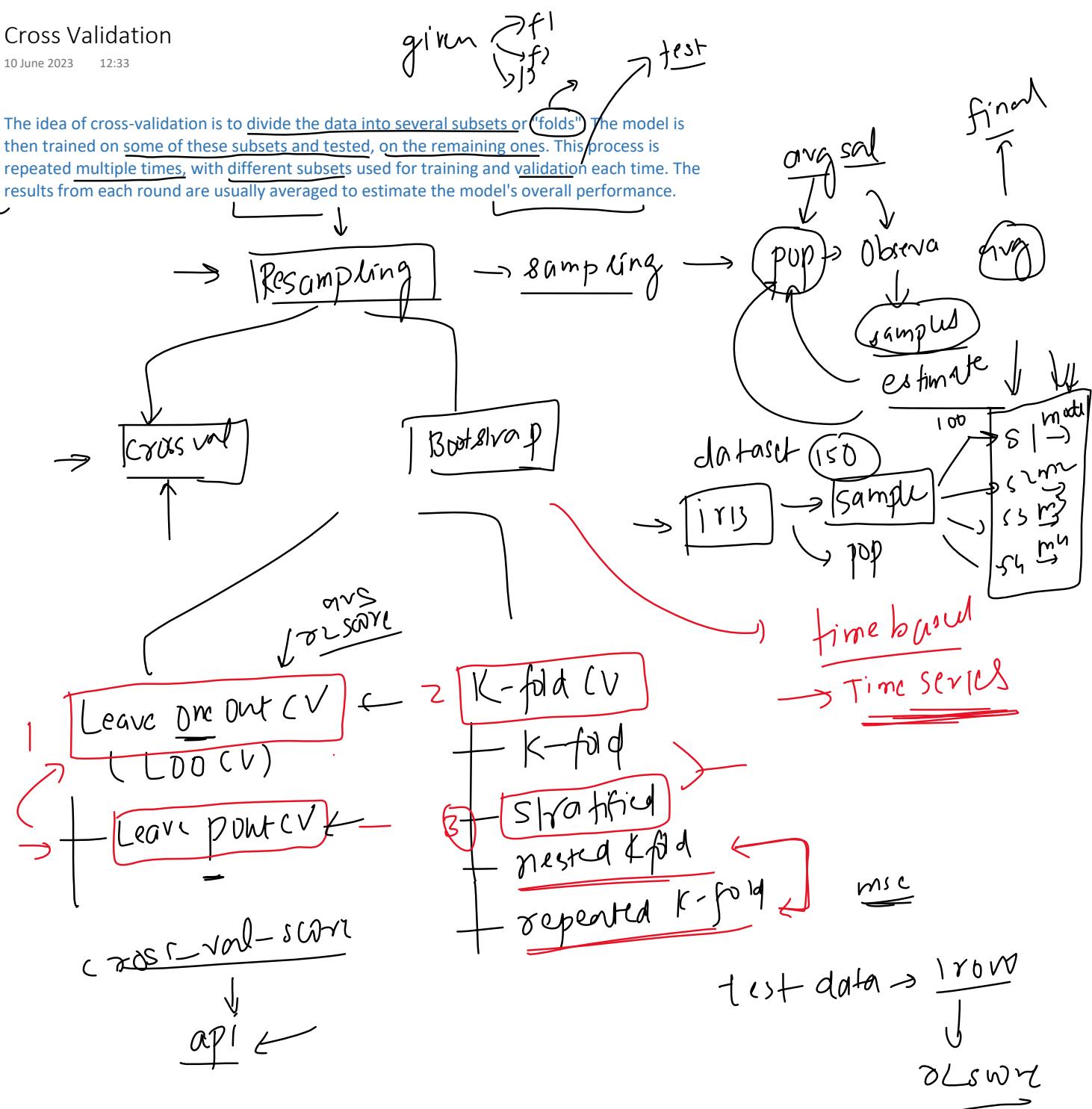
1. Simplicity: The holdout method is straightforward and easy to understand. You simply divide your data into two sets: a training set and a test set. This simplicity makes it appealing, especially for initial exploratory analysis or simple projects.
2. Computational Efficiency: The holdout method is computationally less intensive than methods like k-fold cross-validation. In k-fold cross-validation, you need to train and test your model k times, which can be computationally expensive, especially for large datasets or complex models. With the holdout method, you only train the model once.
3. Large Datasets: For very large datasets, even a small proportion of the data may be sufficient to form a representative test set. In these cases, the holdout method can work quite well.



Cross Validation

10 June 2023 12:33

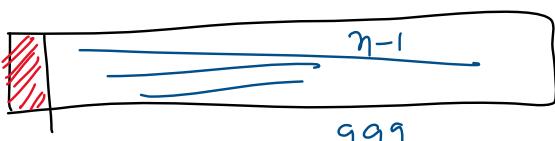
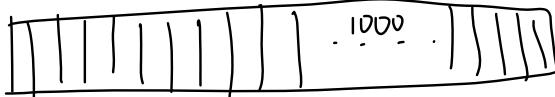
The idea of cross-validation is to divide the data into several subsets or "folds". The model is then trained on some of these subsets and tested on the remaining ones. This process is repeated multiple times, with different subsets used for training and validation each time. The results from each round are usually averaged to estimate the model's overall performance.



Leave One Out Cross Validation (LOOCV)

10 June 2023 12:33

dataset \rightarrow 1000 rows \rightarrow $\frac{\text{models}}{(1000)}$ \rightarrow model evaluation



\rightarrow 999 rows \rightarrow training \rightarrow m_1
1 row \rightarrow test



998 rows \rightarrow m_2
1 row



m_{100}

LOOCV

accuracy

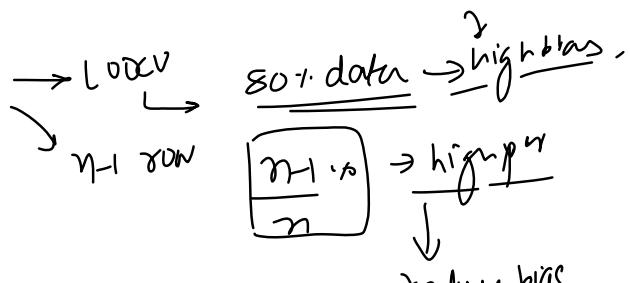
$m_1 + m_2 + \dots + m_{100}$

\downarrow
 $\overline{\text{avg}}$

\downarrow
 $\overline{\text{final accuracy score}}$

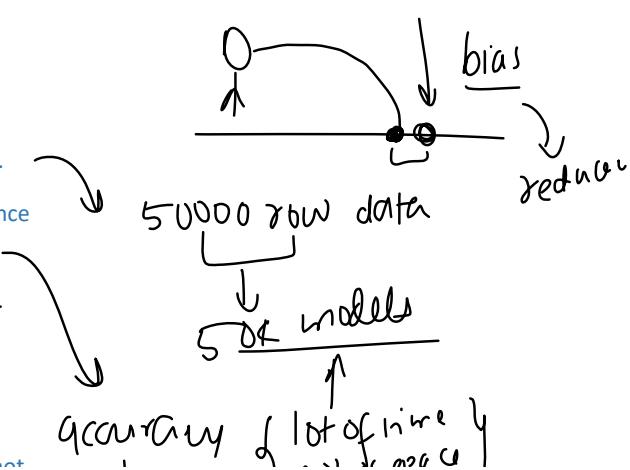
Advantages:

1. Use of Data: LOOCV uses almost all of the data for training, which can be beneficial in situations where the dataset is small and every data point is valuable.
2. Less Bias: Since each iteration of validation is performed on just one data point, LOOCV is less biased than other methods, such as k-fold cross-validation. The validation process is less dependent on the random partitioning of data.
3. No Randomness: There's no randomness in the train/test split, so the evaluation is stable, without variation in the results due to different random splits. (no shuffling)



Disadvantages:

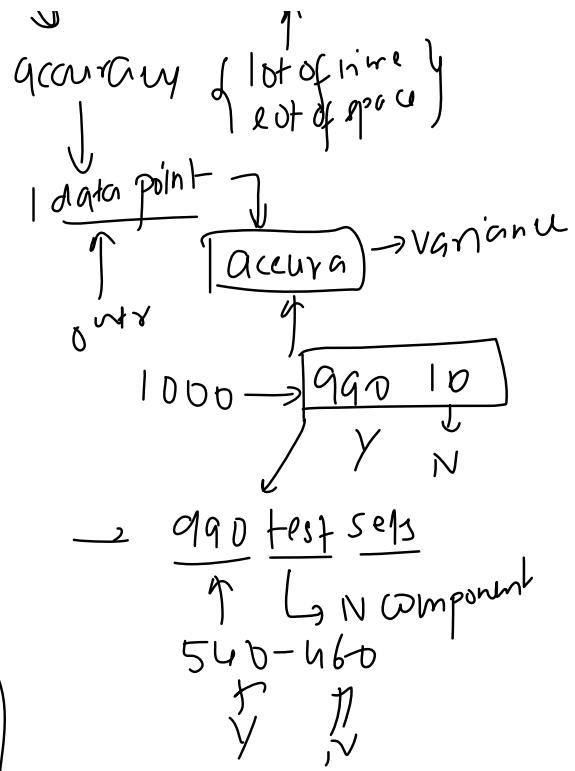
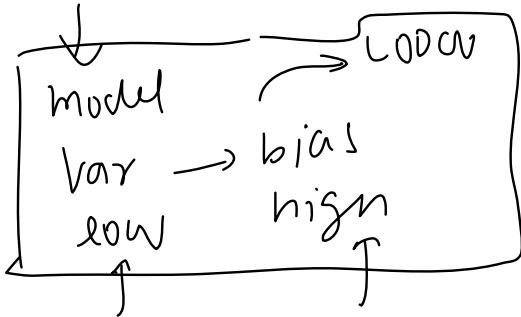
1. Computational Expense: LOOCV requires fitting the model N times, which can be computationally expensive and time-consuming for large datasets.
2. High Variance: LOOCV can lead to higher variance in the model performance since the training sets in all iterations are very similar to each other.
3. Inappropriate Performance Metric: Performance metrics like R^2 are not appropriate to be used with LOOCV as they are not defined when the validation set only has one sample.
4. Not Ideal for Imbalanced Data: In classification problems, if you have imbalanced classes, LOOCV may not provide a reliable estimate of model performance because the single validation sample in each iteration may not

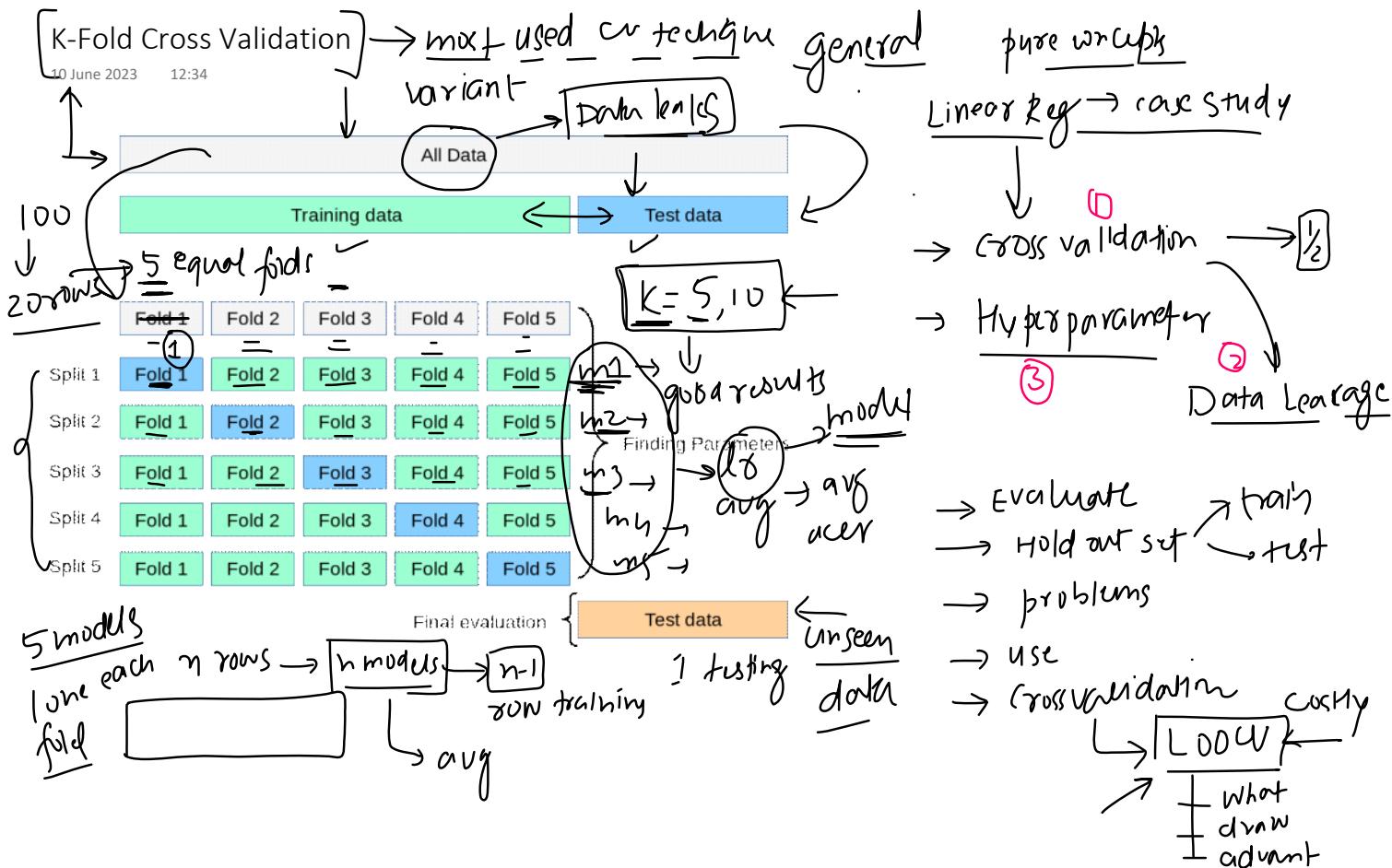


4. Not Ideal for Imbalanced Data: In classification problems, if you have imbalanced classes, LOOCV may not provide a reliable estimate of model performance because the single validation sample in each iteration may not be representative of the overall class distribution.

When to use:

1. Small datasets: LOOCV is most beneficial when you have a limited amount of data. With small datasets, you want to use as much data as possible for training to get a reliable model, which is exactly what LOOCV offers by using all but one data point for training.
2. Balanced datasets: LOOCV might not perform well on imbalanced datasets, especially in classification problems, because the training set might end up missing some classes. Thus, it's more appropriate to use LOOCV when you have a balanced dataset.
3. Need for less biased performance estimate: Since LOOCV uses nearly all the data for training, it gives a less biased estimate of model performance compared to other methods like k-fold cross-validation.





→ K fold $k-1$ fold train
high bias

$n-1$ rows training → bias low

Advantages of K-Fold Cross Validation: ← k -fold → variant

- 1. Reduction of Variance: By averaging over k different partitions, the variance of the performance estimate is reduced. This is beneficial because it means that the performance estimate is less sensitive to the particular random partitioning of the data.
- 2. Computationally Inexpensive: Take less time and space in comparison to LOOCV

Disadvantages of K-Fold Cross Validation:

- 1. Potential for High Bias: If k is too small, there could be high bias if the test set is not representative of the overall population.
- 2. May not work well with Imbalanced Classes: If the data has imbalanced classes, there's a risk that in the partitioning, some of the folds might not contain any samples of the minority class, which can lead to misleading performance metrics.

→ LOOCV → no. of rows / 10000

→ $K=5, 10$ → 5 modules

→ 1 fold → 1 row
multiple → varies
 $1000 \rightarrow 5 \text{ fold}, 200 \text{ rows/fold}$

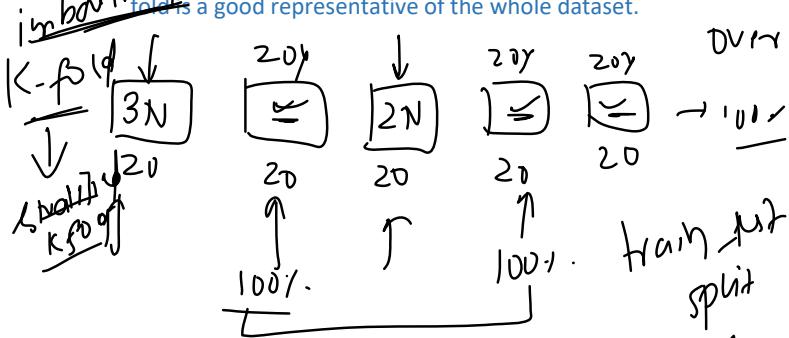
May not work well with imbalanced classes. If the data has imbalanced classes, there's a risk that in the partitioning, some of the folds might not contain any samples of the minority class, which can lead to misleading performance metrics.

Varina K fold

↳ classification
 $\rightarrow 95\% \text{ } 5\%$ \rightarrow 100 rows 5 folds
 $y \quad N$ 

When to use:

- When you have a sufficiently large dataset: K-Fold Cross Validation requires the model to be trained K times, so it can be computationally intensive. However, if your dataset is large enough, this increased computational cost can be justified by the more reliable estimate of model performance.
 - When your data is evenly distributed: K-Fold Cross Validation works best when the data is evenly distributed. If your dataset is imbalanced (i.e., one class has significantly more samples than another), it might be better to use a technique like Stratified K-Fold Cross Validation, which aims to ensure each fold has approximately the same proportion of the whole data.



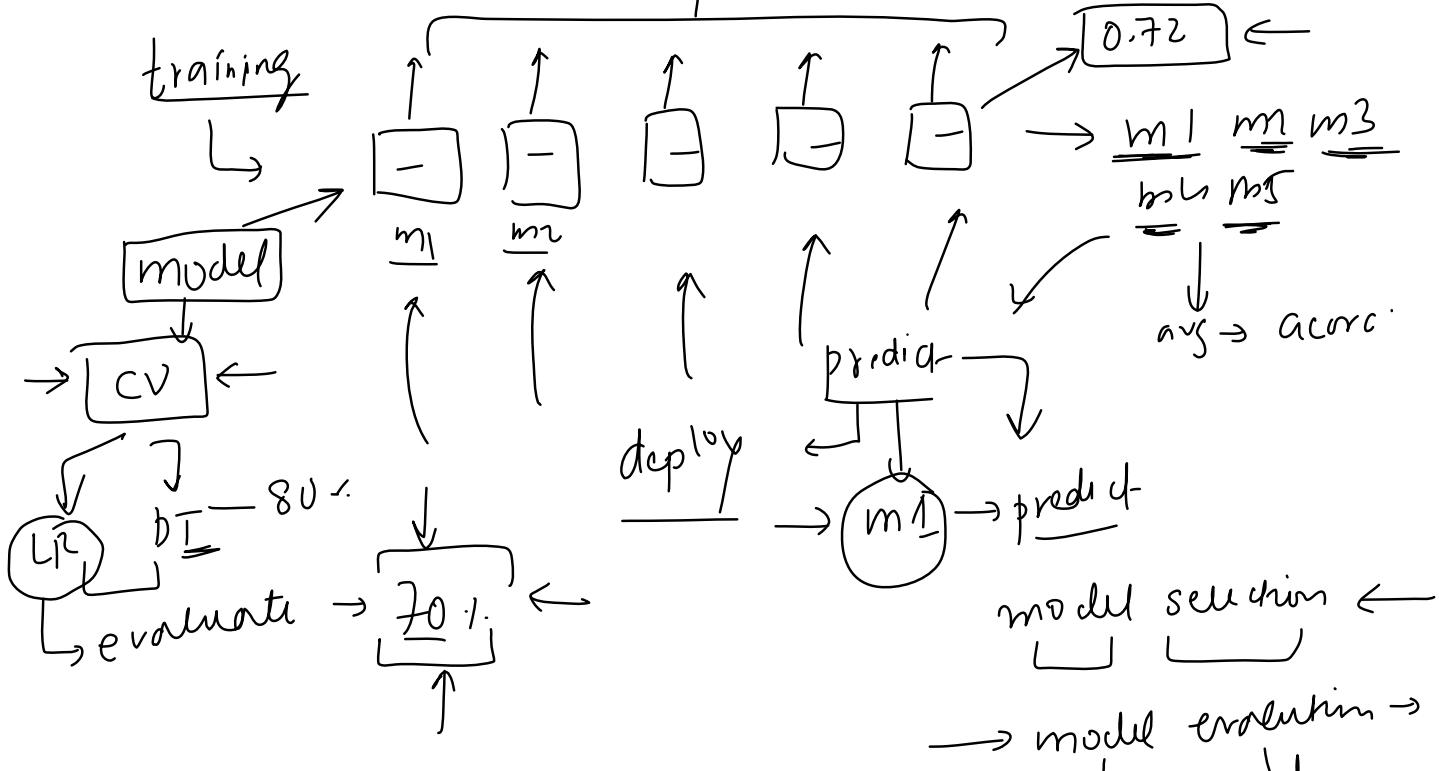
DVR avg 98%

mpg) horsepower

\downarrow poly degree = 1, 2, 3, ..., 6

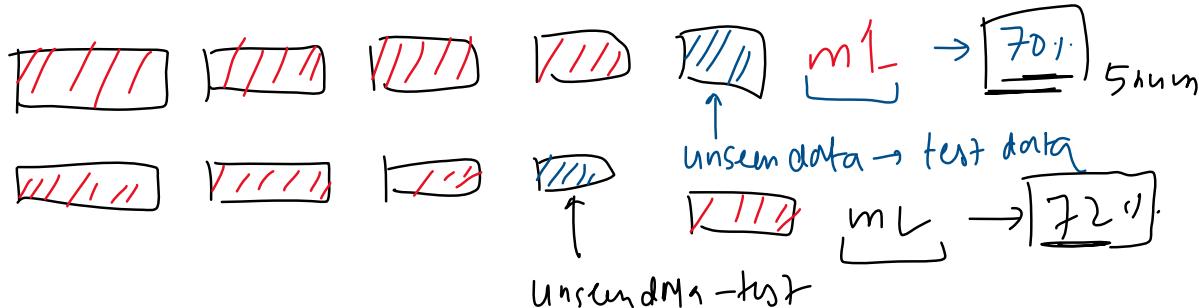
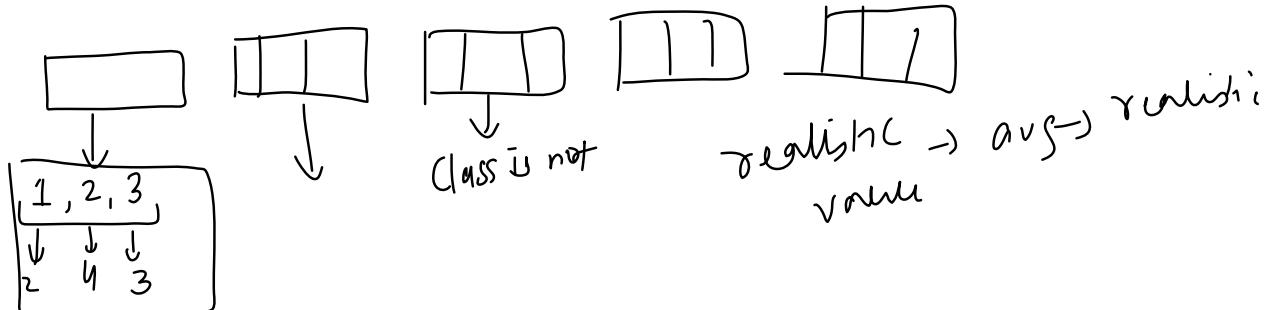
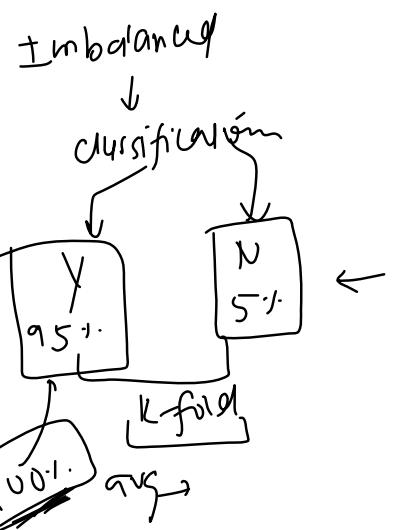
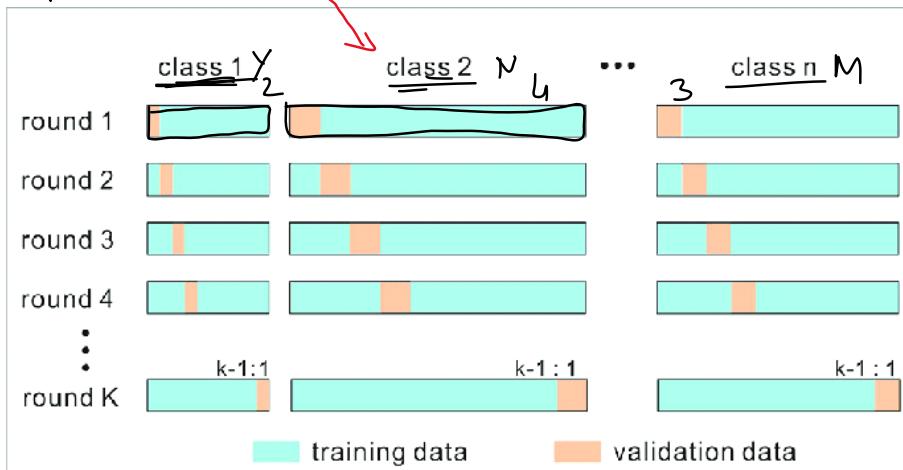
Hold out approach

(random state = 1)



Stratified K-Fold CV

10 June 2023 12:34



m₁ → 70% 5num

m₂ → 72% 5num

m₃ → 71% 5num

m₄ → 73% 5num

m₅ → 74% 5num

m₆ → 75% 5num

m₇ → 76% 5num

m₈ → 77% 5num

m₉ → 78% 5num

m₁₀ → 79% 5num

m₁₁ → 80% 5num

m₁₂ → 81% 5num

m₁₃ → 82% 5num

m₁₄ → 83% 5num

m₁₅ → 84% 5num

m₁₆ → 85% 5num

m₁₇ → 86% 5num

m₁₈ → 87% 5num

m₁₉ → 88% 5num

m₂₀ → 89% 5num

m₂₁ → 90% 5num

m₂₂ → 91% 5num

m₂₃ → 92% 5num

m₂₄ → 93% 5num

m₂₅ → 94% 5num

m₂₆ → 95% 5num

m₂₇ → 96% 5num

m₂₈ → 97% 5num

m₂₉ → 98% 5num

m₃₀ → 99% 5num

m₃₁ → 100% 5num

m₃₂ → 101% 5num

m₃₃ → 102% 5num

m₃₄ → 103% 5num

m₃₅ → 104% 5num

m₃₆ → 105% 5num

m₃₇ → 106% 5num

m₃₈ → 107% 5num

m₃₉ → 108% 5num

m₄₀ → 109% 5num

m₄₁ → 110% 5num

m₄₂ → 111% 5num

m₄₃ → 112% 5num

m₄₄ → 113% 5num

m₄₅ → 114% 5num

m₄₆ → 115% 5num

m₄₇ → 116% 5num

m₄₈ → 117% 5num

m₄₉ → 118% 5num

m₅₀ → 119% 5num

m₅₁ → 120% 5num

m₅₂ → 121% 5num

m₅₃ → 122% 5num

m₅₄ → 123% 5num

m₅₅ → 124% 5num

m₅₆ → 125% 5num

m₅₇ → 126% 5num

m₅₈ → 127% 5num

m₅₉ → 128% 5num

m₆₀ → 129% 5num

m₆₁ → 130% 5num

m₆₂ → 131% 5num

m₆₃ → 132% 5num

m₆₄ → 133% 5num

m₆₅ → 134% 5num

m₆₆ → 135% 5num

m₆₇ → 136% 5num

m₆₈ → 137% 5num

m₆₉ → 138% 5num

m₇₀ → 139% 5num

m₇₁ → 140% 5num

m₇₂ → 141% 5num

m₇₃ → 142% 5num

m₇₄ → 143% 5num

m₇₅ → 144% 5num

m₇₆ → 145% 5num

m₇₇ → 146% 5num

m₇₈ → 147% 5num

m₇₉ → 148% 5num

m₈₀ → 149% 5num

m₈₁ → 150% 5num

m₈₂ → 151% 5num

m₈₃ → 152% 5num

m₈₄ → 153% 5num

m₈₅ → 154% 5num

m₈₆ → 155% 5num

m₈₇ → 156% 5num

m₈₈ → 157% 5num

m₈₉ → 158% 5num

m₉₀ → 159% 5num

m₉₁ → 160% 5num

m₉₂ → 161% 5num

m₉₃ → 162% 5num

m₉₄ → 163% 5num

m₉₅ → 164% 5num

m₉₆ → 165% 5num

m₉₇ → 166% 5num

m₉₈ → 167% 5num

m₉₉ → 168% 5num

m₁₀₀ → 169% 5num

m₁₀₁ → 170% 5num

m₁₀₂ → 171% 5num

m₁₀₃ → 172% 5num

m₁₀₄ → 173% 5num

m₁₀₅ → 174% 5num

m₁₀₆ → 175% 5num

m₁₀₇ → 176% 5num

m₁₀₈ → 177% 5num

m₁₀₉ → 178% 5num

m₁₁₀ → 179% 5num

m₁₁₁ → 180% 5num

m₁₁₂ → 181% 5num

m₁₁₃ → 182% 5num

m₁₁₄ → 183% 5num

m₁₁₅ → 184% 5num

m₁₁₆ → 185% 5num

m₁₁₇ → 186% 5num

m₁₁₈ → 187% 5num

m₁₁₉ → 188% 5num

m₁₂₀ → 189% 5num

m₁₂₁ → 190% 5num

m₁₂₂ → 191% 5num

m₁₂₃ → 192% 5num

m₁₂₄ → 193% 5num

m₁₂₅ → 194% 5num

m₁₂₆ → 195% 5num

m₁₂₇ → 196% 5num

m₁₂₈ → 197% 5num

m₁₂₉ → 198% 5num

m₁₃₀ → 199% 5num

m₁₃₁ → 200% 5num

m₁₃₂ → 201% 5num

m₁₃₃ → 202% 5num

m₁₃₄ → 203% 5num

m₁₃₅ → 204% 5num

m₁₃₆ → 205% 5num

m₁₃₇ → 206% 5num

m₁₃₈ → 207% 5num

m₁₃₉ → 208% 5num

m₁₄₀ → 209% 5num

m₁₄₁ → 210% 5num

m₁₄₂ → 211% 5num

m₁₄₃ → 212% 5num

m₁₄₄ → 213% 5num

m₁₄₅ → 214% 5num

m₁₄₆ → 215% 5num

m₁₄₇ → 216% 5num

m₁₄₈ → 217% 5num

m₁₄₉ → 218% 5num

m₁₅₀ → 219% 5num

m₁₅₁ → 220% 5num

m₁₅₂ → 221% 5num

m₁₅₃ → 222% 5num

m₁₅₄ → 223% 5num

m₁₅₅ → 224% 5num

m₁₅₆ → 225% 5num

m₁₅₇ → 226% 5num

m₁₅₈ → 227% 5num

m₁₅₉ → 228% 5num

m₁₆₀ → 229% 5num

m₁₆₁ → 230% 5num

m₁₆₂ → 231% 5num

m₁₆₃ → 232% 5num

m₁₆₄ → 233% 5num

m₁₆₅ → 234% 5num

m₁₆₆ → 235% 5num

m₁₆₇ → 236% 5num

m₁₆₈ → 237% 5num

m₁₆₉ → 238% 5num

m₁₇₀ → 239% 5num

m₁₇₁ → 240% 5num

m₁₇₂ → 241% 5num

m₁₇₃ → 242% 5num

m₁₇₄ → 243% 5num

m₁₇₅ → 244% 5num

m₁₇₆ → 245% 5num

m₁₇₇ → 246% 5num

m₁₇₈ → 247% 5num

m₁₇₉ → 248% 5num

m₁₈₀ → 249% 5num

m₁₈₁ → 250% 5num

m₁₈₂ → 251% 5num

m₁₈₃ → 252% 5num

m₁₈₄ → 253% 5num

m₁₈₅ → 254% 5num

m₁₈₆ → 255% 5num

m₁₈₇ → 256% 5num

m₁₈₈ → 257% 5num

m₁₈₉ → 258% 5num

m₁₉₀ → 259% 5num

m₁₉₁ → 260% 5num

m₁₉₂ → 261% 5num

m₁₉₃ → 262% 5num

m₁₉₄ → 263% 5num

m₁₉₅ → 264% 5num

m₁₉₆ → 265% 5num

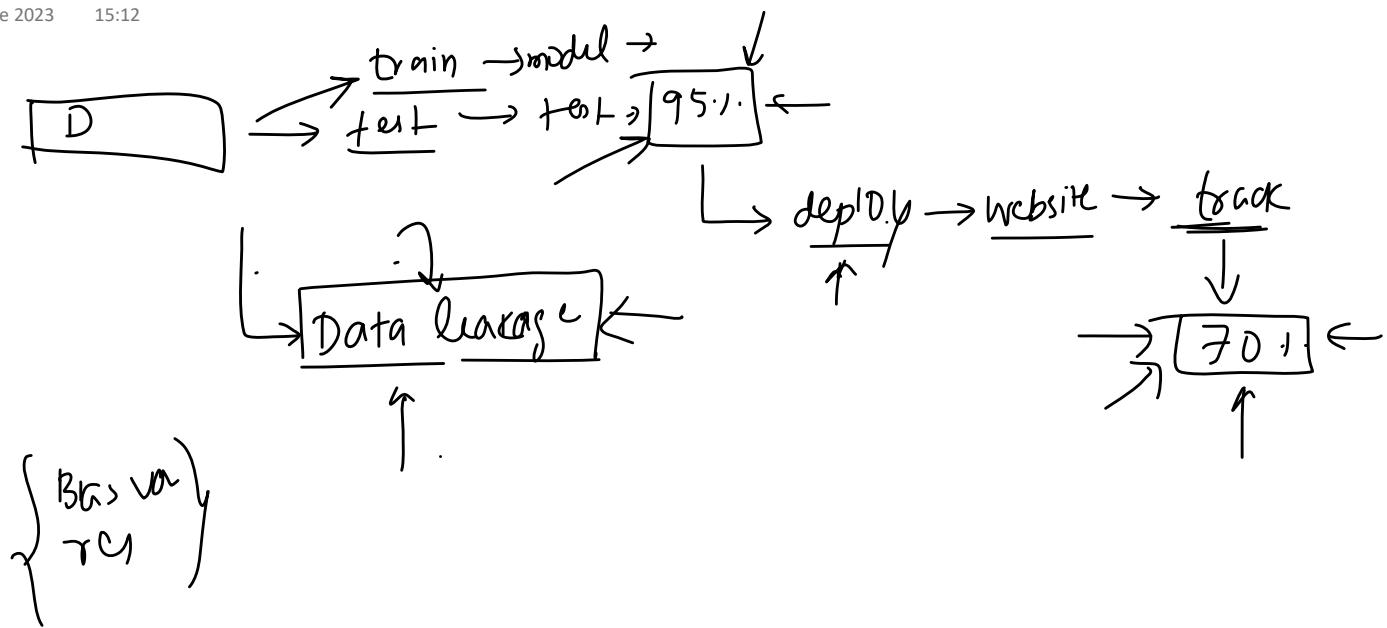
m₁₉₇ → 266% 5num

m₁₉₈ → 267% 5num

m₁₉₉ → 268% 5num

The Problem

12 June 2023 15:12

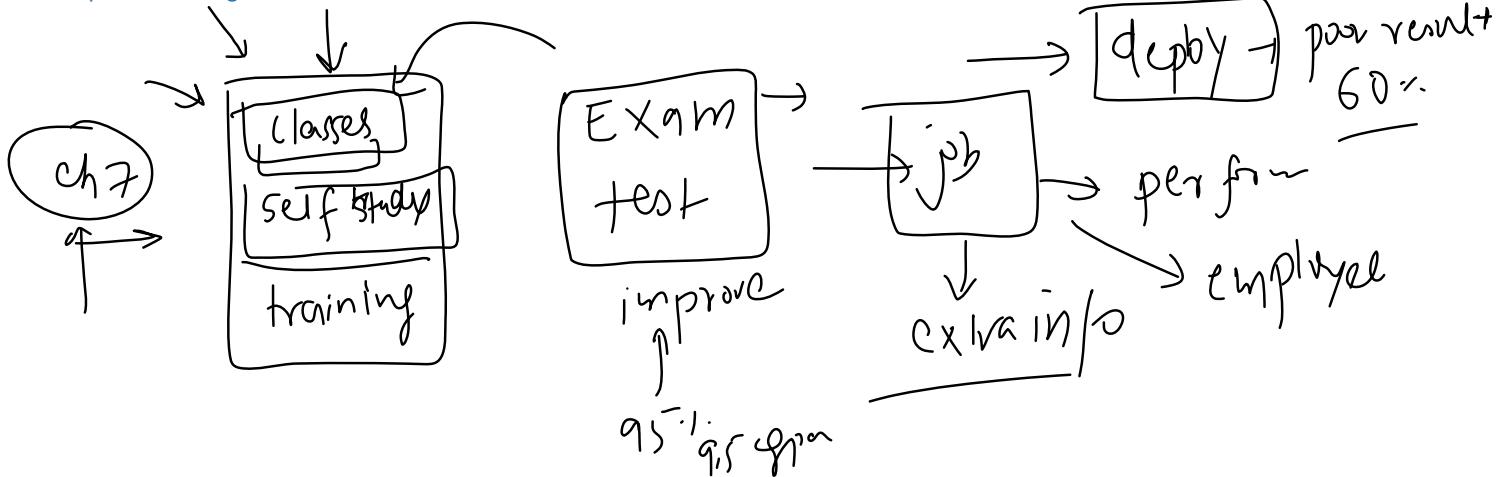
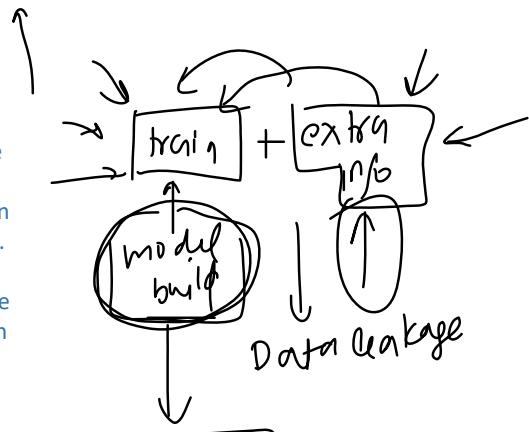


What is Data Leakage?

12 June 2023 15:13

Data leakage, in the context of machine learning and data science, refers to a problem where information from outside the training dataset is used to create the model. This additional information can come in various forms, but the common characteristic is that it is information that the model wouldn't have access to when it's used for prediction in a real-world scenario.

This can lead to overly optimistic performance estimates during training and validation, as the model has access to extra information. However, when the model is deployed in a production environment, that additional information is no longer available, and the performance of the model can drop significantly. This discrepancy is typically a result of mistakes in the experiment design.



Ways in which Data Leakage can occur

12 June 2023 15:13



1. Target Leakage:

Target leakage occurs when your predictors include data that will not be available at the time you make predictions.

2. Multicollinearity with target col

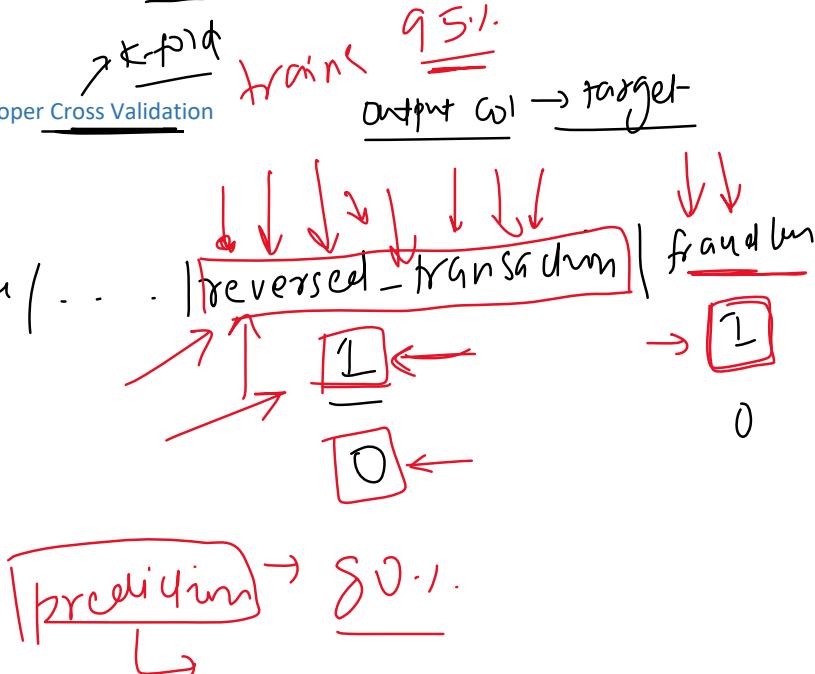
3. Duplicated Data

4. Preprocessing Leakage -> Train test contamination & Improper Cross Validation

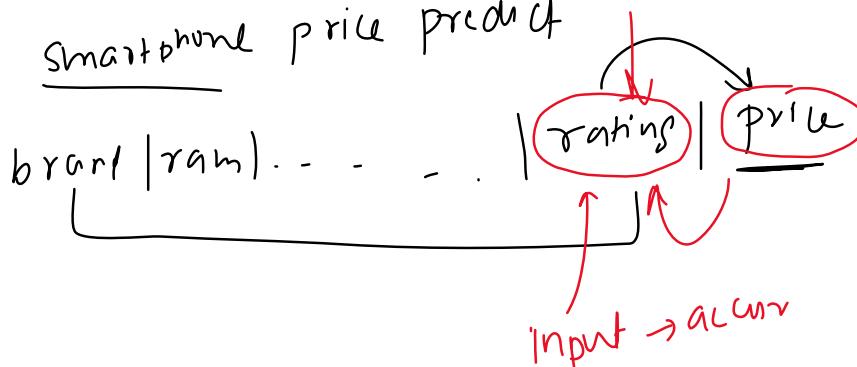
5. Hyperparameter Tuning

edit card fraud detect

<u>value</u>	<u>dark</u>	<u>webshop</u>	<u>...</u>	<u>reversed transaction</u>	<u>fraud bin</u>
<u>500</u>	<u>0</u>	<u>0</u>	<u>...</u>	<u>1</u>	<u>1</u>
<u>normal</u> \rightarrow <u>400</u>	<u>0</u>	<u>0</u>	<u>...</u>	<u>0</u>	<u>0</u>



Smartphone price predict

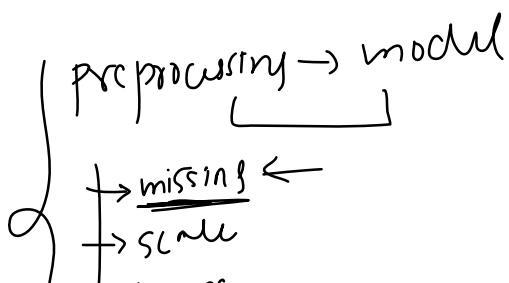
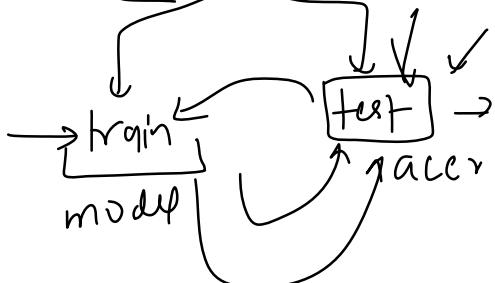


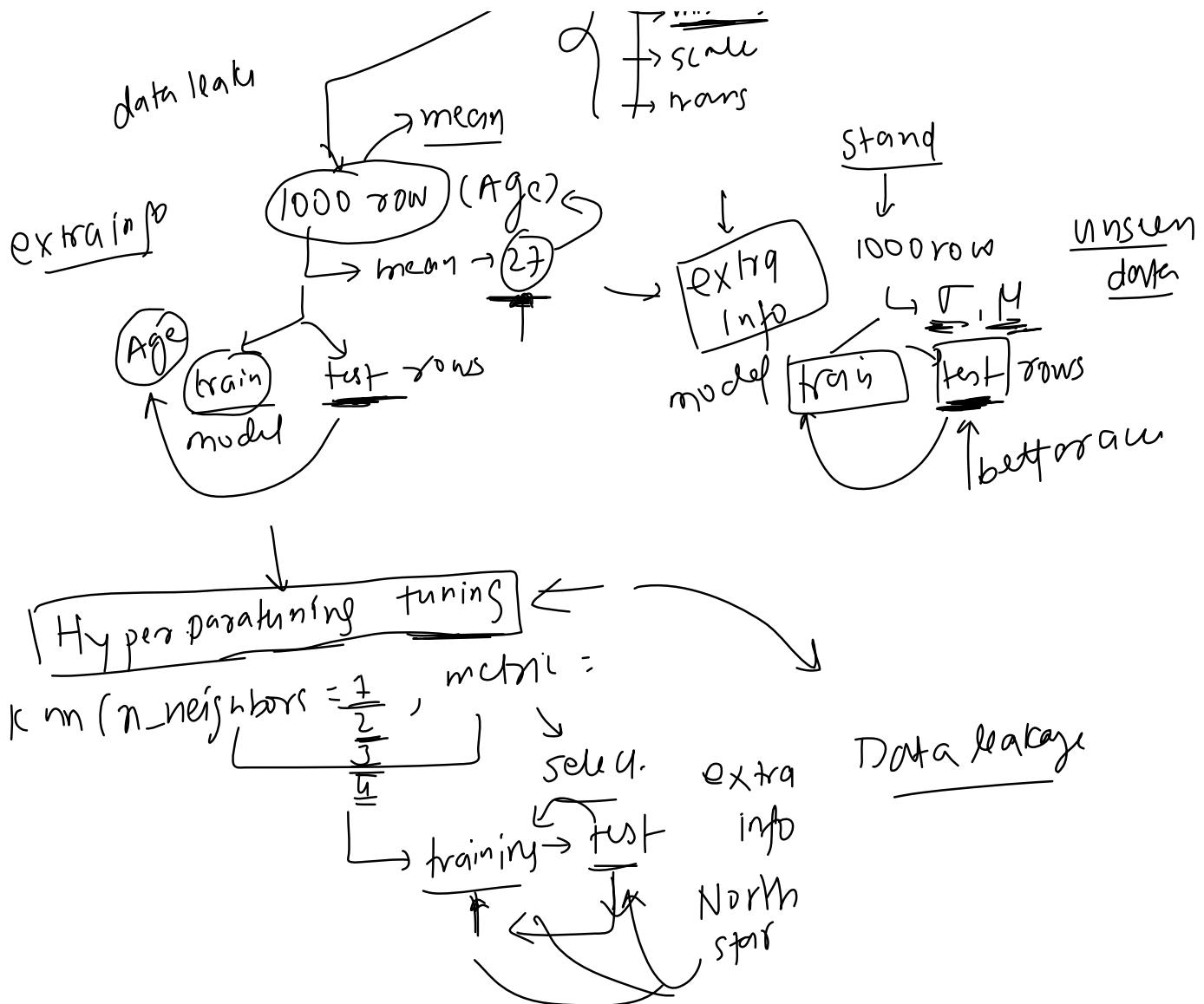
duplicate \rightarrow sort of duplicate

1000 rows

200 row duplca

extra info





How to detect?

12 June 2023 15:13

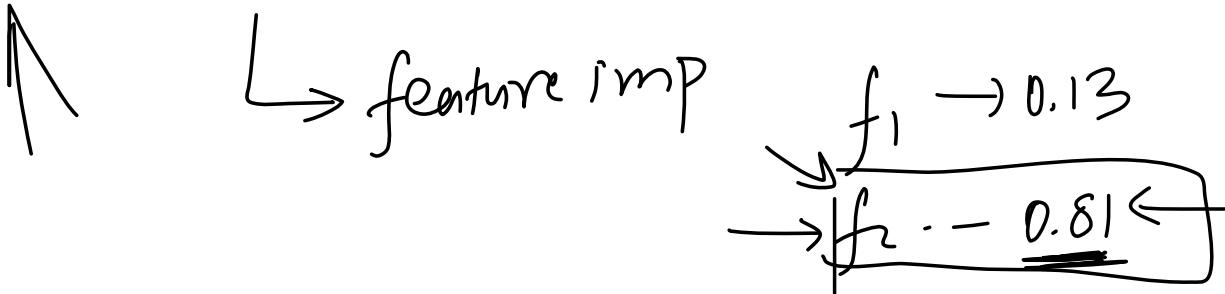
$f_1 f_2 \dots f_3$

1. **Review Your Features:** Carefully review all the features being used to train your model. Do they include any data that wouldn't be available at the time of prediction, or any data that directly or indirectly reveals the target? Such features are common sources of data leakage.

2. **Unexpectedly High Performance:** If your model's performance on the validation or test set is surprisingly good, this could be a sign of data leakage. Most predictive modelling tasks are challenging, and exceptionally high performance could mean that your model has access to information it shouldn't.

3. **Inconsistent Performance Between Training and Unseen Data:** If your model performs significantly better on the training and validation data compared to new, unseen data, this might indicate that there's data leakage.

4. **Model Interpretability:** Interpretable models, or techniques like feature importance, can help understand what the model is learning. If the model places too much importance on a feature that doesn't seem directly related to the output, it could be a sign of leakage.



How to remove Data Leakage

12 June 2023 15:14

train test cross val

1. Understand the Data and the Task: Before starting with any kind of data processing or modelling, it's important to understand the problem, the data, and how the data was collected. You should understand what each feature in your data represents, and whether it would be available at the time of prediction.
2. Careful Feature Selection: Review all the features used in your model. If any feature includes information that wouldn't be available at the time of prediction, or that directly or indirectly gives away the target variable, it should be removed or modified.
3. Proper Data Splitting: Always split your data into training, validation, and testing sets at an early stage of your pipeline, before doing any pre-processing or feature extraction.
4. Pre-processing Inside the Cross-Validation Loop: If you're using techniques like cross-validation, make sure to do any pre-processing inside the cross-validation loop. This ensures that the pre-processing is done separately on each fold of the data, which prevents information from the validation set leaking into the training set.

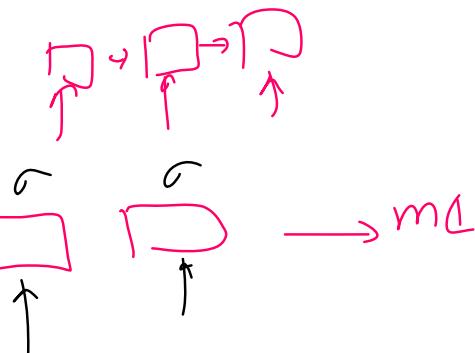
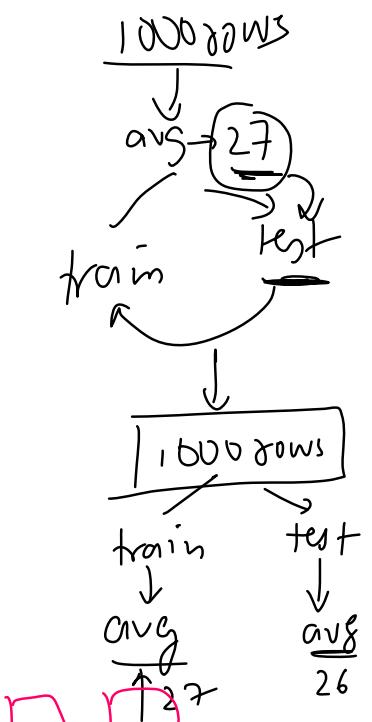
Incorrect way

```
X_normalized = normalize(X) # normalize the whole dataset
cross_val_score(model, X_normalized, y, cv=5) # perform cross-validation
```

↓ Pipeline →

Correct way

```
pipeline = make_pipeline(normalizer, model)
cross_val_score(pipeline, X, y, cv=5) # per
```



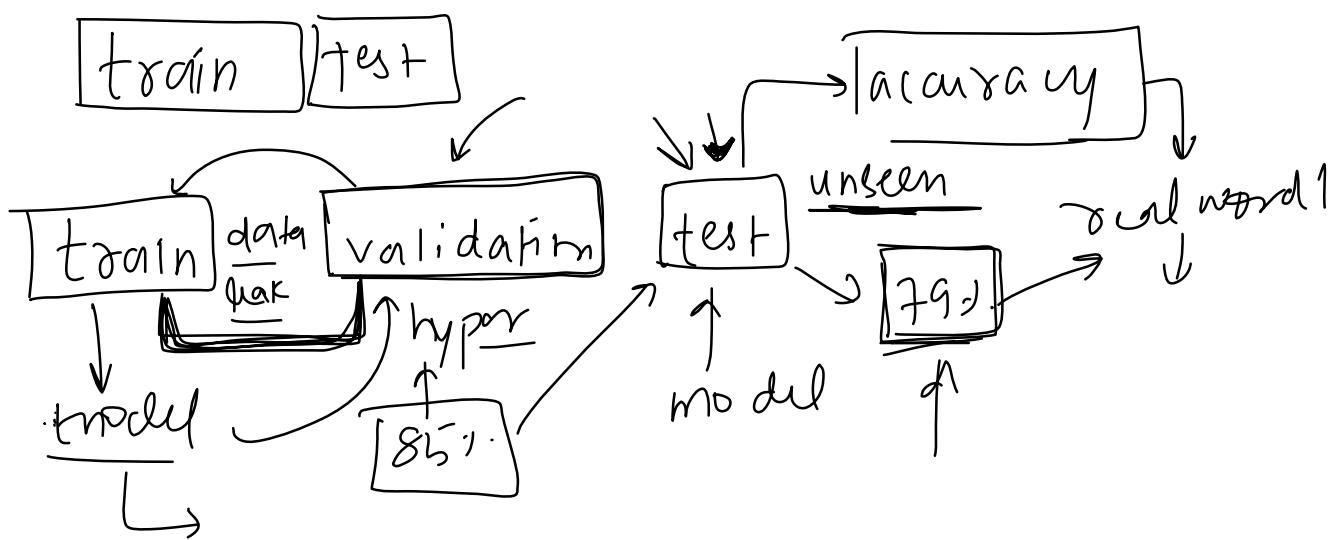
6. Avoid Overlapping Data: If the same individuals, or the same individual, appear in both your training and test sets, this can cause data leakage. It's important to ensure that the training and test sets represent separate, non-overlapping instances.

duplicate
remove
data leakage

Validation Set

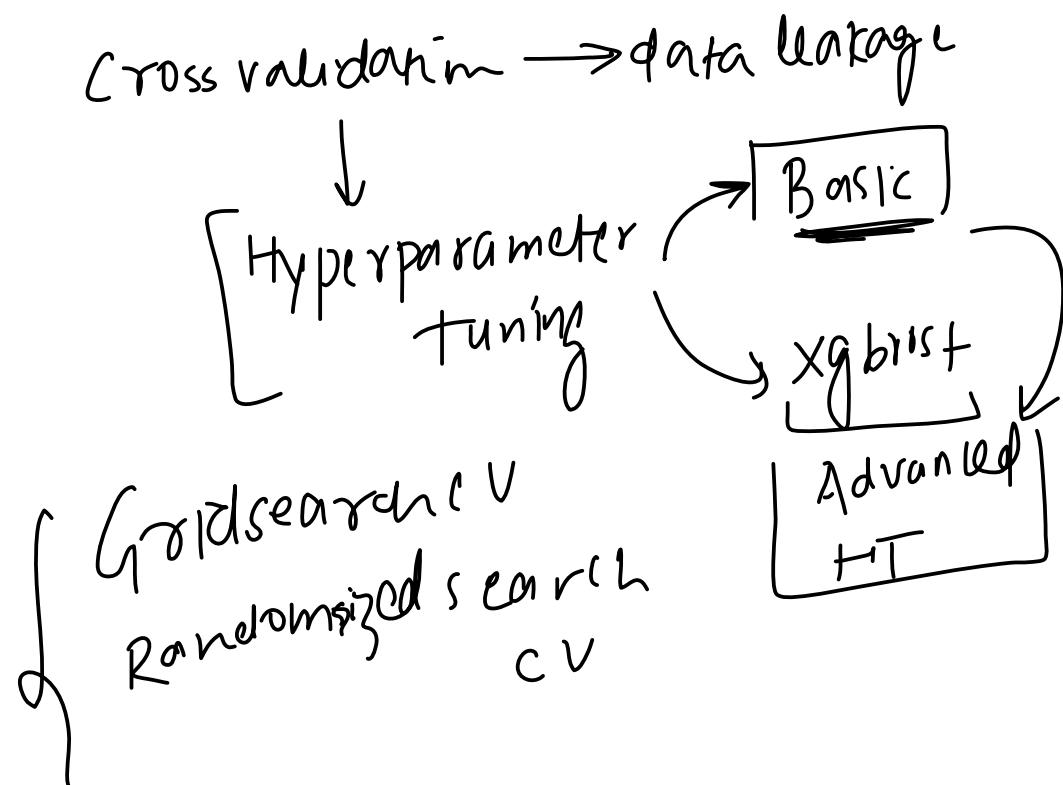
12 June 2023 22:02

Deep learning



Context

14 June 2023 19:12



Parameter Vs [Hyperparameter] Tuning

13 June 2023 10:48

Parameters

Parameters are the internal variables of a model that are learned from the data during the training process. They define the model's representation of the underlying patterns in the data.

For example:

- In a linear regression model, the parameters are the coefficients of the predictors.
- In a neural network, the parameters are the weights and biases of the nodes.
- In a decision tree, the parameters are the split points and split criteria at each node.

The goal of the training process is to find the optimal values for these parameters, which minimize the discrepancy between the model's predictions and the actual outcomes.

Hyperparameters → Setting knobs

In machine learning, hyperparameters are parameters whose values are set before the learning process begins. These parameters are not learned from the data and must be predefined. They help in controlling the learning process and can significantly influence the performance of the model.

- In a neural network, hyperparameters might include the learning rate, the number of layers in the network, or the number of nodes in each layer.
- In a support vector machine, the regularization parameter C or the kernel type can be considered as hyperparameters.
- In a decision tree, the maximum depth of the tree is a hyperparameter.

The best values for hyperparameters often cannot be determined in advance, and must be found through trial and error.

Why the word 'hyper'?

The choice of the word is primarily a naming convention to differentiate between the two types of values (internal parameters and guiding parameters) that influence the behaviour of a machine learning model. It's also a nod to the fact that the role they play is a meta one, in the sense that they control the structural aspects of the learning process itself rather than being part of the direct pattern-finding mission of the model.

linear reg → model

(q2)

coefficients

w₁, w₂, ..

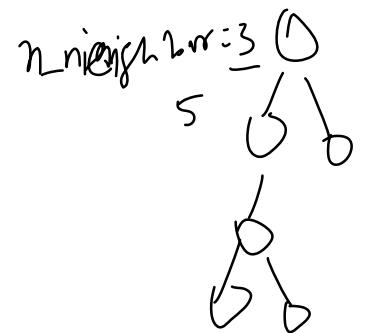
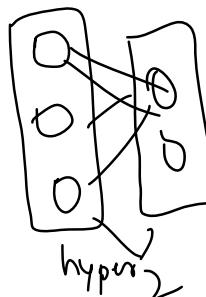
[Hyperparameter]

tuning

weights



3

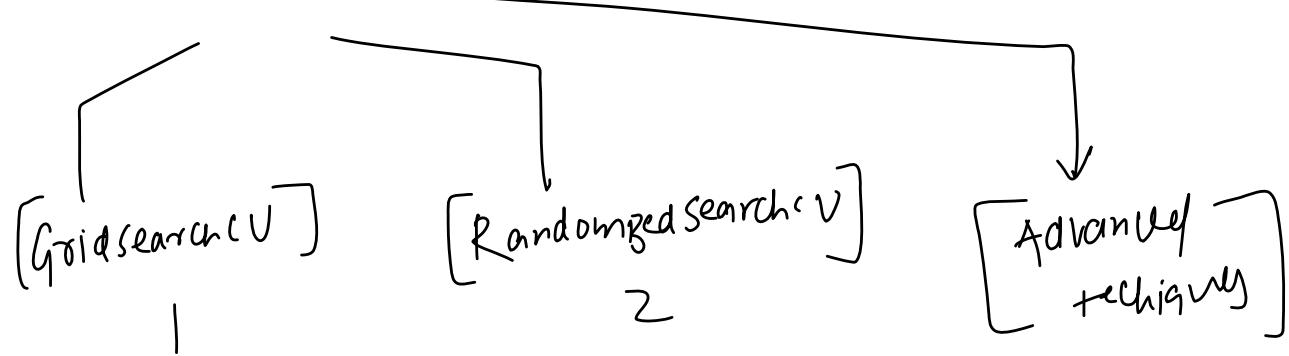


provide

Hyperparameter Tuning

Requirement

13 June 2023 10:48



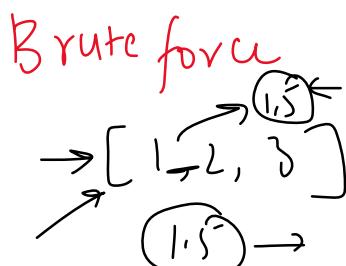
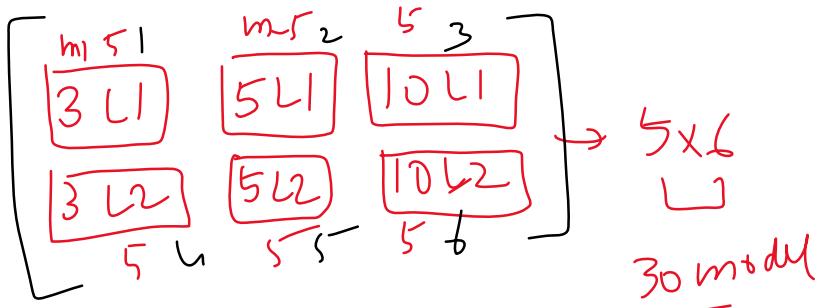
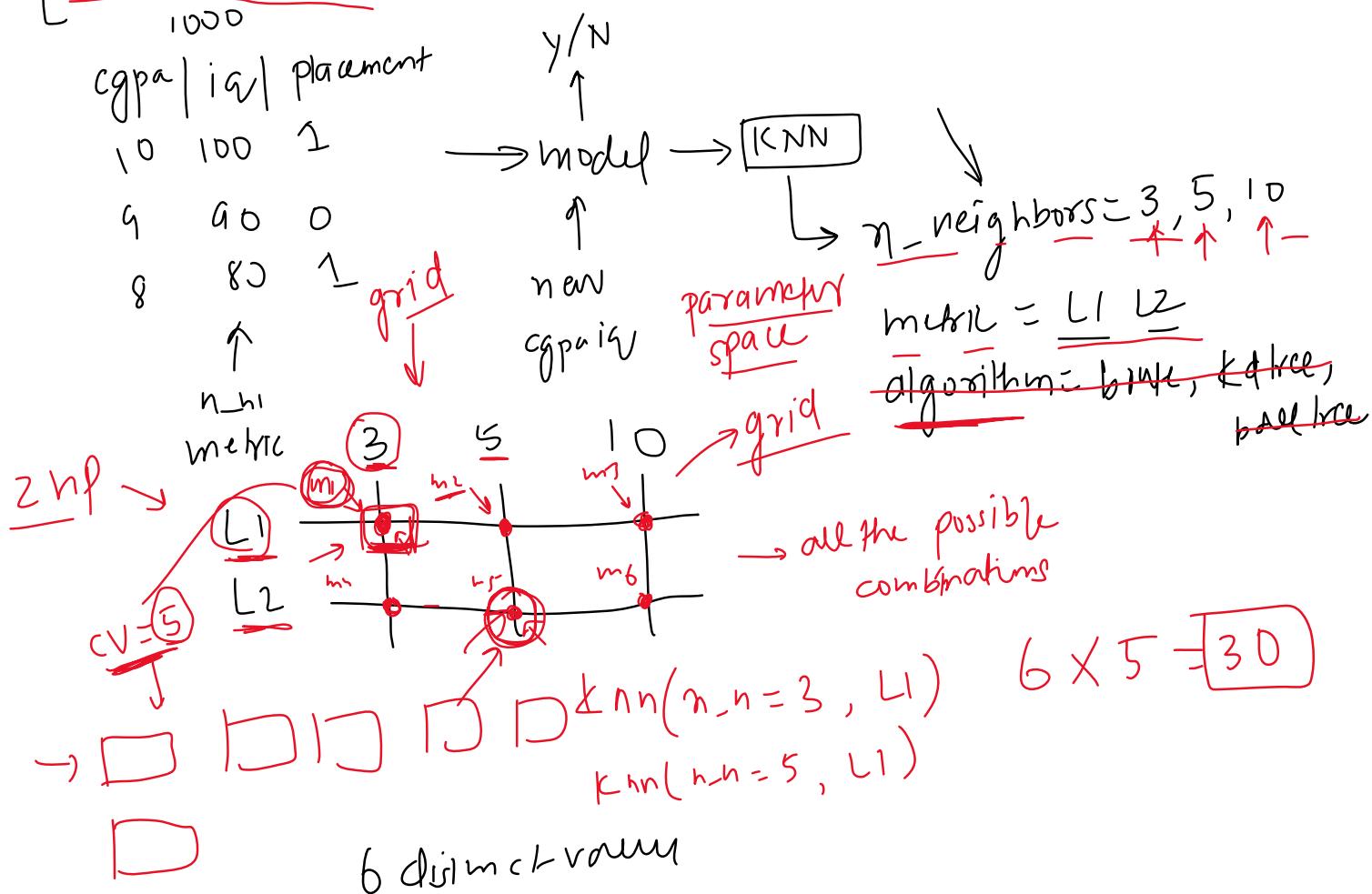
GridSearchCV

13 June 2023 10:49

CV → cross-validation

SD

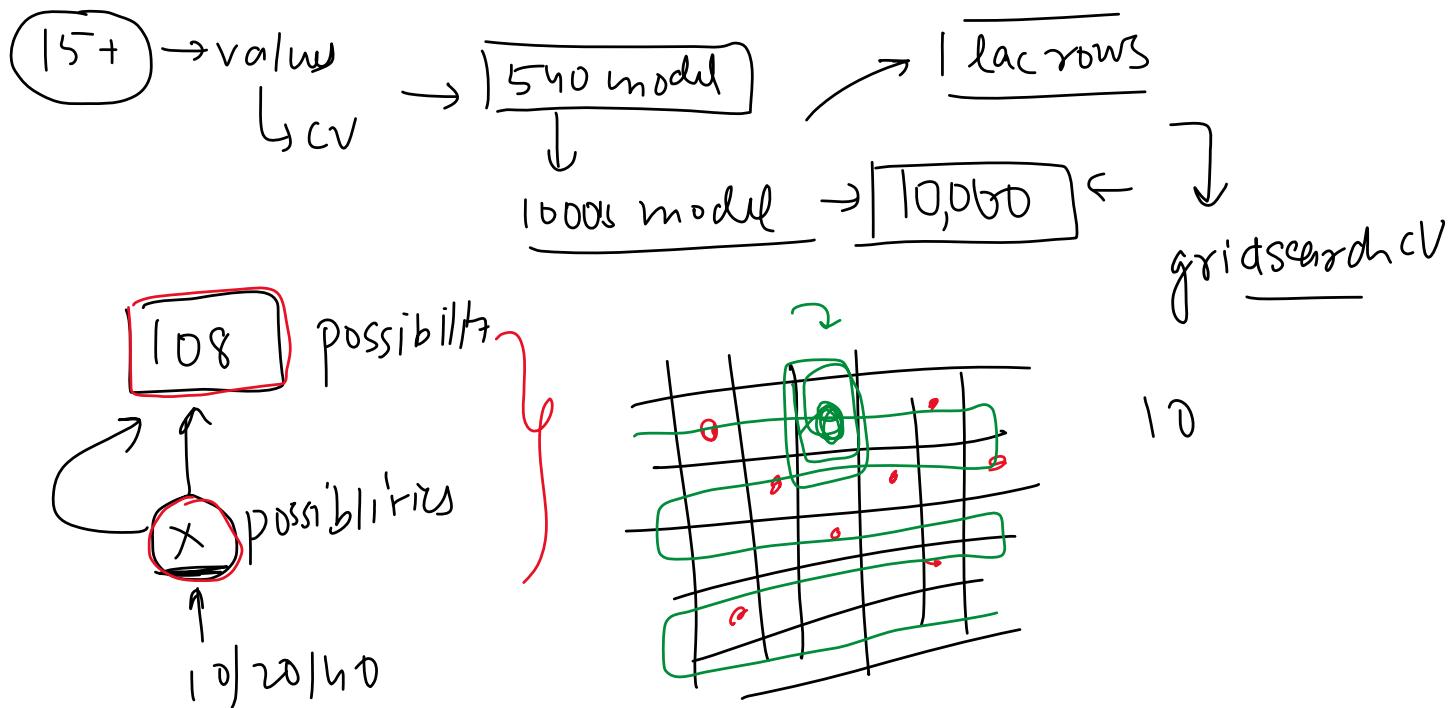
GridSearchCV refers to an algorithm that performs an exhaustive search over a specified grid of hyperparameters, using cross-validation to determine which hyperparameter combination gives the best model performance.



1, 2, 3

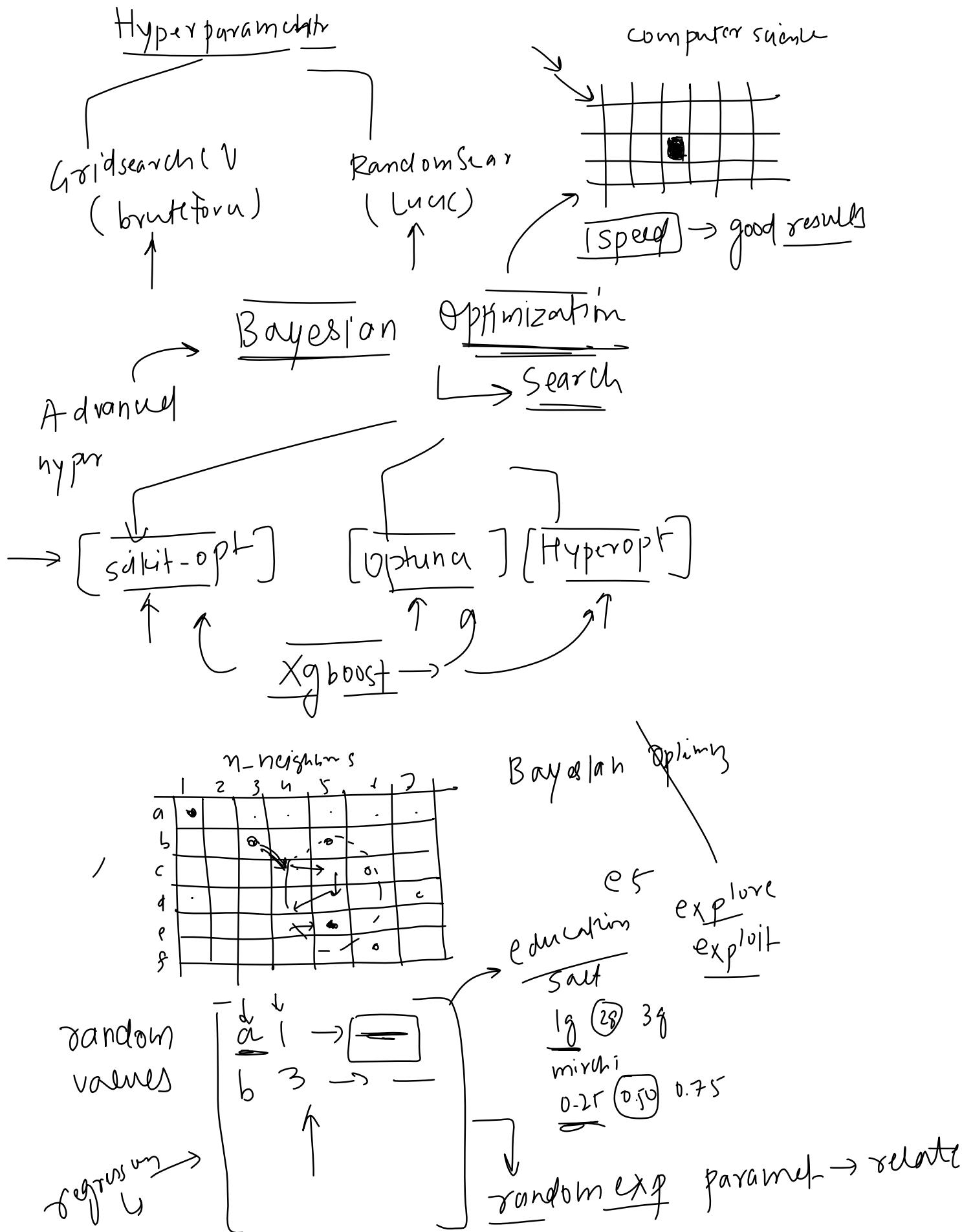
RandomizedSearchCV

13 June 2023 10:49



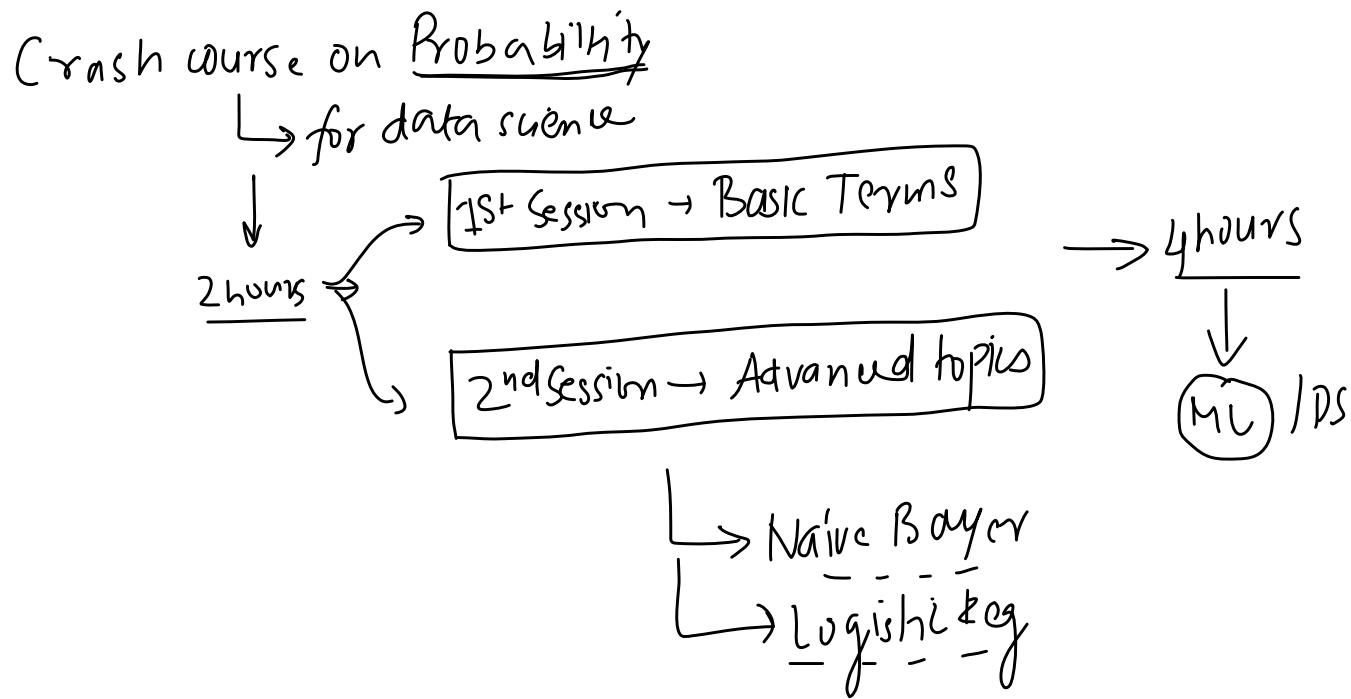
[Can this be improved?]

14 June 2023 07:58



Plan of Attack

16 June 2023 19:46



Terminology

15 June 2023 18:20

(5) terms

1. Random Experiment

An experiment is called random experiment if it satisfies the following two conditions:

- (i) It has more than one possible outcome. ✓
- (ii) It is not possible to predict the outcome in advance ✓

2. Trial →

Trial refers to a single execution of a random experiment. Each trial produces an outcome.

3. Outcome

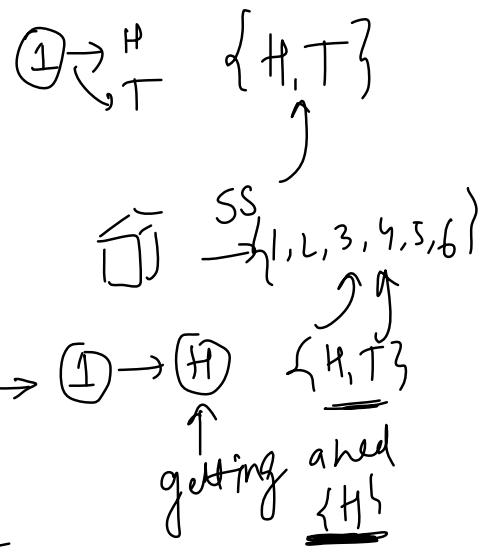
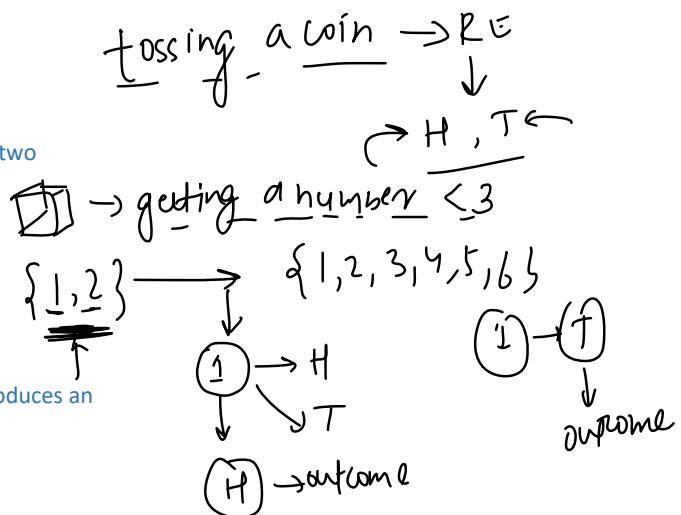
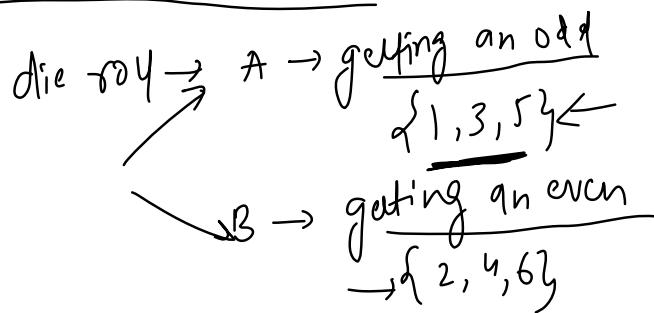
Outcome refers to a single possible result of a trial.

4. Sample Space

Sample Space of a random experiment is the set of all possible outcomes that can occur. Generally, one random experiment will have one set of sample space.

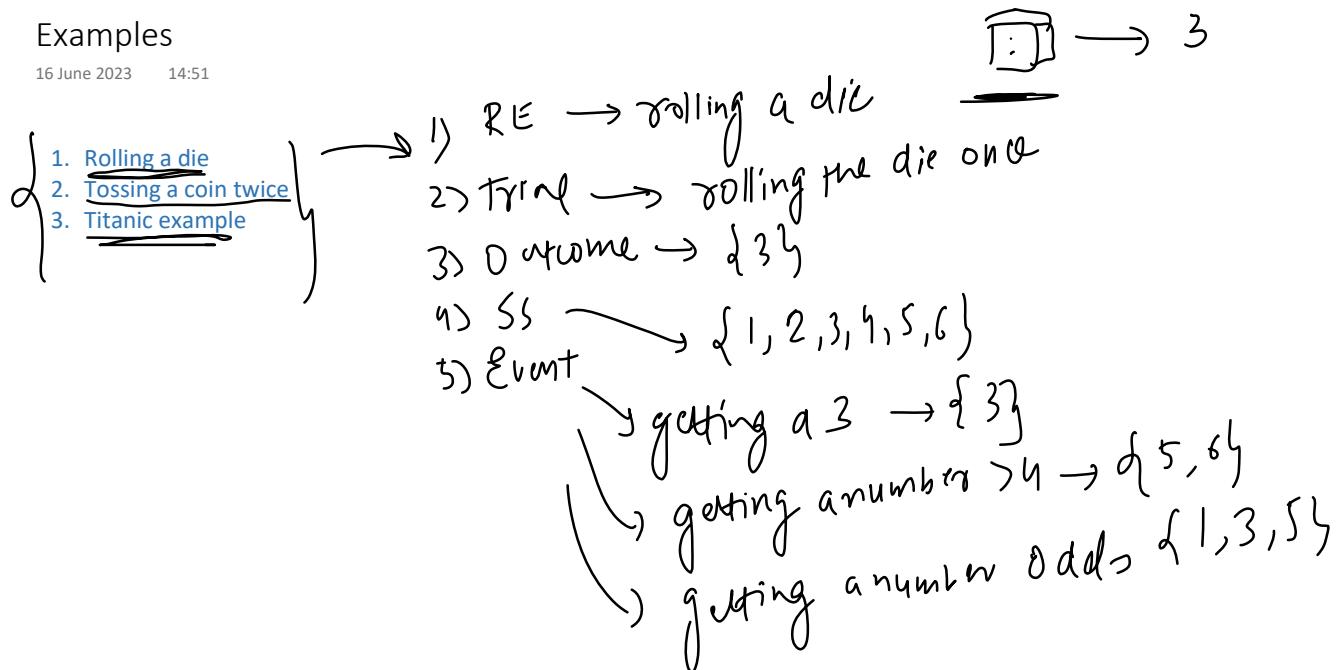
5. Event

Event is a specific set of outcomes from a random experiment or process. Essentially, it's a subset of the sample space. An event can include a single outcome, or it can include multiple outcomes. One random experiments can have multiple events.



Examples

16 June 2023 14:51



①

$\text{RE} \rightarrow$ tossing the coin twice
 $\text{trial} \rightarrow$ tossing the coin twice (once)
 $\text{outcome} \rightarrow \{H, T\}$
 $\rightarrow \text{SS} \rightarrow \{(H,H), (H,T), (T,H), (T,T)\}$
 $\text{Event} \rightarrow$ getting 2 heads $\{\underline{(H,H)}\}$
 $\text{getting at least 1 head } \{\underline{(H,H)}, \underline{(H,T)}, \underline{(T,H)}\}$

1, 2, 3

$\text{Titanic} \rightarrow$ 891 passengers \rightarrow Pclass

$\text{RE} \rightarrow$ randomly drawing out a passenger

$\text{trial} \rightarrow$ and finding its Pclass

$\text{outcome} \rightarrow \{1\}$

$\text{SS} \rightarrow \{1, 2, 3\}$

$\text{Event} \rightarrow A \rightarrow$ the passenger is from $P_{\text{class}}=1$ $\{1\}$

$B \rightarrow$ not from $P_{\text{class}}=2$ $\{\underline{2}, 3\}$

Types of Events

16 June 2023 08:29

2, 4, 6

{ 2 } { 2 } { 1 }
23 { 5, 6 }

1. **Simple Event:** Also known as an elementary event, a simple event is an event that consists of exactly one outcome.

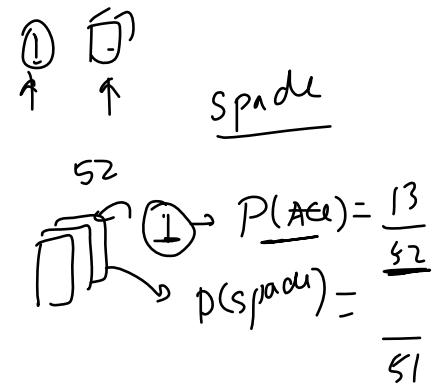
For example, when rolling a fair six-sided die, getting a 3 is a simple event.

2. **Compound Event:** A compound event consists of two or more simple events.

For example, when rolling a die, the event "rolling an odd number" is a compound event because it consists of three simple events: rolling a 1, rolling a 3, or rolling a 5.

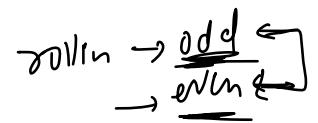
3. **Independent Events:** Two events are independent if the occurrence of one event does not affect the probability of the occurrence of the other event.

For example, if you flip a coin and roll a die, the outcome of the coin flip does not affect the outcome of the die roll.



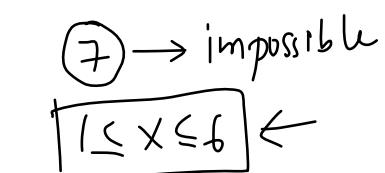
4. **Dependent Events:** Events are dependent if the occurrence of one event does affect the probability of the occurrence of the other event.

For example, if you draw two cards from a deck without replacement, the outcome of the first draw affects the outcome of the second draw because there are fewer cards left in the deck.



5. **Mutually Exclusive Events:** Two events are mutually exclusive (or disjoint) if they cannot both occur at the same time.

For example, when rolling a die, the events "roll a 2" and "roll a 4" are mutually exclusive because a single roll of the die cannot result in both a 2 and a 4.



6. **Exhaustive Events:** A set of events is exhaustive if at least one of the events must occur when the experiment is performed.

For example, when rolling a die, the events "roll an even number" and "roll an odd number" are exhaustive because one or the other must occur on any roll.

7. **Impossible event and Certain Event**

↓ ↑

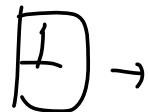
What is Probability

15 June 2023 18:14

In simplest terms, probability is a measure of the likelihood that a particular event will occur. It is a fundamental concept in statistics and is used to make predictions and informed decisions in a wide range of disciplines, including science, engineering, medicine, economics, and social sciences.

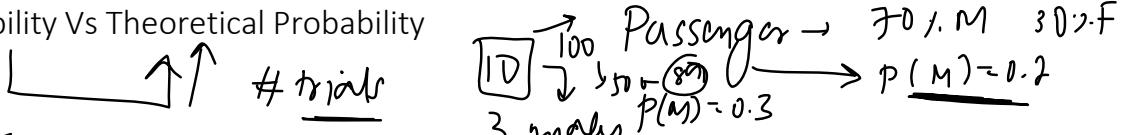
Probability is usually expressed as a number between 0 and 1, inclusive:

- A probability of 0 means that an event will not happen.
- A probability of 1 means that an event will certainly happen.
- A probability of 0.5 means that an event will happen half the time (or that it is as likely to happen as not to happen).



Empirical Probability Vs Theoretical Probability

15 June 2023 18:15



Empirical probability, also known as experimental probability, is a probability measure that is based on observed data, rather than theoretical assumptions. It's calculated as the ratio of the number of times a particular event occurs to the total number of trials.

$$P(H) = \frac{55}{100}$$

A. Suppose that, in our 100 tosses, we get heads 55 times and tails 45 times. What is the empirical probability of getting a head?

H → 55

Trials → 100

B. Let's say you have a bag with 50 marbles. Out of these 50 marbles, 20 are red, 15 are blue, and 15 are green. You start to draw marbles one at a time, replacing the marble back into the bag after each draw. After 200 draws, you find that you've drawn a red marble 80 times, a blue marble 70 times, and a green marble 50 times. What is the empirical probability of getting a red marble?

↑ 30 die

$$\begin{array}{c} 200 \text{ red} \\ \rightarrow \frac{80}{200} \\ \downarrow \end{array}$$

$$\{1, 2, 3, 4, 5, 6\}$$

roll → 3

$$\begin{array}{c} 1 \\ \hline 6 \end{array} \rightarrow P(1)$$

$$\begin{array}{c} 1 \\ \hline 2 \end{array} \rightarrow P(1)$$

Theoretical (or classical) probability is used when each outcome in a sample space is equally likely to occur. If we denote an event of interest as Event A, we calculate the theoretical probability of that event as:

Theoretical Probability of Event A = Number of Favourable Outcomes (that is, outcomes in Event A) / Total Number of Outcomes in the Sample Space

A. Consider a scenario of tossing a fair coin 3 times. Find the probability of getting exactly 2 heads.

B. Consider a scenario of rolling 2 dice. What is the probability of getting a sum = 7

No. of trials → infinite tosses

$$\{H, T\}$$

Empirical prob → theoretical

$$\begin{array}{c} 100 \rightarrow 3 \rightarrow 0.3 \\ \downarrow \\ 1000 \end{array}$$

$$\begin{array}{c} 45H \rightarrow 0.45 \\ 470 \rightarrow 0.47 \end{array}$$

Random Variable] → misleading
 15 June 2023 18:16 → function

In the context of probability theory, a random variable is a function that maps the outcomes of a random process (known as the sample space) to a set of real numbers.

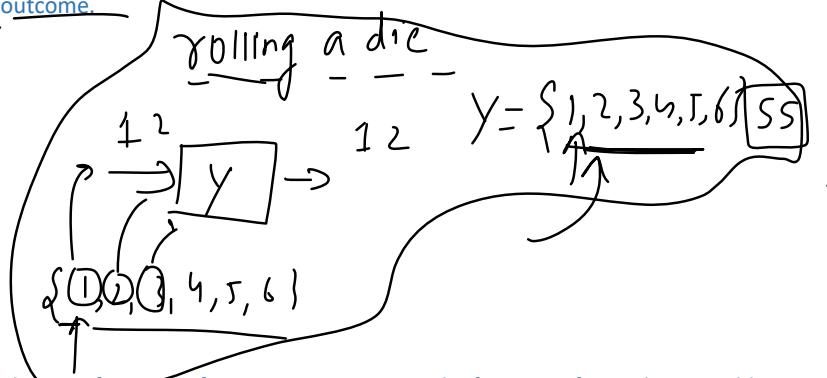
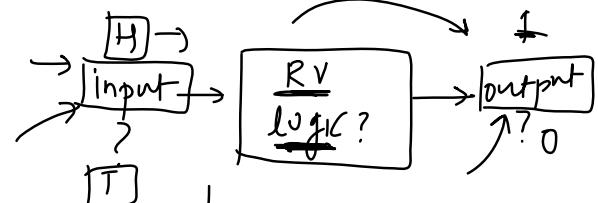
Discrete RV
Continuous RV

$$X = \{1, 0\}$$

H T

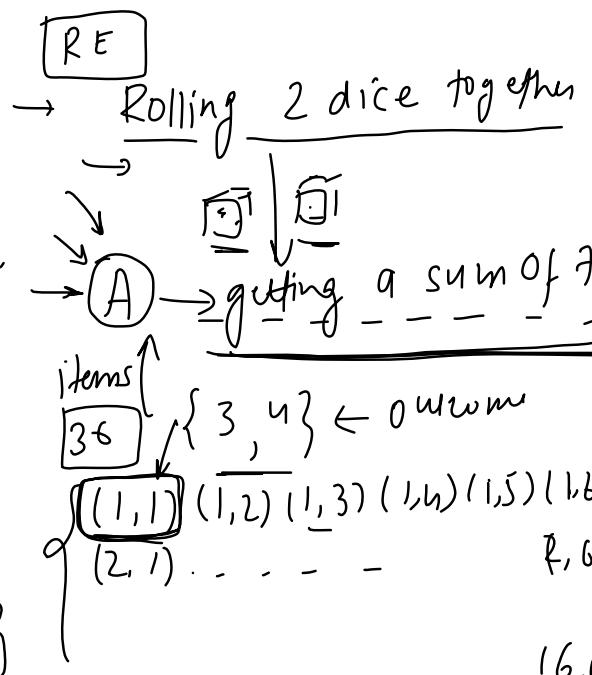
[Input: The input to the function is an outcome from the sample space of a random process.

input → logic → output



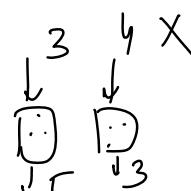
The transformation from input to output in the function of a random variable is determined by how we choose to define the random variable.

And the choice of how to define a random variable often depends on the specific aspects of the random process (or event) that you're interested in studying.



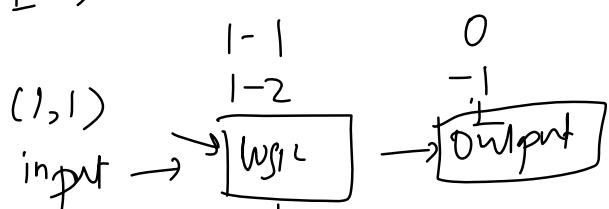
$$X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

↑ PMF →

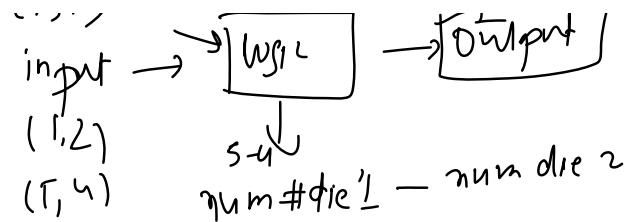


RE → rolling 2 dice
 → A where the num on die 1 > num on die 2

$$SS = \{(1,1), (1,2), (4,5), (5,4), (1,1), (1,1)\}$$



$$SS = \sum (1, 1) \text{ to } (6, 6) - \\ 4, 5, 5, 4$$

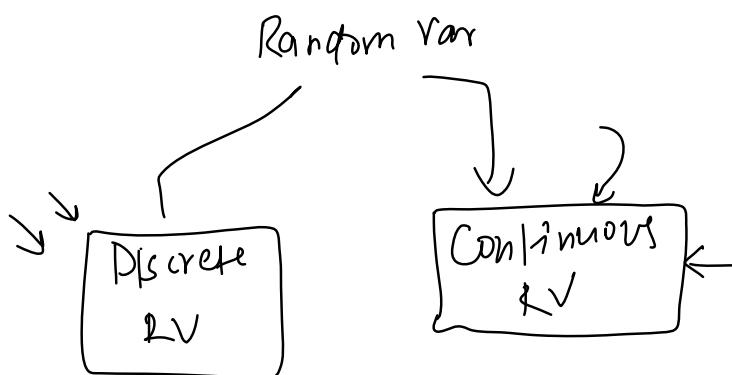


* random \rightarrow fvc + event

0 \rightarrow Sun
 -ve \rightarrow not sun

Output \rightarrow 1.5
 1.32

Discrete



RE \rightarrow cga \rightarrow 0-1D

{ 0-1D }

{ 0-1D } \rightarrow logic \rightarrow cont.
8.35 8.35

Probability Distribution of a Random Variable (Discrete)

15 June 2023 18:16

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values.

- 1) Toss
- 2) 1 die
- 3) 2 dice

X	1	0
$P(X)$	$\frac{1}{2}$	$\frac{1}{2}$

Probability dist of a random var

Toss a coin
 $\{H, T\}$

$$\hookrightarrow RV \rightarrow X = \{1, 0\}$$

$$P(X=1) = \frac{1}{2} \quad P(X=0) = \frac{1}{2}$$

rolling a die
 $S = \{1, 2, 3, 4, 5, 6\}$

X	1	2	3	4	5	6
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\hookrightarrow RV \rightarrow \{1, 2, 3, 4, 5, 6\}$$

probability dist of random variable

(input)

Sample space

Output

rolling 2 dice

$S = 36$ items

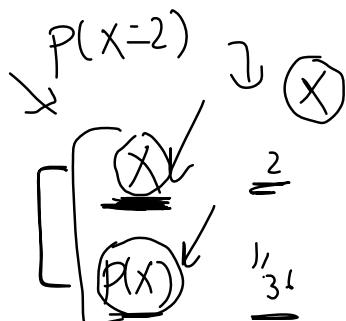
$$\downarrow \quad RV \rightarrow \text{sym}$$

1 2

(a,b)	1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$$X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$



$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{1}{36}$
----------------	----------------	----------------	----------------	---------------	----------------	---------------	----------------	---------------	----------------

Probability Dist Function

$$X \rightarrow P(X) \rightarrow (Y, f(x))$$

100 dice roll

$$100 \rightarrow 60^o$$

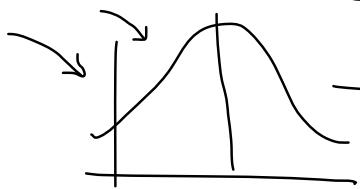
$$P(X) \rightarrow Y = X^2 + 5x + 3$$

contin
RV

$$X \rightarrow P(X) \rightarrow \boxed{Y = f(X)} \rightarrow P(X) \rightarrow Y = \underline{x^2 + 5x + 3}$$

PDF

mathematical $X = 5/36$



a lot of tiny \otimes

PDF / PMF
 contm discrete
 ↳ normal

PDF

Density (PPF)
 (continuous)

mass (PMP)

(discrete)

Mean of a Random Variable

15 June 2023 18:17

$$\frac{21}{6} = 3.5$$

Mean of X

multiple trials (0)

trials (0)

1000 dice

Avg →

mean (X) ↓

$\frac{3+4+2+5+...}{1000} = \square$

rolling a die

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

mean

$$= 3.5$$

↑

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

↓

avg →

↓

mean (X) ↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

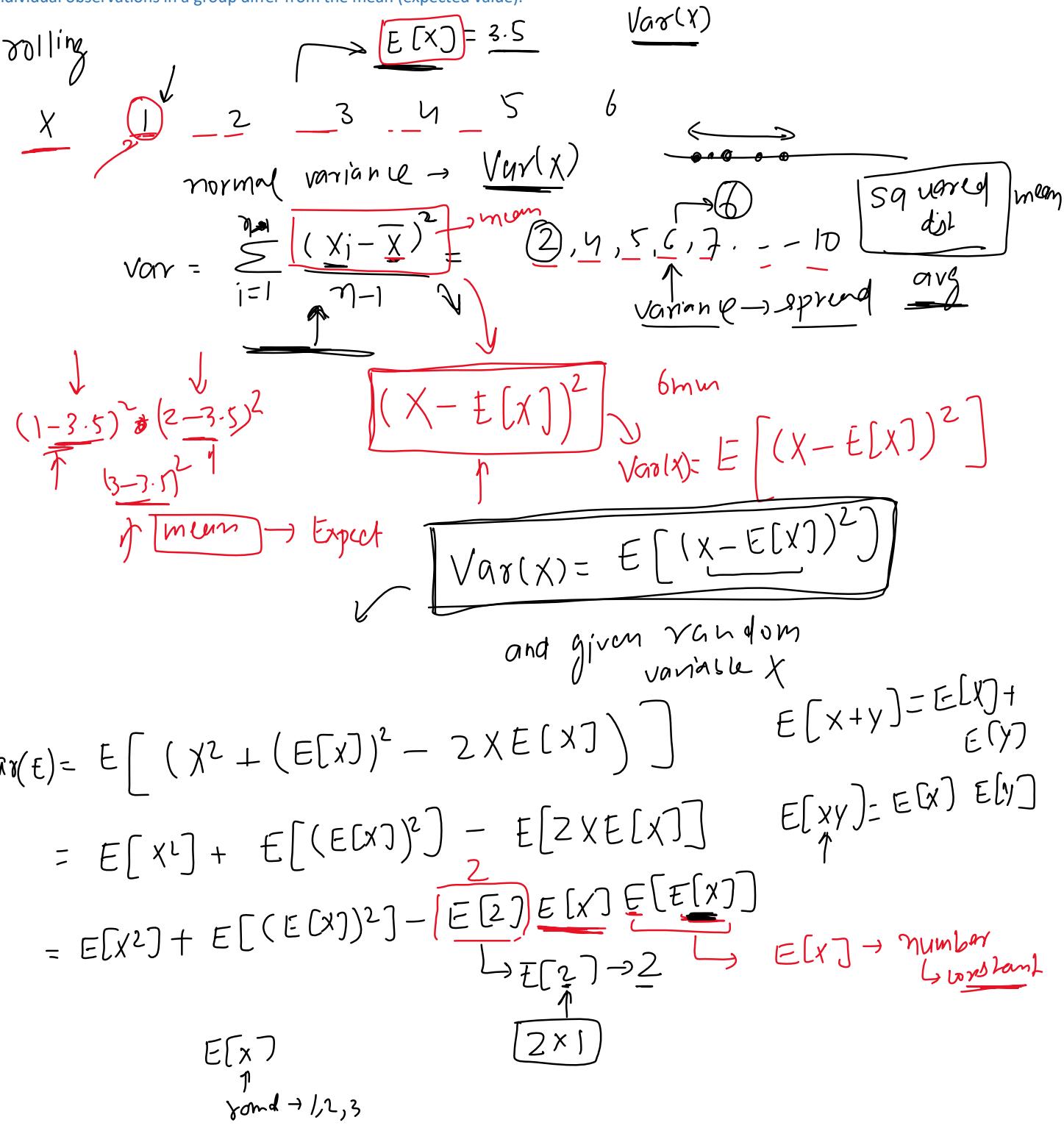
↓

↓

Variance of a Random Variable

15 June 2023 18:17

The variance of a random variable is a statistical measurement that describes how much individual observations in a group differ from the mean (expected value).



$$= E[X^2] + E[(E[X])^2] - 2 E[X] E[X]$$

$$= E[X^2] + E[(E[X])^2] - 2(E[X])^2$$

$$= E[X^2] + (E[(E[X])^2]) - 2(E[X])^2$$

$$= E[X^2] + (E[(E[X])^2]) - 2(E[X])^2$$

$$= E[X^2] + \cancel{(E[X])^2} - 2(E[X])^2$$

$$\boxed{Var(x) = E[X^2] - (E[X])^2} \rightarrow \boxed{\text{cont}} \quad \boxed{\text{discrete}}$$

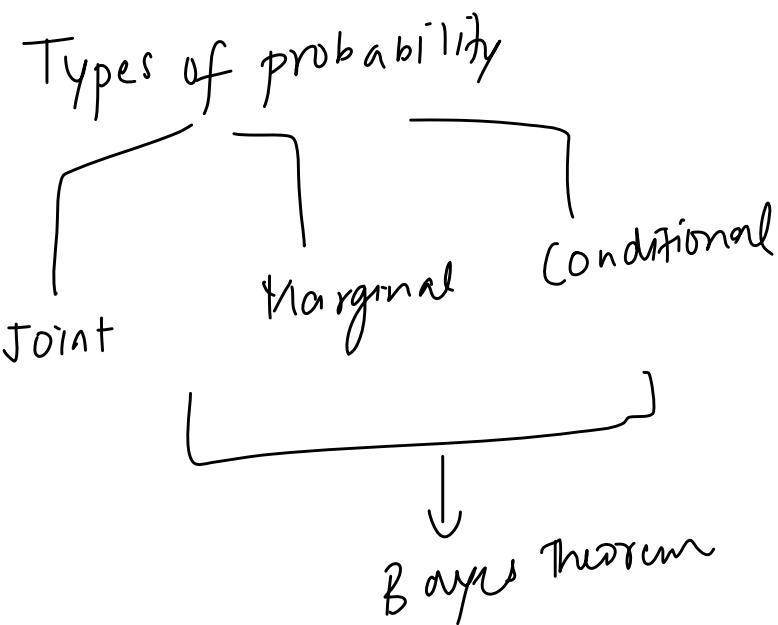
$$\boxed{Var(x) = E[(X - E[X])^2]}$$

Recap

21 June 2023 16:30

Terms

- + CXP
- + trial
- + Outcome
- + sample
- + Events



Venn Diagrams in Probability

20 June 2023 15:07

Set Theory



Probability

$$\{1, 2, 3, 4, 5, 6\}$$



Ω (sample set)

$$P(\Omega) = 1$$

rolling a die

Event $A \rightarrow$ getting a num ≥ 4

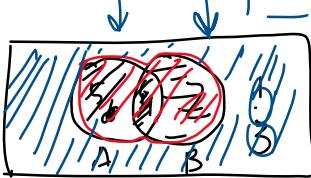
$$P(A) = \frac{3}{6} = \frac{1}{2}$$

$$P(A)' = \frac{1}{2}$$

$$P(\Omega) - P(A)$$

$$A \rightarrow \text{num} \geq 4 \{4, 5, 6\}$$

$$B \rightarrow \text{even num} \{2, 4, 6\}$$



$$\{1, 2, 3, 4, 5, 6\}$$

$$P(A \cap B) = \frac{2}{6}$$

$$P(\Omega) = 1$$

$$P(A) = \frac{3}{6} = \frac{1}{2} \quad P(B) = \frac{1}{2}$$

$$P(A \cup B)' = \frac{2}{6}$$

$$P(A \cup B) = \frac{5}{6}$$

$$P(\Omega) - P(A \cup B)$$

$$1 - \frac{2}{3} = \boxed{\frac{1}{3}}$$

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

$$= \frac{1}{3} + \frac{1}{3} -$$

Contingency Tables in Probability

20 June 2023 15:07

experiment \rightarrow rolling a die

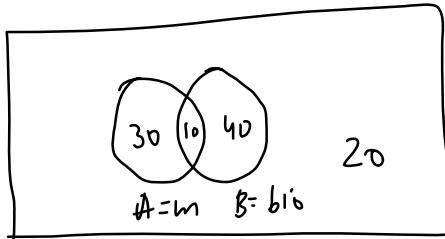
- $A \rightarrow \geq 4$ X
- $B \rightarrow \text{even number}$ Y

	even	not even
> 4	2	1
< 4	1	2

100 students

- $40 \rightarrow$ only bio
- $30 \rightarrow$ only math
- $10 \rightarrow$ both

	math	not math
bio	10	40
not bio	30	20



Joint Probability

15 June 2023 18:22

random var

Let's say we have two random variables X and Y . The joint probability of X and Y , denoted as $P(X = x, Y = y)$, is the probability that X takes the value x and Y takes the value y at the same time.

titanic

Let X be a random variable associated with the Pclass of a passenger
Let Y be a random variable associated with the survival status of a passenger

Pclass	1	2	3
Survived	0 → 80	97	372 /891
	1	136	87 119

contingency table

random var
 $X = 1$
 $X = 2$
 $X = 3$
 $X \rightarrow \{1, 2, 3\}$

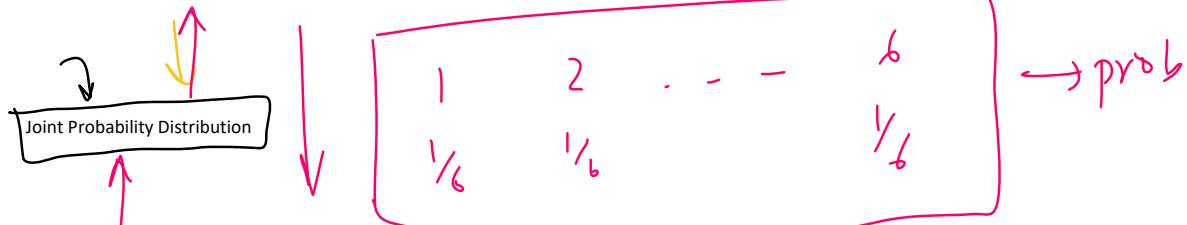
$Y = 0$
 $Y = 1$
 $Y \rightarrow \{0, 1\}$

total passengers

Pclass	1	2	3
Survived	0 → 0.089787	0.108866	0.417508
	1	0.152637	0.097643

$$P(X=1, Y=0)$$

$$P(X=1, Y=0)$$



$$X=1, Y=0 \quad X=2, Y=0 \quad \dots \quad X=3, Y=1$$

$$\downarrow \quad \downarrow \quad \downarrow$$

$$P(x) \quad P(x) \quad \dots \quad P(x)$$

Marginal Probability / Simple / Unconditional

15 June 2023 21:15

Marginal probability refers to the probability of an event occurring irrespective of the outcome of some other event. When dealing with random variables, the marginal probability of a random variable is simply the probability of that variable taking a certain value, regardless of the values of other variables.

Let X be a random variable associated with the Pclass of a passenger
Let Y be a random variable associated with the survival status of a passenger

Pclass	1	2	3	All
Survived	0	80 97 372	549	891
	1	136 87 119	342	891
All	216 184 491	891		

marginal prob

Pclass	1	2	3	All
Survived	0	0.089787 0.108866	0.417508	0.616162
	1	0.152637 0.097643	0.133558	0.383838
All	0.242424 0.206510	0.551066	1.000000	

$y = 0, 1$

$p(x)$

$$p(y=0) = 0.61 \\ p(y=1) = 0.38$$

Marginal Probability Distribution
Also known as unconditional probability $P(A)$

$$p(X=1) = 0.24 \\ p(X=2) = 0.20 \\ p(X=3) = 0.55$$

$p(y)$

$X = 1, 2, 3$
marginal prob dist of Y

marginal prob dist of X

Conditional Probability

15 June 2023 18:15

$$A \leftarrow \boxed{B} \quad P(A|B)$$

$$P(A|B)$$

Conditional probability is a measure of the probability of an event occurring, given that another event has already occurred. If the event of interest is A and event B has already occurred, the conditional probability of A given B is usually written as $P(A|B)$.

$$P(B|A)$$

Three unbiased coins are tossed. What is the conditional probability that at least two coins show heads, given that at least one coin shows heads?

$$\left\{ \begin{array}{l} \text{of } \underline{\text{HHH}}, \underline{\text{HHT}}, \underline{\text{HTH}}, \underline{\text{THT}} \\ \quad \underline{\text{HTT}}, \underline{\text{THT}}, \underline{\text{TTH}}, \underline{\text{TTT}} \end{array} \right\} \rightarrow A \rightarrow \text{at least 2 heads} \\ \rightarrow B \rightarrow \text{at least } \underline{1 \text{ head}} \\ \rightarrow 7 \text{ outcomes} \quad \frac{4}{7} = P(A|B)$$

Two fair six-sided dice are rolled. What is the conditional probability that the sum of the numbers rolled is 7, given that the first die shows an odd number?

		dice 2					
		2	3	4	5	6	7
dice 1		3	4	5	6	7	8
→	→	4	5	6	7	8	9
→	→	5	6	7	8	9	10
→	→	6	7	8	9	10	11
→	→	7	8	9	10	11	12

$$\begin{array}{l} A \rightarrow \text{sum} = 7 \\ B \rightarrow \text{die 1} \rightarrow \text{odd} \end{array}$$

$$P(A|B) = \frac{3}{18} = \frac{1}{6}$$

Two fair six-sided dice are rolled, denoted as D1 and D2. What is the conditional probability that D1 equals 2, given that the sum of D1 and D2 is less than or equal to 5?

		D2					
		2	3	4	5	6	7
D1		2	3	4	5	6	7
→	→	2	3	4	5	6	7

$$\begin{array}{l} A \rightarrow D1 = 2 \\ B \rightarrow D1 + D2 \leq 5 \end{array}$$

1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12
7	8	9	10	11	12	

$$\frac{P_1 + P_2}{B} \leq \frac{5}{10}$$

$$P(A|B) = \frac{3}{10}$$

Formula for Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

joint prob A and B

marginal prob of B

Pclass	1	2	3	
Survived	0	80	97	372
1	136	87	119	381

Pclass	1	2	3	
Survived	0	0.089787	0.108866	0.417508
1	0.152637	0.097643	0.133558	

$$P(Y=0 | X=3) = \frac{P(Y=0 \cap X=3)}{P(X)} \rightarrow = \frac{0.41}{0.54} = 0.75$$

$$P(Y=0 \cap X=2) = \frac{0.10}{0.19} = 0.52 \rightarrow \left[\frac{0.08}{0.23} \right] = 0.34$$

$$P(X=3 \mid Y=0)$$

Intuition behind the Conditional Probability Formula

20 June 2023 20:06

The intuition behind the formula for conditional probability, $P(A | B) = P(A \cap B) / P(B)$, is based on the concept of reducing our sample space.

The denominator, $P(B)$, is the probability of event B occurring. When we say we want to know the probability of A given B, we're effectively saying that B has occurred and therefore B is our new "universe" or sample space. So we're not considering cases when B didn't occur anymore, and we're normalizing by the probability of B.

The numerator, $P(A \cap B)$, is the joint probability of A and B, meaning both A and B occur. So in the context of our new universe where B has occurred, $P(A \cap B)$ represents the cases where A also occurs.

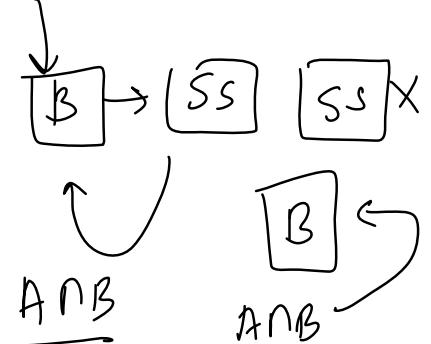
By dividing the joint probability $P(A \cap B)$ by the probability of B ($P(B)$), we effectively find the proportion of times that A occurs given that B has occurred.

In summary, the conditional probability of A given B is just the joint probability of A and B happening (A and B together in the "universe" where everything happens), normalized by the probability of B (the new "universe" where only B happens).

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

definition

$$\frac{A}{B}$$



Independent Vs Mutually Exclusive Events

15 June 2023 18:16

Independent events are events where the occurrence of one event does not affect the occurrence of another.

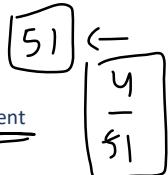
Examples

1. Flipping a coin and rolling a die
2. Drawing a card with replacement

Dependent events are events where the occurrence of one event does affect the occurrence of another.

Examples

1. Drawing a card without replacement



$$\rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)}$$

independent

Mutually Exclusive Events

Mutually exclusive events are events that cannot both occur at the same time. In other words, if one event occurs, the other cannot.



Examples

1. Flipping a coin $\rightarrow [H, T]$
2. Rolling a die

even / odd

$$P(A \cap B) = 0$$

↳ 0

$$P(A|B) = 0$$

$$\rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)}$$

	If statistically independent	If mutually exclusive
$P(A B) =$	$P(A)$	0
$P(B A) =$	$P(B)$	0
$P(A \cap B) =$	$P(A)P(B)$	0

independent events

$$\rightarrow P(A|B) = \frac{P(A)}{P(B)}$$

$$\rightarrow P(B|A) = \frac{P(B)}{P(A)}$$

$$\rightarrow P(A \cap B) = P(A)P(B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)}$$

Bayes Theorem

15 June 2023 18:16

→ Bayesian Statistics ←

Bayes' Theorem is a fundamental concept in the field of probability and statistics that describes how to update the probabilities of hypotheses when given evidence. It's used in a wide variety of fields, including machine learning, statistics, and game theory.

The theorem is named after Thomas Bayes, who introduced an early version of the rule in his work on probability theory. Bayes' Theorem provides a way to revise existing predictions or theories (update probabilities) given new evidence.

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

$$\rightarrow P(A|B) = P(A) \times \frac{P(B|A)}{P(B)}$$

posterior prior likelihood marginal / evidence

Mathematical Proof

conditional Prob

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad A \cap B = B \cap A$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \Rightarrow P(B \cap A) = P(A \cap B) = P(B|A) P(A)$$

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \rightarrow \underline{\text{Bayes Theorem}}$$

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

gender survived

M	→	0
M	→	0
F	→	1
F	→	0
M	→	1

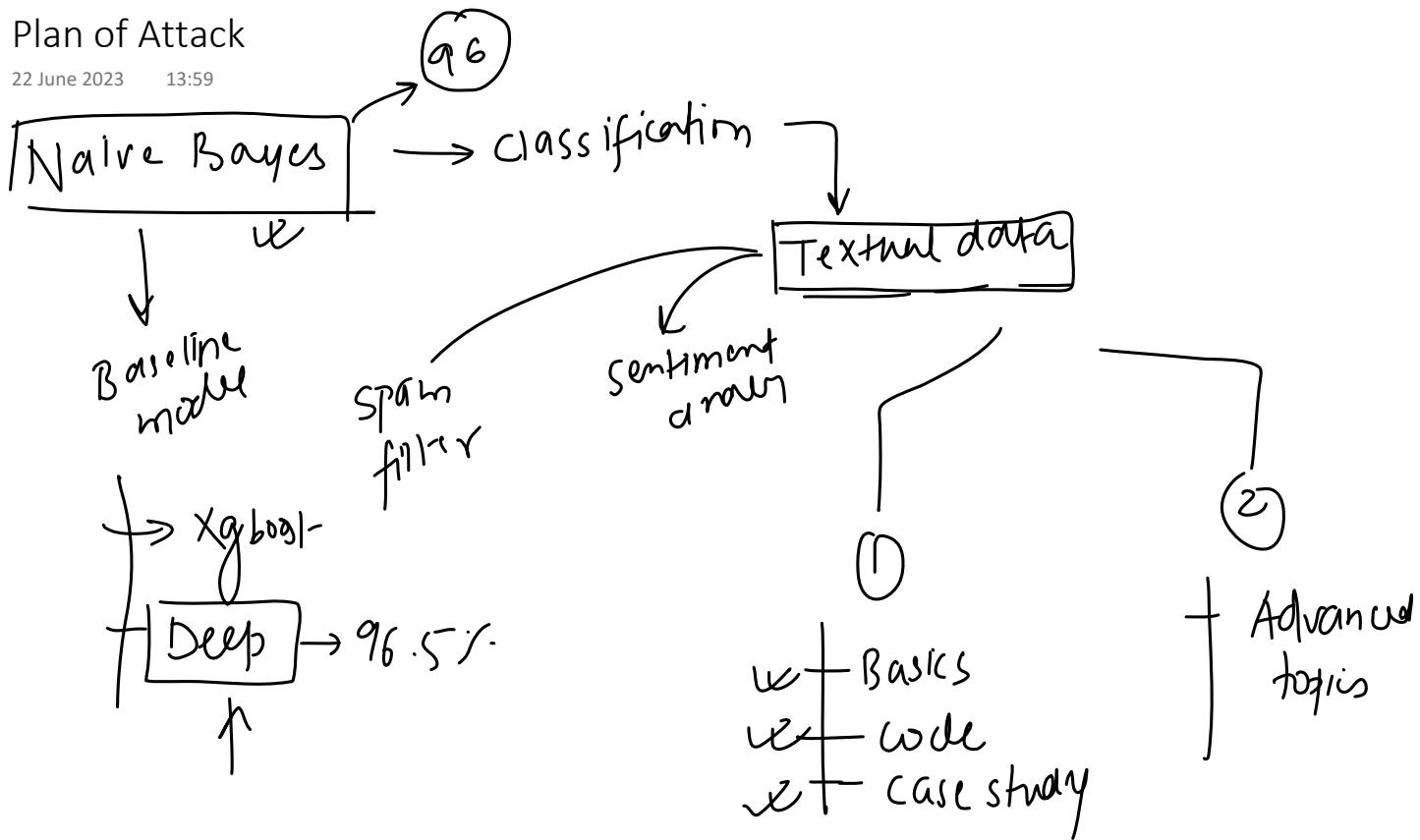
$$P(0|M) = \frac{P(M|0) P(0)}{P(M)} \rightarrow \frac{3/5}{3/5} = 1$$

$$P(1|M) = \frac{P(M|1) P(1)}{P(M)} \rightarrow \frac{2/5}{3/5} = \frac{2}{3}$$

$$\begin{aligned}
 & P(M|I) = \frac{P(M|I)P(I)}{P(M)} \\
 & P(0|M) = \frac{2}{3} \quad \checkmark \\
 & P(1|M) = \frac{1}{3} \quad \checkmark \\
 & \text{dead} \Rightarrow \text{alive} \quad \frac{\frac{1}{2} \times \frac{2}{5}}{\frac{3}{5}} = \frac{1}{3}
 \end{aligned}$$

Plan of Attack

22 June 2023 13:59



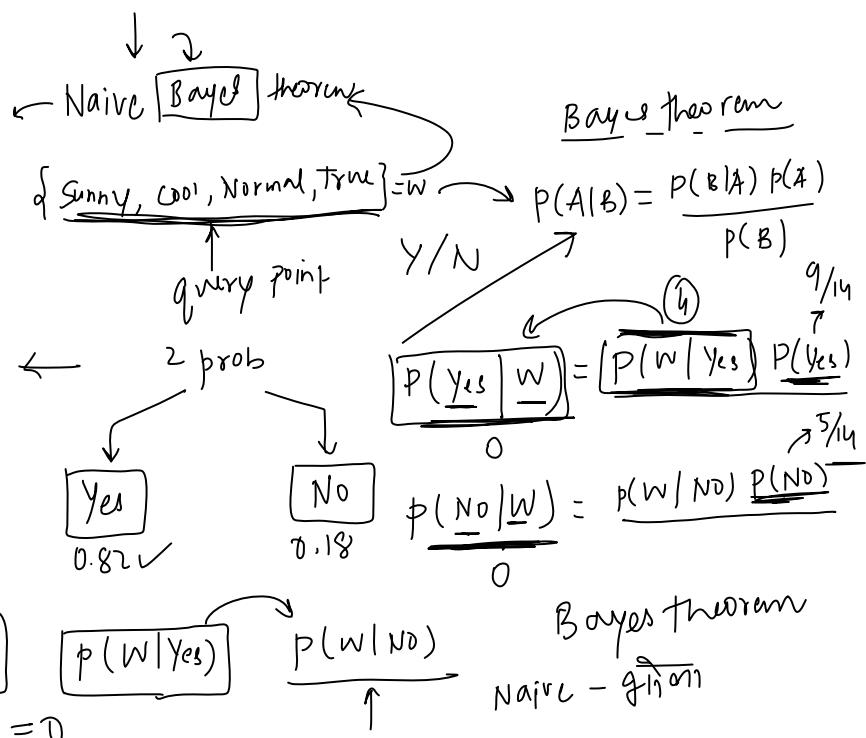
Intuition

22 June 2023 13:59

D)
14
days

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No ✗
Sunny	Hot	High	True	No ✗
Overcast	Hot	High	False	Yes ✓
Rainy	Mild	High	False	Yes ✓
Rainy	Cool	Normal	False	Yes ✓
Rainy	Cool	Normal	True	No ✗
Overcast	Cool	Normal	True	Yes ✓
Sunny	Mild	High	False	No ✗
Sunny	Cool	Normal	False	Yes ✓
Rainy	Mild	Normal	False	Yes ✓
Sunny	Mild	Normal	True	Yes ✓
Overcast	Mild	High	True	Yes ✓
Overcast	Hot	Normal	False	Yes ✓
Rainy	Mild	High	True	No ✗

Play Tennis Toy
binary classification



~~$P(\underline{\text{sunny}} \cap \underline{\text{cool}} \cap \underline{\text{normal}} \cap \underline{\text{true}} | Yes)$~~

$P(\underline{\text{sunny}} \cap \underline{\text{cool}} \cap \underline{\text{normal}} \cap \underline{\text{true}} | No) = 0$

$P(\underline{\text{sunny}} | Yes) \quad P(\underline{\text{cool}} | Yes) \quad P(\underline{\text{normal}} | Yes) \quad P(\underline{\text{true}} | Yes)$

$\frac{2}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{3}{9} = \frac{9}{14}$

$p(\underline{\text{Sunny}} | No) \quad p(\underline{\text{cool}} | No) \quad p(\underline{\text{normal}} | No) \quad p(\underline{\text{true}} | No) \times \frac{5}{14}$

Mathematical Formulation

22 June 2023 13:59

$x_1 \ x_2 \ \dots \ x_n \ y$ multiclass classification $\underbrace{1, 2, 3, \dots, k}_{k \text{ classes}}$

$y \in N$

$\rightarrow \langle x_1, x_2, x_3, \dots, x_n \rangle \rightarrow \text{naive bayes} \quad k \text{ prob}$

$$\left\{ \begin{array}{l} p(y_1 | x_T) = p(x_T | y_1) p(y_1) \\ p(y_2 | x_T) \rightsquigarrow p(x_T | y_2) p(y_2) \\ \vdots \\ p(y_k | x_T) \rightsquigarrow p(x_T | y_k) p(y_k) \end{array} \right.$$

$$x_T = \langle x_1, x_2, \dots, x_n \rangle$$

$$p(y_k | x_T) = p(x_T | y_k) p(y_k)$$

$$= p(\underbrace{x_1 \cap x_2 \cap x_3 \cap \dots \cap x_n}_{n} | y_k) p(y_k)$$

$$= p(\overbrace{x_1, x_2, x_3, \dots, x_n}^A | \overbrace{y_k}^B) p(y_k)$$

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

$$= p(x_1, x_2, x_3, \dots, x_n, y_k) \ p(y_k)$$

$$p(A \cap B) = p(A|B) p(B)$$

$$= p(\overbrace{x_1}^A \cap \overbrace{x_2, x_3, \dots, x_n, y_k}^B)$$

$$= p(x_1 | x_2, x_3, \dots, x_n, y_k) p(x_2, x_3, \dots, x_n, y_k)$$

$$p(x_2 | x_3, x_4, \dots, x_n, y_k) p(x_3 | x_4, \dots, x_n, y_k)$$

$$= p(x_1 | x_2, x_3, \dots, x_n, y_k) p(x_2 | x_3, x_4, \dots, y_k) \dots p(x_{n-1} | x_n, y_k) p(x_n | y_k) p(y_k)$$

Naive assumption \rightarrow features are independent of each other

$$\underbrace{x_1}_{\text{A}} \ \underbrace{x_2}_{\text{B}} \ \underbrace{x_3}_{\text{C}} \ \dots \ \underbrace{x_n}_{\text{D}}$$

$$p(A|B) = p(A)$$

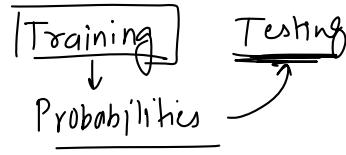
$$p(A|B \cap C) = p(A \cap B \cap C) = \frac{p(A \cap B \cap C)}{p(B \cap C)}$$

Code

22 June 2023 14:00

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

2

No, YesSunny, Hot, High, FalseDictionary

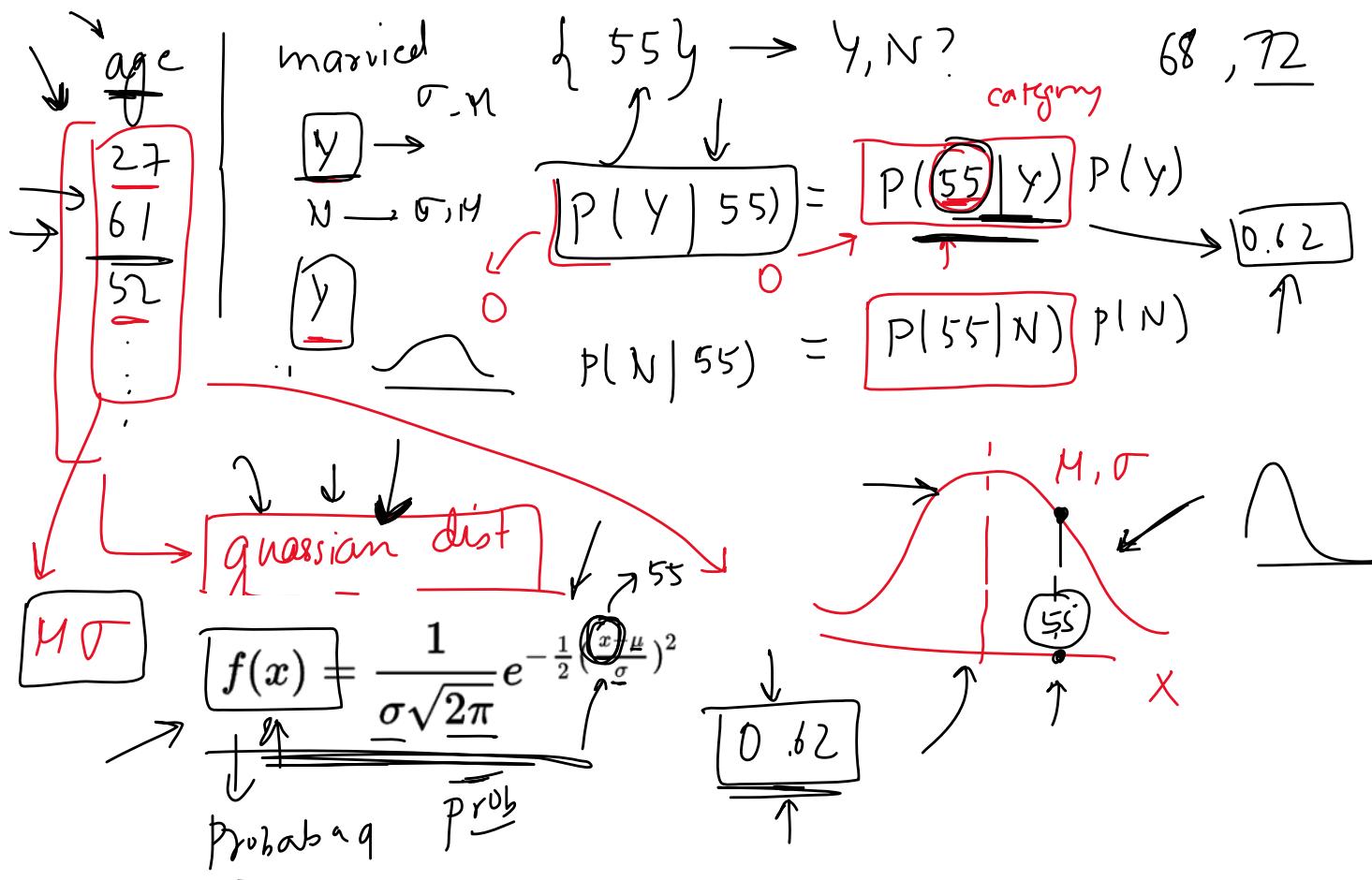
Outlook → Sunny | Overcast | Rainy

$$[3 \times 2] = 6 \text{ Prob} \quad 3 \times 2 \quad 2 \times 2$$

$P(\text{Sunny} \text{No})$	$P(\text{Hot} \text{Y})$	$P(\text{H} \text{Y})$
$P(\text{Sunny} \text{N})$	$P(\text{H} \text{N})$	$P(\text{H} \text{N})$
$P(\text{Overcast} \text{Yes})$	$P(\text{Mild} \text{Y})$	$P(\text{Mild} \text{Y})$
$P(\text{Overcast} \text{N})$	$P(\text{Mild} \text{N})$	$P(\text{Mild} \text{N})$
$P(\text{Rainy} \text{Yes})$	$P(\text{Cool} \text{Y})$	$P(\text{Cool} \text{Y})$
$P(\text{Rainy} \text{N})$	$P(\text{Cool} \text{N})$	$P(\text{Cool} \text{N})$

How Naïve Bayes handles numerical data?

22 June 2023 16:51



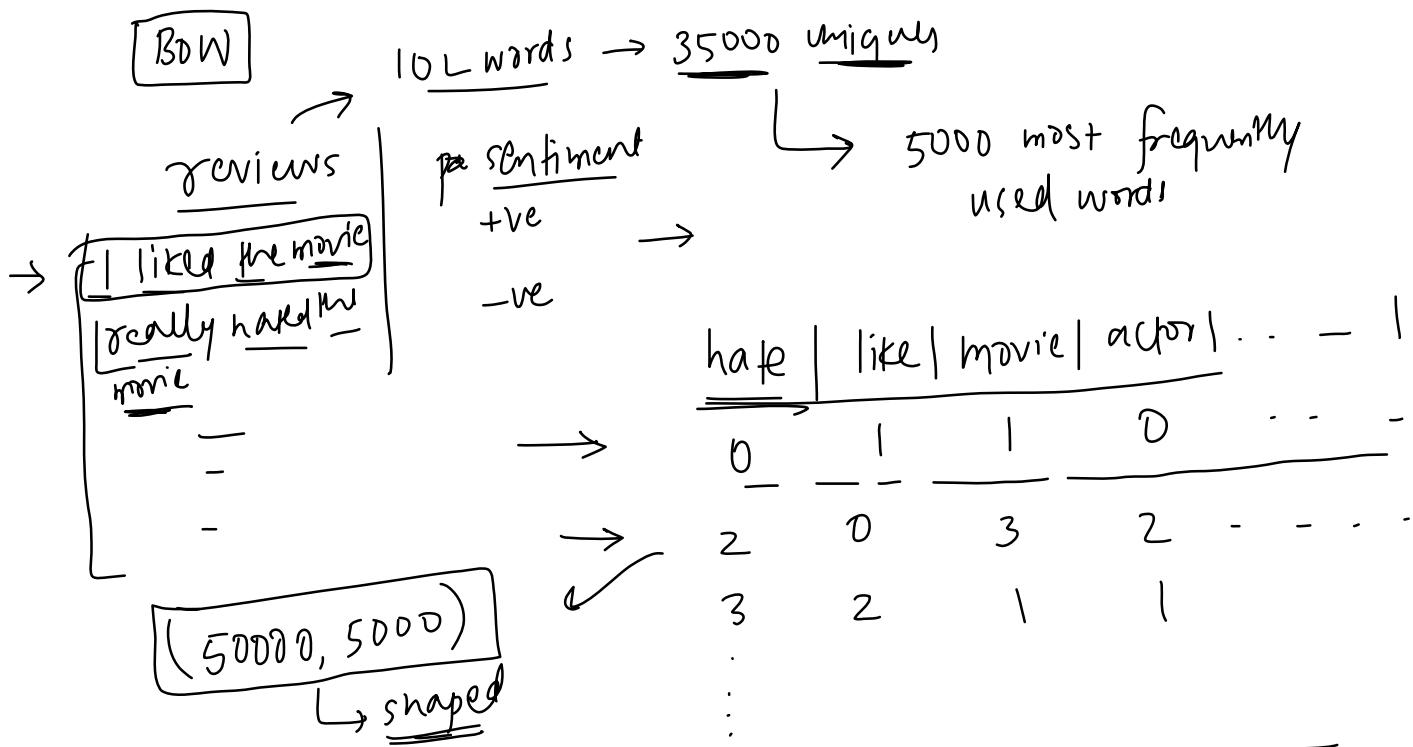
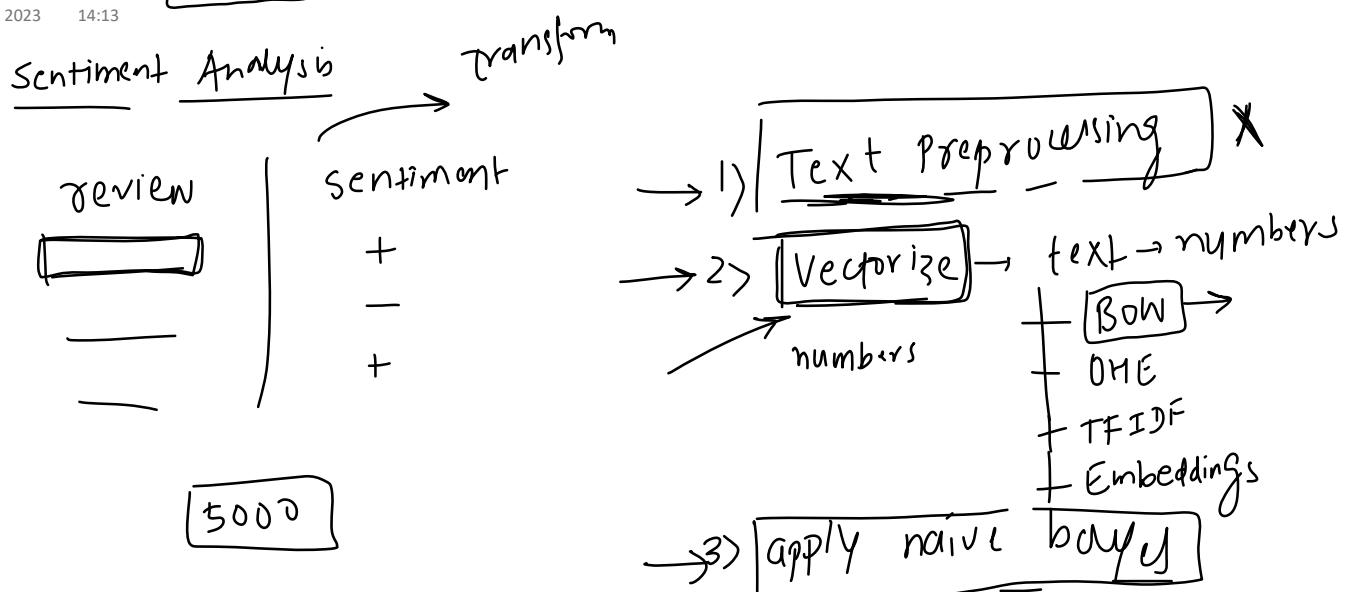
What is data is not Gaussian?

22 June 2023 16:52

1. Data Transformation: Depending on the nature of your data, you could apply a transformation to make it more normally distributed. Common transformations include the logarithm, square root, and reciprocal transformations.
2. Alternative Distributions: If you know or suspect that your data follow a specific non-normal distribution (e.g., exponential, Poisson, etc.), you can modify the Naïve Bayes algorithm to assume that specific distribution when calculating the likelihoods.
3. Discretization: You can turn your continuous data into categorical data by binning the values. There are various ways to decide on the bins, including equal width bins, equal frequency bins, or using a more sophisticated method like k-means clustering. Once your data is binned, you can use the standard Multinomial or Bernoulli Naïve Bayes methods.
→ Try out
binning
4. Kernel Density Estimation: A non-parametric way to estimate the probability density function of a random variable. Kernel density estimation can be used when the distribution is unknown.
5. Use other models: If none of the above options work well, it may be best to consider a different classification algorithm that doesn't make strong assumptions about the distributions of the features, such as Decision Trees, Random Forests, or Support Vector Machines

Naïve Bayes on Text Data

22 June 2023 14:13



Adj - - - cat movie

→ numbers

$$[0 \ 2 \ 1 \ - \ - \ - \ 5] \rightarrow \begin{pmatrix} 1,500 \\ \text{vector} \end{pmatrix}$$

→ +ve

→ -ve

$$P(+ve | \text{hate}=0, \text{like}=2, \dots) = 0.37$$

$$P(-ve | \text{hate}=0, \text{like}=2, \dots) = 0.23$$

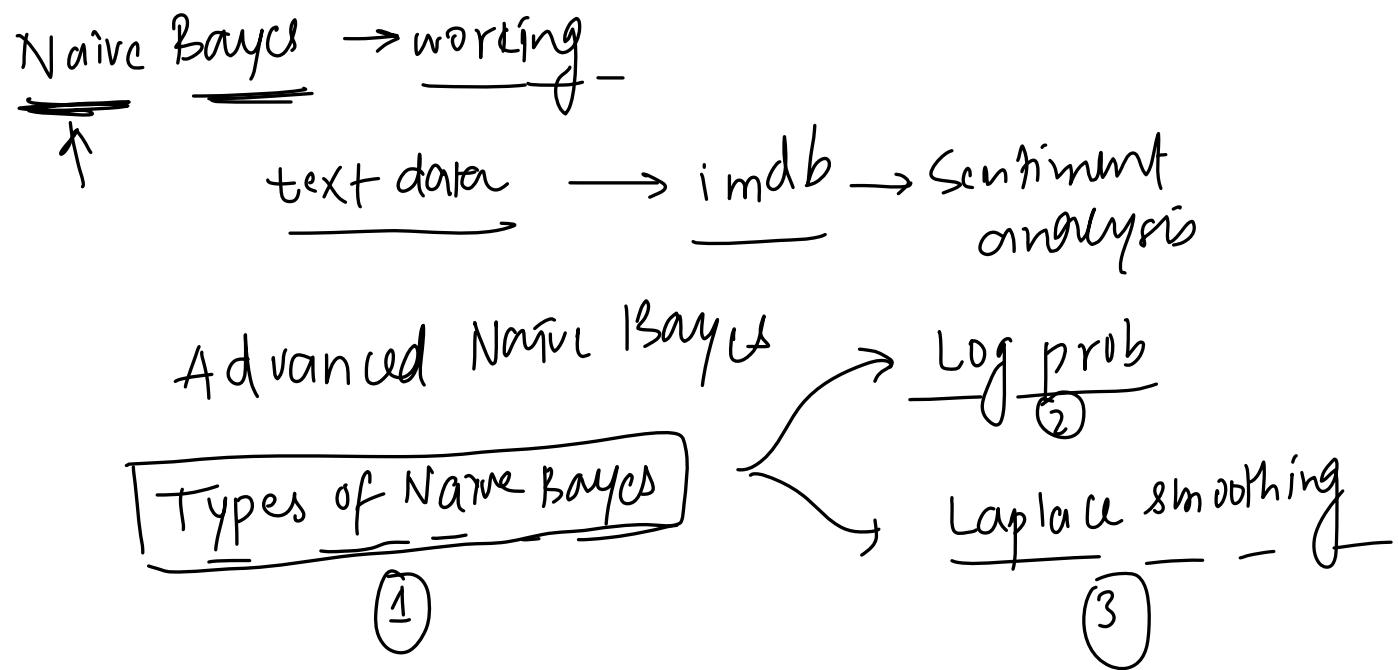
+ve
positive
sentiment

$$P(\text{hate}=0 | +ve) P(\text{like}=2 | +ve) \dots$$

↓
p

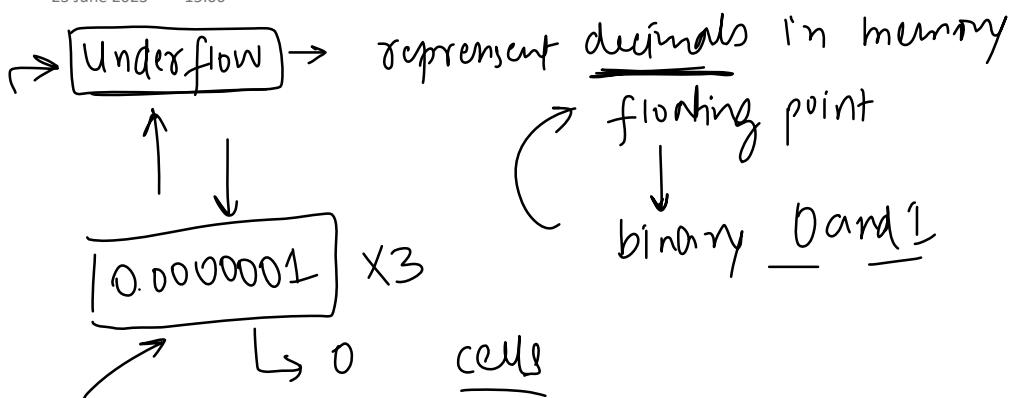
Recap

24 June 2023 09:32



[Numerical Stability]

23 June 2023 15:06



2500
100 wbs

$$0 = 0$$

cgpa | iq

placement

$$\{8.1, 81\} \rightarrow y|N$$

$$P(y|8.1, 81) = P(y) P(8.1|y) P(81|y)$$

$0 < x < 1$

$$\log(p_y) + \log(p(81|y)) + \log(p(8.1|y)) + \dots$$

$0.1 \times 0.3 \times 0.6 \times 0.6$

$$= \log_{\text{prob}} 0.0000$$

log probabilities

$$\log(p(a)p(b)p(c)\dots)$$

$$\log(ab) = \log a + \log b$$

$$\log(0.5) \rightarrow y = -153$$

$N = \lceil -135 \rceil$

N class ✓

$$\log(0.3 \times 0.5 \times 0.7)$$

$n + \log(10,0)$

$$\log(0.5) + \log(10.5) + \log(10.7)$$
$$\text{---} \quad \text{---} \quad \text{---}$$
$$-1.2 \quad -0.7 \quad -0.3 = \boxed{-2.2}$$

(N)

What is Underflow in Computing

24 June 2023 07:47

Underflow is a condition that can occur in computing when a number nears zero and the computer can no longer store it accurately in memory using floating-point representation. It happens when a calculated result is a smaller absolute value than the computer can actually represent.

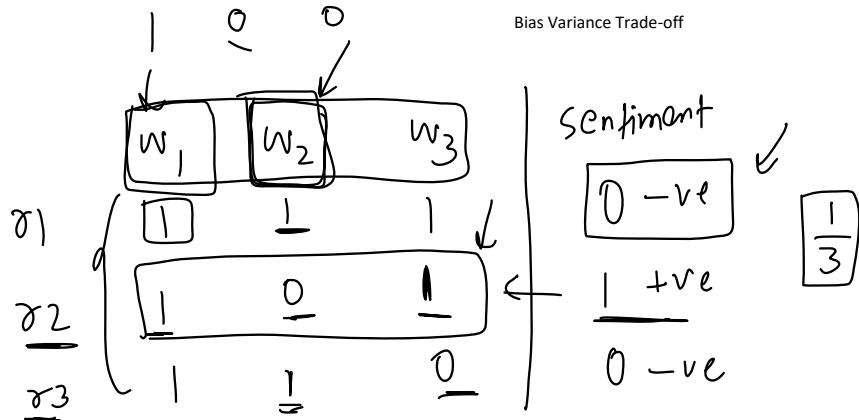
Most computers use a form of representation called floating-point to represent real numbers. This representation has a certain precision limit, and it can only represent numbers between a certain minimum and maximum value. If a number is too close to zero (but not zero), it might be smaller than the smallest representable positive number in the machine's floating-point representation. When an operation on such small numbers is performed, the machine might round the result to zero, leading to a loss of precision.

Underflow can be a problem in certain domains, such as machine learning, where calculations often involve probabilities. Probabilities are positive numbers that can be very close to zero. When multiplying many small probabilities together, the result can underflow. One common way to avoid underflow in such scenarios is to perform calculations in the log domain, where addition and subtraction are used instead of multiplication and division, thereby maintaining higher numerical precision.

Laplace Additive Smoothing

23 June 2023 15:13

	<u>binary bow</u>	<u>sentiment</u>
<u>review</u>	<u>w₁, w₂, w₃</u>	0
	<u>w₁, w₃, w₃</u>	1
	<u>w₂, w₂, w₁</u>	0



$\text{log } \rightarrow w_1, w_1, w_1 \rightarrow +ve, -ve$

$$w_0 = \frac{P(+ve | \text{log})}{P(-ve | \text{log})} = \frac{P(+ve)}{P(-ve)} \cdot \frac{P(w_1=1 | +ve)}{P(w_1=0 | -ve)} \cdot \frac{P(w_2=0 | +ve)}{P(w_2=1 | -ve)} \cdot \frac{P(w_3=0 | +ve)}{P(w_3=1 | -ve)}$$

$$0 = P(-ve | \text{log}) = \frac{2/3}{2/3} \cdot \frac{2/2}{2/2} \cdot \frac{0}{0} \cdot \frac{1/2}{1/2} \rightarrow 0.0000001 \text{ epsilon}$$

$\log = \log(0) = \text{undefined}$

$\alpha = \frac{1 + \alpha}{3 + n\alpha}$ - default

$$\frac{\alpha}{n\alpha} = \frac{1}{3}$$

$$\frac{0 + 1}{1 + 2(1)} = \frac{1}{3}$$

$$0.1 \rightarrow 100000$$

Bias Variance Tradeoff

$\text{prob} \rightarrow 0 \rightarrow$

Bias Variance Tradeoff

$$P(_) = \frac{+ \alpha}{n \alpha}$$

hyper

$$0 \rightarrow 0.00001$$

$$P(\underline{\quad})$$

↳ Laplace smo

$$= \frac{+\alpha}{+\eta\alpha} \rightarrow$$

flexibility → control
bias and variance

of

high bias

$\alpha \rightarrow$ low bias

high variance $\alpha \rightarrow$ low variance

$\alpha = \underline{\text{small}}$

$$\boxed{\alpha = 0}$$

min

min = 0

$$0.063$$

$$0.02$$

$$\frac{100}{N}$$

$$\boxed{f_1 | f_2 | \dots | y}$$

$$\begin{array}{c} 500 \\ \downarrow \\ \text{rows} \\ N = 100 \end{array}$$

$$\underline{p(y|x)} = \underline{p(y)}$$

$$0$$

$$\frac{0}{500}$$

$$\underline{\eta\alpha}$$

$$\underline{n}$$

$$-500 \leftarrow \underline{\text{small}}$$

$$-500 \leftarrow \underline{\text{small}}$$

$$0 \leftarrow$$

$$p(CN|X)$$

$$\frac{1}{500}$$

$$\frac{1}{500}$$

$$\alpha = 0, 0.01, 0.007$$

high variance

overfitting

high variance

$\alpha = \text{Very high}$

$$= 1000$$

$$\frac{1+1000}{500+2000} = \frac{1001}{2500} = \frac{101}{250} = \frac{1}{2} = \boxed{\frac{1}{2}}$$

$$= 10000$$

$$\alpha = 100000$$

$$\frac{1+10000}{500+20000} = \frac{10001}{20500} \approx \boxed{\frac{1}{2}}$$

$$\frac{100001}{200050} = \boxed{\frac{1}{2}}$$

$$\boxed{\frac{1}{2}}$$

$$\alpha \uparrow$$

$$\boxed{\frac{1}{2}}$$

$$\frac{+\alpha}{+\eta\alpha} = \boxed{\frac{1}{2}}$$

$$\alpha = 1$$

$$\boxed{y}$$

$$\boxed{k_2}$$

$$\boxed{p(y)}$$

$$\boxed{\frac{1}{2}}$$

$$\boxed{p(f_1|y)}$$

$$\boxed{\frac{1}{2}}$$

$$\boxed{p(f_2|y)}$$

$$\dots$$

$$\dots$$

$$\boxed{p(f_n|y)}$$

$$\dots$$

$$\boxed{p(f_m|y)}$$

same

$$p(y|x) = \boxed{p(N)}$$

$$p(N|x) = \frac{P(N)}{\frac{P(g_1|N)}{k_1} + \frac{P(g_2|N)}{k_2} + \dots}$$

data \rightarrow
 y
 N
underfitting

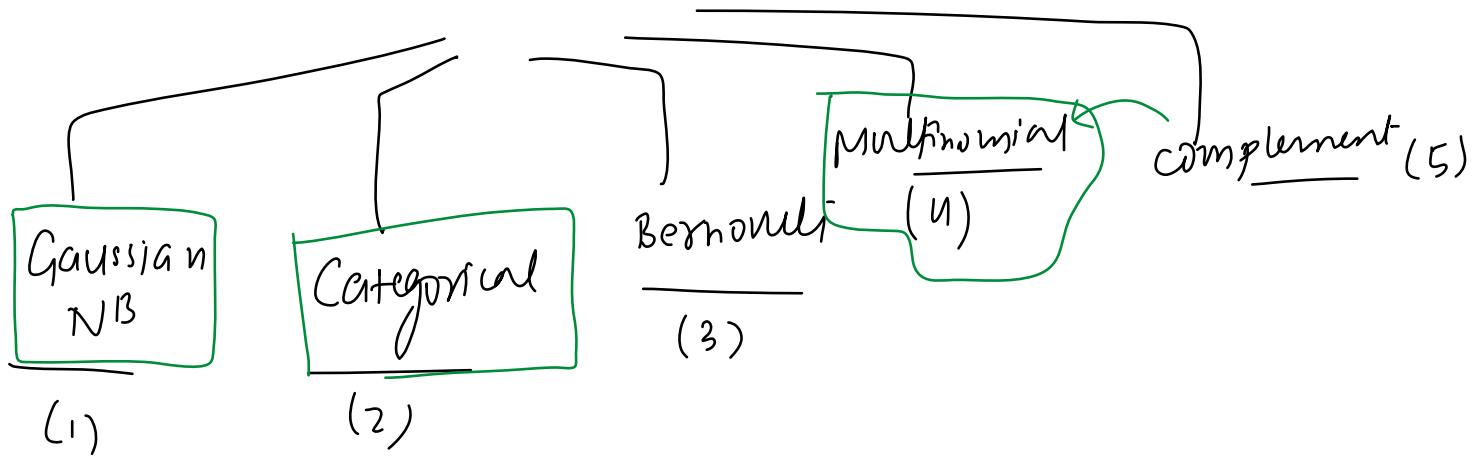
$X \rightarrow \text{array}$
 $\alpha = \text{hyperparameter}$
tune

$\alpha \uparrow$ high
lead to high bias
or
underfitting

$\alpha \downarrow$ low = 0
high variance
or
overfitting

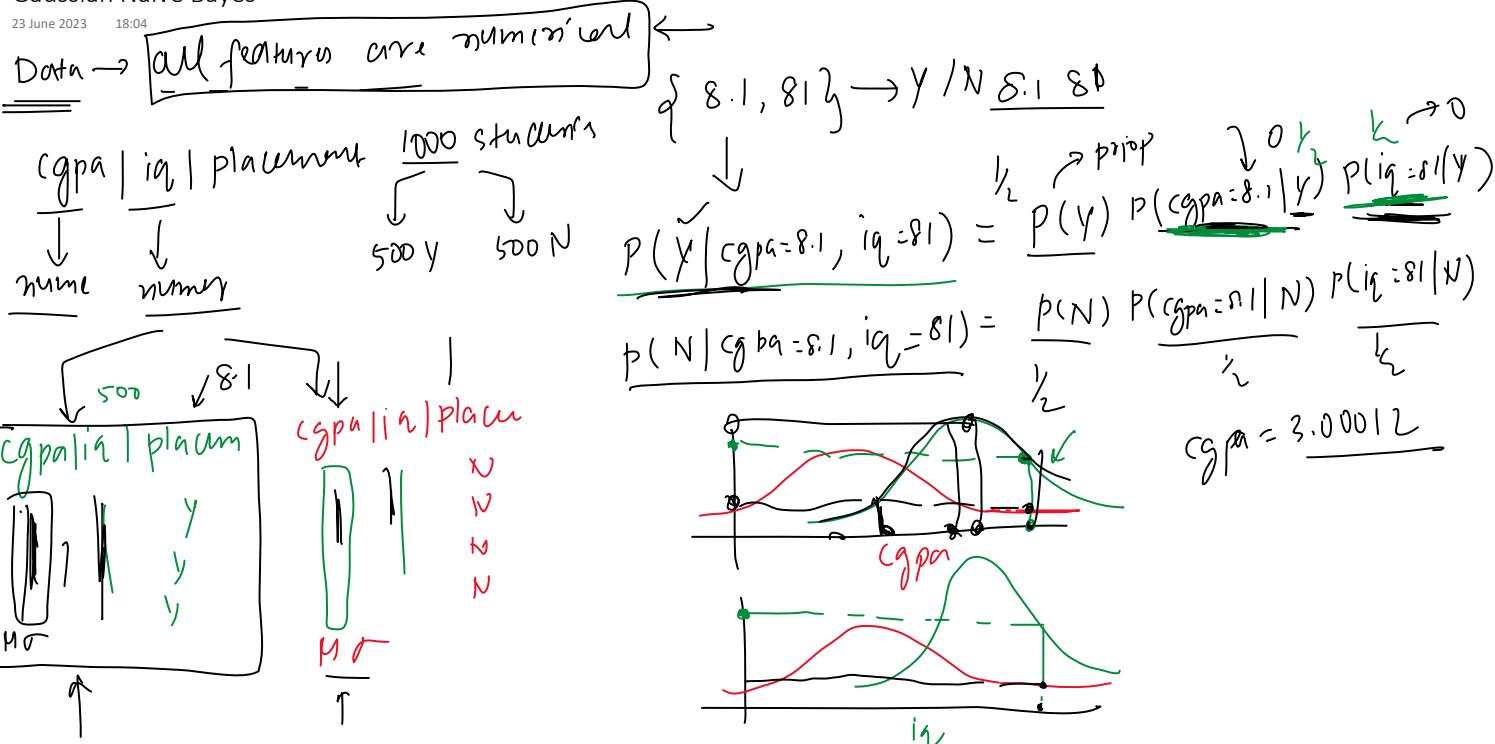
Types of Naïve Bayes

24 June 2023 07:48



Gaussian Naïve Bayes

23 June 2023 18:04



Question on Gaussian NB

24 June 2023 16:17

Why Laplace Additive Smoothing not applied on Gaussian Naïve Bayes?

Categorical Naïve Bayes

29 June 2023 15:22

Categorical Naïve Bayes is a variant of the Naïve Bayes algorithm designed specifically to handle categorical data.

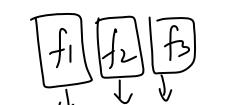
Data - all features are categorical

$x_q \{ \text{sunny, Hot, High, False} \}$ \downarrow
 y/N \rightarrow likelihood

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No ✓
Sunny	Hot	High	True	No ✓
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No ✓
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No ✓
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No ✓

\downarrow \downarrow

Data every feature should follow multinomial distribution



gender	p_{class}			survived
	f_1	f_2	f_3	
M	p_1			1
F	p_2			0
F	p_3			0

$$\{ M, P_3 \} \xrightarrow{\frac{1}{2}}$$

1
0
0

$$P(1 | \{ M, P_3 \})$$

1
0
0

$$P(0 | \{ M, P_3 \})$$

$$P(M | 1)$$

$$P(M | 0)$$

$$P(P_3 | 1)$$

$$P(P_3 | 0)$$

$$\frac{392}{891}$$

$$\frac{213}{891}$$

$$\frac{175}{891}$$

$$\frac{490}{891}$$

$$\frac{392}{891}$$

$$\frac{213}{891}$$

✓ 

$$p(M|D) = \frac{2 + \alpha}{3 + n\alpha} \rightarrow n=2$$

α → no. of categories

Question on Categorical NB

24 June 2023 14:40

Let's say I have 4 features in my dataset 2 of them are categorical like gender and is married and 2 are numerical like age and height. Can I apply Categorical Naïve Bayes?

1. Transform numerical features into categorical ones: You can discretize numerical features by binning them into different categories. For example, you could create an "age group" feature that bins age into categories like "0-18", "19-35", "36-50", "51+". This allows you to treat the numerical feature as categorical, so you can use Categorical Naive Bayes. However, you should be aware that this may lead to loss of information, as binning reduces the granularity of the data.
2. Use a mixed Naive Bayes model: These models can handle both numerical and categorical data by making different assumptions for different types of features. For instance, numerical features could be modelled using a Gaussian distribution while categorical features could be modelled using a multinomial or categorical distribution. You might need to look for a different library or implement it yourself.
3. Use another type of model: Some machine learning models can naturally handle mixed data types. Decision trees and their ensemble variants (like random forests and gradient boosted trees) are capable of handling both numerical and categorical features without requiring any explicit feature transformation.

Multinomial Naïve Bayes

23 June 2023 15:08

Multinomial Naive Bayes is a variant of the Naive Bayes algorithm that is particularly suited for classification tasks involving discrete features, such as text classification where features correspond to word counts or frequencies within the documents.

$P(w|c) = (T_{\{c,w\}} + 1) / (text_c + B)$

Here, '`T_{c,w}`' is the count of word '`w`' in class '`c`', '`text_c`' is the total count of words in class '`c`', and '`B`' is the size of the vocabulary.

f_1	f_2	f_3	f_4	y
2	1	4	3	0
1	2	10	11	1
0	5	6	11	1
9	1	2	3	0

Table 13.1: Data for parameter estimation examples.

	docID	words in document	in $c = \text{China}$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

BOW ↗ binary
non-binary (count)

movies review | sentiment

→ great -
fun -
- -

1 0 1

→ BOW
discrete

great fun epic

→ 2(1) 0 0
0 3(1) 2(1)

1 1 1

{count} bow multinomial Naive Bayes

fractives → Tf-idf

$$B = n d$$

Table 13.1: Data for parameter estimation examples.

	docID	words in document	in $c = \text{China?}$
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(w|c) = (T_{c,w} + 1) / (text_c + B)$$

Here, ' $T_{c,w}$ ' is the count of word ' w ' in class ' c ', ' $text_c$ ' is the total count of words in class ' c ', and ' B ' is the size of the vocabulary.

$$\begin{array}{ccccccc}
 d_2 & (2) & 0 & (1) & 0 & 0 & \\
 d_3 & (1) & 0 & 0 & 1 & 1 & \\
 d_4 & 1 & 0 & 0 & 1 & 1 & \\
 d_5 & 3 & 0 & 0 & 1 & 1 & \\
 \end{array}$$

$P(\text{margin} | y)^0 = 1$

$P(\text{Chinese} | y)$ $\frac{5}{8}$
 $\frac{0 + \alpha}{8 + \eta_d}$ $\alpha = 1$
 $\frac{0 + 1}{8 + 6 \times 1} = \frac{1}{14}$
 size of the vocabulary

$$\frac{P(y | \text{chinese} = 3, \text{bei} = 0, \text{sha} = 0, \text{mac} = 0, \text{torc} = 1, \text{jap} = 1)}{P(\text{Chinese} | y)^3 P(\text{bei} | y)^0} =$$

$$P(N | \text{chinese} = 3, \text{bei} = 0, \text{sha} = 0, \text{mac} = 0, \text{torc} = 1, \text{jap} = 1)$$

$$P(y) \xrightarrow{3/11} P(\text{Chinese} | y)^3 P(\text{bei} | y)^0 \dots P(\text{Japan} | y)^1$$

$$P(N) = \frac{1+1}{3+6(1)} = \frac{2}{9}$$

$$P(N) = \frac{1}{9} \quad \left. \begin{array}{l} \text{multinomial} \\ \text{distribution} \end{array} \right\} \downarrow$$

	docID	words in document	in $c = \text{China?}$
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	

$$\left. \begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \right\} \xrightarrow{\text{discrete/fractions}} \left. \begin{array}{l} 3 \\ 2 \\ 0 \end{array} \right\} \xrightarrow{\text{BOW}} \left. \begin{array}{l} 0 \\ 0 \\ 0 \end{array} \right\}$$

	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	Output
d1	2	1	0	0	0	0	y
d2	2	0	1	0	0	0	y
d3	1	-	0	0	1	0	y

$$\frac{5}{8} = \frac{\text{total words}}{8}$$

multinomial distribution

d_2	2	-	0	1	0	0	0	y	y
d_3	1	-	0	0	1	0	0	y	
d_4	1	-	0	0	0	1	1	N	
d_5	(3)	-	0	0	0	(1)	(1)	?	

mwm...
distribution

$$\left(\frac{0}{8} \right) \frac{0}{8}$$

$$p(y) \left[p(y) \left| \text{chin=3, bei=0, sha=0, mac=0, tok=1, Jap=1} \right. \right] = \\ p(y) \underbrace{p(\text{chinese}=3|y)}_{\frac{5}{8}} \underbrace{p(\text{bei}=0|y)}_{\frac{5}{8}} \underbrace{p(\text{sha}=0|y)}_{\frac{5}{8}} \underbrace{p(\text{mac}=0|y)}_{\frac{5}{8}} \underbrace{p(\text{tok}=1|y)}_{\frac{0}{8}} \underbrace{p(\text{Jap}=1|y)}_{\frac{0}{8}}$$

Chinese Chinese Chinese Tokyo Japan
 \downarrow \rightarrow
 $p(\text{chinese}|y)$ $p(\text{chinese}|y)$ $p(\text{chinese}|y)$ $p(\text{tokyo}|y)$ $p(\text{Japan}|y)$

$$\left(\frac{5}{8} \right) \quad \left(\frac{5}{8} \right) \quad \left(\frac{5}{8} \right) \quad \left(\frac{0}{8} \right) \quad \left(\frac{0}{8} \right)$$

$$\frac{5!}{3! \cdot 1! \cdot 1! \cdot 1!}$$

$$\rightarrow \left[\frac{3}{4} \left(\frac{5}{8} \right)^3 \left(\frac{0}{8} \right)^1 \left(\frac{0}{8} \right)^1 \right] =$$

$$\frac{0+\alpha}{8+\frac{n}{2}\alpha} (6)$$

of features

Bernoulli Naïve Bayes

23 June 2023 15:12

BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, Boolean) variable.

Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a BernoulliNB instance may binarize its input (depending on the binarize parameter).

3 categories

f_1	f_2
1	0
0	0
0	1
1	0

Table 13.1: Data for parameter estimation examples.

	docID	words in document	in c = China?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

→ Bag of words → binary

6 words bernoulli distributed

	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	output
d ₁	1	1	0	0	0	0	y
d ₂	1	0	1	0	0	0	y
d ₃	1	0	0	1	0	0	y
d ₄	1	0	0	0	1	1	N
d ₅	1	0	0	0	1	1	?

$$\begin{bmatrix} 0 \\ 3 \end{bmatrix} \rightarrow$$

$$3 \text{ out } y \\ p(y) = 3/5$$

$$\frac{1}{3}$$

$$\frac{0}{3} = 0$$

$$P(y | \text{Chinese}=1, \text{Bei}=0, \text{Sha}=0, \text{Mac}=0, \text{Tok}=1, \text{Jap}=1) = k=0$$

$$P(N | \text{chi}=1, \text{bei}=0, \text{sha}=0, \text{ma}=0, \text{tok}=1, \text{jap}=1)$$

$$p_k + (1-p)^{1-k} (p)$$

$$p(y) p(\text{Chinese}=1|y) p(\text{Bei}=0|y) p(\text{Sha}=0|y) p(\text{Mac}=0|y) p(\text{Tok}=1|y) p(\text{Jap}=1|y)$$

$$\begin{aligned} & (1-p) \\ & \downarrow \\ & 2/3 \end{aligned}$$

$$\begin{aligned} & p \\ & \downarrow \\ & 1/3 \end{aligned}$$

$$P(\text{Chinese}=1|y) = 2/3$$

$$\boxed{\frac{3}{3}} = 1$$

$$p(X=k) = p^k + (1-p)^{1-k}$$

$$k=0, 1$$

$$p(X=1) = \frac{p}{p+1}$$

(Chinese)

p(1)

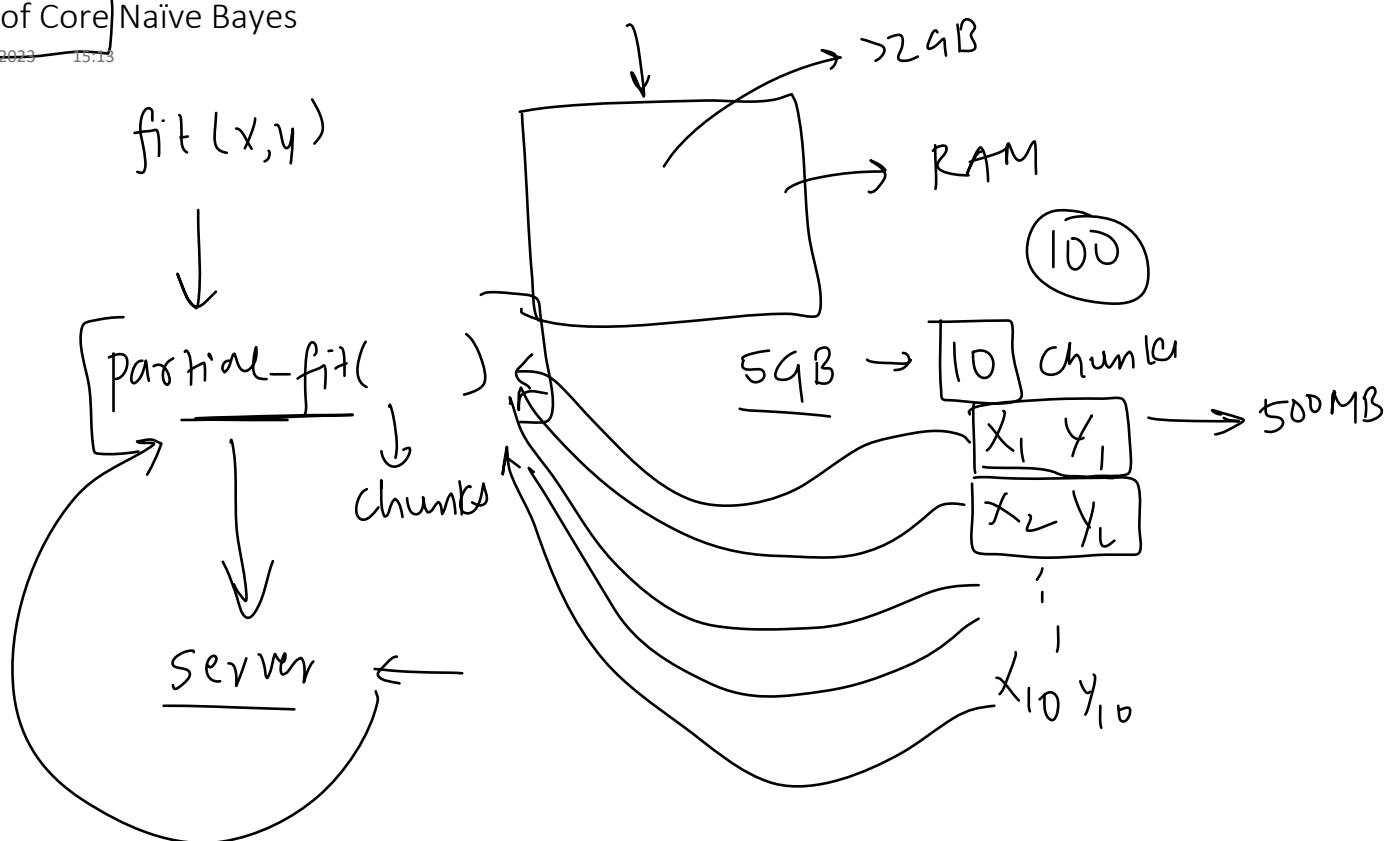
$$\begin{aligned} & D + \alpha \rightarrow 4 \\ & 3 + n\alpha \downarrow \\ & \rightarrow (2) \end{aligned}$$

Complement Naïve Bayes

23 June 2023 15:12

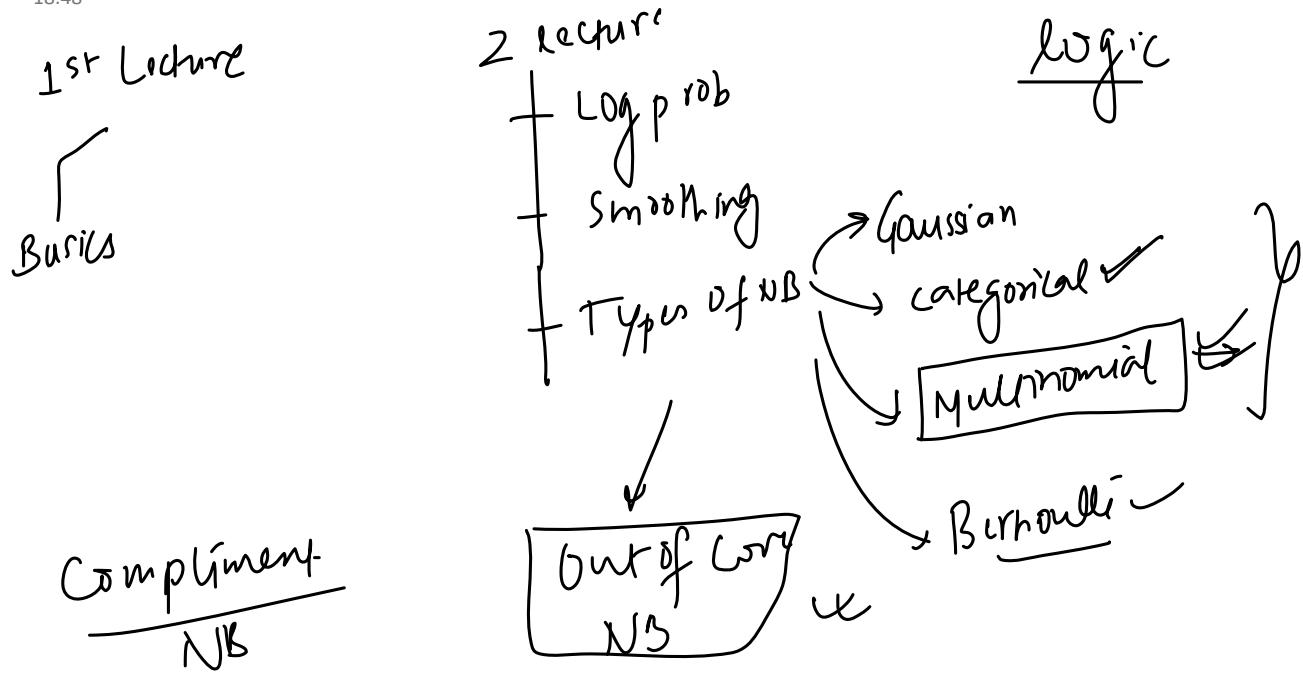
Out of Core Naïve Bayes

23 June 2023 15:13



Recap

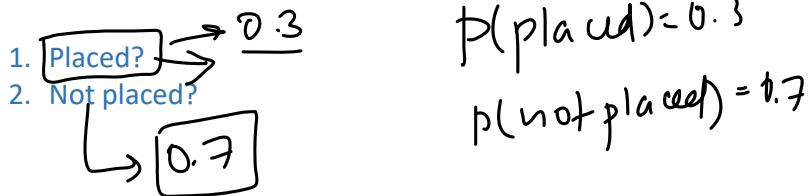
27 June 2023 18:48



Probability Distributions

27 June 2023 08:03

S An engineering college has a placement rate of 0.3, meaning that any given student has a 0.3 chance of getting placed through campus recruitment. If you randomly select a student what is the probability that the student is:



$$\text{Bernoulli} \rightarrow (X) = \{0, 1\}$$

1 trial 1-place
 0-not place

$$P(X=k) = \frac{p^k}{k!} + \frac{(1-p)^k}{k!}$$

$k=0, 1$ p = prob of getting 1
 $(1-p)$ \textcircled{P}

An engineering college has a placement rate of 0.3, meaning that any given student has a 0.3 chance of getting placed through campus recruitment. If you randomly select 10 students, what is the probability that:

1. 9 out of 10 students get placed? →
2. 3 out of 10 students get placed?

$$\underline{n \text{ trials}} = 1$$

Binomial distribution

$$n=10$$

\downarrow
Bernoulli

$$\overbrace{N \xrightarrow{\quad} P \xrightarrow{\quad} P}^{\text{Initial}} \xrightarrow{\quad} P \xrightarrow{\quad} P$$

$$10 \times (0.3)^9 (0.7)^1$$

Small

10 x 9.00 00 small per cu

10x small prw

$$p(X) = \binom{n}{k} p^k (1-p)^{n-k}$$

pmf

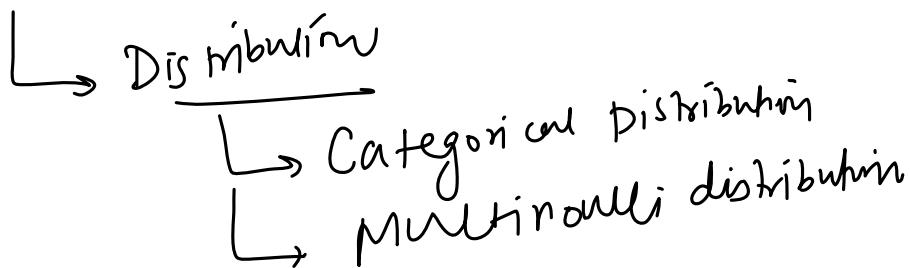
{ An engineering college has a placement system where any given student has a 0.3 chance of getting placed through campus recruitment, a 0.05 chance of opting out of the placement process, and a 0.65 chance of trying but not getting placed. If you randomly select a student, what is the probability that the student:

trial = 1
n = 1

- 1. Gets placed?
- 2. Doesn't get placed but doesn't opt out either? $\rightarrow 0.65)$
- 3. opts out of placement?

{ $p(\text{placed}) = 0.3$
 $p(\text{optout}) = 0.05$
 $p(\text{not placed}) = 0.65$

not binary
categorical
more than 2 categories



{ An engineering college has a placement system where any given student has a 0.3 chance of getting placed through campus recruitment, a 0.05 chance of opting out of the placement process, and a 0.65 chance of trying but not getting placed. If you randomly select 10 students, what is the probability that:

n = 16

- 1. 3 students get placed, 1 student opts out of placement, and 6 students try but do not get placed?
- 2. No student gets placed, 2 students opt out of placement, and 8 students try but do not get placed?

$$p(\text{placed}) = 0.3$$

$$\rightarrow p(\text{not placed}) = 0.65$$

$$P(\text{Placed}) = \dots$$

$$P(\text{Drown}) = 0.65$$

$$P(\text{not placed}) = 0.35$$

3 placed + 0 placed = 6 not placed

$k = \text{no. of cases}$

$\frac{n!}{n_1! n_2! n_3! \dots n_k!}$

$(0.3)^3 (0.65)^1 (0.35)^6$

$\frac{10!}{3! 1! 6!}$

$\frac{10!}{2! 8!} (0.3)^0 (0.65)^2 (0.35)^8$

Multinomial distribution

The diagram illustrates the derivation of the multinomial coefficient. It shows the total number of ways to arrange 10 items (3 P's, 1 D, 6 N's) as $10!$, and then divides this by the product of factorials of the counts of each category: $3! 1! 6!$. Arrows point from the labels "3 placed" and "0 placed" to the terms $(0.3)^3 (0.65)^1$ and $(0.35)^6$ respectively. Another arrow points from the term $(0.3)^0 (0.65)^2 (0.35)^8$ to the label "Multinomial distribution".

Definitions

27 June 2023 08:35

The Bernoulli distribution is a discrete probability distribution that models the outcomes of a binary random variable.

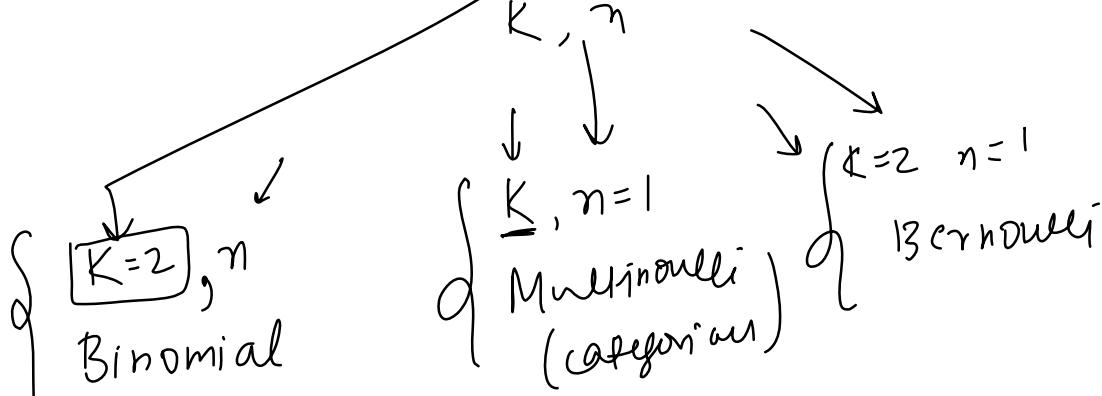
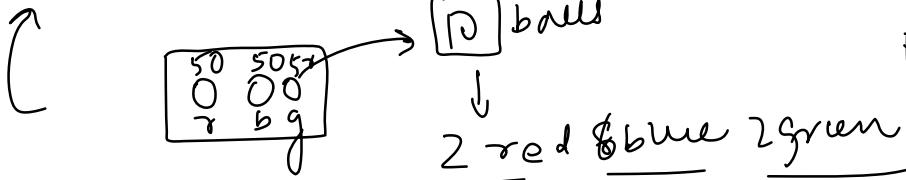
The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials.



The categorical distribution is a discrete probability distribution that models the probabilities of different outcomes in a categorical or discrete random variable.
Unlike the Bernoulli or binomial distributions that deal with binary outcomes, the categorical distribution accommodates multiple categories or outcomes. Each category has an associated probability, and the sum of the probabilities for all categories is equal to 1.

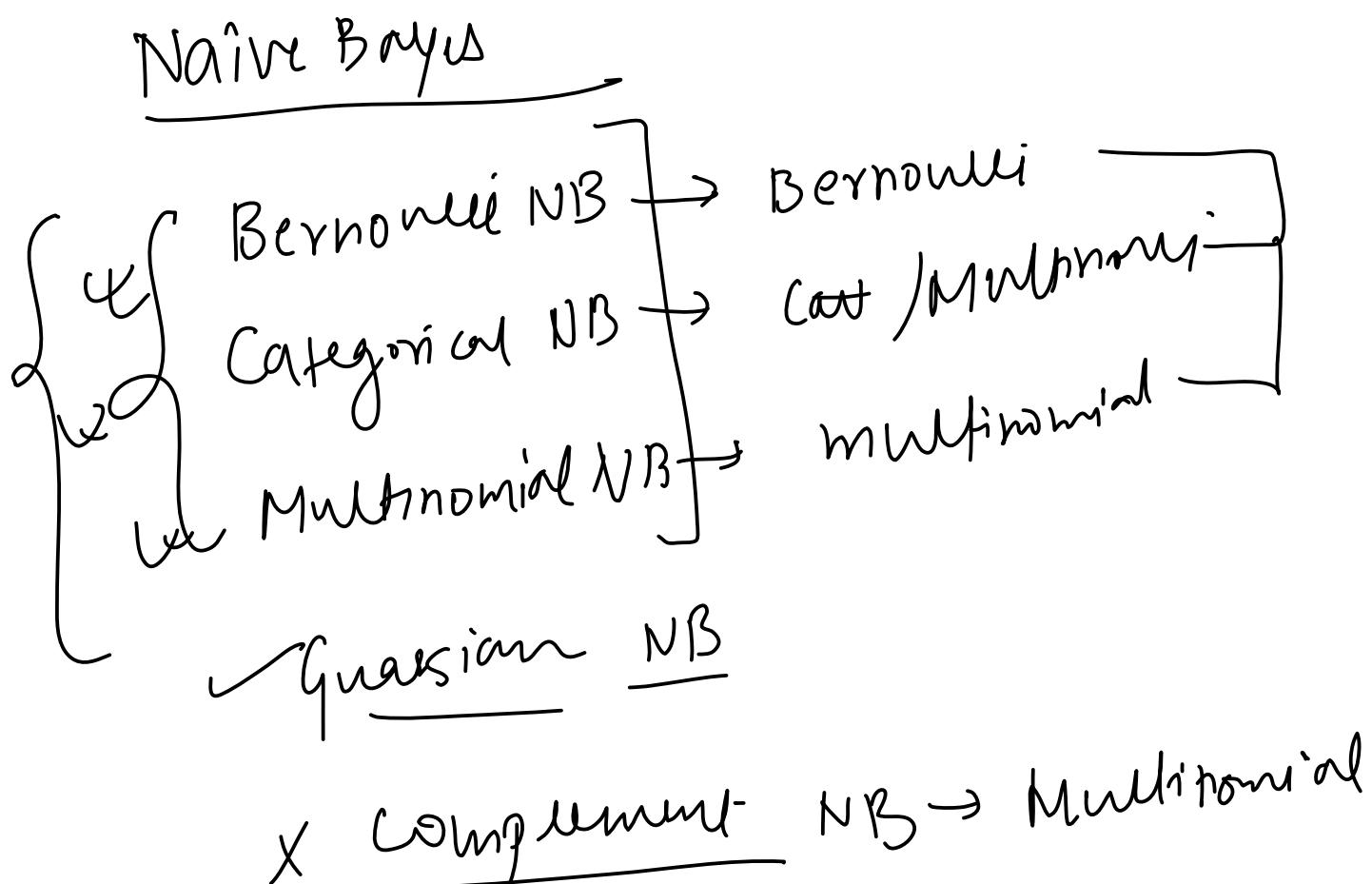
$$P(X=k) = P_k$$

Multinomial distribution allows us to calculate the probability of observing a specific count or combination of counts for each category in a fixed number of trials.



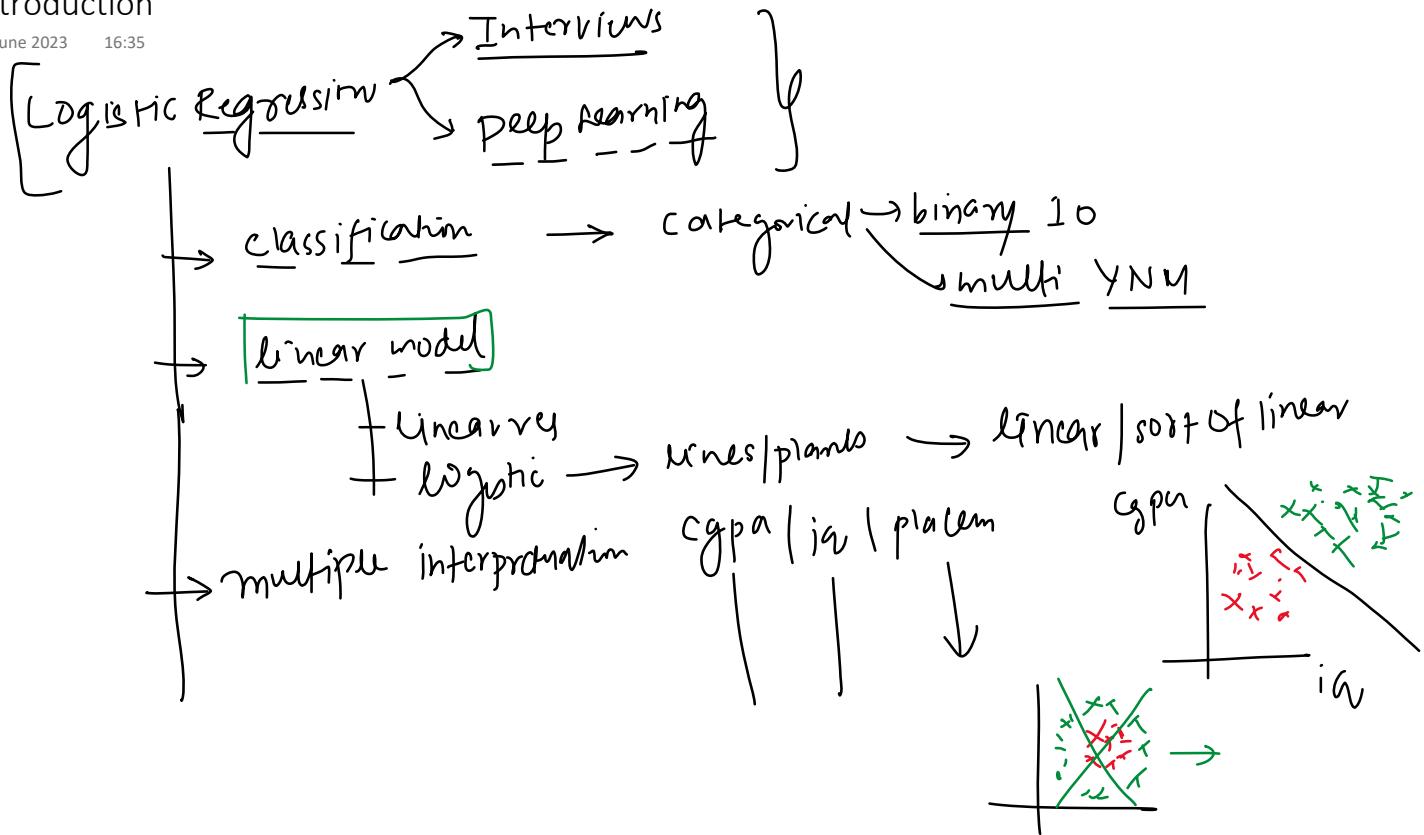
Why did I teach these now?

27 June 2023 08:44



Introduction

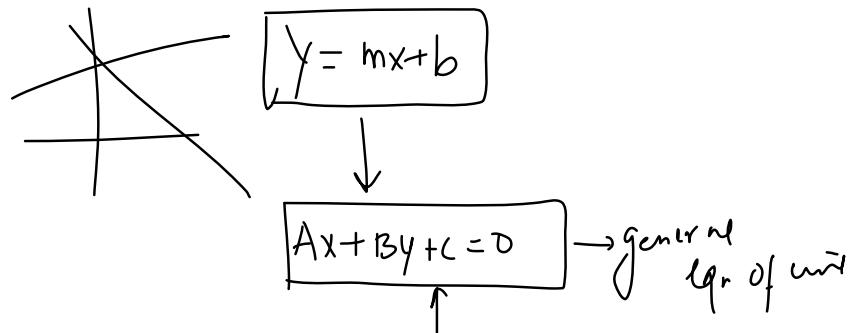
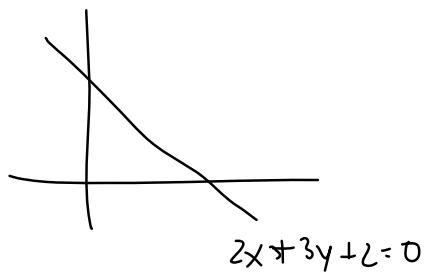
28 June 2023 16:35



Some Basic Geometry

28 June 2023 16:41

- Every line has a positive side and a negative side.



- How to find out if a given point lies on a given line?

$$y_1 + 3y_2 + 5 = 0 \quad (5, 2)$$

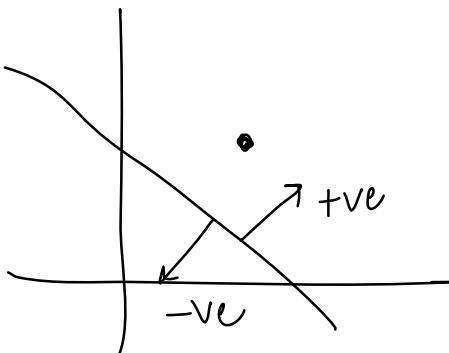
$$5 + 3(2) + 5 = 0 \quad (0)$$

$$Ax_1 + By_1 + C = 0$$

$$(x_1, y_1) \rightarrow (x_1, y_1, 1)$$

$$Ax_1 + By_1 + C(1) = 0$$

- How to find out if a given point is on the positive side of the line or the negative side of the line.

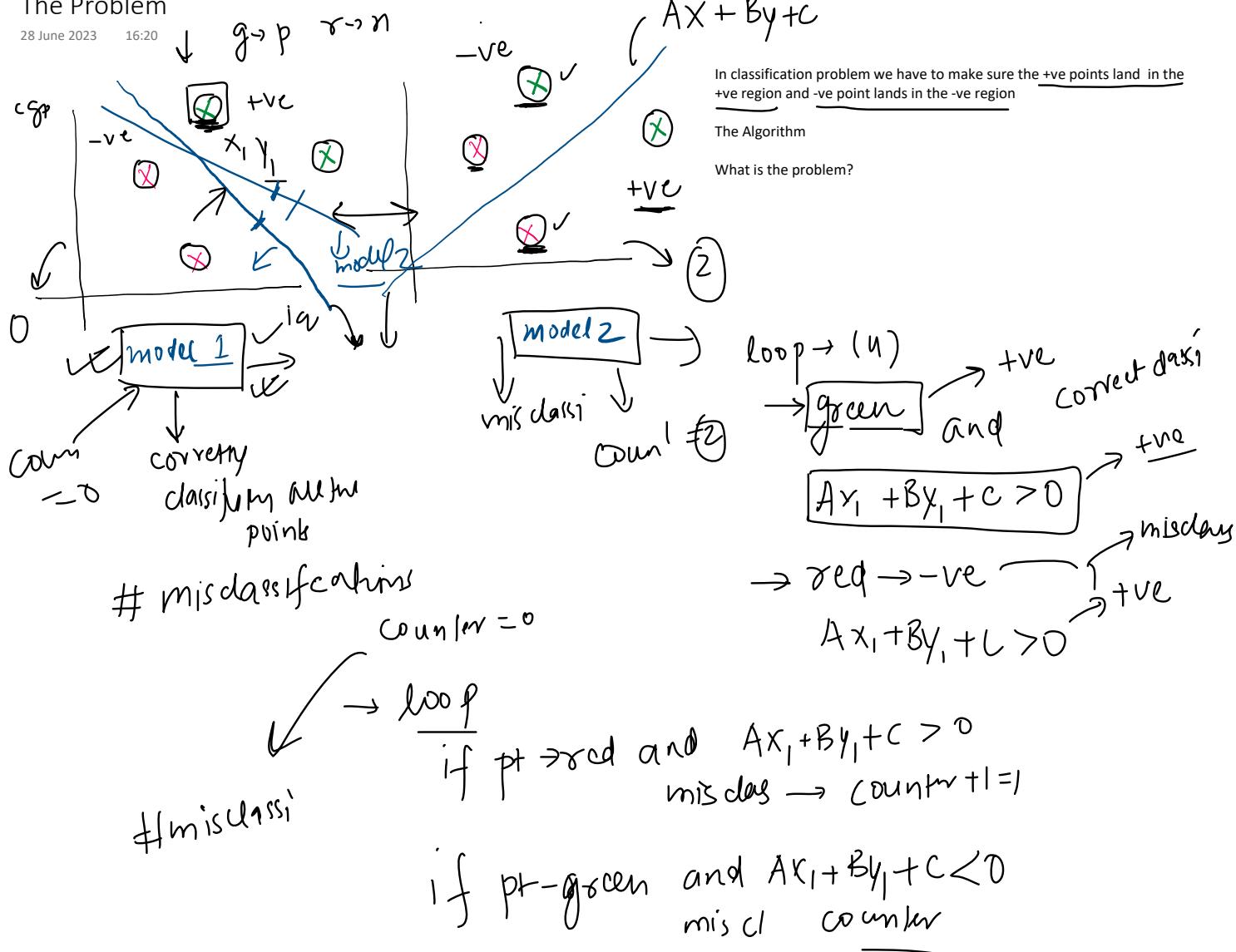


$$Ax_1 + By_1 + C > 0 \rightarrow +ve \text{ region}$$

$$Ax_1 + By_1 + C < 0 \rightarrow -ve \text{ region}$$

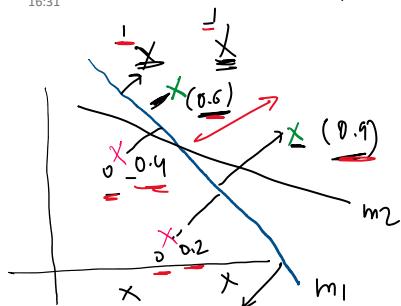
The Problem

28 June 2023 16:20



New Problem Formulation

28 June 2023 16:31

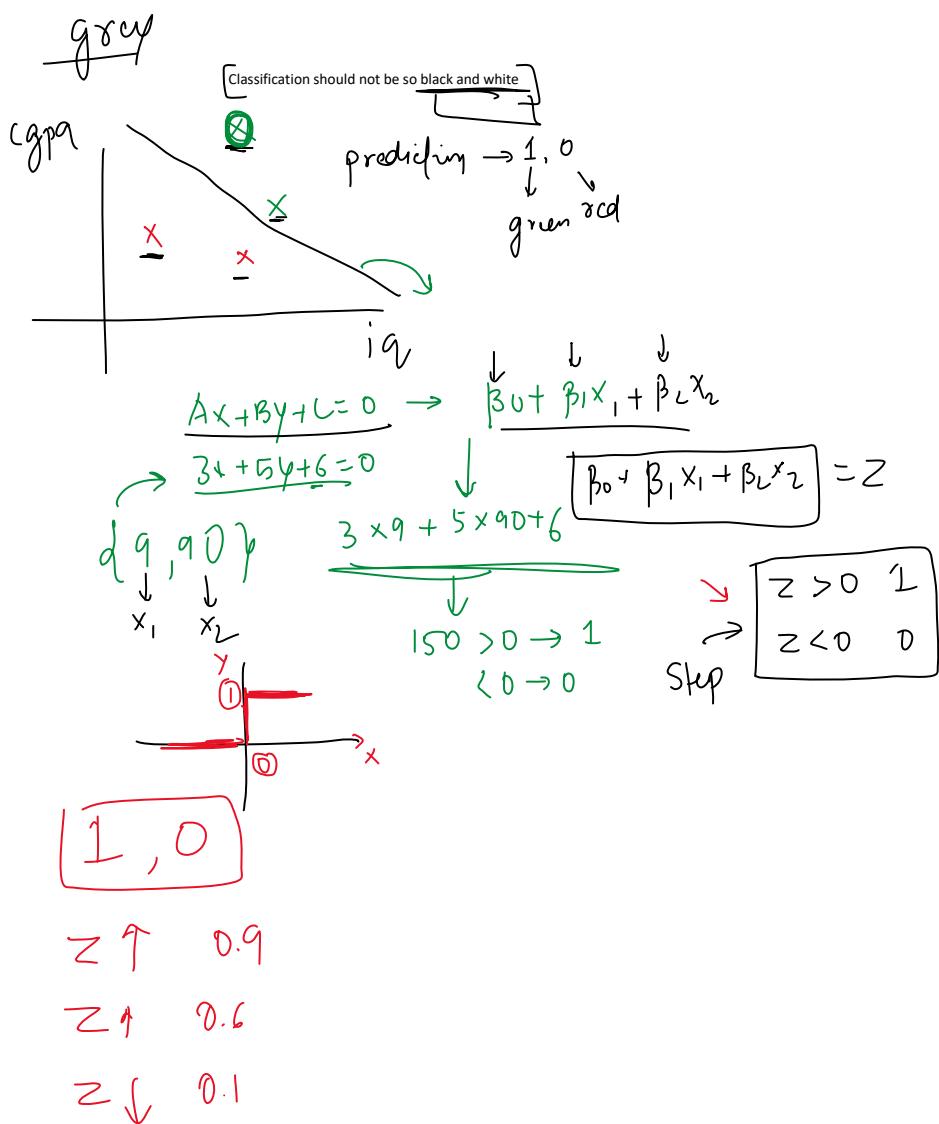


$$\begin{array}{c|cc|c} \text{iq} & \text{cgpa} & p & \text{Ind} \\ \hline 6 & 60 & 0 \\ 4 & 40 & 0 \\ 8 & 80 & 1 \\ \hline 9 & 90 & 1 \end{array} \rightarrow$$

$$w \cdot z \rightarrow \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

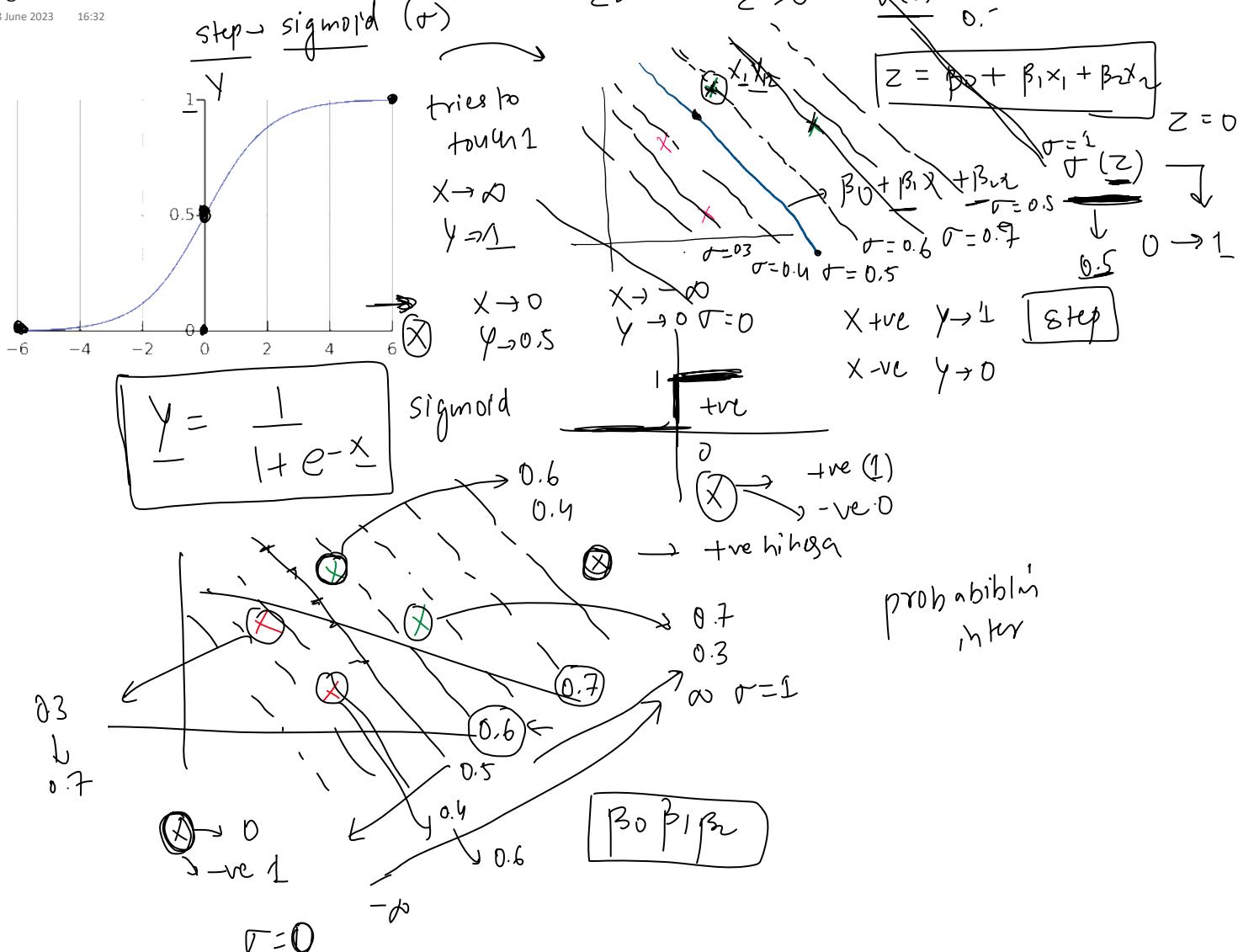
\rightarrow Step(z) $\rightarrow 0, 1$

\times



Sigmoid Function

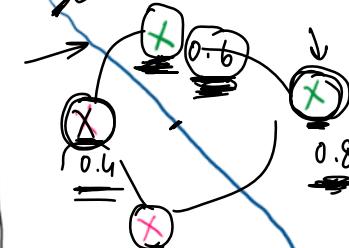
28 June 2023 16:32



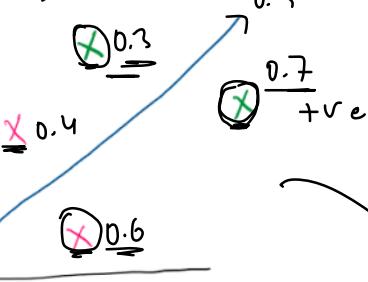
Maximum Likelihood

28 June 2023 16:32

0.20 (1)



(2) -ve 0.5



The likelihood function is the product of the predicted probabilities for the actual class of each observation

model 1
0.3
likelihood

$$0.8 \times 0.6 \times 0.2 \times 0.6$$

$$0.20$$

$$0.7 \times 0.4 \times 0.3 \times 0.6$$

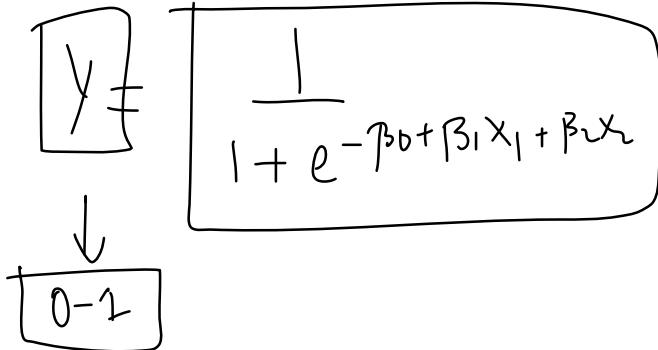
$$0.05$$

maximum
↳ best line
↳ logisitic ln

maximum
↳ better model

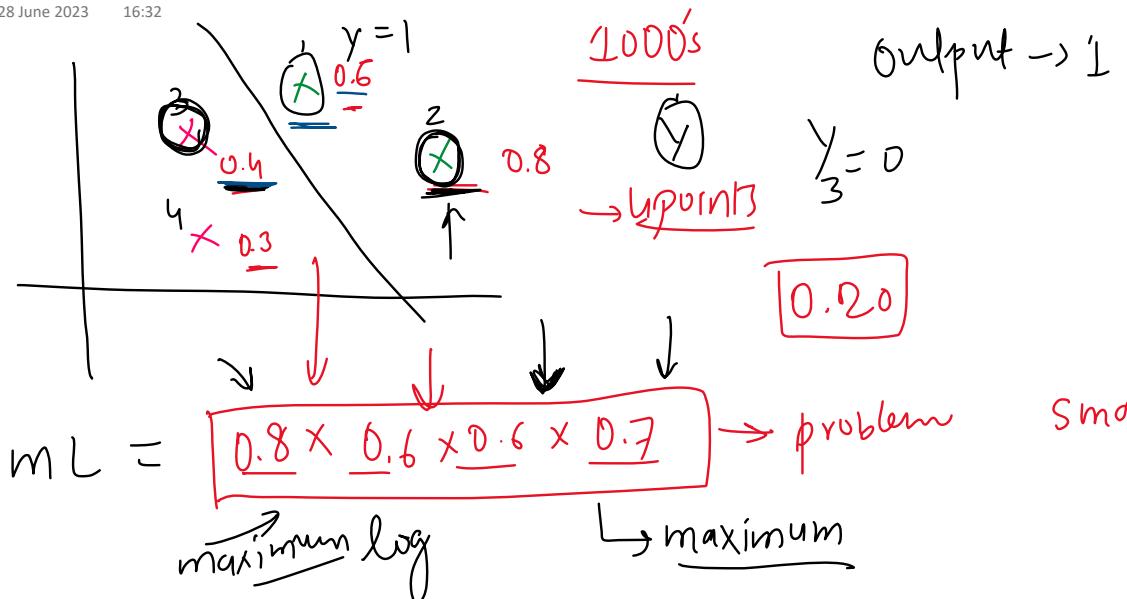
$$Y = \frac{1}{1 + e^{-X}}$$

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



Log Loss

28 June 2023 16:32



$$\begin{aligned}
 \log(m_l) &= \log(0.8 \times 0.6 \times 0.6 \times 0.7) \\
 &= \underline{\log 0.8} + \underline{\log 0.6} + \underline{\log 0.6} + \underline{\log 0.7} \\
 &= \boxed{-\log 0.8 - \log 0.6 - \log 0.6 - \log 0.7} \\
 &\xrightarrow{\text{minimize}} 0.1 \rightarrow 1
 \end{aligned}$$

log (0-1)
 -ve number
 positive minimum

$$-\log(\hat{y}_1) + \cancel{\log(\hat{y}_2)} - \log(\underline{\hat{y}_3}) - \log(\underline{\hat{y}_4})$$

$$\hat{y}_1 = \tau(z) \rightarrow (p) \rightarrow \text{getting green}$$

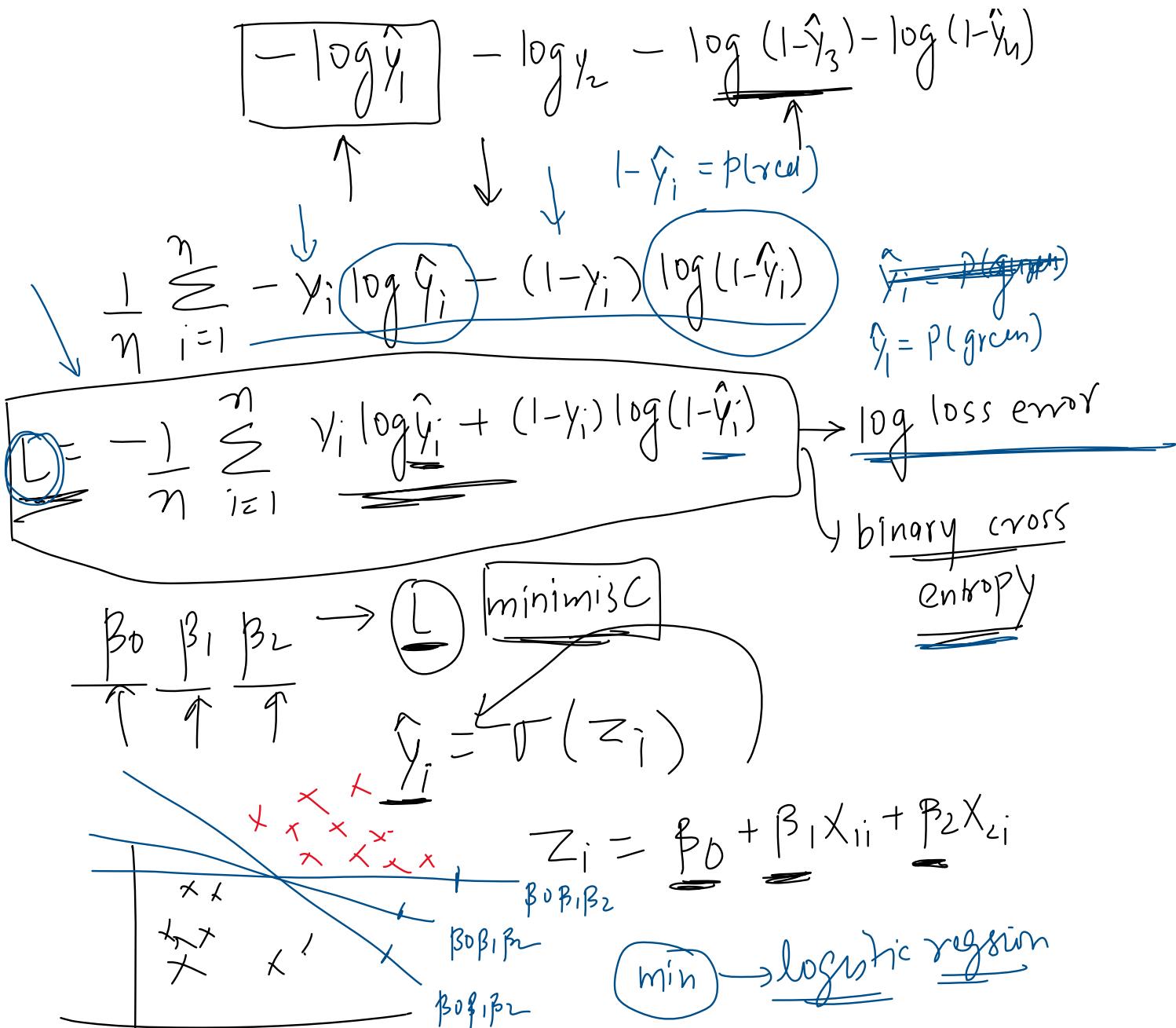
$\hookrightarrow z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$$\begin{matrix} \hat{Y}_1 \rightarrow P_1 \\ \hat{Y}_2 \rightarrow P_2 \end{matrix} \quad \left. \right\} \text{getting green}$$

$$-\hat{y}_1 \log \hat{y}_1 - (1-\hat{y}_1) \log (1-\hat{y}_1)$$

$$(1-\sigma) \log(1-\hat{Y}_3)$$

$$[-\log \hat{y}_1] - \log y_2 - \log (1-\hat{y}_3) - \log (1-\hat{y}_4)$$



Gradient Descent

28 June 2023 16:32

\min

$$L = -\frac{1}{n} \sum_{i=1}^n y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

Closed form solution
mse →
 $\beta_0 \beta_1 \beta_2$

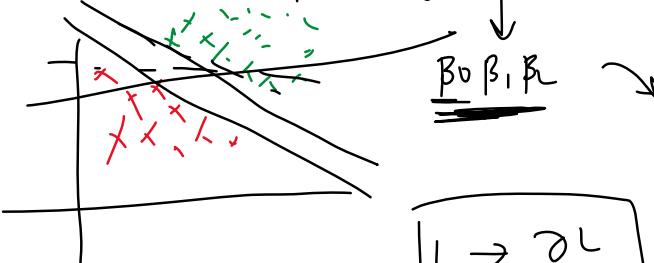
x_1	x_2	y	\hat{y}_i
1	2	0	0.37
1	2	1	0.37
0	0	0	0.89

$\hat{y}_i \rightarrow p(\text{green})$
 $(1-\hat{y}_i) \rightarrow p(\text{red})$

$(100) = n$

$$1 - 0 = 1$$

$$1 - 1 = 0$$



Gradient Descent

$$\beta_0 \beta_1 \beta_2 \rightarrow L \min$$

$$\beta_0 = \beta_0 - \eta \frac{\partial L}{\partial \beta_0} \rightarrow \text{gradient}$$

learning rate

$$\beta_1 = \beta_1 - \eta \frac{\partial L}{\partial \beta_1}$$

$$\beta_2 = \beta_2 - \eta \frac{\partial L}{\partial \beta_2}$$

$$\frac{\partial L}{\partial \beta_0} \quad \frac{\partial L}{\partial \beta_1} \quad \frac{\partial L}{\partial \beta_2}$$

$$\hat{y}_i = p(\text{green})$$

$$\frac{\partial L}{\partial \beta_1} = -\frac{1}{n} \sum_{i=1}^n y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i) \quad n \rightarrow \text{points}$$

$\frac{\partial}{\partial x} \sigma(x) \rightarrow \sigma(x)[1-\sigma(x)] \quad \log x \rightarrow \frac{1}{x}$ 1 point

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \left[-y \log \hat{y} - (1-y) \log(1-\hat{y}) \right]$$

$$\hat{y} = p(\text{green})$$

$$-y \sigma(z)[1-\sigma(z)] x_1$$

$$\sigma(z) \downarrow \underline{\beta_1}$$

$$\frac{-y}{\hat{y}} \underset{\approx}{=} \sigma(z) [1 - \sigma(z)] x_1$$

$$\nabla L \leftarrow \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\text{sig} = \sigma(z)(1-\sigma(z))$$

$$\boxed{\frac{-y}{\hat{y}} \underset{\approx}{=} (\hat{y}(1-\hat{y})) x_1}$$

$$\boxed{-y(1-\hat{y})x_1}$$

$$\frac{\partial \hat{y}}{\partial \beta_1} \quad \frac{\partial \sigma(z)}{\partial \beta_1}$$

$$z \rightarrow \underline{\beta_1 x_1}$$

$$\frac{\partial}{\partial \beta_1} - (1-y) \log(\underline{1-\hat{y}}) = \odot \frac{(1-y) \hat{y}(1-\hat{y}) x_1}{1-\hat{y}}$$

$$\boxed{(1-y) \hat{y} x_1}$$

$$-y(1-\hat{y})x_1 + (1-y)\hat{y}x_1$$

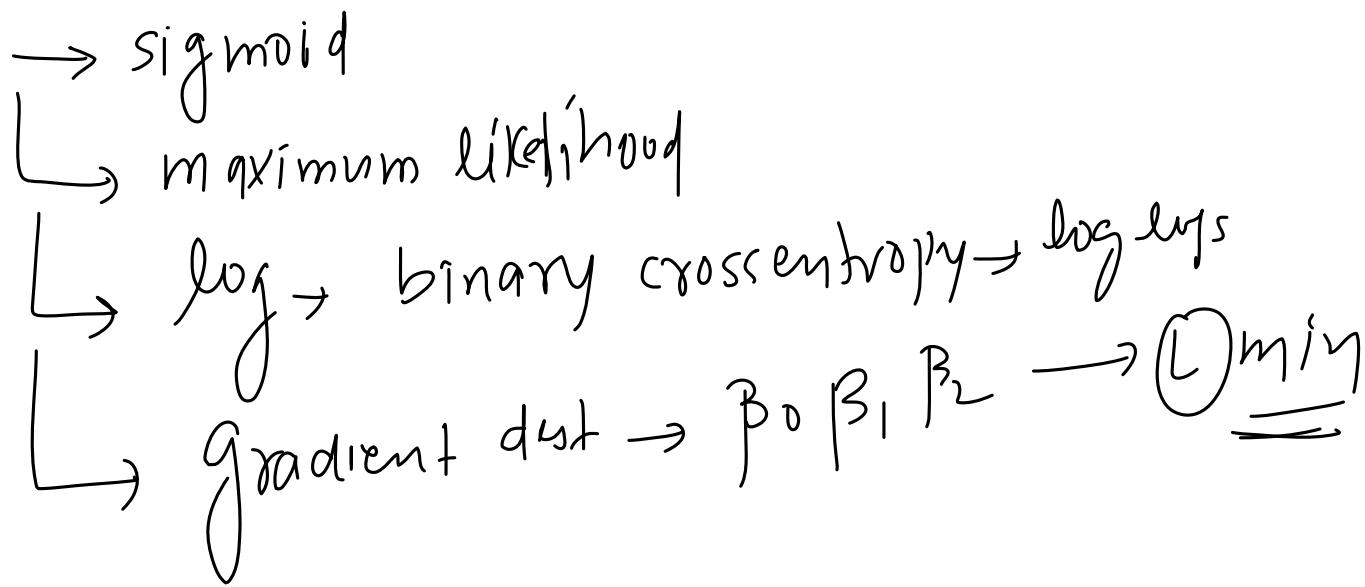
$$\boxed{[-y + y/\cancel{\hat{y}} + \cancel{\hat{y}} - y/\cancel{\hat{y}}] x_1}$$

$$\boxed{\frac{\partial L}{\partial \beta_1} \underset{\approx}{=} (\hat{y}_i - y_i) x_{1i}} \rightarrow \frac{\partial L}{\partial \beta_2} = (\hat{y} - y) x_2$$

$$\frac{\partial L}{\partial \beta_0} = (\hat{y} - y) \rightarrow$$

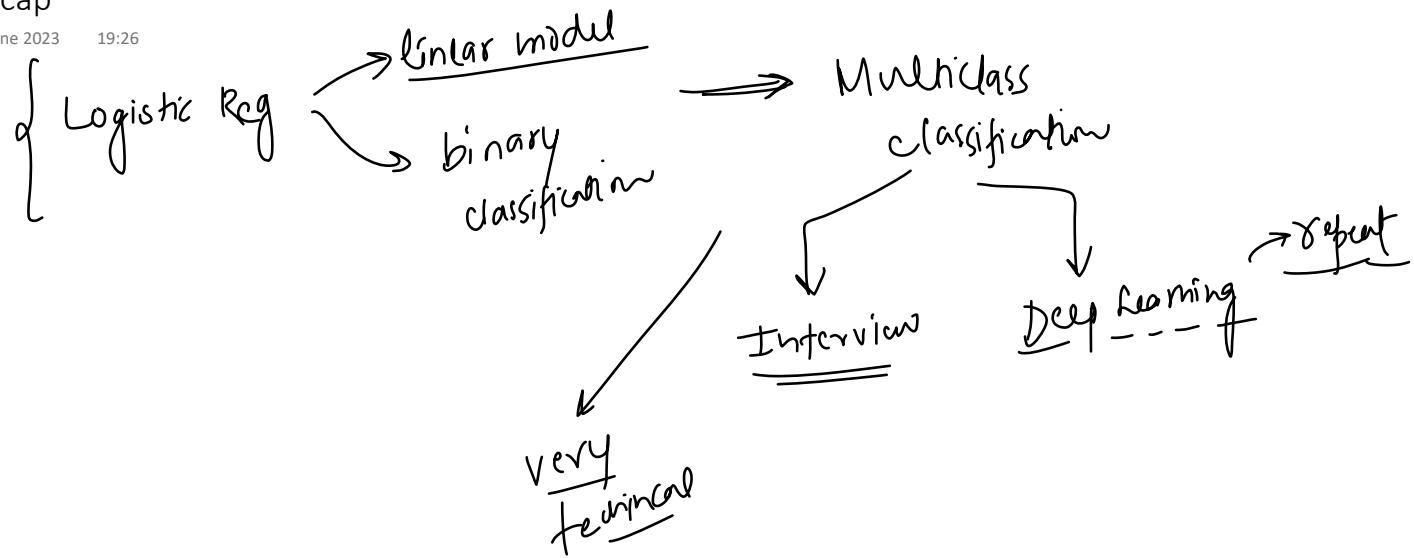
Summary

28 June 2023 16:35



Recap

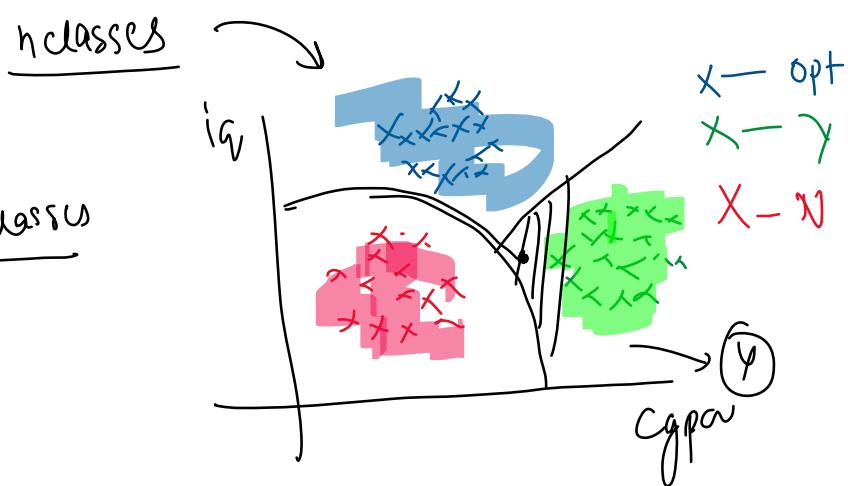
30 June 2023 19:26



What is Multiclass Classification

30 June 2023 14:16

category	1	2	placement
q	9	90	+
6	60	+	N
7	70	OPT	
	<u>6.5</u>	<u>65</u>	



How to Logistic Regression handles Multiclass Classification Problems

30 June 2023 17:12

binary classification

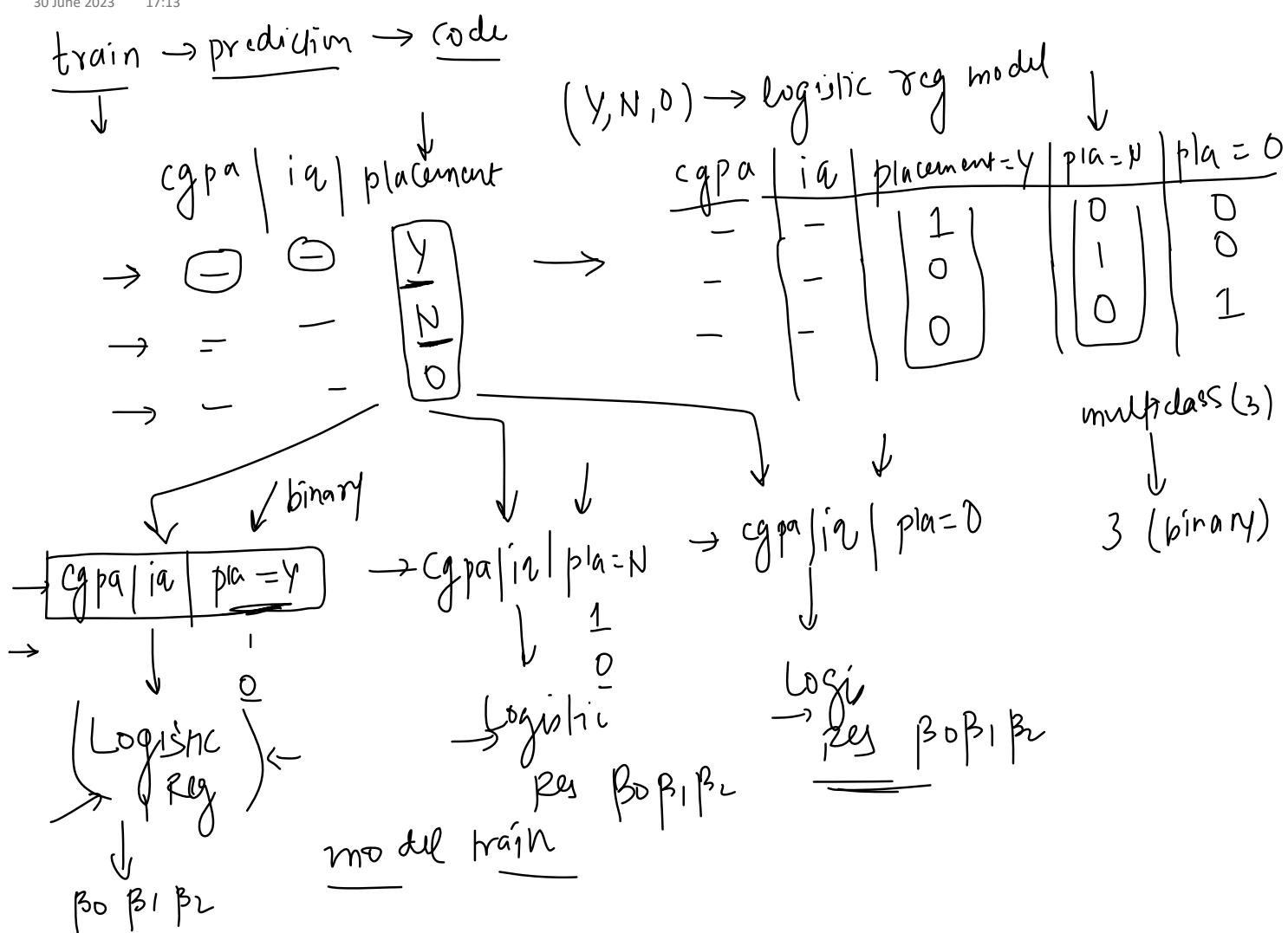
→ OVR (one vs rest) → OVA (one vs all)

→ Multinomial LR → Softmax Reg

\hookrightarrow K models \rightarrow K prob \rightarrow normalize

OVR Approach

30 June 2023 17:13



Prediction $\{6.5, 65\} \rightarrow Y, N, 0$

$p(Y) = \frac{0.6}{0.6 + 0.3 + 0.5} = 0.42$
 $p(N) = \frac{0.3}{0.6 + 0.3 + 0.5} = 0.21$
 $p(O) = \frac{0.5}{0.6 + 0.3 + 0.5} = 0.35$

Efficient \rightarrow large dataset
 with high number

v
n =



» efficient \rightarrow curve narrow
has high number
of classes

Code

30 June 2023 17:13

SoftMax Function
30 June 2023 18:00

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

softmax LR

Sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

← binary

prob

0-1

$$\underline{z_1} \quad \underline{z_2} \quad \underline{z_3}$$

$$\sigma(z_1) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\sigma(z_1) + \sigma(z_2) + \sigma(z_3) = 1$$

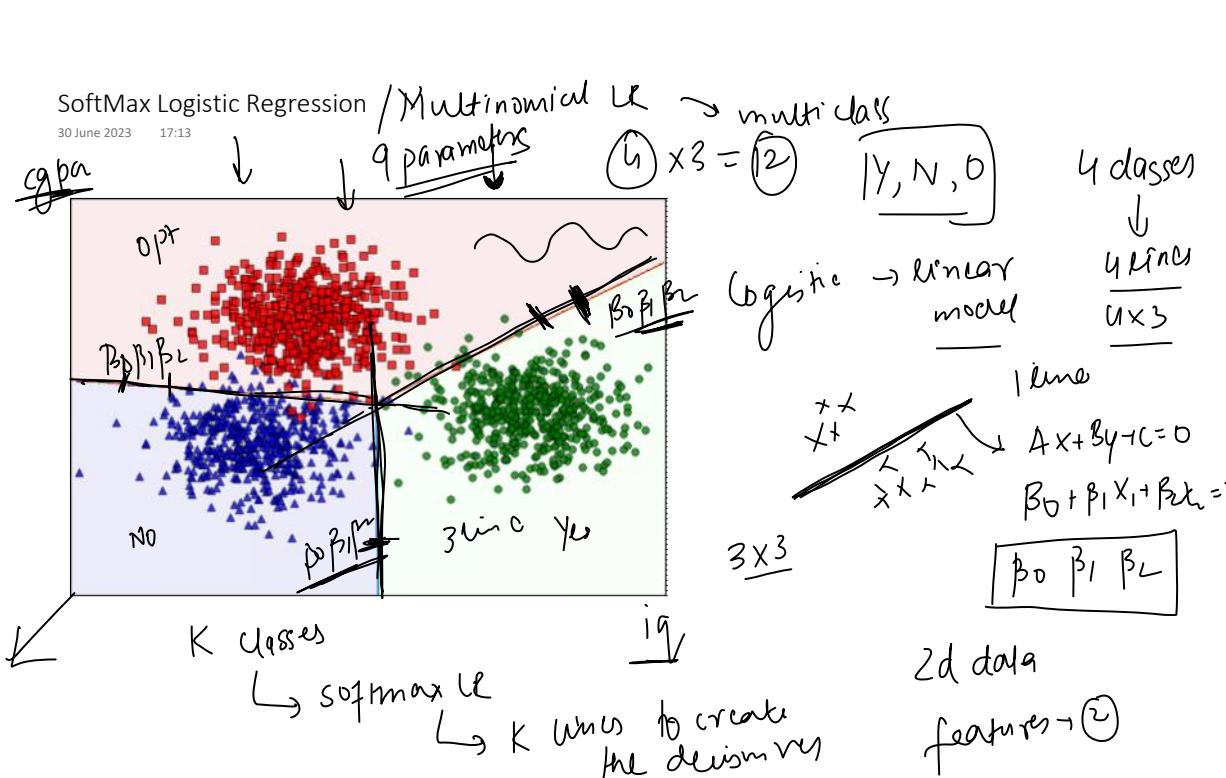
↑
multiclass

$$\sigma(z_2) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\sigma(z_3) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

SoftMax Logistic Regression / Multinomial LR \rightarrow multi class

30 June 2023 17:13

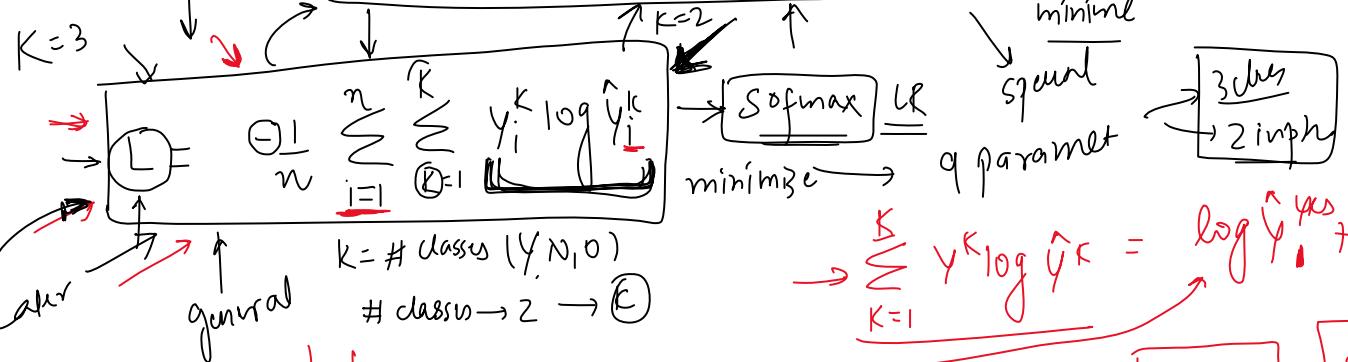


training \rightarrow prediction \rightarrow code

cgpa	iq	placment = Y	pla = N	pla = O
-	-X	1	0	0
-	-N	0	1	0
-	-O	0	0	1

loss function \rightarrow

$$\text{binary} \rightarrow L = -\frac{1}{n} \sum_{i=1}^n y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$



cgpa	iq	plac = Y	pla = N	pla = O
-	-	1	✓ 0	✓ 0
-	-	0	✓ 1	✓ 0
-	-	0	✓ 0	✓ 1

$\text{loss} = - \sum_{i=1}^n \sum_{k=1}^K y_i^k \log(\hat{y}_i^k)$ (1)

output softmax row number y-value
 minimize $L = \log(\hat{y}_1) + \log(\frac{\hat{y}_2}{\beta_0 \beta_1 \beta_2}) + \log(\frac{\hat{y}_3^{\text{opt}}}{\beta_0 \beta_1 \beta_2})$

$\hat{y}_1, \hat{y}_2, \hat{y}_3^{\text{opt}}$ \rightarrow softmax \rightarrow random \rightarrow sigmoid \rightarrow $0, 1$

$\hat{y}_i \rightarrow \text{Output of logistic}$ $\hat{y}_i = \sigma(z_i)$
 $z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

$\hat{y}_{\text{yes}} = \sigma(z_{\text{yes}}) = \frac{e^{z_{\text{yes}}}}{e^{z_{\text{yes}}} + e^{z_{\text{no}}} + e^{z_{\text{opt}}}}$ $\hat{y}_{\text{no}} = \frac{e^{z_{\text{no}}}}{e^{z_{\text{yes}}} + e^{z_{\text{no}}} + e^{z_{\text{opt}}}}$
 $\hat{y}_{\text{opt}} = \frac{e^{z_{\text{opt}}}}{e^{z_{\text{yes}}} + e^{z_{\text{no}}} + e^{z_{\text{opt}}}}$

$z_{\text{yes}} = \beta_0 + \beta_1 8 + \beta_2 80$

$z_{\text{no}} = \beta_0^{(2)} + \beta_1^{(2)} 8 + \beta_2^{(2)} 80$

$z_{\text{opt}} = \beta_0^{(3)} + \beta_1^{(3)} 8 + \beta_2^{(3)} 80$

$\beta_0' = \beta_0' - \eta \frac{\partial L}{\partial \beta_0}$ $\eta \text{ different}$
 $\eta \in [0, 1]$

$m \times n$
 1000

$$\begin{aligned}
 & \text{L} = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K y_i^k \log(\hat{y}_i^k) \quad \boxed{k=2} \rightarrow \text{binary cross} \\
 & \text{grad des} \\
 & \text{cgpai | iq | place} \\
 & \begin{array}{ll} y_i^1 & y_i^0 \\ 1 & 0 \\ 0 & 1 \\ \hline \end{array} \quad \begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array} \\
 & \begin{array}{ll} y_i^1 & y_i^0 \\ 1 & 0 \\ 0 & 1 \\ \hline \end{array} \quad \begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array} \\
 & \begin{array}{ll} y_i^1 & y_i^0 \\ 1 & 0 \\ 0 & 1 \\ \hline \end{array} \quad \begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array} \\
 & \begin{array}{ll} y_i^1 & y_i^0 \\ 1 & 0 \\ 0 & 1 \\ \hline \end{array} \quad \begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array} \\
 & \boxed{\begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array}} = \boxed{\begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array}} \\
 & \boxed{\begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array}} = \boxed{\begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array}} \\
 & \boxed{\begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array}} = \boxed{\begin{array}{l} \text{y}_i^1 \\ \text{y}_i^0 \end{array}}
 \end{aligned}$$

$$\begin{array}{l}
 \text{grad des} \\
 \begin{array}{l}
 \min' \rightarrow \boxed{\beta_0^1 \beta_1^1 \beta_2^1} \\
 \min \rightarrow \boxed{\beta_0^2 \beta_1^2 \beta_2^2} \\
 \min^3 \rightarrow \boxed{\beta_0^3 \beta_1^3 \beta_2^3}
 \end{array} \\
 \downarrow \text{predict} \\
 \boxed{6.5, 65}
 \end{array}$$

Prediction

$$\text{Softmax} \rightarrow \sigma(z_1) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} = 0.3 \quad \sigma(z_3) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}} = \underline{0.5}$$

$$\sigma(z_2) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} = 0.2$$

Code

30 June 2023 17:13

When to use what?

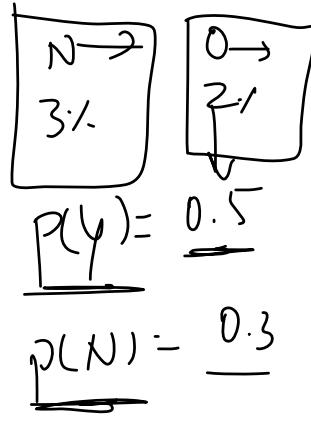
30 June 2023 17:13

Use One-vs-Rest (OVR) when:

1. Classes are Non-Mutually Exclusive: OVR is appropriate if an instance can belong to more than one class, as each classifier provides an independent probability for each class.
2. Dealing with Imbalanced Data: OVR might perform better when class distribution is highly imbalanced since each class gets a dedicated model.

Use Multinomial Logistic Regression (SoftMax Regression) when:

1. Computational Efficiency is Required: Softmax Regression is generally more efficient for large datasets and a high number of classes.
2. Classes are Mutually Exclusive: SoftMax Regression is a good choice when each instance can only belong to one class. The SoftMax function provides a set of probabilities that sum to 1, fitting well with mutually exclusive classes.
3. Interpretability is Important: The probabilities output by SoftMax Regression are more interpretable than those from OVR, as they always sum to 1. This can make model predictions easier to explain.

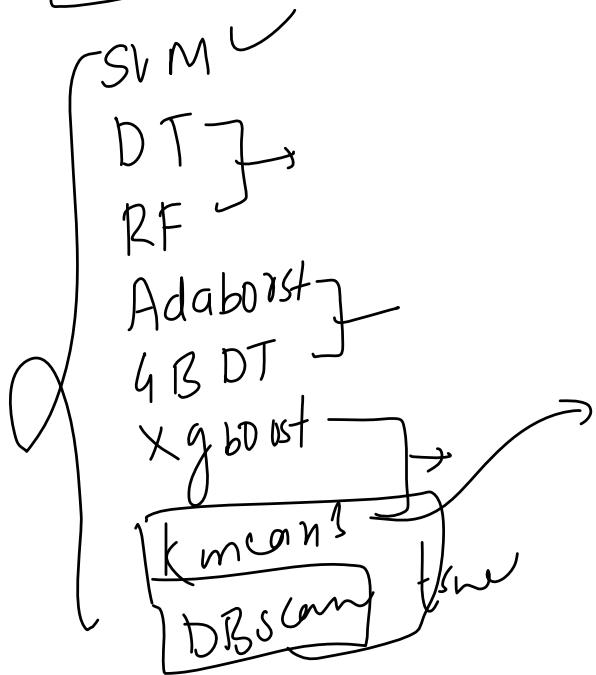


Tasks

30 June 2023 18:19

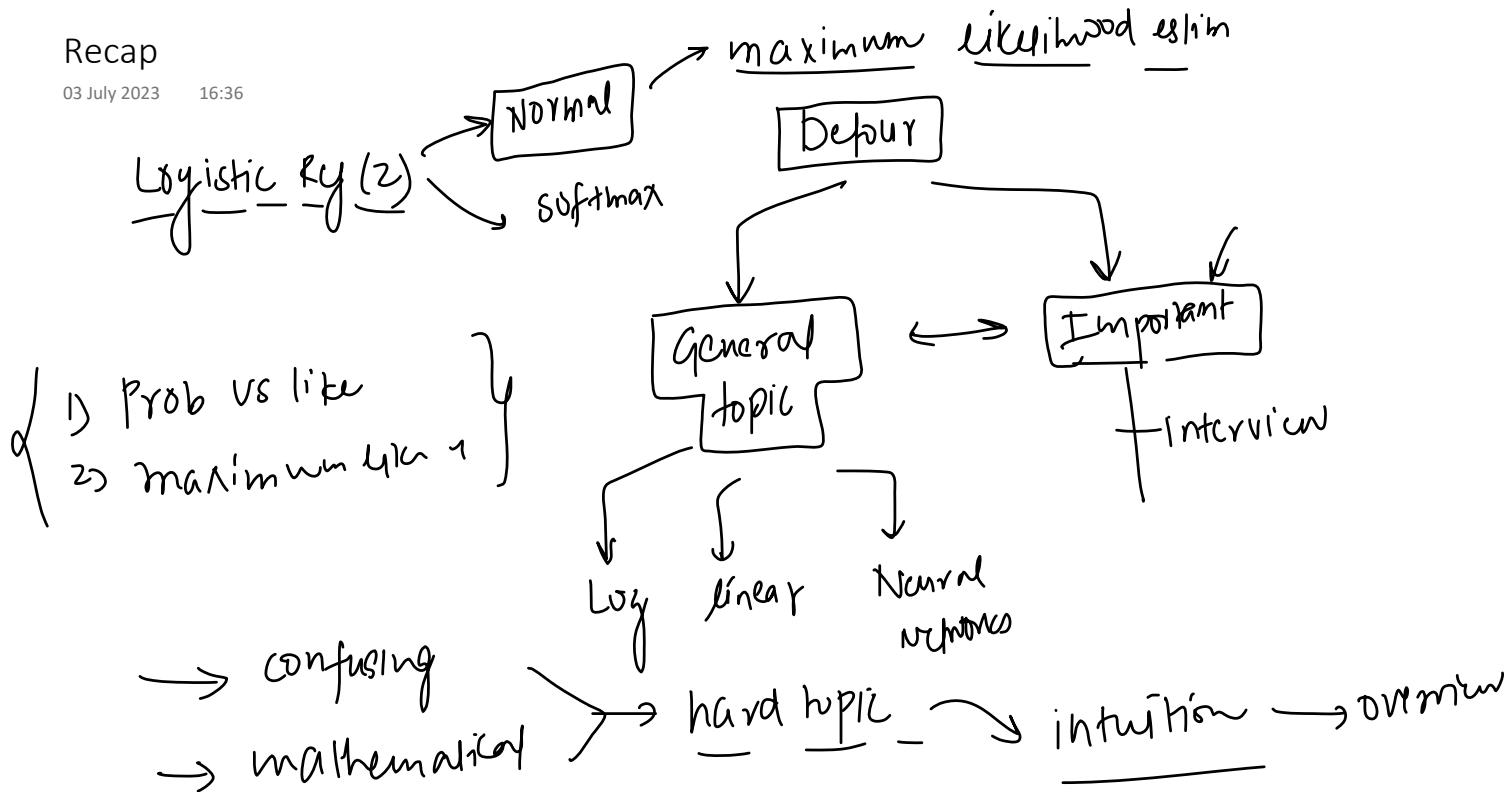
1. Derive sigmoid from softmax
2. Derive binary cross entropy from categorical cross entropy
3. Find the derivative of Softmax Function ←
4. Find the gradients of cross entropy error

$$\sigma(z) = \sigma(z)(1 - \sigma(z))$$



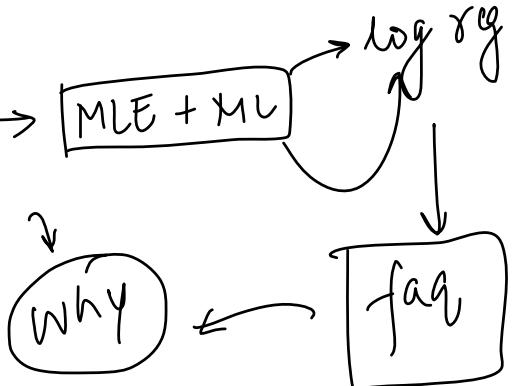
Recap

03 July 2023 16:36



Plan of attack

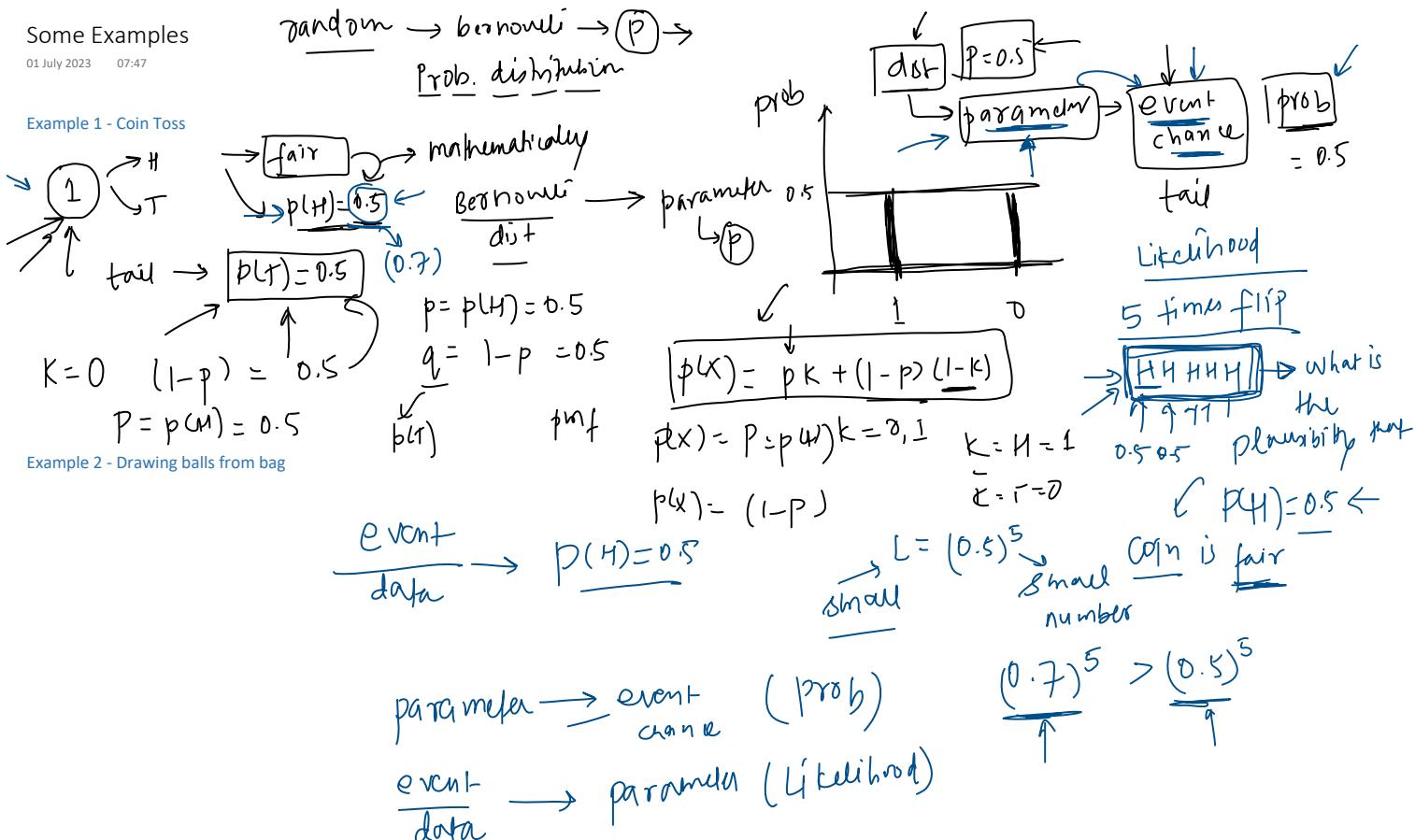
MLE \rightarrow Likelihood



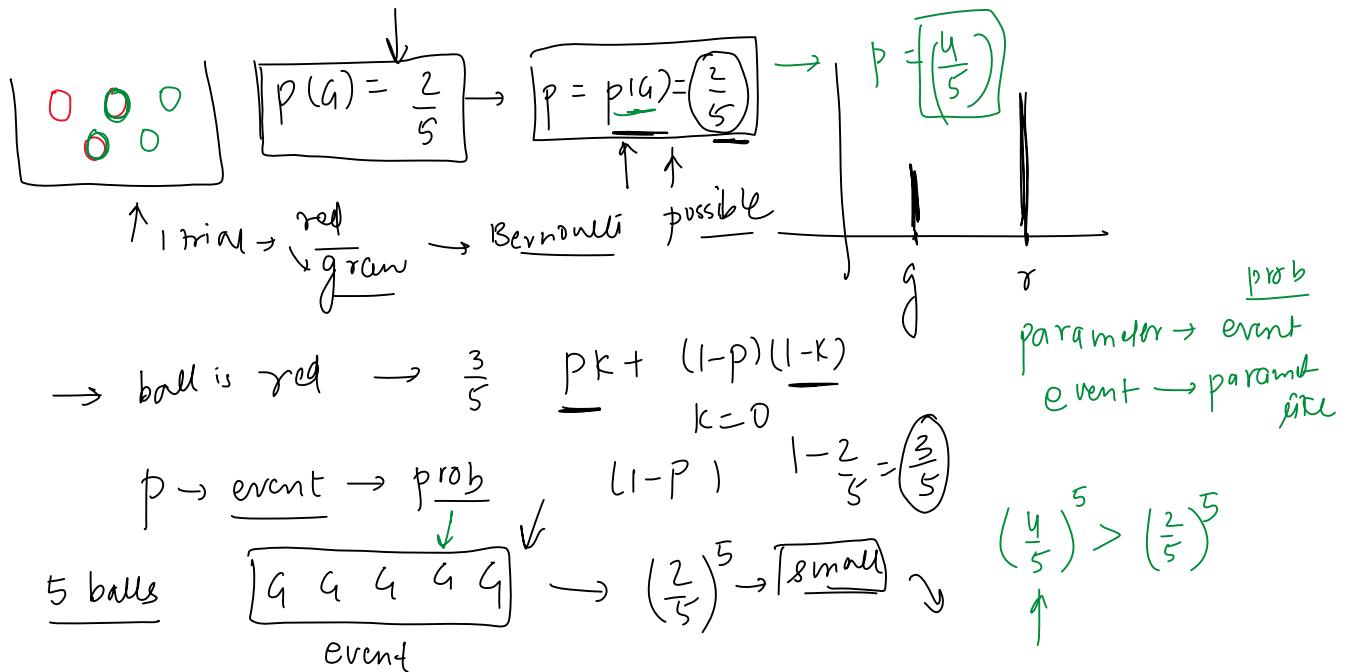
Some Examples

01 July 2023 07:47

Example 1 - Coin Toss



Example 3 - Normal Distribution



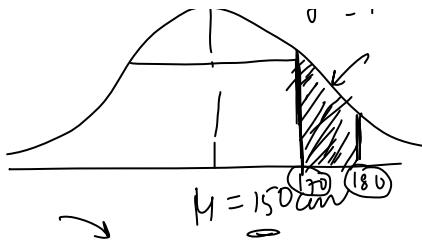
3) Discrete → bernoulli
↳ continuous



\hookrightarrow continuous

Normal distribution

heights



$$\text{dist} \rightarrow (\mu, \sigma) \rightarrow P(m < X < 180) = 0.82$$

$$\begin{array}{c} \xrightarrow{\text{dist}} \xrightarrow{\mu} \\ \xrightarrow{\sigma} \boxed{100} \text{ cm} \rightarrow N(\mu, \sigma^2) \end{array}$$

$\boxed{130}$

$$L(\mu, \sigma | X=100) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

$$\xrightarrow{\text{very small}} \underline{1.47 \times 10^{-7}}$$

$$\begin{aligned} \mu &= 150 \\ \sigma &= 10 \end{aligned}$$

parameter \rightarrow event (prob)

event \rightarrow parameter (Maximum)

data ——————

Probability Vs Likelihood

01 July 2023 09:37

Probability: This is a measure of the chance that a certain event will occur out of all possible events. It's usually presented as a ratio or fraction, and it ranges from 0 (meaning the event will not happen) to 1 (meaning the event is certain to happen).

Likelihood: In statistical context, likelihood is a function that measures the plausibility of a particular parameter value given some observed data. It quantifies how well a specific outcome supports specific parameter values.

More Definitions

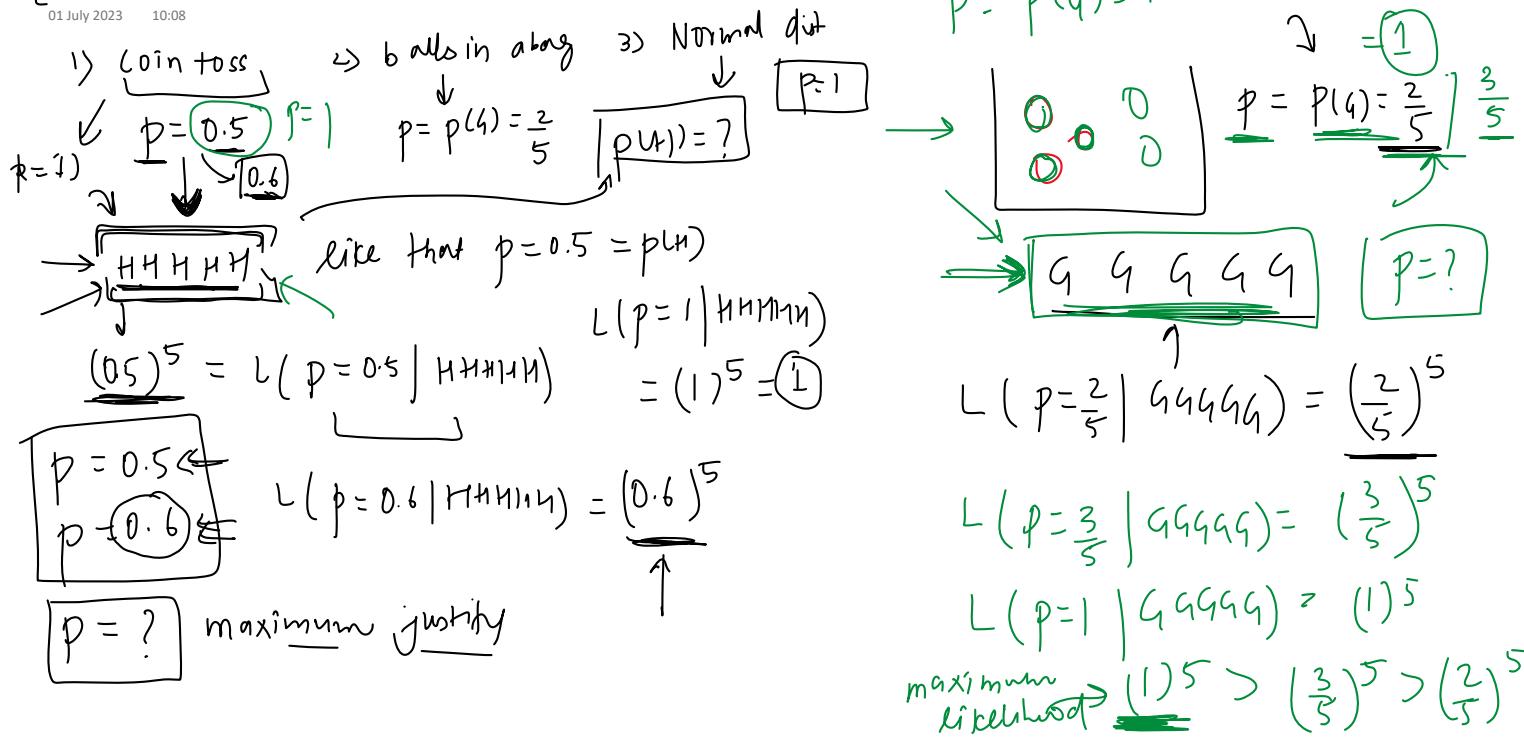
A probability quantifies how often you observe a certain outcome of a test, given a certain understanding of the underlying data.

A likelihood quantifies how good one's model is, given a set of data that's been observed.

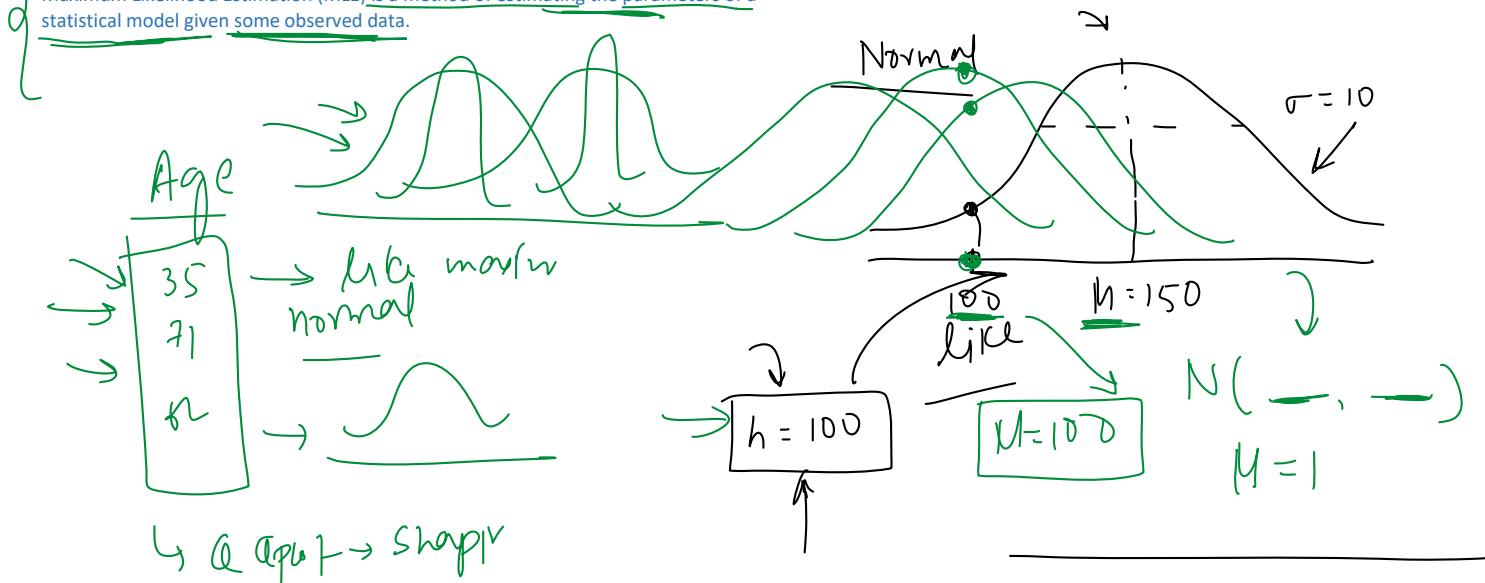
Probabilities describe test outcomes, while likelihoods describe models.

[Maximum Likelihood Estimation]

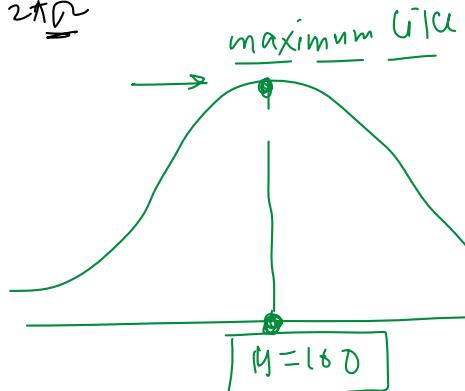
01 July 2023 10:08



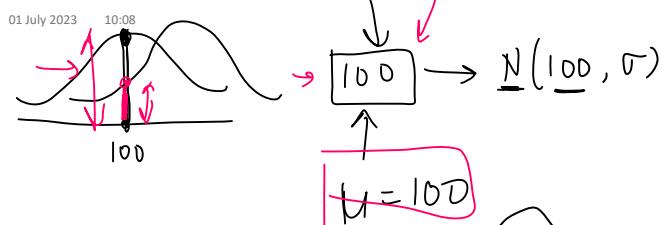
Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given some observed data.



$$L(\mu=150, \sigma=10 | X=100) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(100-150)^2}{2\sigma^2}}$$



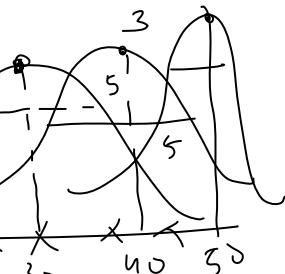
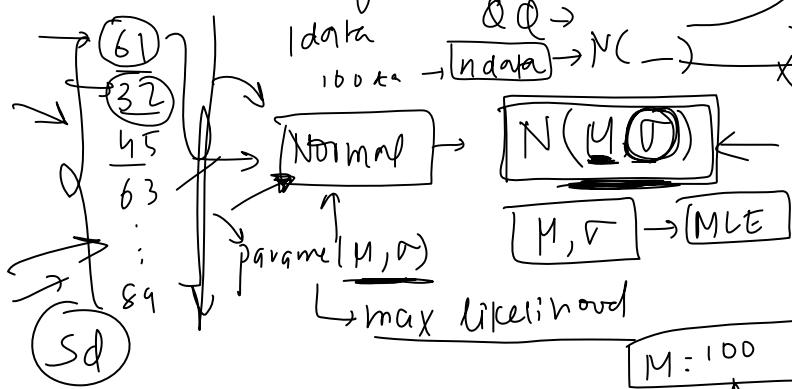
MLE for Normal Distribution



$\max_{a, b, c}$

$\log a, \log b, \log c$

dataset = 100 Age



$\left. \begin{array}{l} \mu = \text{mean of all} \\ \text{observed data points} \end{array} \right\}$

$$L(\mu, \sigma | x_1, x_2, x_3, \dots, x_n) \quad \underline{\mu = 100, \sigma = 20}$$

$$L(\mu = 100, \sigma = 10 | x_1 = 61, x_2 = 32, \dots, x_n = 89) = 0.89$$

$$L(\mu = 100, \sigma = 20 | \quad \quad \quad) = 0.79$$

$$\log(a \cdot b) = \log a + \log b$$

log

independent

$$L(\mu, \sigma | x_1, x_2, x_3, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \right) \times \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \right) \times \dots \times \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \right)$$

$$L(\mu, \sigma | x_1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}$$

log-likelihood

$$\log(L(\mu, \sigma | x_1, x_2, \dots, x_n)) = \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \right) + \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \right) + \dots + \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \right)$$

$$\log(ab) =$$

$$\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \right) = \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}$$

$$\log a^b = b \log a$$

$$\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) = -\frac{1}{2} \log(2\pi\sigma^2) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2)$$

$$\log a^b = b \log a$$

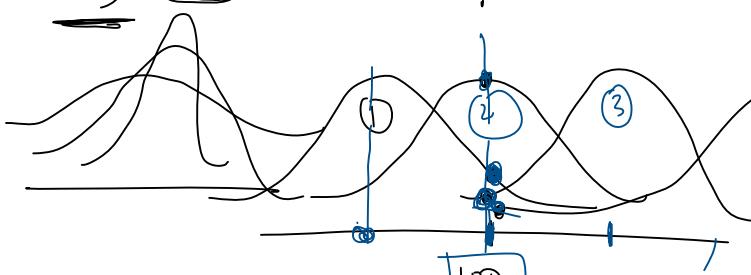
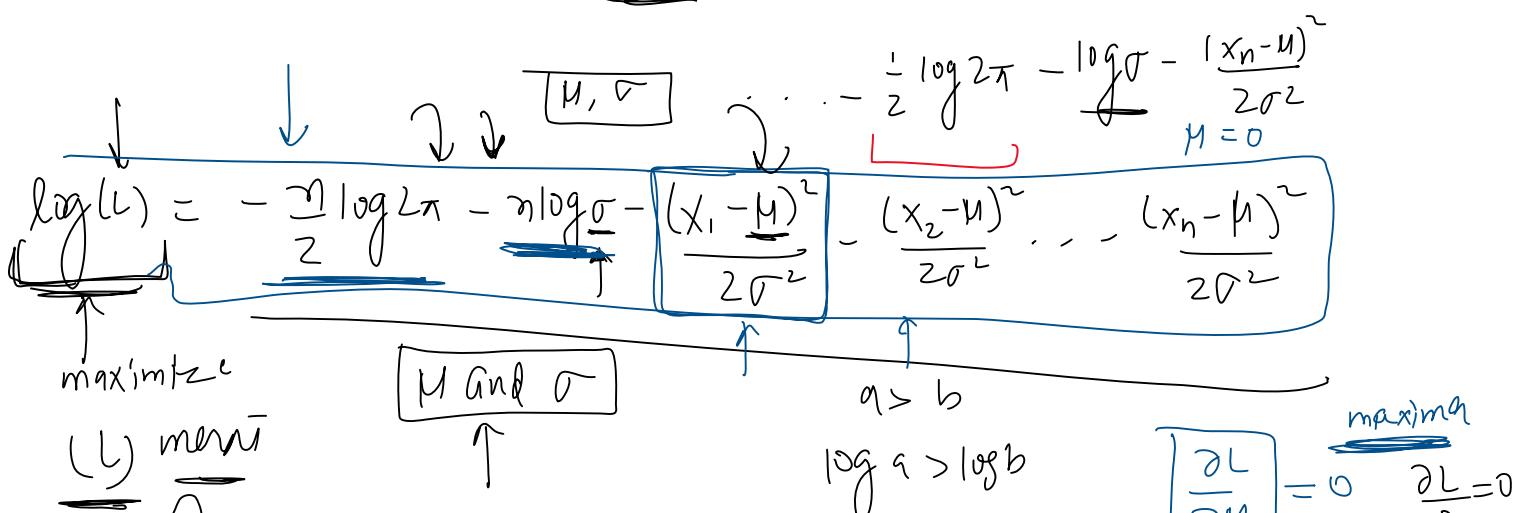
$$\underbrace{\log (2\pi\sigma^2)^{-\frac{1}{2}}}_{\text{constant}} - \frac{(x_1 - \mu)^2}{2\sigma^2}$$

$$-\frac{1}{2} \log (2\pi\sigma^2) - \frac{(x_1 - \mu)^2}{2\sigma^2}$$

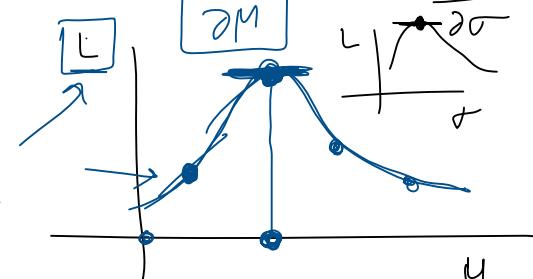
$$= -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{(x_1 - \mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{(x_1 - \mu)^2}{2\sigma^2}$$

$$-\frac{1}{2} \log 2\pi - \frac{\log \sigma^2}{2} - \frac{(x_1 - \mu)^2}{2\sigma^2} - \frac{1}{2} \log 2\pi - \frac{\log \sigma^2}{2} - \frac{(x_2 - \mu)^2}{2\sigma^2}$$



$$\log a > \log b$$



$$\frac{\partial \log(L)}{\partial \mu} = \frac{1}{2} \frac{(x_1 - \mu)}{\sigma^2} + \frac{(x_2 - \mu)}{\sigma^2} + \frac{(x_3 - \mu)}{\sigma^2} + \dots + \frac{(x_n - \mu)}{\sigma^2} = 0$$

$$(x_1 - \mu) + (x_2 - \mu) + \dots + (x_n - \mu) = 0$$

$$x_1 + x_2 + x_3 + \dots + x_n - n\mu = 0$$

$$n\mu = x_1 + x_2 + \dots + x_n$$

$$\boxed{\frac{H = \frac{x_1 + x_2 + \dots + x_n}{n}}{n}}$$

$$\boxed{\sigma^{-2}} \rightarrow \boxed{-2\sigma^3} = \boxed{\frac{-2}{\sigma^3}} \quad \log \sigma = \boxed{\frac{1}{\sigma}} \quad \boxed{\sqrt{\sigma^{-2}}} = \boxed{-2\sigma^{-3}}$$

$$\log(L) = -\frac{n}{2} \log(2\pi) - \frac{n \log \sigma}{2} - \boxed{\frac{(x_1 - \mu)^2}{2\sigma^2}} - \frac{(x_2 - \mu)^2}{2\sigma^2} - \dots - \frac{(x_n - \mu)^2}{2\sigma^2}$$

$$\frac{\partial \log(L)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{(x_1 - \mu)^2}{2\sigma^3} + \frac{(x_2 - \mu)^2}{\sigma^3} + \dots + \frac{(x_n - \mu)^2}{\sigma^3} = 0$$

$$\frac{(x_1 - \mu)^2}{\sigma^3} + \left[\frac{(x_2 - \mu)^2}{\sigma^3} + \dots + \frac{(x_n - \mu)^2}{\sigma^3} \right] \frac{n}{\sigma^3}$$

$$\frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2} + \dots + \frac{(x_n - \mu)^2}{\sigma^2} = n$$

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} \quad \text{Variance}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} \quad \boxed{sd}$$

$$\boxed{50, 63, 65, \dots, 100} \rightarrow \text{Normal} \\ \rightarrow N(\text{avg}, \text{std})$$

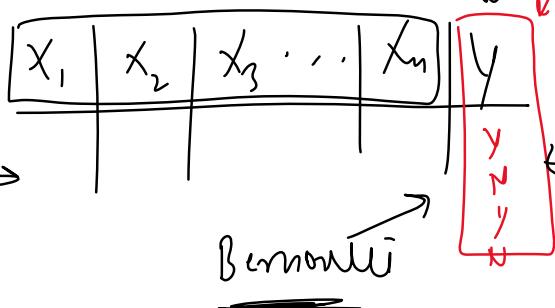


→ $27, 28, 31, 35, \dots, 100$
distribution → MLE

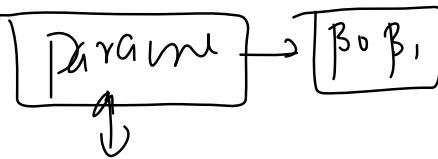
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

y, N, opt

- 1) assume some dist
- 2) use a likelihood
- 3) find the value for parameters that maximizes that likeli



- 1) find out the dist of $y | x$
- 2) decide to apply a ml model parametric in nature → logistic regression



non-parametric → decision rule

- 3) You randomly decide some values $\beta_0 \beta_1 \beta_2 \dots \beta_n$

→ 4) select a likelihood function

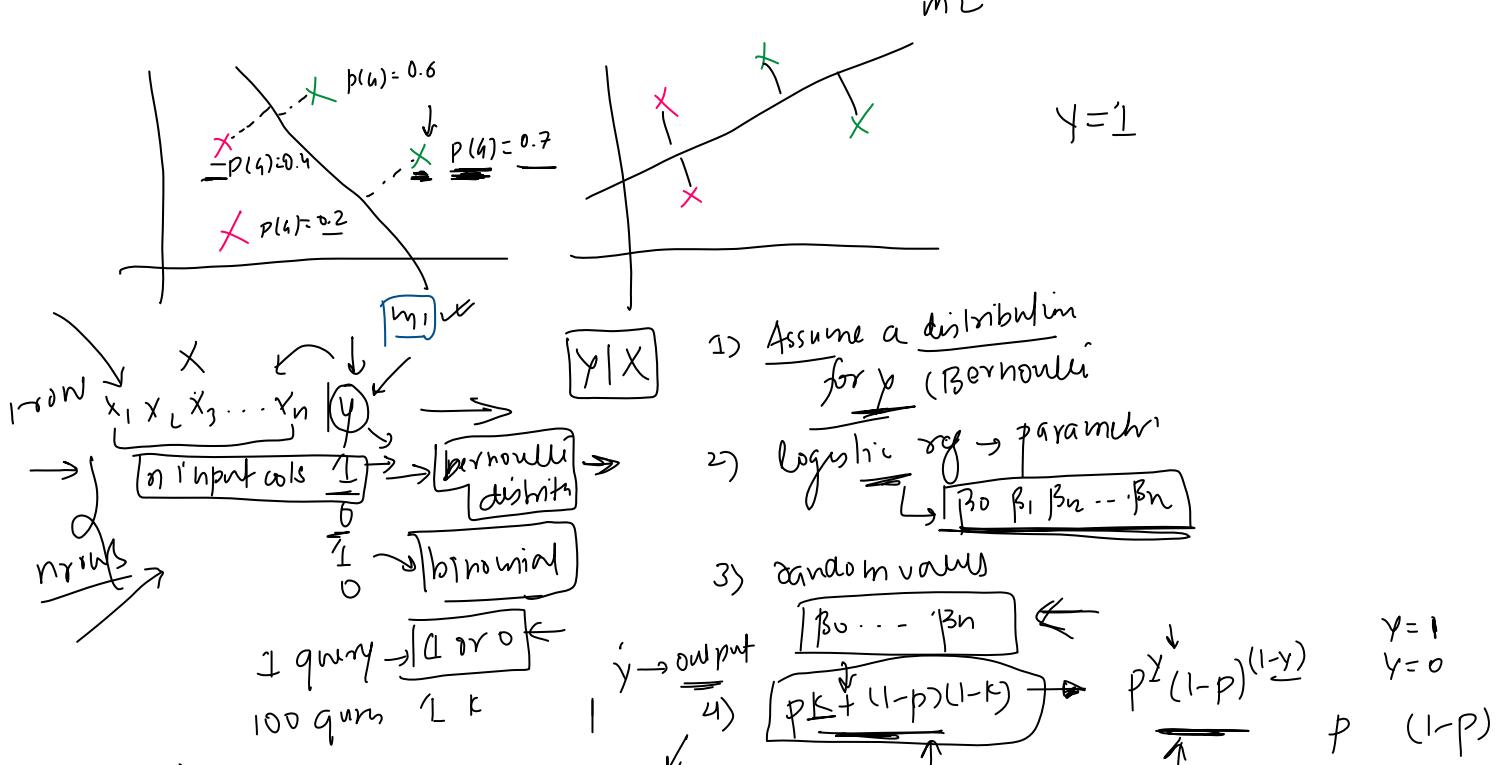
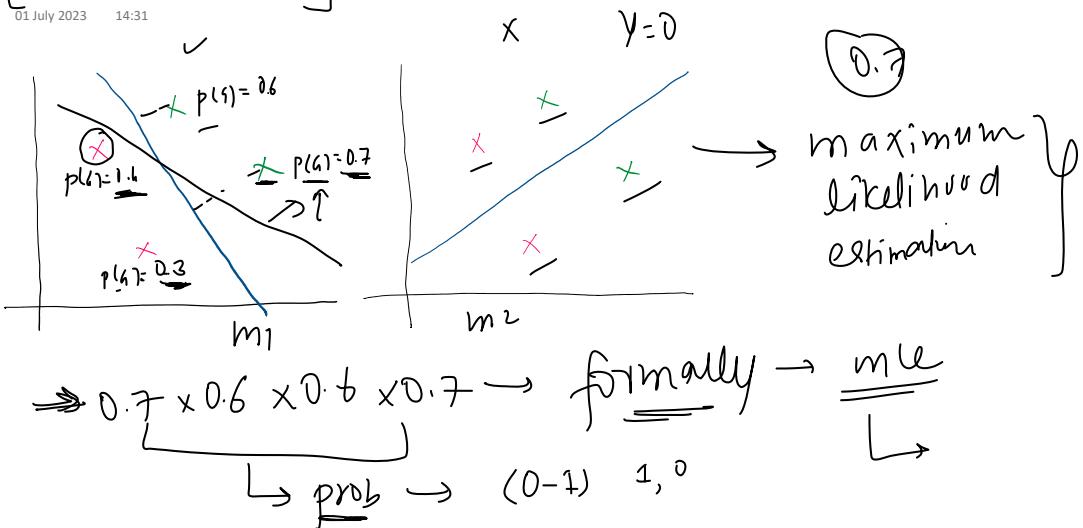
$$\text{pmf} = pK + (1-p)(1-K)$$

5) values of $\beta_0 \beta_1 \beta_2 \dots \beta_n$

↳ maximum value

MLE in Logistic Regression

01 July 2023 14:31



$$P = \frac{p^{(1)}}{p^{(0)}}$$

$$1-P = \frac{p^{(0)}}{p^{(1)}}$$

$$L(y|x; \beta) =$$

$\gamma|x \rightarrow \text{parameters } \beta$

all y_i 's are independent

$$L(y|x; \beta) = (p_1^{y_1} (1-p_1)^{1-y_1}) \times (p_2^{y_2} (1-p_2)^{1-y_2}) \times \dots \times (p_n^{y_n} (1-p_n)^{1-y_n})$$

$$L(y|x; \beta) = \prod_{i=1}^n p_i^{y_i} (1-p_i)^{1-y_i}$$

log ab
bernoilli dist
 $\rightarrow L(p^n)$

$$\log(L) = \sum_{i=1}^n \log(p_i^{y_i} (1-p_i)^{1-y_i}) = \sum_{i=1}^n \underbrace{\log p_i}_{a} + \underbrace{\log(1-p_i)}_{b}^{1-y_i}$$

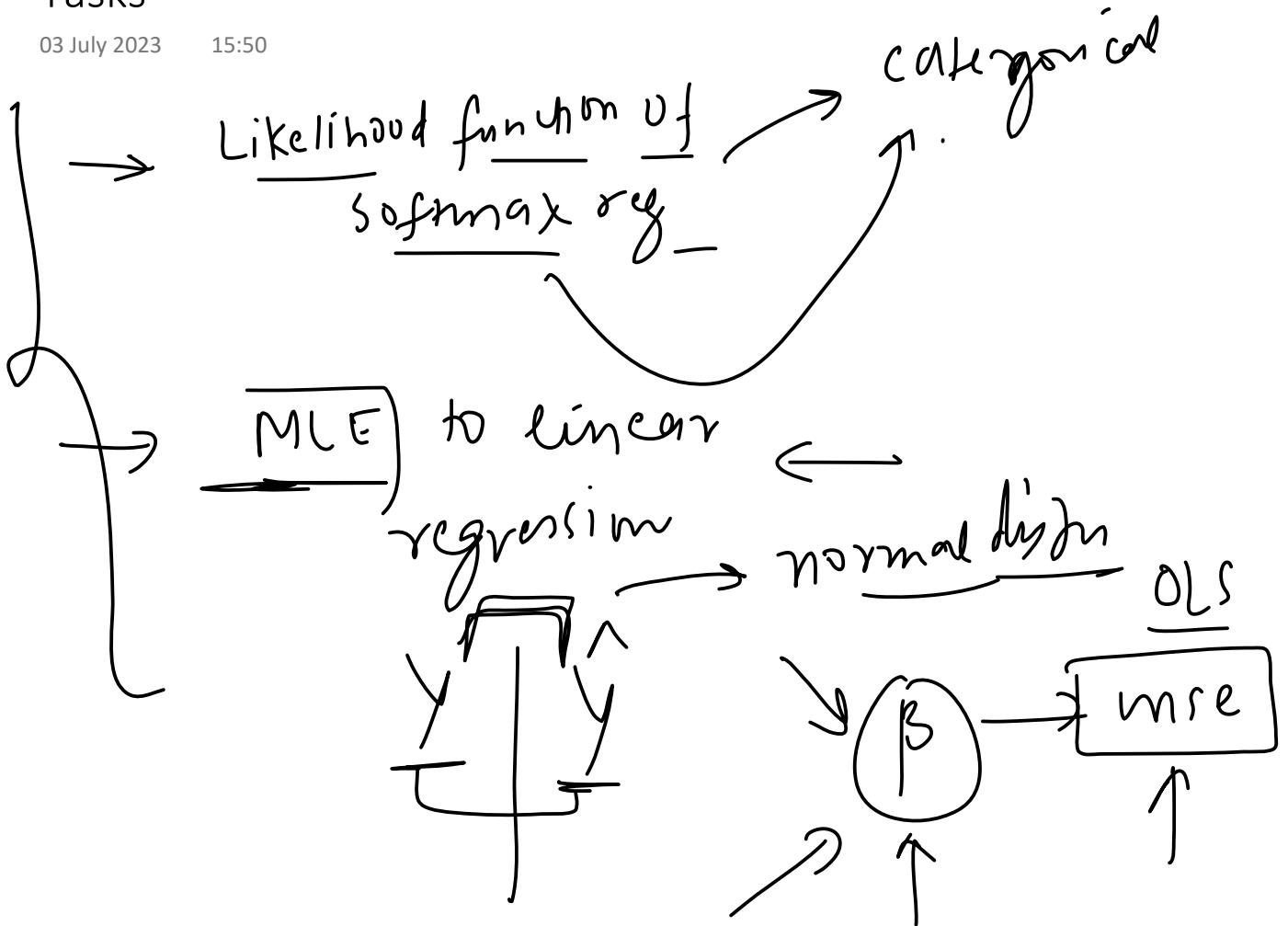
$\log \text{like}$
 \uparrow
 \max

$$= \sum_{i=1}^n y_i \log p_i + (1-y_i) \log (1-p_i)$$

↓ maximize
 $\hookrightarrow \beta$
 $\circlearrowleft \beta$ gradient technique

Tasks

03 July 2023 15:50



Some Important Questions

01 July 2023 17:15

1. Is MLE a general concept applicable to all machine learning algorithms?

No

given data
parameters
↓
require

Maximum Likelihood Estimation (MLE) is a general statistical concept that can be applied to many machine learning algorithms, particularly those that are parametric (i.e., defined by a set of parameters), but it's not applicable to all machine learning algorithms.

MLE is commonly used in algorithms such as linear regression, logistic regression, and neural networks, among others. These algorithms use MLE to find the optimal values of the parameters that best fit the training data.

However, there are some machine learning algorithms that don't rely on MLE. For example:

1. Non-parametric methods: Some machine learning methods, such as k-Nearest Neighbors (k-NN) and Decision Trees, are non-parametric and do not make strong assumptions about the underlying data distribution. These methods don't have a fixed set of parameters that can be optimized using MLE.
2. Unsupervised learning algorithms: Some unsupervised learning algorithms, like K-means clustering, use different objective functions, not necessarily tied to a probability distribution.
3. Reinforcement Learning: Reinforcement Learning methods generally don't use MLE, as they are more focused on learning from rewards and punishments over a sequence of actions rather than fitting to a specific data distribution.

parametric
|
+ Linear
+ neural networks

2. How is MLE related to the concept of loss functions?

In machine learning, a loss function measures how well a model's predictions align with the actual values. The goal of training a machine learning model is often to find the model parameters that minimize the loss function.

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model to maximize the likelihood function, which is conceptually similar to minimizing a loss function. In fact, for many common models, minimizing the loss function is equivalent to maximizing the likelihood function.

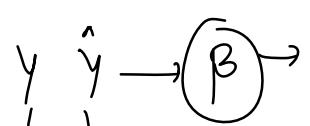
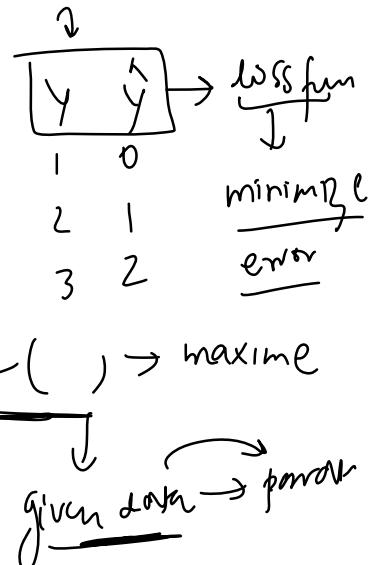
MLE and the concept of loss functions in machine learning are closely related. Many common loss functions can be derived from the principle of maximum likelihood estimation under certain assumptions about the data or the model. By minimizing these loss functions, we're effectively performing maximum likelihood estimation.

3. Then why does loss function exist, why don't we maximize Likelihood

The confusion arises from the fact that we're using two different perspectives to look at the same problem.

In many machine learning algorithms, the aim is to minimize the difference between the predicted and actual values, and this is typically represented by a loss function. When we talk about minimizing the loss function, it's essentially the same as saying we're trying to find the best model parameters that give us the closest predictions to the actual values.

On the other hand, when we look at the problem from a statistical perspective, we talk in terms of maximizing the likelihood of seeing the observed data given the model parameters. This is represented by a likelihood function.



terms of maximizing the likelihood of seeing the observed data given the model parameters. This is represented by a likelihood function.

For many models, these two perspectives are equivalent - minimizing the loss function is the same as maximizing the likelihood function. In fact, many common loss functions can be derived from the principle of MLE under certain assumptions about the data.

So why do we often talk about minimizing the loss function instead of maximizing the likelihood? There are a few reasons:

1. Computational reasons: It's often easier and more computationally efficient to minimize a loss function than to maximize a likelihood function. This is particularly true when working with complex models like neural networks.
2. Generalization: The concept of a loss function is more general and can be applied to a wider range of problems. Not all machine learning problems can be framed in terms of maximizing a likelihood. For example, many non-parametric methods and unsupervised learning algorithms don't involve likelihoods.
3. Flexibility: Loss functions can be easily customized to the specific needs of a problem. For instance, we might want to give more weight to certain types of errors, or we might want to use a loss function that is robust to outliers.

In summary, while the concepts of loss function minimization and maximum likelihood estimation are closely related and often equivalent, the concept of a loss function is more flexible and computationally convenient, which is why it's more commonly used in the machine learning community.

4. Then why study about maximum likelihood at all?

The study of Maximum Likelihood Estimation (MLE) is essential for several reasons, despite the prevalence of loss functions in machine learning:

1. Statistical Foundation: MLE provides a strong statistical foundation for understanding machine learning models. It gives a principled way of deriving the loss functions used in many common machine learning algorithms, and it helps us understand why these loss functions work and under what assumptions.
2. Interpretability: The MLE framework gives us a way to interpret our model parameters. The MLEs are the parameters that make the observed data most likely under our model, which can be a powerful way of understanding what our model has learned.
3. Model Comparison: MLE gives us a way to compare different models on the same dataset. This can be done using tools like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which are based on the likelihood function and can help us choose the best model for our data.
4. Generalization to Other Methods: MLE is a specific case of more general methods, like Expectation-Maximization and Bayesian inference, which are used in more complex statistical modelling. Understanding MLE can provide a stepping stone to these more advanced topics.
5. Deeper Understanding: Lastly, understanding MLE can give us a deeper understanding of our models, leading to better intuition, better model selection, and ultimately, better performance on our machine learning tasks.

In short, while you can often get by with a practical understanding of loss functions and optimization algorithms in applied machine learning, understanding MLE can be extremely valuable for gaining a deeper understanding of how and why these models work.

Assumptions of Logistic Regression

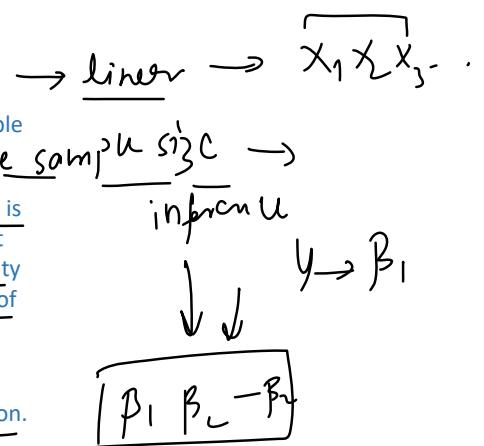
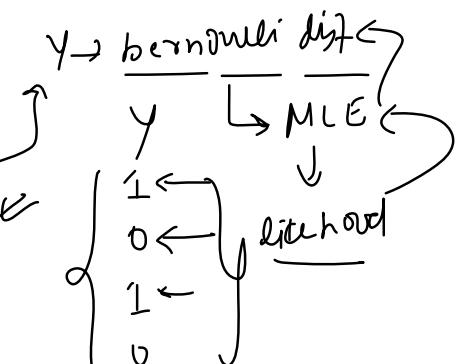
05 July 2023 07:21

log odd \leftrightarrow logit

Logistic regression, like other statistical methods, relies on certain assumptions. Here are the main assumptions of logistic regression:

1. Binary Logistic Regression requires the dependent variable to be binary: That means the outcome variable must have two possible outcomes, such as "yes" vs "no", "success" vs "failure", "spam" vs "not spam", etc.
2. Independence of observations: The observations should be independent of each other. In other words, the outcome of one instance should not affect the outcome of another.
3. Linearity of independent variables and log odds: Although logistic regression does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
4. Absence of multicollinearity: The independent variables should not be too highly correlated with each other, a condition known as multicollinearity. While not a strict assumption, multicollinearity can be a problem because it can make the model unstable and difficult to interpret.
5. Large sample size: Logistic regression requires a large sample size. A general guideline is that you need at least 10 cases with the least frequent outcome for each independent variable. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is 0.10, then you would need a minimum sample size of 500 ($10^5 / 0.10$).

Note that violating these assumptions doesn't mean you can't or shouldn't use logistic regression, but it may impact the validity of the results and you should proceed with caution.



		10x5					Y
		x_1	x_2	x_3	x_4	x_5	
5×10	0.10	580					Y
		<u>rows</u>					

$1 \rightarrow 90\%$
 $0 \rightarrow 10\%$
 $(0) \rightarrow 10 \cdot 0 \quad 90 \cdot 1$

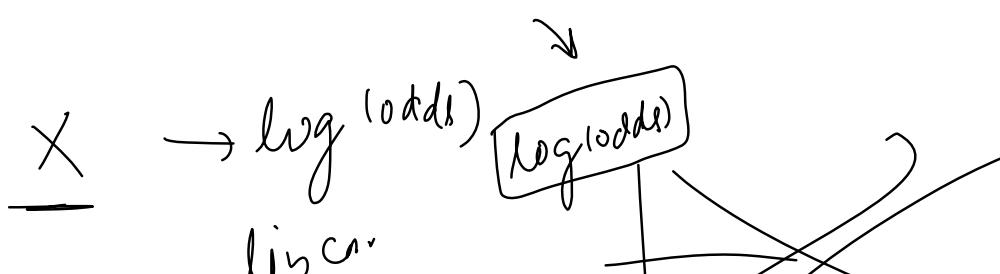
linear reg \rightarrow



$$Y = \beta X \leftarrow Y = \beta X$$

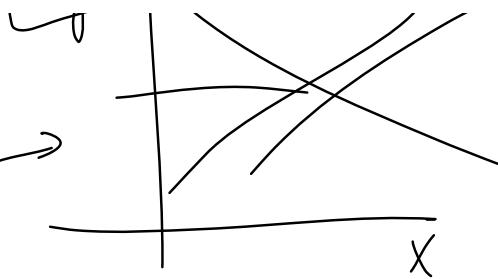
linear \downarrow
non linear \downarrow

$$Y = \frac{1}{1 + e^{-\beta X}}$$



U

Odds \rightarrow log odds



$$\log \left(\frac{P}{1-P} \right) = \beta X$$

Odds and Log(Odds)

05 July 2023 07:20

prob →

Odds: The odds of an event is the ratio of the probability of the event happening (P) to the probability of the event not happening (1-P). It's a way of expressing the likelihood of an event. If the odds are greater than 1, the event is more likely to happen than not, and vice versa.

$$\text{Odds} \rightarrow 3 \text{ odds} \rightarrow \left[\frac{1}{6} \right] \quad \left[\frac{5}{6} \right]$$

log(Odds) → logR

$$\text{Odds} = \frac{P}{1-P}$$

$$\text{Odds} \rightarrow \frac{\frac{1}{6}}{\frac{5}{6}} = \left[\frac{1}{5} \right] = 0.2 \quad \boxed{1 \text{ in } 5}$$

P → prob of success

0-1

$$\checkmark \checkmark \checkmark \checkmark \times \rightarrow \frac{4}{5} / \frac{1}{5} \rightarrow \left[\frac{4}{1} \right] \rightarrow \boxed{4}$$

$$\rightarrow \boxed{0 \rightarrow \infty} \rightarrow$$

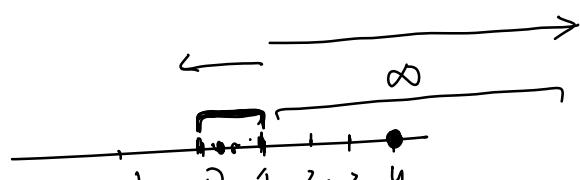
$$\cancel{\times} \times \times \times \rightarrow \frac{1}{5} = \frac{1}{4} \rightarrow 0.25 \quad \frac{1}{9} = \frac{0}{9}$$

$$\checkmark \checkmark \checkmark \times \rightarrow \frac{4}{1} \quad 39 \quad \underline{399} \quad \underline{3999}$$

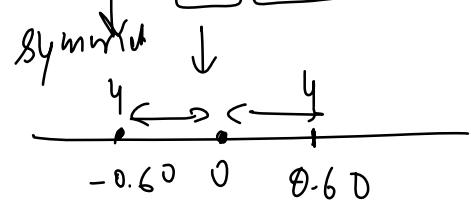
log(Odds)

log_e(Odds) →

Odds



$$\times \checkmark \checkmark \checkmark \rightarrow \frac{4}{1} \rightarrow \boxed{4} \rightarrow$$



$$\checkmark \times \times \times \rightarrow \frac{1}{4} \quad 0.25$$

Another Interpretation of Logistic Regression

05 July 2023 07:21

$\log(\text{odds}) \rightarrow \text{LR}$

$$\log(\text{odds}) = \log\left(\frac{P}{1-P}\right)$$

cgpaiiq		place	Y
8	80		
-	-	1	0
-	-	0	0

$$\begin{aligned} \hat{y} &\rightarrow p \\ 0.93 &\leftarrow 0.37 \\ 0.21 & \end{aligned}$$

$$\beta_0 = 0 \quad \beta_1 = 1 \quad \beta_2 = 2$$

$$p(1) \rightarrow p$$

$$\hat{y} = \underline{p(1)} = \underline{p} = \frac{1}{1 + e^{-\underline{\beta X}}}$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$[0 + 1 \times 8 + 2 \times 80]$$

$$p = \frac{1}{1 + e^{-\beta X}} \Rightarrow \frac{1 + e^{-\beta X}}{1} = \frac{1}{p}$$

$$e^{-\beta X} = \frac{1}{p} - 1 \Rightarrow e^{-\beta X} = \frac{1-p}{p}$$

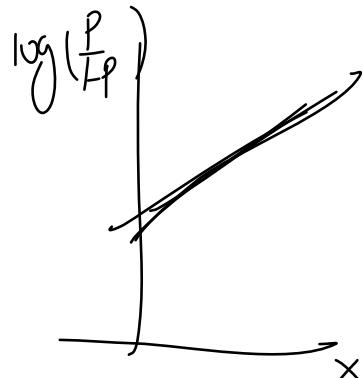
$$\frac{1}{e^{\beta X}} = \frac{1-p}{p} \Rightarrow \frac{p}{1-p} = e^{\beta X}$$

$$\log\left(\frac{p}{1-p}\right) = \log(e^{\beta X})$$

$$\log\left(\frac{p}{1-p}\right) = \beta X$$

$$\log\left(\frac{p}{1-p}\right) = \beta X$$

↑
ln or odds



log odds



$$\log\left(\frac{\hat{y}}{1-\hat{y}}\right) \rightarrow X$$

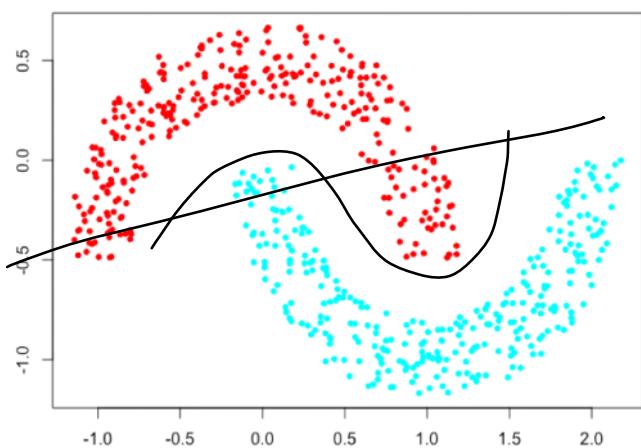
↳ linear

$$\hat{y} \rightarrow X$$

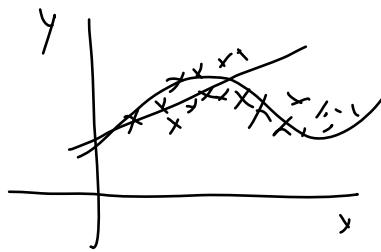
↳ Nonlinear

Polynomial Features

05 July 2023 10:18



$\left\{ \begin{array}{l} \text{deg} \rightarrow \text{incre} \rightarrow \text{overfit} \\ \text{small} \rightarrow \text{underfit} \end{array} \right.$



Polynomial features

$\rightarrow x \rightarrow \text{degree} \rightarrow 2$

$$\begin{matrix} x^0 & | & x^1 & | & x^2 \\ \uparrow & & \uparrow & & \uparrow \\ \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_0 \end{matrix}$$

$\rightarrow \text{draw non-linear decision bound}$

$$x_1 \quad x_2 \quad \text{degree}=2$$

$$(x_1^0 \quad x_1^1 \quad x_1^2 \quad x_2^0 \quad x_2^1 \quad x_2^2)$$

Regularization in Logistic Regression

05 July 2023 07:22

train - w ↓
test ↑

Regularization is a technique used in machine learning models to prevent overfitting, which occurs when a model learns the noise along with the underlying pattern in the training data. Overfitting leads to poor generalization performance when the model is exposed to unseen data.

In the context of linear models like linear regression and logistic regression, regularization works by adding a penalty term to the loss function that the model tries to minimize. This penalty term discourages the model from assigning too much importance to any single feature, which helps to prevent overfitting.

The most common types of regularization in linear models are L1 and L2 regularization:

1. L1 regularization (Lasso Regression): This technique adds a penalty term equal to the absolute value of the magnitude of the coefficients. Mathematically, it's represented as the sum of the absolute values of the weights ($\|w\|_1$). This can lead to sparse models, where some feature weights can become exactly zero. This property makes L1 regularization useful for feature selection.
2. L2 regularization (Ridge Regression): This technique adds a penalty term equal to the square of the magnitude of the coefficients. Mathematically, it's represented as the sum of the squared values of the weights ($\|w\|_2^2$). L2 regularization tends to spread the weight values more evenly across features, leading to smaller, but non-zero, weights.

me

$$\sum (y_i - \hat{y}_i)^2 + \lambda \|w\|^2$$

There's also Elastic Net regularization, which is a combination of L1 and L2 regularization. The contribution of each type can be controlled with a separate hyperparameter.

In all these techniques, the amount of regularization to apply is controlled by a hyperparameter, often denoted as λ (lambda). Higher values of λ mean more regularization, leading to simpler models that might underfit the data. Lower values of λ mean less regularization, leading to more complex models that might overfit the data. The optimal value of λ is typically found through cross-validation.

$$L = -\frac{1}{n} \sum_{i=1}^n y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i) + \lambda \frac{\|\beta\|_2^2}{2} \quad \text{L2 reg}$$
$$+ \lambda \|\beta\|_1 \rightarrow \text{L1 reg}$$
$$C = \frac{1}{\lambda}$$

$\lambda \rightarrow \text{hyper}$
 $\lambda \downarrow \text{increas}$
 $\lambda \downarrow \text{underfitt}$
 $\lambda \downarrow \text{small}$
 $\lambda \downarrow \text{overf..}$
 $\lambda \rightarrow \text{big values}$
 $\lambda \downarrow \text{hard}$

Hyperparameters

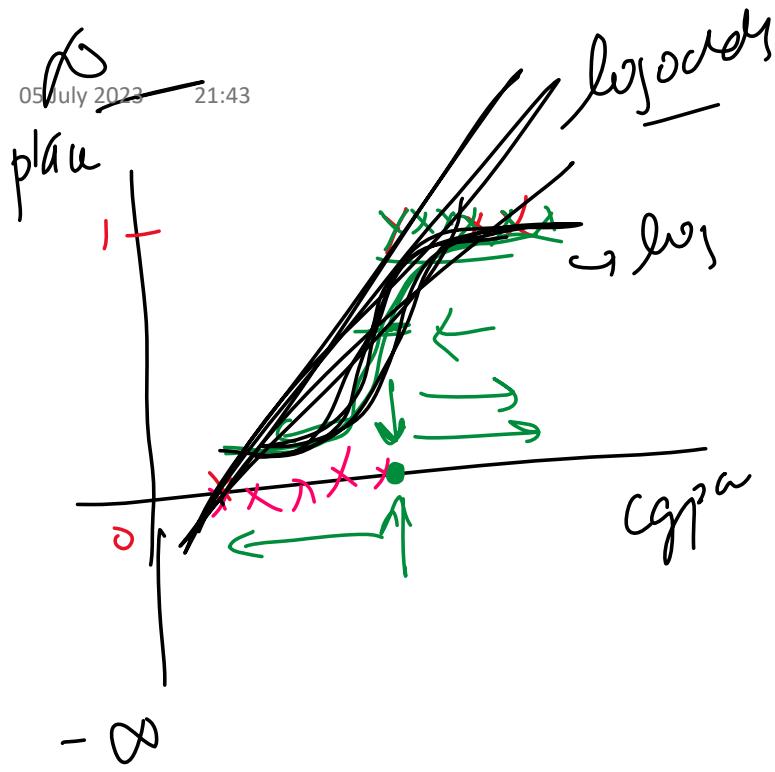
05 July 2023 10:18

Tasks

05 July 2023 16:28

1. Find the solution for regularized loss function
2. Apply hyper parameter tuning to a real world dataset

analytic \rightarrow World's first \leftarrow
↑



Copy placement	
8	1
6	1
4	0
9	1
,	
,	
,	

Lspn

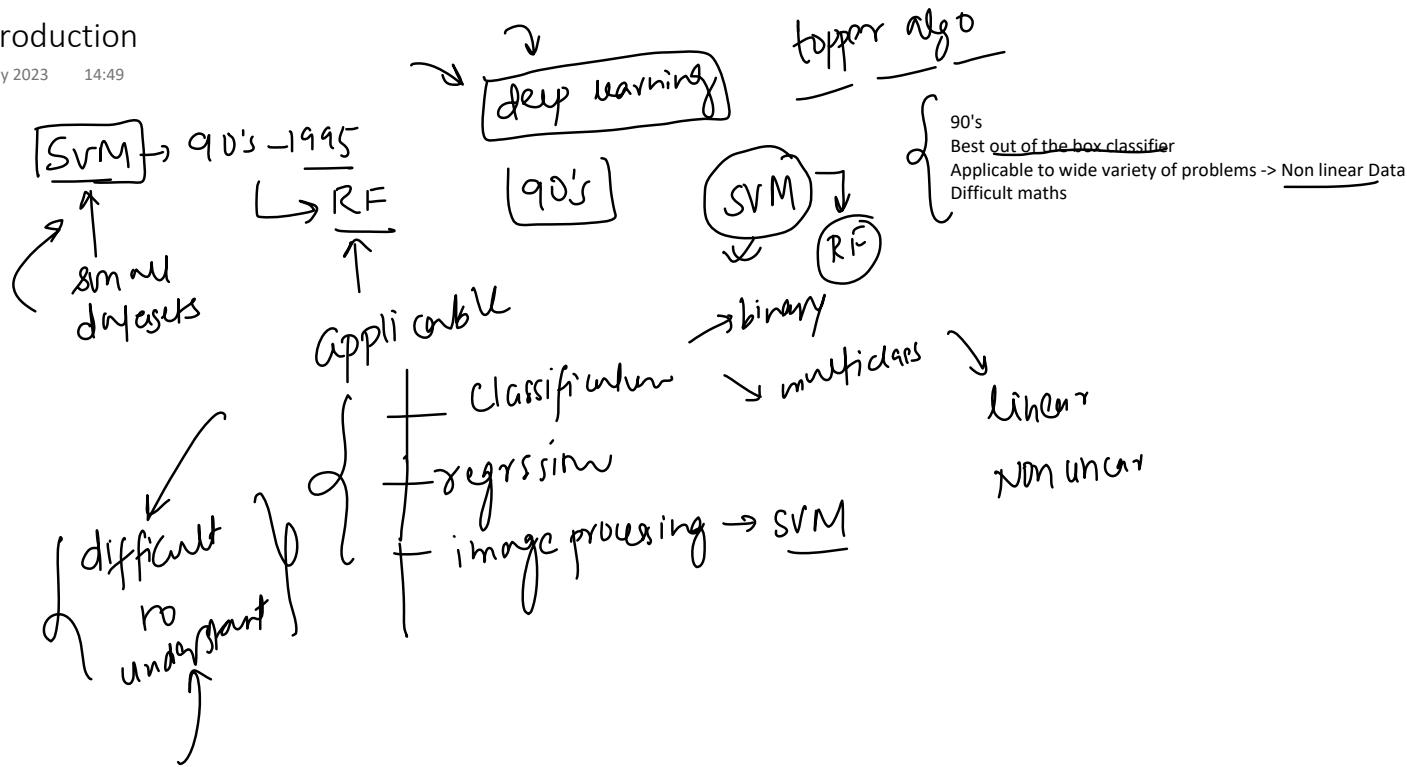
CSP
IG
Chr

$$\text{Zn} \quad \left\{ \begin{array}{l} \text{CGPa} \\ \text{IV} \\ \hline \text{12/m} \\ \text{plane} \end{array} \right.$$

✓✓✓✓

Introduction

06 July 2023 14:49



Plan of Attack

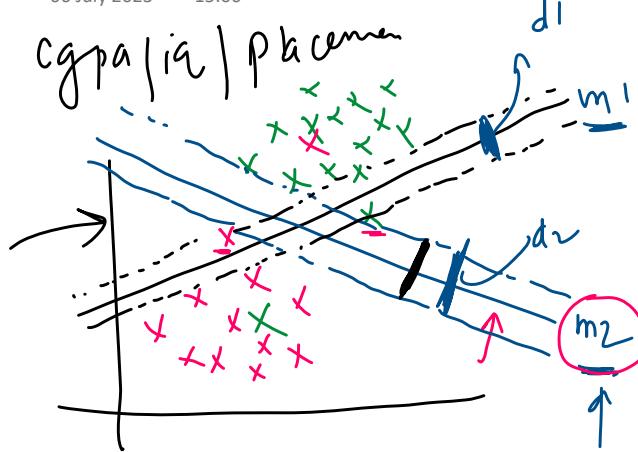
06 July 2023 15:06

- 1) Maximal margin classifier \rightarrow SVM \rightarrow Hard margin SVM
- 2) Soft margin SVM \rightarrow Support vector classifier \rightarrow SVC \rightarrow linear datasets
- 3) SVM \rightarrow kernels \rightarrow non-linear
- 4) SVM for multiclass support
- 5) SVR

[Maximal Margin] Classifier

06 July 2023 15:06

SVM Hard margin



m_1 or m_2

Basic requirement

$d_2 > d_1$

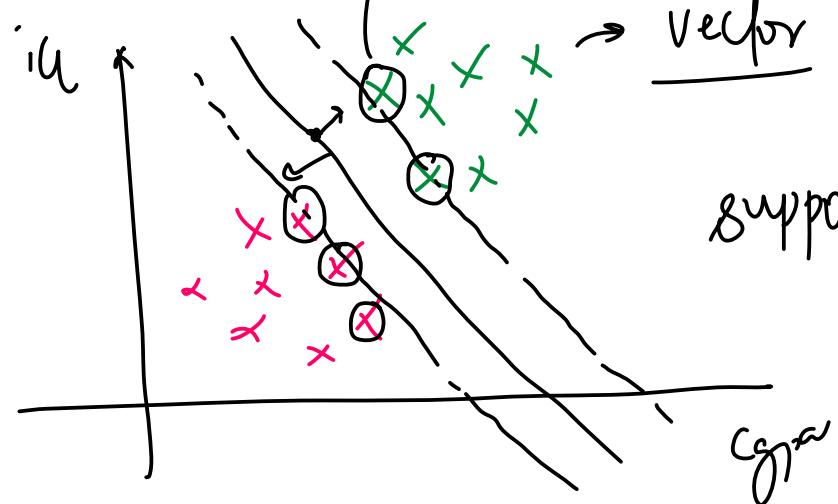
Linearly separable

Support Vectors
06 July 2023 15:16

SVM

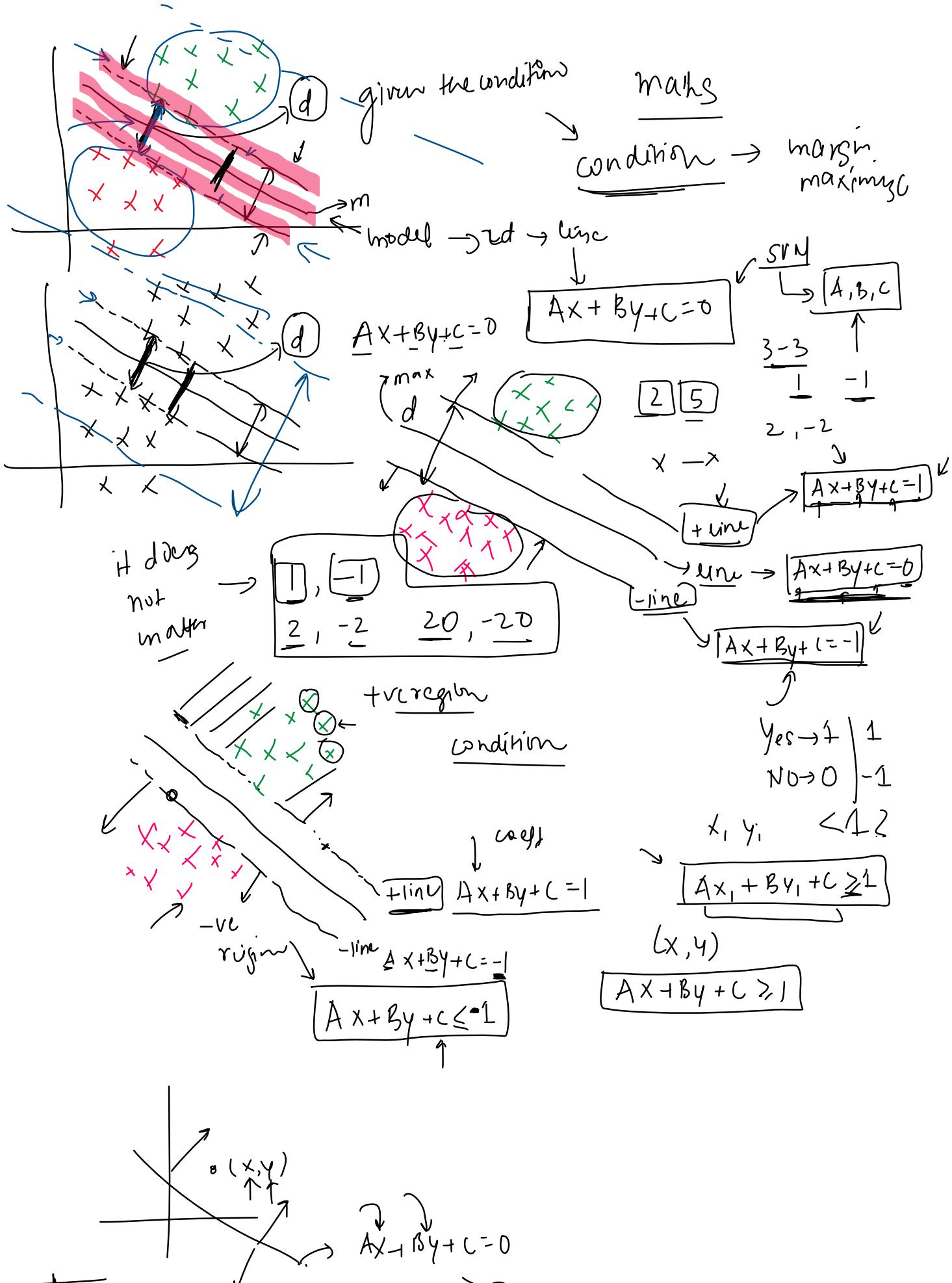
vector

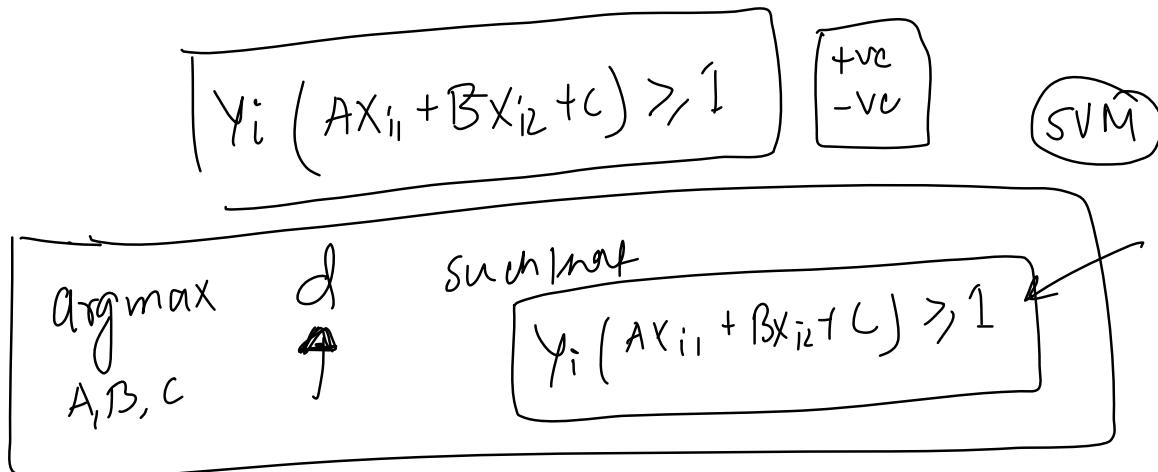
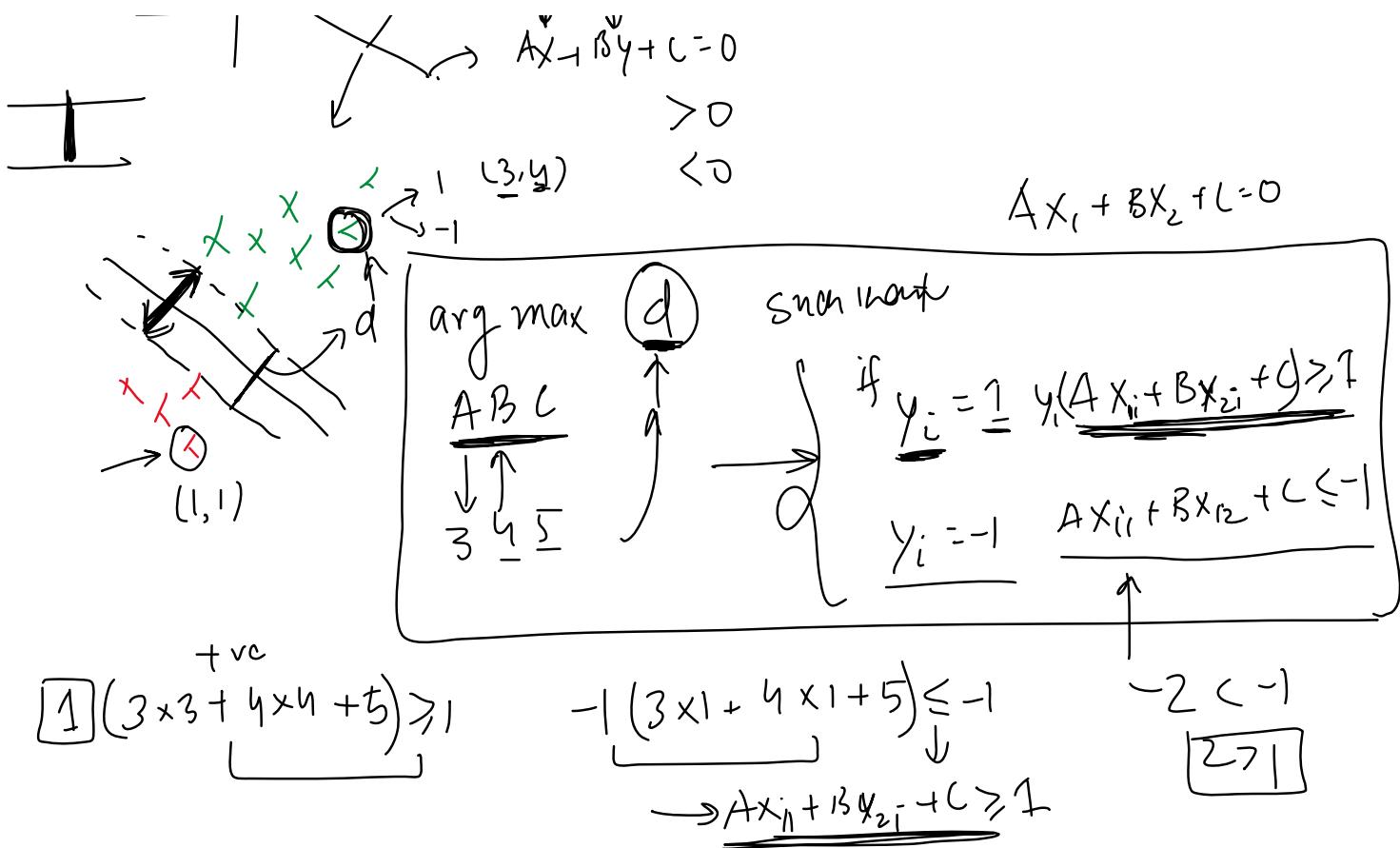
cgpalig | plan
ret



Mathematical Formulation

06 July 2023 15:07





$Ax + By + C = 1 \rightarrow Ax + By + C - 1 = 0$
 $ax + by + c_1 = 0$
 $ax + by + c_2 = 0$
 \downarrow
 $Ax + By + C + 1 = 0$
 $d = \frac{|C_1 - C_2|}{\sqrt{a^2 + b^2}}$
 $d = \frac{|d + 1 - (-1)|}{\sqrt{A^2 + B^2}} = \frac{2}{\sqrt{A^2 + B^2}} = \phi$

$$\begin{array}{c}
 \text{argmax}_{A, B, C} \frac{1}{\sqrt{A^2 + B^2}} \quad \text{given } \left\{ \begin{array}{l} Y_i (Ax_{i1} + Bx_{i2} + C) > 1 \\ \vdots \end{array} \right. \\
 \text{argmin}_{A, B, C} \frac{\sqrt{A^2 + B^2}}{2} \quad \text{given}
 \end{array}$$

How to solve this?

06 July 2023 16:36

constrained
optimization
problem

constraint

Quadratic Programming

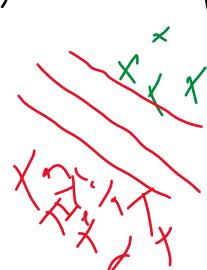
argmax
 $A B C$

$$\frac{2}{\sqrt{A^2 + B^2}}$$

given {

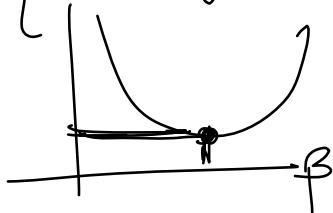
$$y_i (Ax_{i1} + Bx_{i2} + C) \geq 0$$

linear
program

$$(P) L = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \min$$


$$(Q) L = -\frac{1}{n} \sum_{i=1}^n y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i)$$

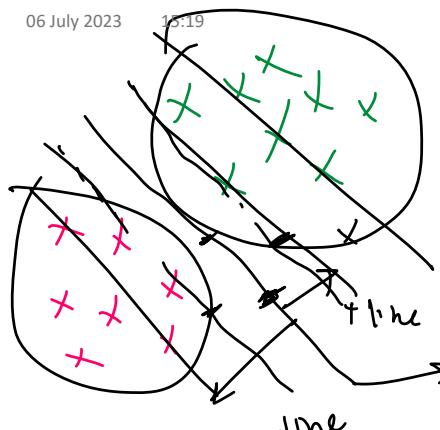
min L gradient



Prediction

06 July 2023

18:19



(8, 80)

$A \times 8 + B \times 80 + C \geq 0 \rightarrow \text{placement}$

$< 0 \rightarrow \text{no placement}$

$$\underline{Ax + By + C = 0}$$

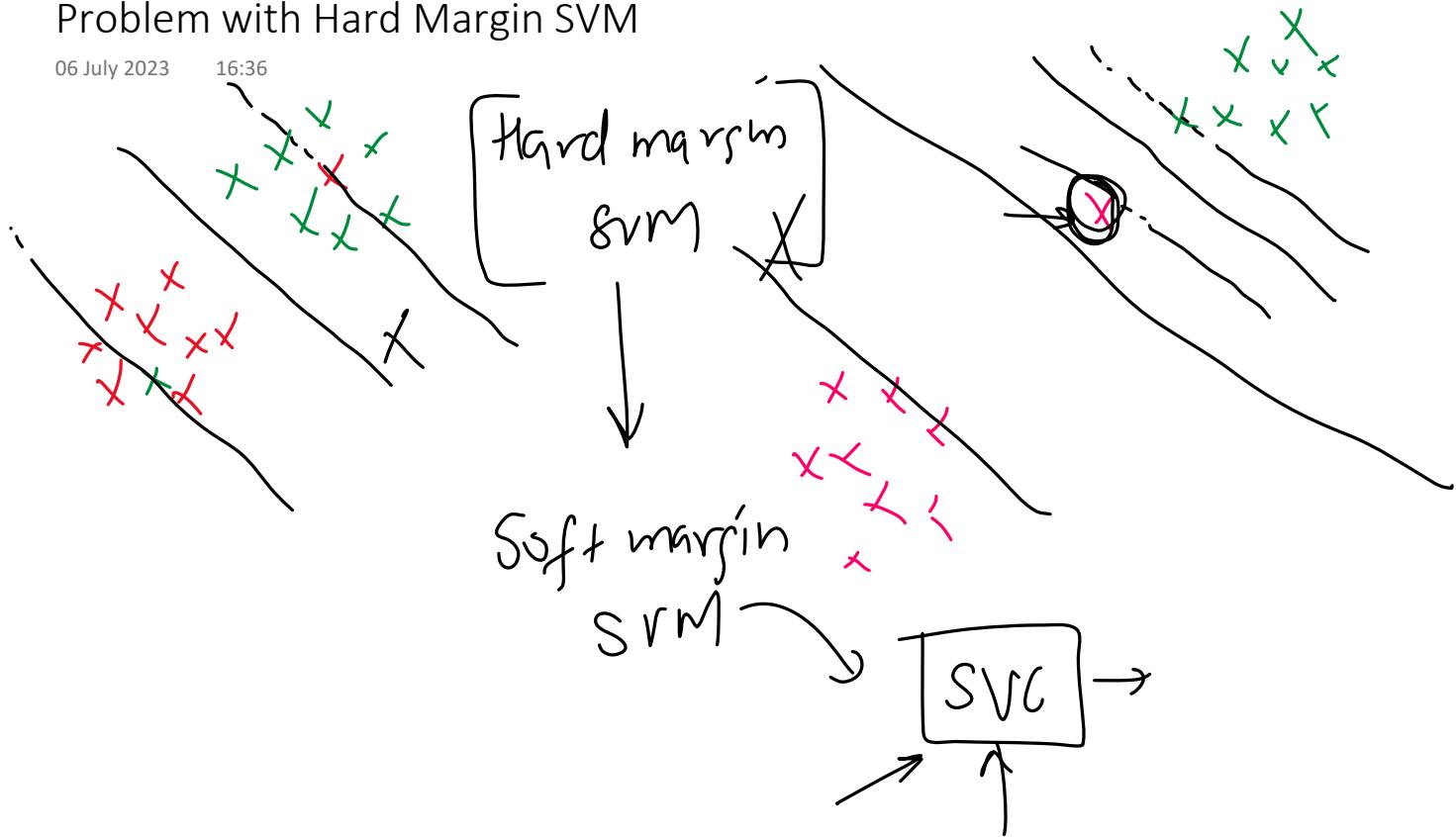
Hard margin SVM

Code

06 July 2023 15:07

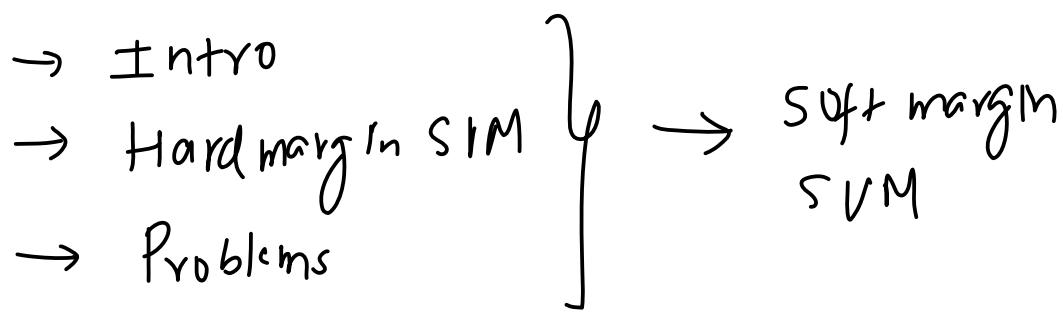
Problem with Hard Margin SVM

06 July 2023 16:36



Recap

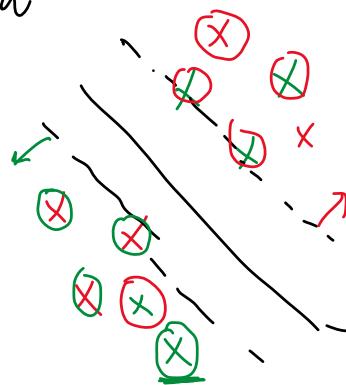
10 July 2023 14:58



Problems with Hard Margin SVM \rightarrow Soft margin SVM

08 July 2023 08:07

\rightarrow linear data



\rightarrow big prob

$$Ax + By + C = 0$$

x_{1i}	x_{2i}	y_i
x_{11}	x_{21}	y_1
x_{12}	x_{22}	y_2
x_{13}	x_{23}	y_3

$\underset{A, B, C}{\operatorname{argmax}}$

$$\frac{2}{\sqrt{A^2 + B^2}}$$

such that

$$y_i (A x_{1i} + B x_{2i} + C) \geq 1$$

for all x_i

Hard margin

\hookrightarrow softn \rightarrow Soft margin SVM

[Slack Variable] $\rightarrow \xi \rightarrow \text{misclassification score} \downarrow$
 08 July 2023 10:19 \rightarrow Hinge loss \leftarrow

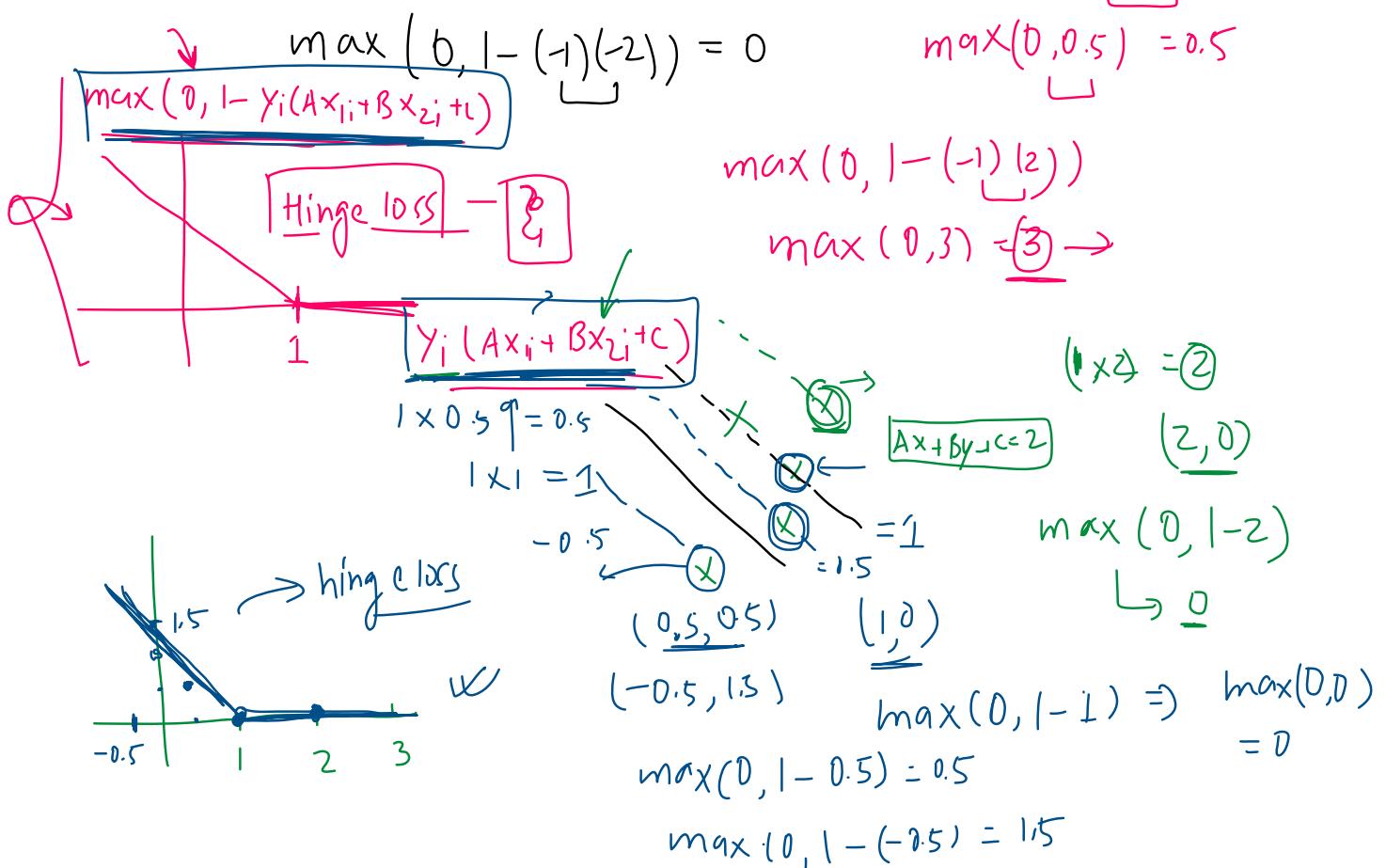
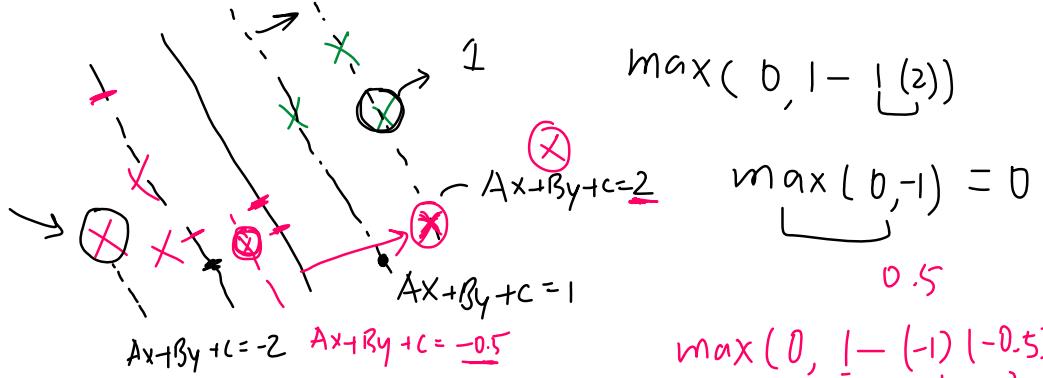
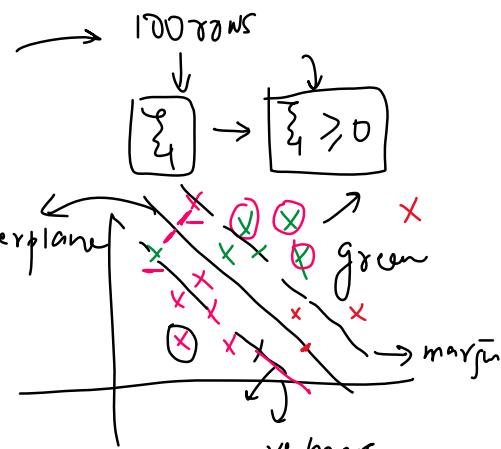
The concept of slack variables was introduced by Vladimir Vapnik in 1995 and is used in the formulation of the "soft-margin" SVM to handle cases where data is not linearly separable, or when one allows for some degree of error in classification.

Mathematically, for each data point i , a slack variable $\xi_i \geq 0$ is introduced. The slack variable ξ_i measures the degree of misclassification of the data point x_i .

- $\xi_i = 0$ if x_i is on the correct side of the margin.
- $0 < \xi_i < 1$ if x_i is on the correct side of the hyperplane but on the wrong side of the margin.
- $\xi_i \geq 1$ if x_i is on the wrong side of the hyperplane, i.e., it is misclassified.

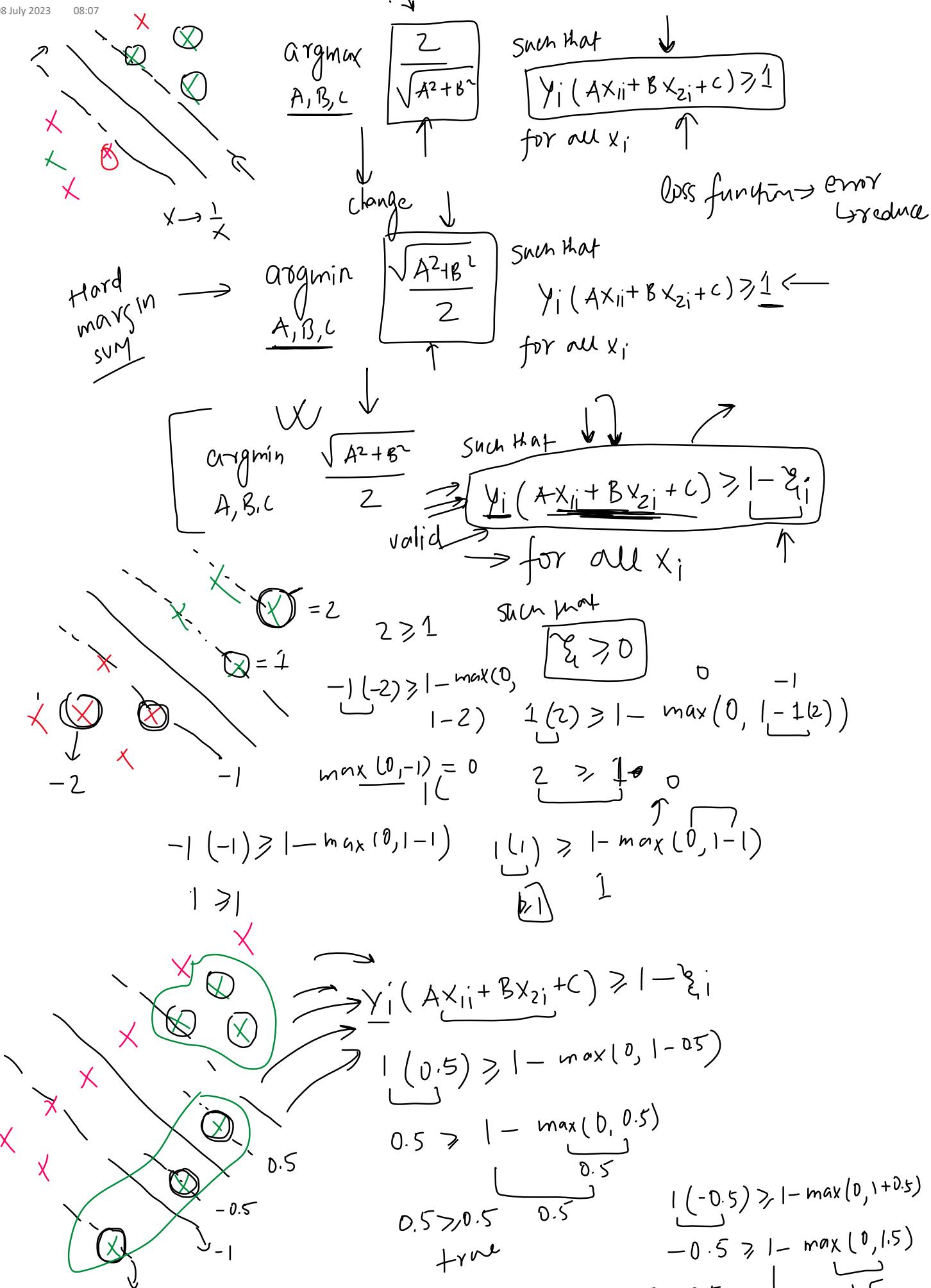
Hinge loss
slack

$$= \max(0, 1 - \underbrace{y_i(Ax_{ii} + Bx_{2i} + C)}_{\text{Hinge loss}})$$



Soft Margin SVM

08 July 2023 08:07



$$\text{true}$$

$$-0.5 \geq 1 - \max(0, 1.5)$$

$$-0.5 \geq 0.5$$

$$-0.5$$

$\boxed{-2 \geq -2}$ true

$$1 (-2) \geq 1 - \max(0, 1+2)$$

$$-2 \geq 1 - 3$$

\rightarrow

$$y_i(Ax_{1i} + Bx_{2i} + C) \geq 1 - \xi_i$$

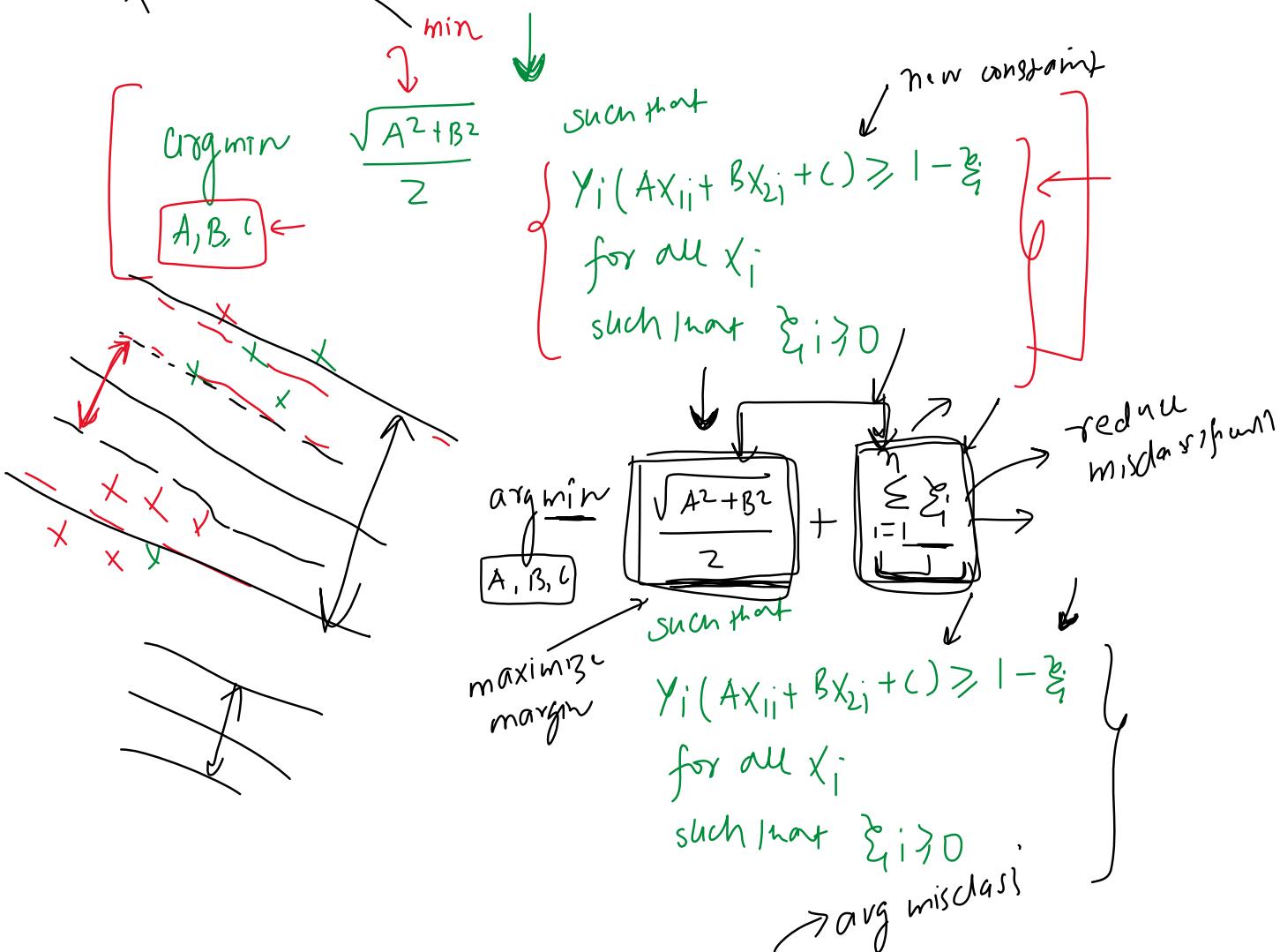
it is allowing all the points

this is no more a constraint

All true condition

$\times \quad \otimes \quad \otimes \quad \times$

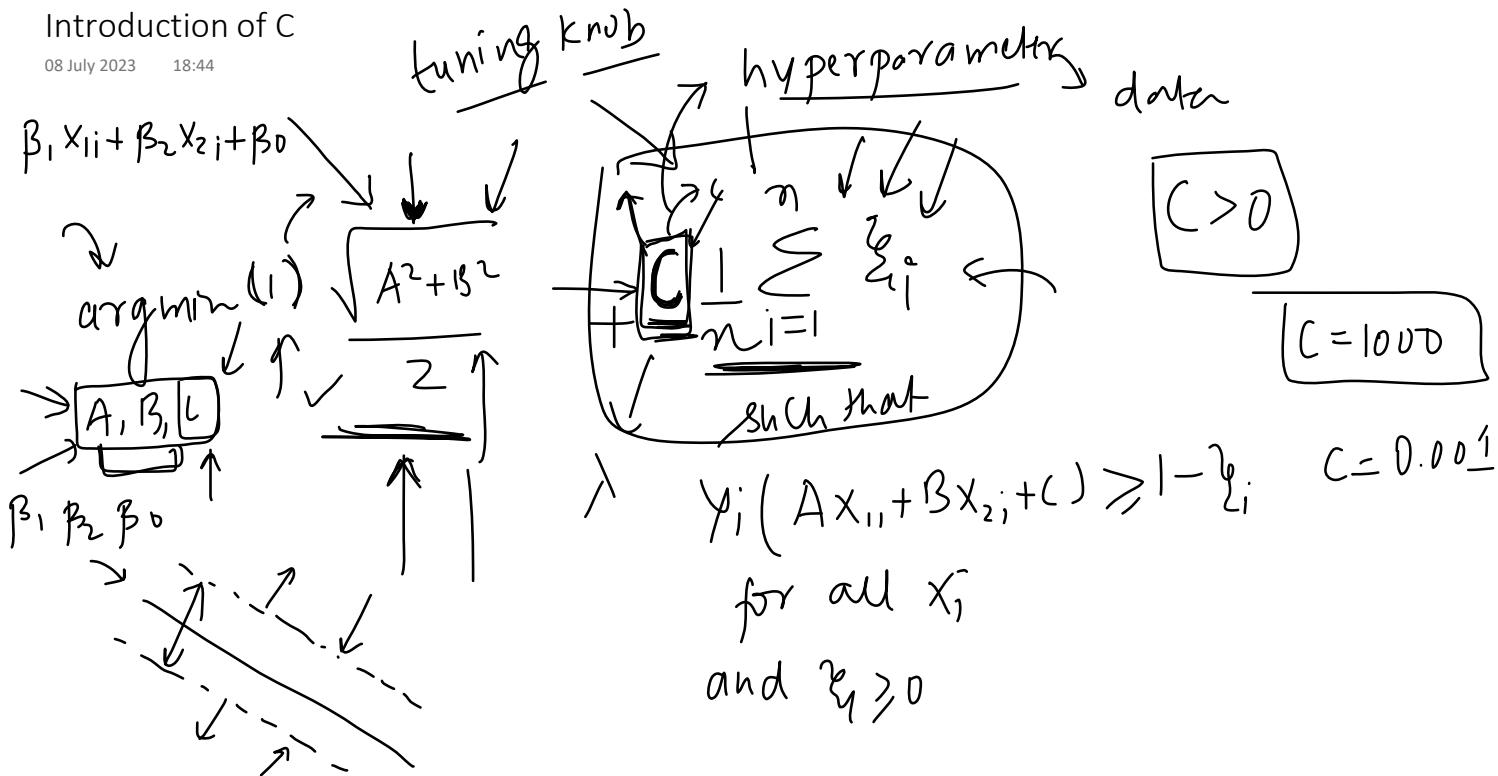
hard margin \hookrightarrow flexibility



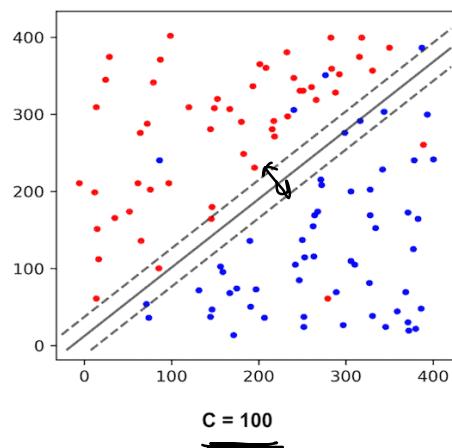
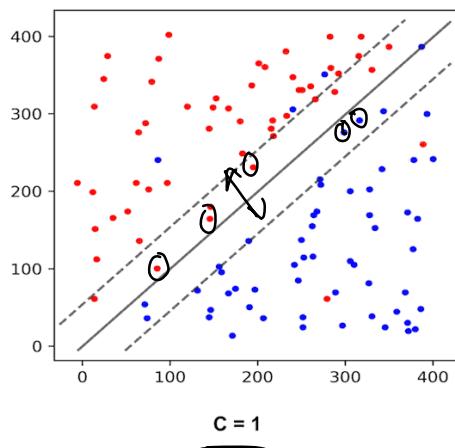
$$\begin{aligned}
 & \text{argmin}_{A, B, C} \quad \frac{\sqrt{A^2 + B^2}}{2} + \frac{1}{n} \sum_{i=1}^n \xi_i \\
 & \text{such that} \\
 & y_i(Ax_{1i} + Bx_{2i} + C) \geq 1 - \xi_i \\
 & \text{for all } x_i \\
 & \text{and } \xi_i \geq 0
 \end{aligned}$$

Introduction of C

08 July 2023 18:44



SVM Parameter C



Bias Variance Tradeoff

08 July 2023 08:07

$$\underset{A, B, C}{\operatorname{argmin}} \frac{\sqrt{A^2 + B^2}}{2} + \frac{C}{n} \sum_{i=1}^n \hat{y}_i$$

$\rightarrow C$ high \rightarrow overfitting (low bias high variance)

$\rightarrow C$ low \rightarrow underfitting (high bias low variance)



Code Example

08 July 2023 08:08

Relationship with Logistic Regression

11 July 2023 19:49

$$\underset{\beta_0, \beta_1, \beta_2}{\operatorname{argmin}} \quad$$

$$\frac{\sqrt{\beta_1^2 + \beta_2^2}}{2} + C \frac{1}{n} \sum_{i=1}^n \sum_{j_i}$$

loss function
hinge loss
constraint loss function

$$\underline{\text{L2 reg}}$$

$$\underset{\beta_0, \beta_1, \beta_2}{\operatorname{argmin}} \quad$$

$$\log \text{loss} + \lambda \left(\frac{\sqrt{\beta_1^2 + \beta_2^2}}{2} \right)$$

regu
 $\|\beta\|^2$
regu

$$C \propto \frac{1}{\lambda}$$

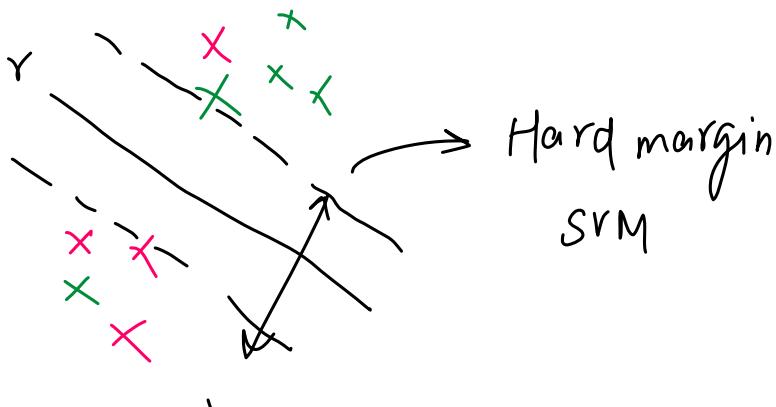
$$\lambda \propto \frac{1}{C}$$

$$-\frac{1}{n} \sum_{i=1}^n y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

Problem with SVC

14 July 2023 09:15

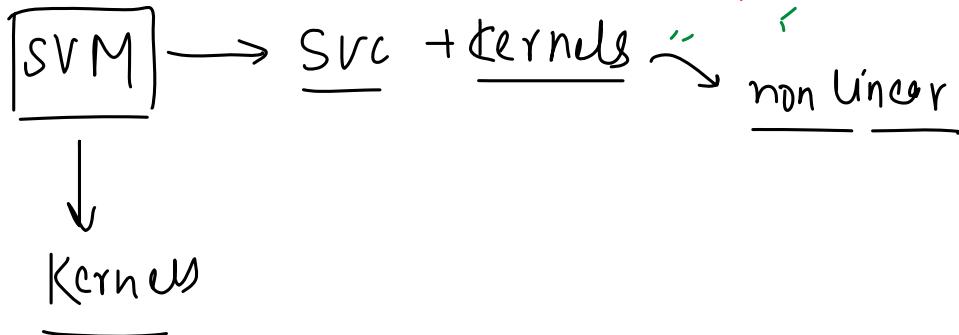
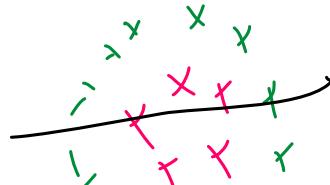
✓ Maximal margin classifier



Hard margin
SVM

✓ SVC { soft margin SVM }

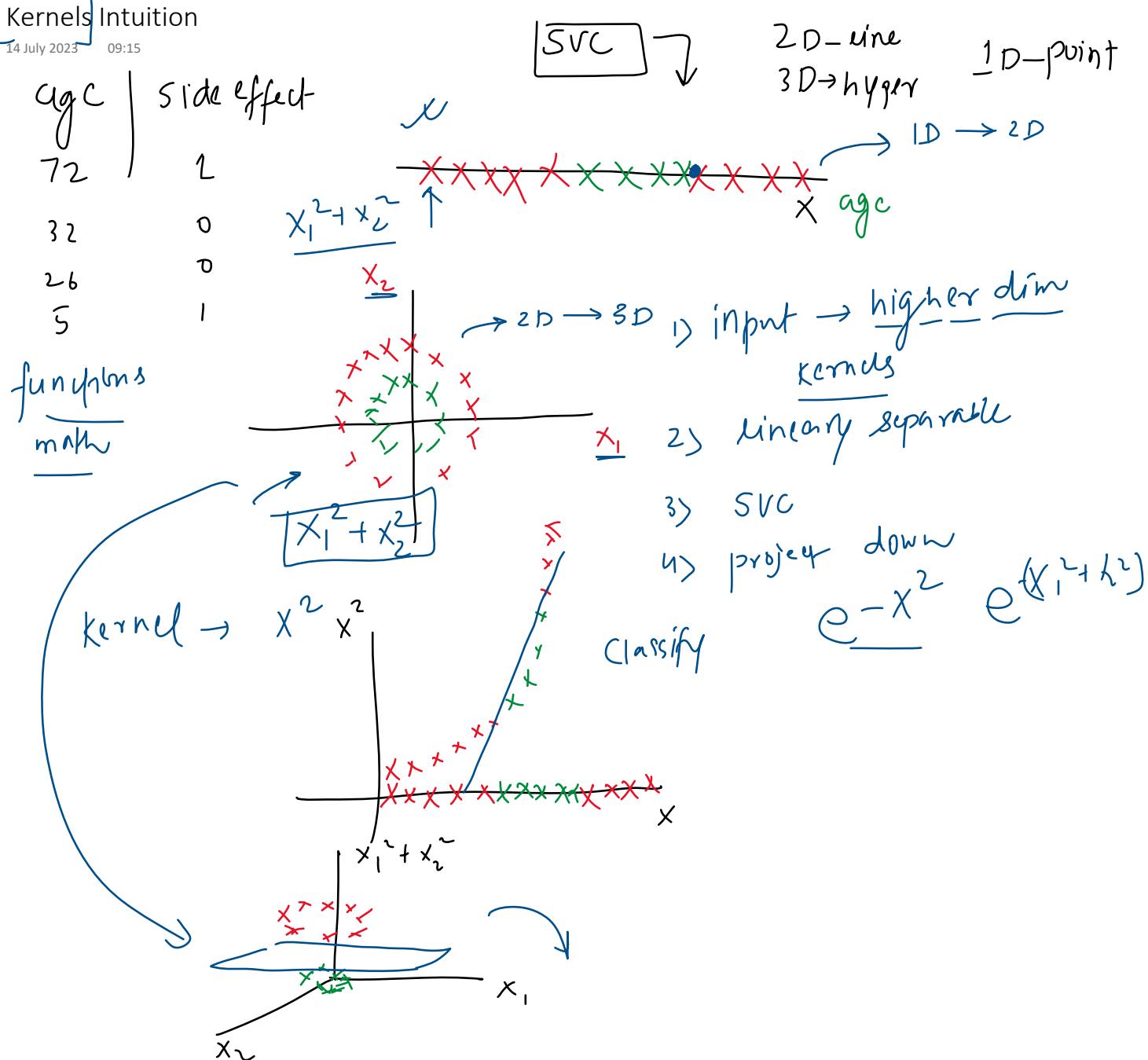
linear dataset



Kernels Intuition

14 July 2023

09:15

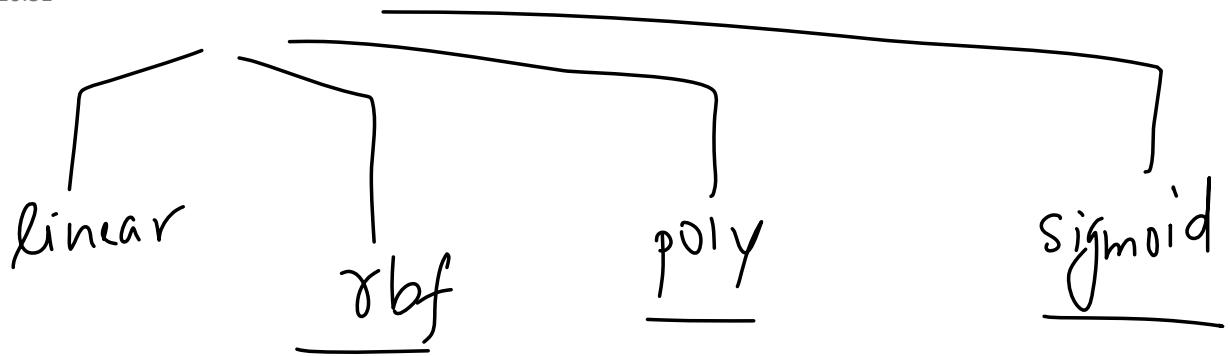


Code

14 July 2023 10:31

Types of Kernels

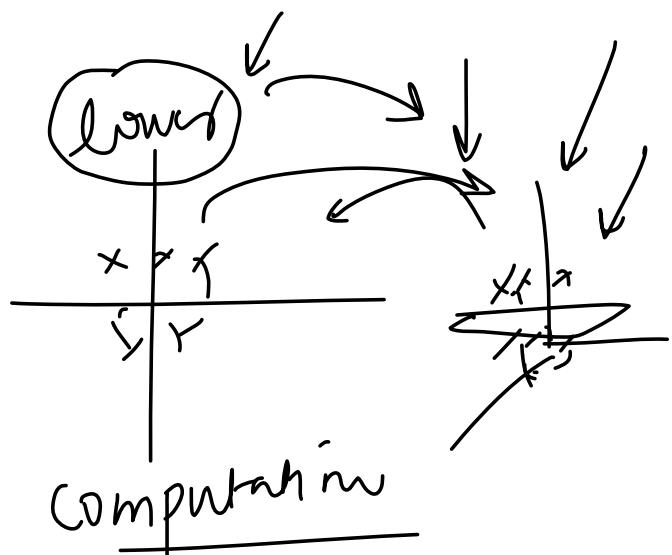
14 July 2023 10:31



Why is it called Trick?

14 July 2023 10:32

Kernel trick?
↳ fast

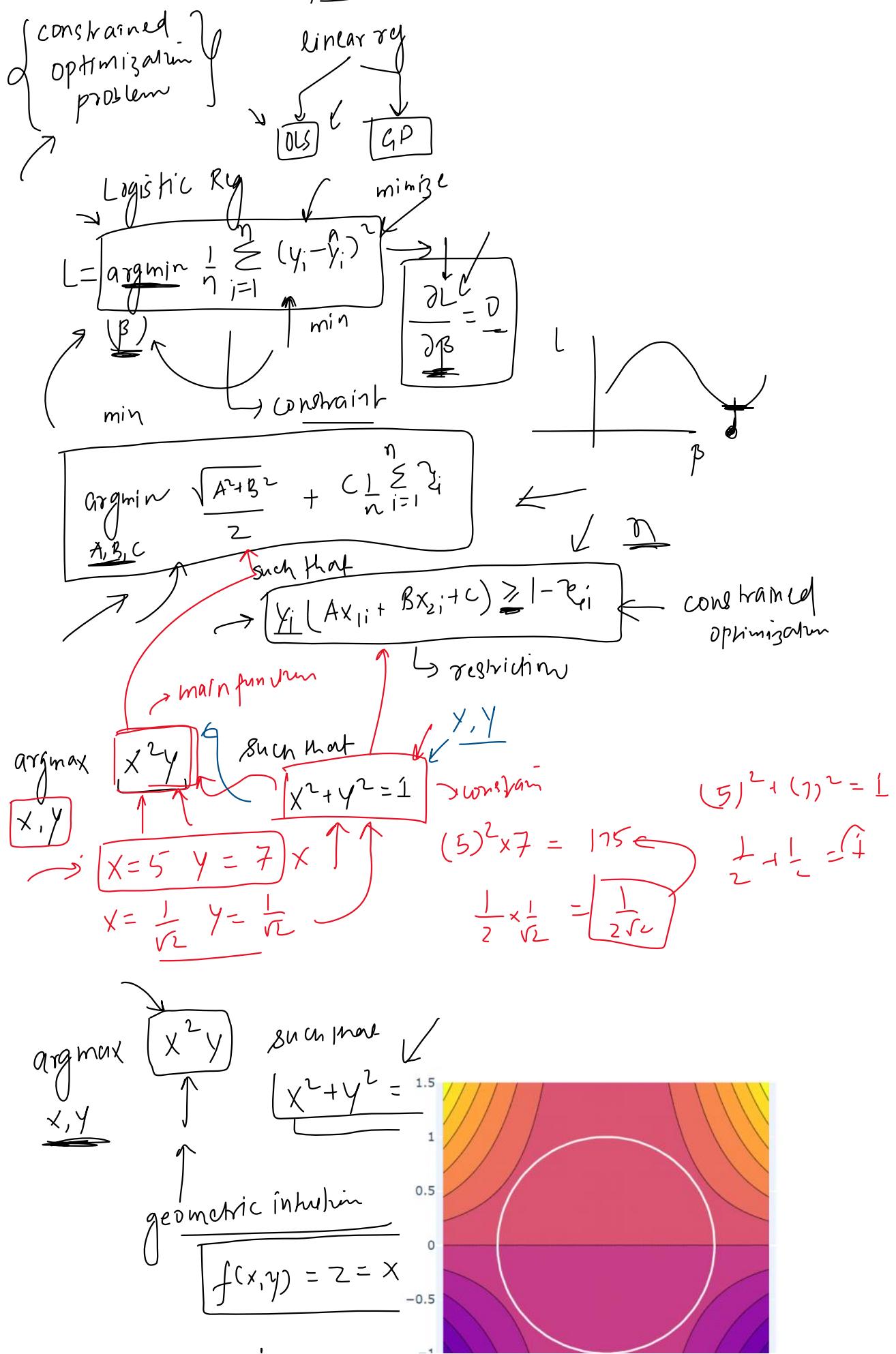


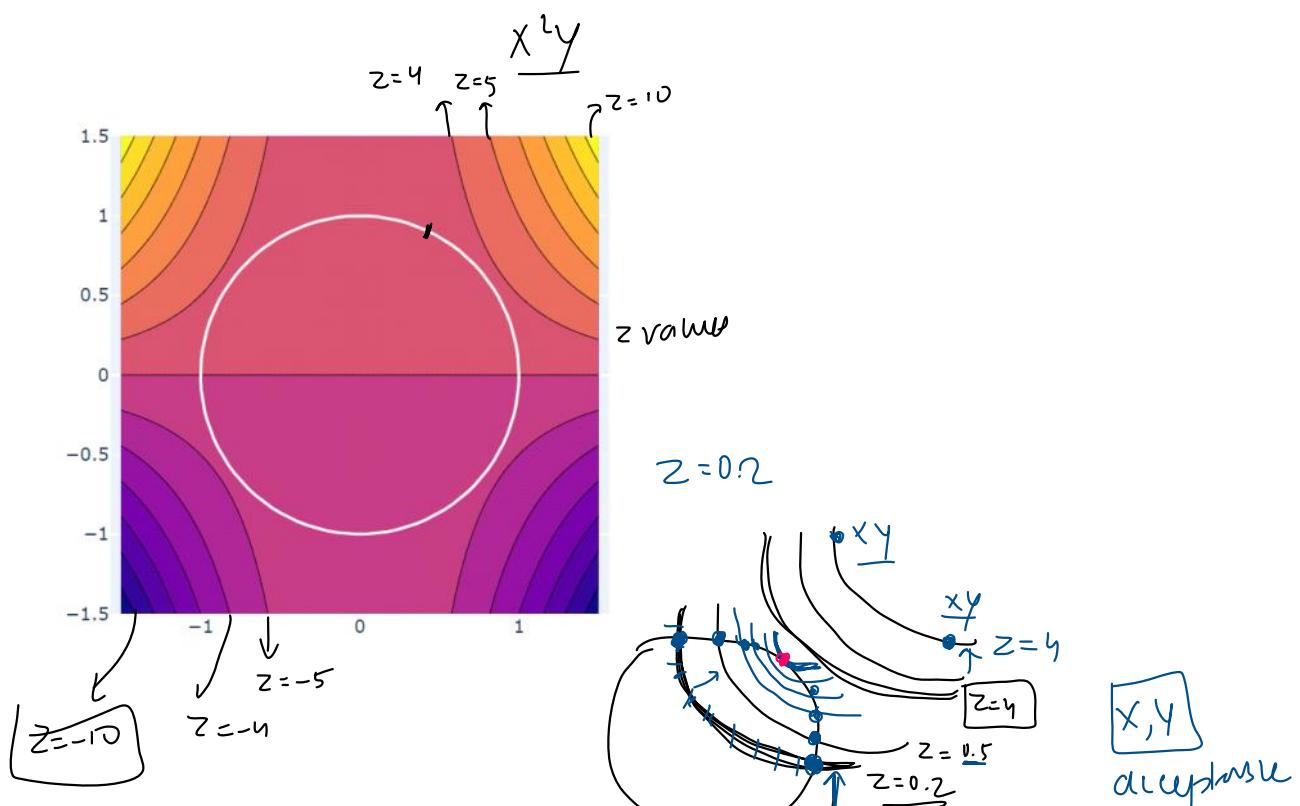
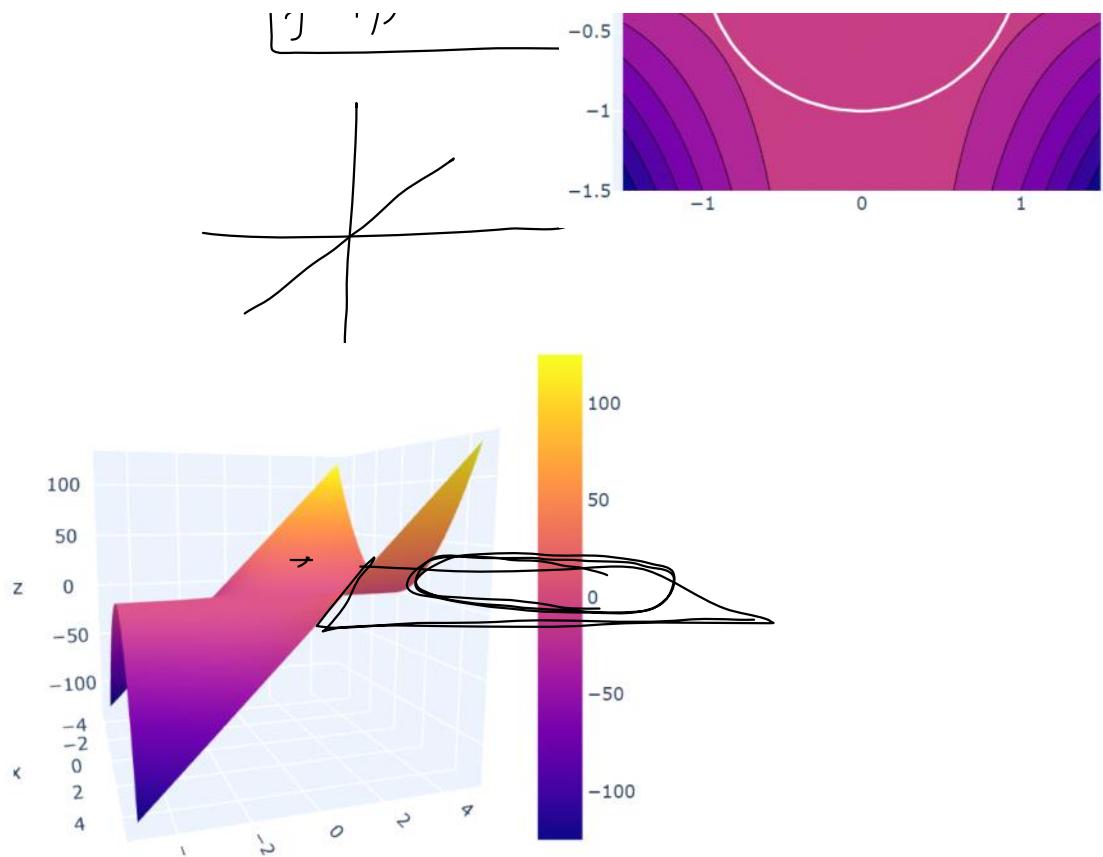
The gradient of a function at a point is a vector that points in the direction of the steepest ascent of the function at that point. The magnitude (or length) of the gradient vector is equal to the rate of increase of the function in that direction.

Contour lines (or level sets) of a function are curves that connect points where the function has the same value. For a 2D function, these are like the lines of constant altitude on a topographic map.

There are a few key relationships between gradients and contour lines:

1. The gradient at a point is perpendicular (or orthogonal) to the contour line passing through that point. This is because the contour line represents the direction of no change in the function value, while the gradient represents the direction of maximum change.
2. The gradient points in the direction where the function increases most rapidly. If you were to walk along the contour line (where the function value doesn't change), the direction you'd need to go to start climbing as steeply as possible is the direction of the gradient.
3. The magnitude of the gradient (how long the gradient vector is) indicates how steeply the function is increasing. If the contour lines are close together, that means the function is changing rapidly, so the gradient is large. If the contour lines are far apart, the function is changing slowly, so the gradient is small.

optimization prob



$$z = -10$$

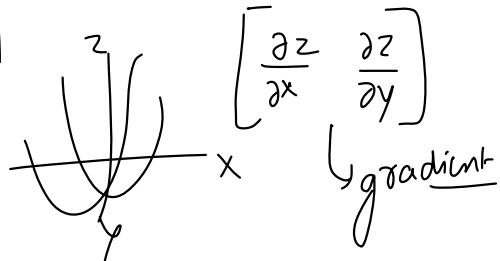
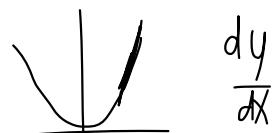
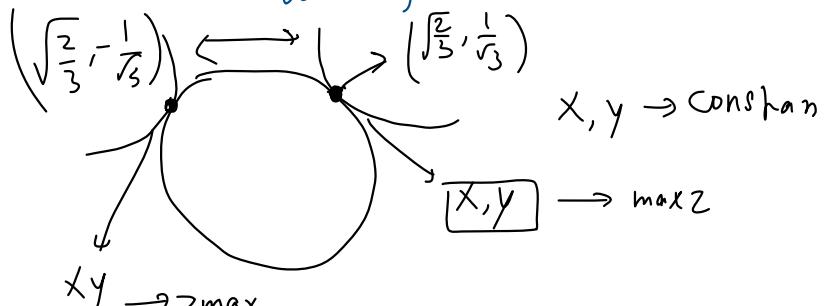
$$z = -u$$

$$f(x, y) = x^2 y$$

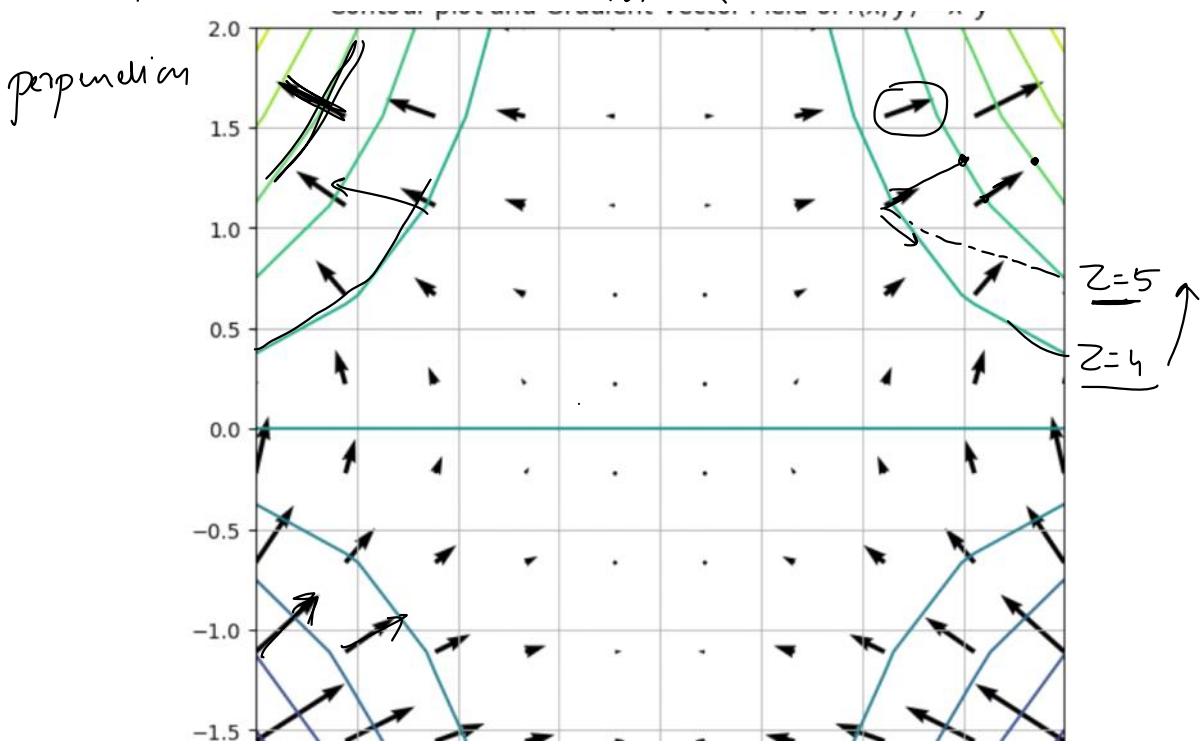
constraint: $x^2 + y^2 = 1$

(x, y) always true

$z = 0.2 \rightarrow$ highest poss' value of z (x, y)



direction of maximum ascent / change

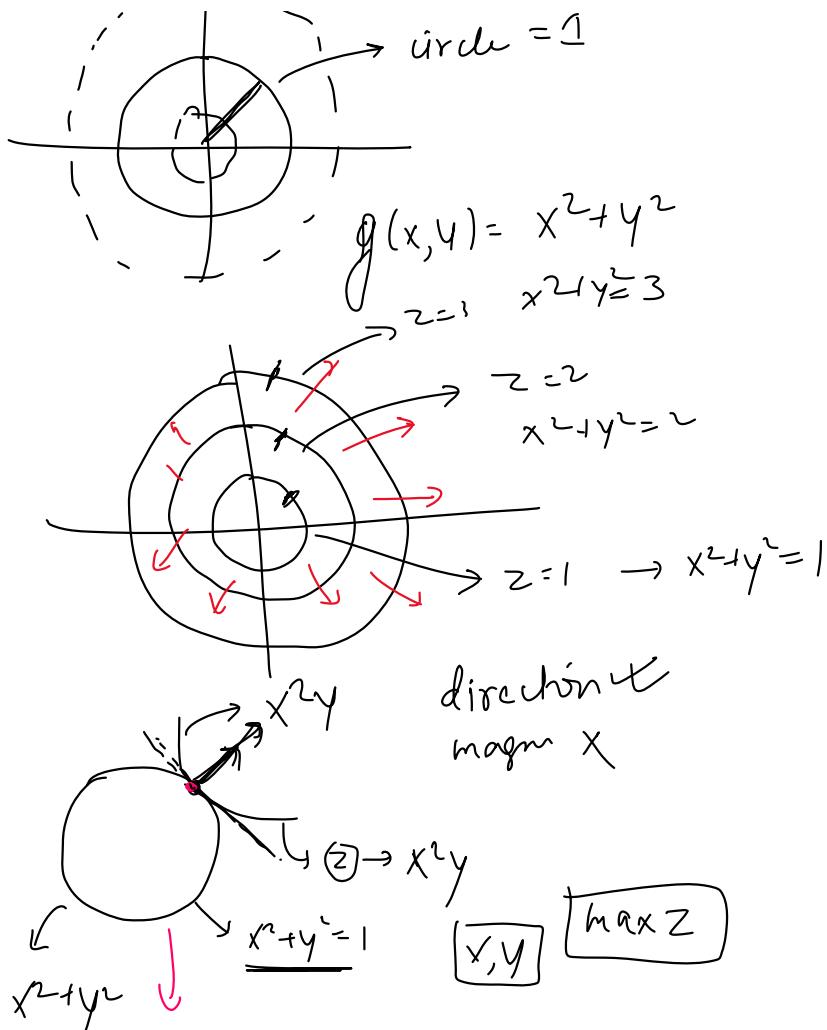


$$x^2 + y^2 = 1$$

$$x^2 + y^2 = 2$$

$$x^2 + y^2 = 0.5$$





$$\nabla f(x,y) = \lambda \nabla g(x,y)$$

$f(x,y) = x^2y$ $\boxed{\nabla f(x,y) = \lambda \nabla g(x,y)}$ $x^2 + y^2 = 1$ $\boxed{x,y}$ $\boxed{\max z}$

$\nabla f(x,y) = \lambda \nabla g(x,y)$ may
 $\nabla f(x,y) \quad \nabla g(x,y)$ $\hookrightarrow \text{scalar}$
 $\nabla f(x,y) \quad \nabla g(x,y)$ $\hookrightarrow \text{Lagrange's multiplier}$

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \quad \begin{aligned} x^2y &\rightarrow \frac{\partial f}{\partial x} = 2xy \\ &\rightarrow \frac{\partial f}{\partial y} = x^2 \\ x^2 + y^2 &\rightarrow \frac{\partial g}{\partial x} = 2x \end{aligned}$$

$\begin{bmatrix} 2xy \\ x^2 \end{bmatrix}$

$$\nabla g(x, y) = \begin{bmatrix} 2xy \\ x^2 \end{bmatrix}$$

$$x^2 + y^2 \rightarrow \frac{\partial g}{\partial x} = 2x$$

$$\frac{\partial g}{\partial y} = 2y$$

$$\nabla g(x, y) = \begin{bmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\begin{bmatrix} 2xy \\ x^2 \end{bmatrix} = \lambda \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\begin{cases} 2xy = \lambda 2x \\ x^2 = \lambda 2y \end{cases} \quad \begin{array}{l} \text{---(1)} \\ \text{---(2)} \end{array}$$

$x, y, \lambda \nearrow$

$$x^2 + y^2 = 1 \quad \text{---(3)}$$

$\boxed{y = \lambda}$

$$2y^2 + y^2 = 1$$

$$3y^2 = 1 \quad y^2 = \pm \frac{1}{\sqrt{3}}$$

$$x^2 + \frac{1}{3} = 1$$

$$x^2 = 1 - \frac{1}{3}$$

$$x^2 = \frac{2}{3} \quad x = \pm \sqrt{\frac{2}{3}}$$

$$x = \sqrt{\frac{2}{3}}, -\sqrt{\frac{2}{3}}$$

$\geq \rightarrow$ maximum

$$y = \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}$$

$$\left(\sqrt{\frac{2}{3}}, \frac{1}{\sqrt{3}} \right) \quad \left(\sqrt{\frac{2}{3}}, -\frac{1}{\sqrt{3}} \right) \quad \left(-\sqrt{\frac{2}{3}}, \frac{1}{\sqrt{3}} \right) \quad \left(-\sqrt{\frac{2}{3}}, -\frac{1}{\sqrt{3}} \right)$$

$\uparrow \quad \uparrow \quad \uparrow \quad \uparrow$

$Z = x^2y$ $\rightarrow \frac{2}{3} \frac{1}{\sqrt{3}} = \frac{2}{3\sqrt{3}}$
 $x, y \rightarrow z \max$ given $x^2 + y^2 = 1$
 $\left(\frac{\sqrt{2}}{3}\right)^2 + \left(\frac{1}{\sqrt{3}}\right)^2 = \frac{2}{3} + \frac{1}{3} = \frac{3}{3} = 1$

Lagrangian multipliers \rightarrow SRM

$\begin{bmatrix} \text{argmax}_{x,y} & x^2y \\ \text{ST} & x^2 + y^2 = 1 \end{bmatrix} \rightarrow \boxed{\nabla f(x,y) = \lambda g(x,y)}$

$L(x,y,\lambda) = \text{argmax}_{x,y} f(x,y) - \lambda(g(x,y) - 1)$

$\frac{\partial L}{\partial x} = 0$ $\frac{\partial L}{\partial y} = 0$ $\frac{\partial L}{\partial \lambda} = 0$
 $\frac{\partial f}{\partial x} - \lambda \frac{\partial g}{\partial x} = 0$
 $\frac{\partial f}{\partial y} - \lambda \frac{\partial g}{\partial y} = 0$
 $x^2 - \lambda 2y = 0$
 $x^2 = \lambda 2y$

$\sum_{i=1}^n (x_i - y_i)$
 $(x_1 - y_1) + (x_2 - y_2)$

$\begin{bmatrix} \text{argmax}_{x,y} & x^2y \\ \text{ST} & x^2 + y^2 = 1 \end{bmatrix}$ \rightarrow opt
 norm - \rightarrow computer

\downarrow
 Optm-

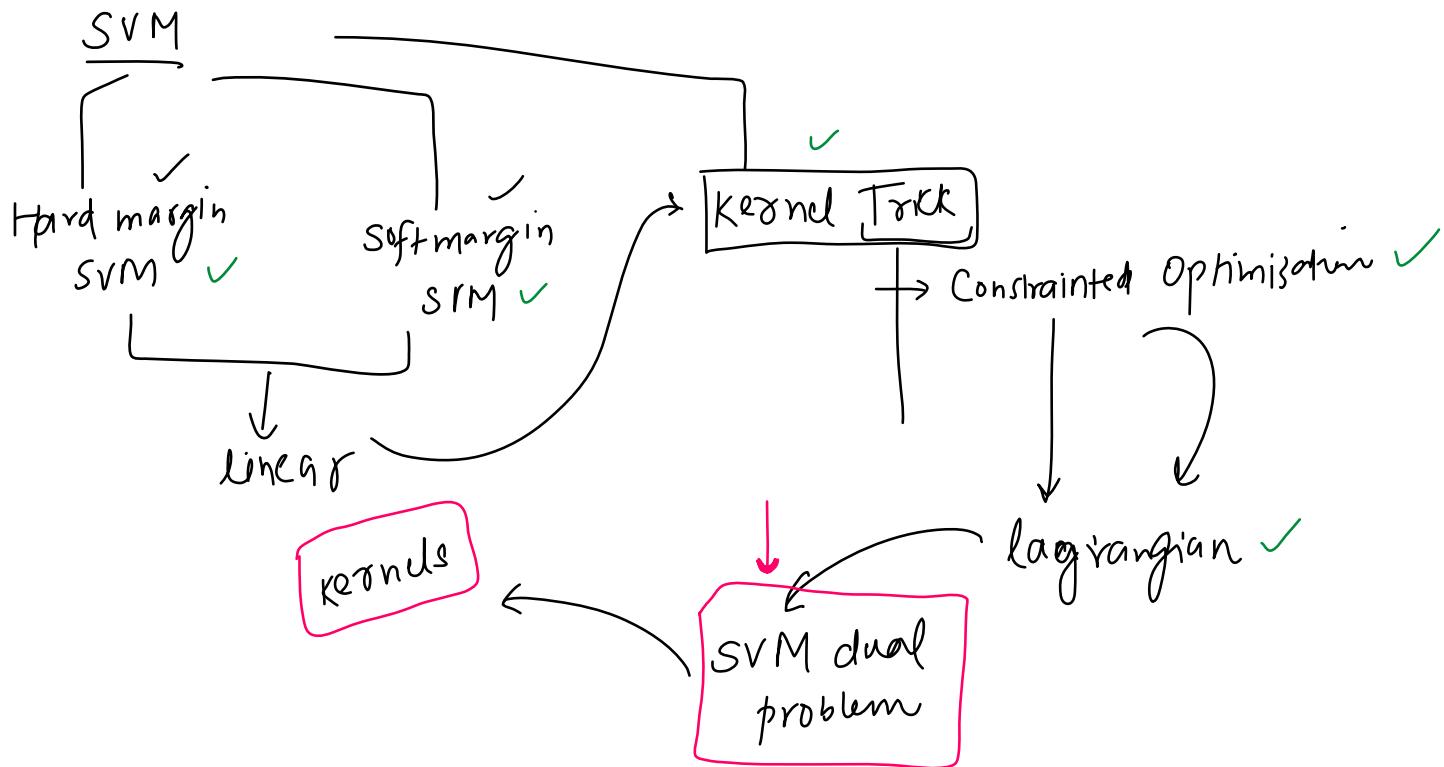
$$L(x, y, \alpha) = \underset{\alpha}{\operatorname{arg\,max}} \quad x^T y - \lambda (x^2 + y^2 - 1)$$

$x^T y$ \rightarrow $w^T x$
 $x^2 + y^2 - 1$ \rightarrow $w^T w$

large margin \rightarrow SVM dual problem

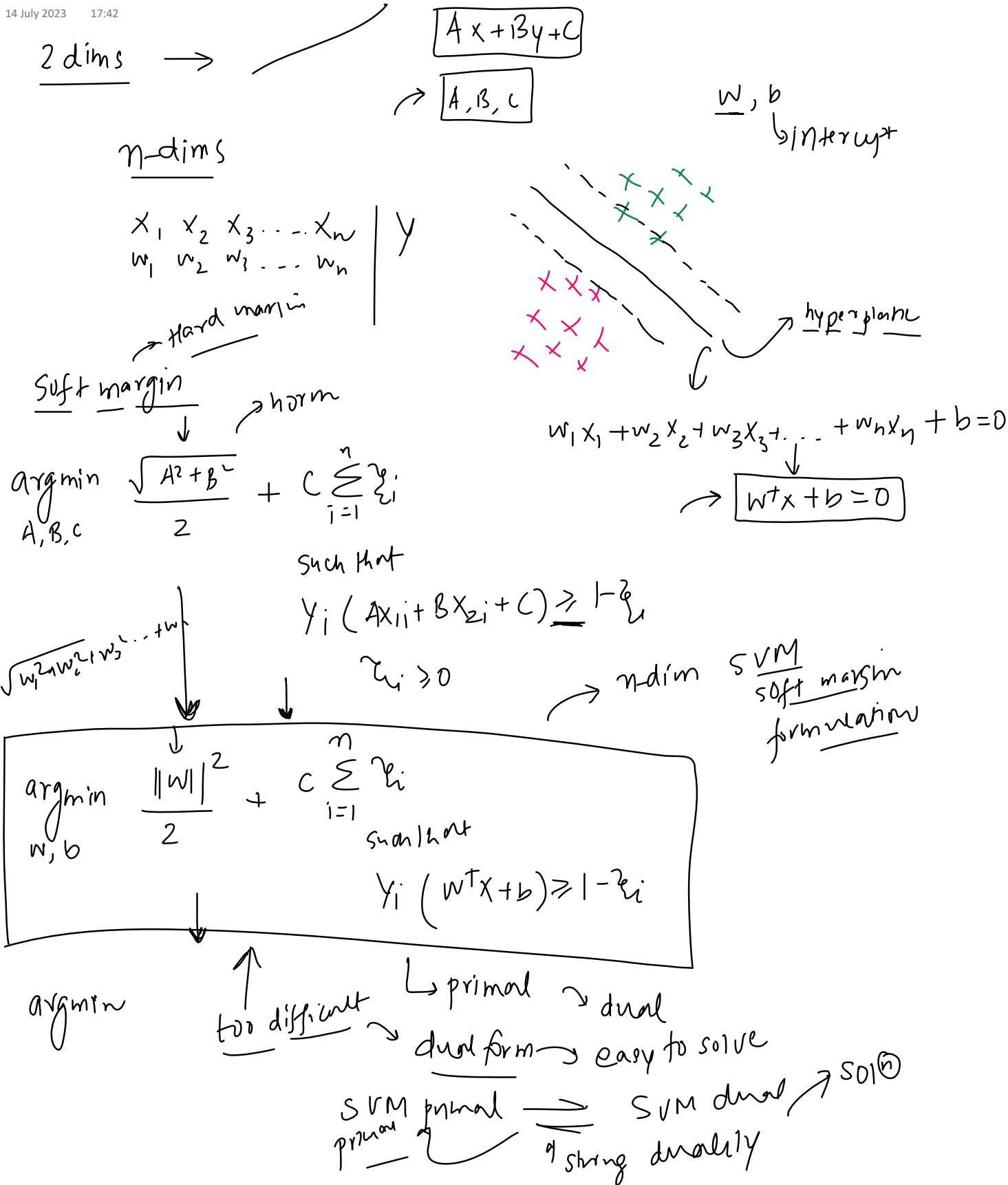
Recap

16 July 2023 00:02



SVM in n Dimensions

14 July 2023 17:42



Constrained Optimization Problems with Inequality

16 July 2023 00:02

Constrained optimization
problem equality

$$\text{maximize } f(x, y) = \underline{x^2 y}$$

$$x, y \text{ such that } x^2 + y^2 = 1$$

SVM

Karush Kuhn Tucker Conditions (KKT conditions)

16 July 2023 11:55

They generalize the method of Lagrange multipliers to handle inequality constraints. In the context of support vector machines (SVMs) and many other optimization problems, the KKT conditions play a key role in deriving the dual problem from the primal problem.

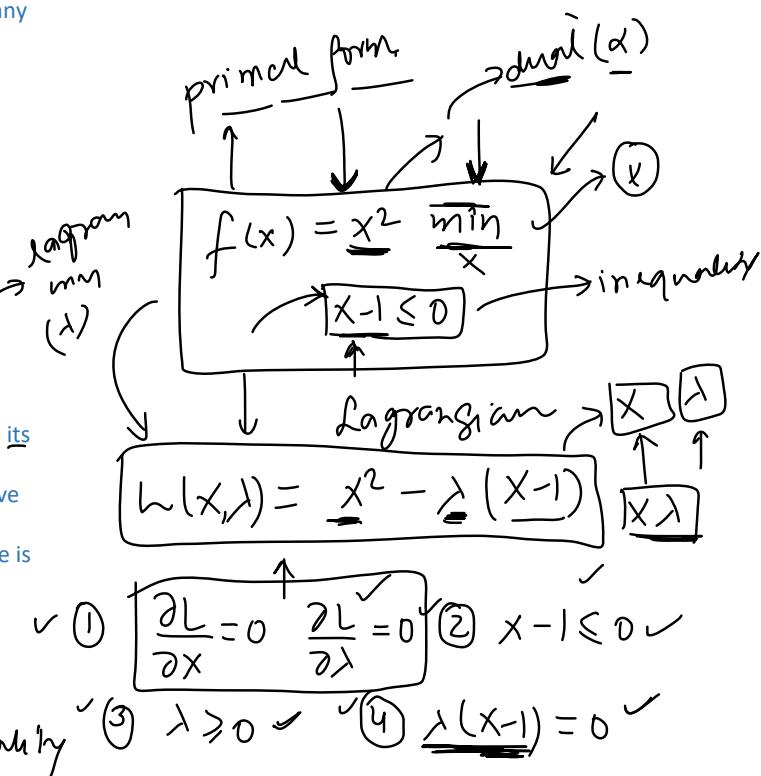
The KKT conditions are:

- Stationarity**: The derivative of the Lagrangian with respect to the primal variables, the dual variables associated with inequality constraints, and the dual variables associated with equality constraints are all zero.
- Primal feasibility**: All the primal constraints are satisfied.
- Dual feasibility**: All the dual variables associated with inequality constraints are nonnegative.
- Complementary slackness**: The product of each dual variable and its associated inequality constraint is zero. This means that at the optimal solution, for each constraint, either the constraint is active (equality holds) and the dual variable can be nonzero, or the constraint is inactive (strict inequality holds) and the dual variable is zero.

$$\min_x f(x) = x^2 \text{ such that } x - 1 \leq 0$$

Inequality

SVM dual problem



Example

16 July 2023 14:52

$$f(x, y) = \underline{x^2 + y^2} \quad \text{minimize}$$

subject to $\begin{cases} xy \\ x+y-1 \leq 0 \end{cases}$

$$\begin{aligned} \frac{\partial L}{\partial \lambda} &= \underline{x^2 + y^2 - \lambda x - \lambda y + \lambda} \\ &= \cdot - x - y = 0 \\ &\quad - x = y \quad (0 \leq 0) \\ &0.5 + 0.5 - 1 \end{aligned}$$

Lagrangian

$$L(x, y, \lambda) = \underline{x^2 + y^2 - \lambda(x+y-1)}$$

$$\frac{\partial L}{\partial x} = 0 \quad \frac{\partial L}{\partial y} = 0 \quad \frac{\partial L}{\partial \lambda} = 0$$

② $\underline{x+y-1 \leq 0} \checkmark$

③ $\underline{\lambda \geq 0} \checkmark$

④ $\underline{\lambda(x+y-1) = 0} \checkmark$

$$\begin{array}{lcl} 2x - \lambda = 0 & 2y - 1 = 0 & -x - y + 1 = 0 \\ \lambda = 2x & \lambda = 2y & x - y \\ x = \frac{\lambda}{2} & y = \frac{\lambda}{2} & \downarrow 0.5 \\ \hookrightarrow 0.5 & \downarrow 0.5 & \end{array}$$

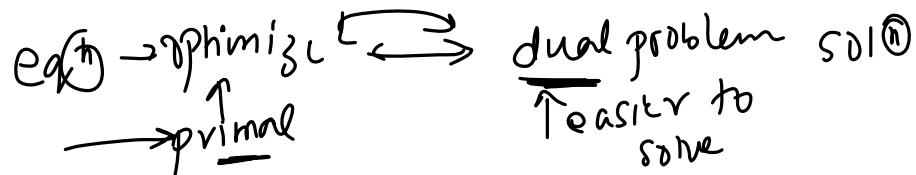
$$\frac{\lambda}{2} + \frac{\lambda}{2} = 1 \quad \boxed{\lambda = 1}$$

$$\boxed{x = 0.5 \quad y = 0.5 \quad \lambda = 1} \rightarrow \text{Lagrange}$$

$$\rightarrow \boxed{x = 0.5 \quad y = 0.5}$$

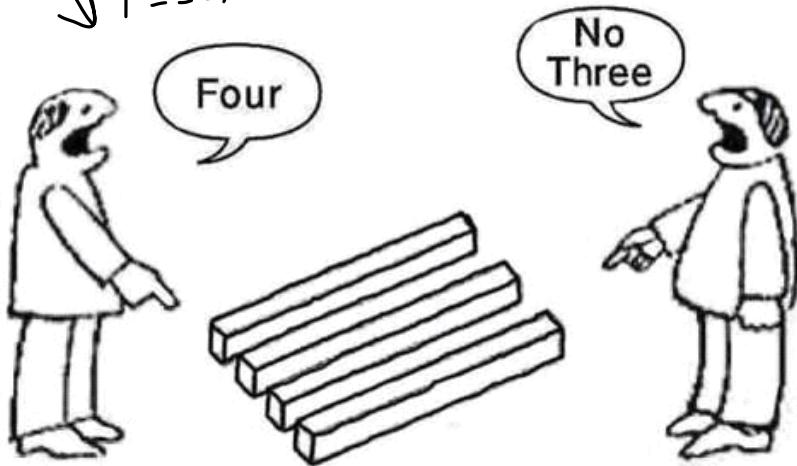
Concept of Duality

16 July 2023 08:46



The duality principle is fundamental in optimization theory. It provides a powerful tool for solving difficult or complex optimization problems by transforming them into simpler or easier-to-solve problems. The solution to the dual problem provides a lower bound on the solution of the primal problem. If strong duality holds (i.e., the optimal values of the primal and dual problems are equal), then solving the dual problem can directly give the solution to the primal problem.

perspective



The primal problem is the original optimization problem that you are trying to solve. It involves finding the minimum or maximum of a particular objective function, subject to certain constraints.

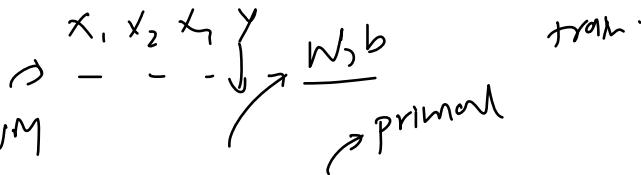
The dual problem is a related optimization problem that is derived from the primal problem. It provides a lower or upper bound on the solution to the primal problem.

\downarrow
SVM (w, b)
 $\overline{\text{derivative}} \rightarrow \text{primal} \rightarrow \text{dual}$

SVM Dual Problem

15 July 2023 23:58

Hard margin SVM



The primal form of hard margin SVM is given by:

$$\begin{aligned} & \underset{\mathbf{w}, b}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, n \end{aligned}$$

$\mathbf{w}, b \rightarrow \text{primal variable}$

$$\begin{aligned} & \underset{\mathbf{a}}{\text{maximize}} \quad \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j a_i a_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ & \text{subject to} \quad \begin{cases} a_i \geq 0, & i = 1, \dots, n \\ \sum_{i=1}^n a_i y_i = 0 \end{cases} \end{aligned}$$

$\mathbf{a} \rightarrow \text{dual variable}$

$\mathbf{w}, b \rightarrow \text{dual form hard margin}$

Soft margin SVM

Hard margin SVM
Soft margin SVM

The primal form of soft margin SVM is given by:

$$\begin{aligned} & \underset{\mathbf{w}, b, \xi}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ & \text{subject to} \quad \begin{cases} y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, & i = 1, \dots, n \\ \xi_i \geq 0, & i = 1, \dots, n \end{cases} \end{aligned}$$

$$\begin{aligned} & \underset{\mathbf{a}}{\text{maximize}} \quad \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j a_i a_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ & \text{subject to} \quad \begin{cases} 0 \leq a_i \leq C, & i = 1, \dots, n \\ \sum_{i=1}^n a_i y_i = 0 \end{cases} \end{aligned}$$

$a \rightarrow \text{dual}$

Dual Problem Derivation

16 July 2023 01:14

Primal form

$$\underset{w, b}{\operatorname{argmin}} \frac{\|w\|^2}{2}$$

$$\begin{array}{c} n \\ \text{rows} \\ \downarrow \end{array} \quad \begin{array}{c} x_1 & x_2 & x_3 & \cdots & x_m | y \\ \hline \hline \end{array}$$

such that

$$y_i(w^T x_i + b) \geq 1 \quad \forall i$$

for every row
by constraint

$$L(w, b, \alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^n \alpha_i [y_i(w^T x_i + b) - 1] - \dots$$



$$L(w, b, \alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^n \alpha_i [y_i(w^T x_i + b) - 1]$$

$$\frac{\partial L}{\partial w} = 0 \quad \frac{\partial L}{\partial b} = 0$$

$$= \frac{\|w\|^2}{2} - \sum_{i=1}^n \alpha_i y_i w^T x_i + \alpha_i b - \alpha_i$$

$$L(w, b, \alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^n \alpha_i y_i w^T x_i - \underbrace{\sum_{i=1}^n \alpha_i y_i b}_{\uparrow} + \sum_{i=1}^n \alpha_i$$

$$\frac{\partial L}{\partial w} = \frac{\|w\|^2}{2} - \sum_{i=1}^n \alpha_i y_i x_i = 0$$

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad \rightarrow \textcircled{1}$$

$$\frac{\partial L}{\partial b} = - \sum_{i=1}^n \alpha_i y_i = 0$$

$$\sum_{i=1}^n \alpha_i y_i = 0$$

$$\begin{aligned}
 L(w, b, \alpha_i) &= \frac{\|w\|^2}{2} - \underbrace{\sum_{i=1}^n \alpha_i y_i w \cdot x_i}_{\text{---}} - \underbrace{\frac{\sum_{i=1}^n \alpha_i y_i b}{n}}_{\text{---}} + \underbrace{\sum_{i=1}^n \alpha_i}_{\text{---}} \\
 &= \frac{1}{2} \left(\sum_{i=1}^n \alpha_i y_i x_i \right) \left(\sum_{j=1}^n \alpha_j y_j x_j \right) - \left(\sum_{i=1}^n \alpha_i y_i x_i \right) \left(\sum_{j=1}^n \alpha_j y_j x_j \right) + \sum_{i=1}^n \alpha_i \\
 &\quad - \frac{1}{2} \sum_{i=1}^n \alpha_i y_i x_i \sum_{j=1}^n \alpha_j y_j x_j + \sum_{i=1}^n \alpha_i \\
 &\quad \downarrow (\alpha_i) \quad \downarrow w, b \quad \downarrow x \\
 &\boxed{\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)} \rightarrow \text{dual form}
 \end{aligned}$$

$$\begin{array}{l}
 \boxed{\alpha_i \geq 0} \\
 \sum_{i=1}^n \alpha_i y_i = 0
 \end{array}
 \quad
 \begin{array}{l}
 \downarrow \\
 \alpha_i \\
 \uparrow w, b
 \end{array}
 \quad
 \begin{array}{l}
 \text{if } y_i = 1 \\
 \text{if } y_i = -1
 \end{array}$$

$$\boxed{\text{maximize}_{\alpha_i}} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j x_i \alpha_j (x_i \cdot x_j) \quad \xrightarrow{\text{dot product}}$$

$$\begin{array}{l}
 \min \rightarrow \max \quad \alpha_i \rightarrow \# \text{ rows} \quad \text{Lagrange} \\
 \downarrow \\
 \left\{ \begin{array}{l} \min \rightarrow \max \\ \max \rightarrow \min \end{array} \right\}
 \end{array}$$

Observations

15 July 2023 23:59

- 1) easy to solve
- 2) kernel friendly

The primal form of hard margin SVM is given by:

$$\begin{aligned} & \text{minimize}_{\mathbf{w}, b} \quad \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \quad i = 1, \dots, n \end{aligned}$$

- ★ 1. $\alpha_i > 0$ only for support vectors → the equation is not as dangerous as it seems
- ★ 2. Dot product

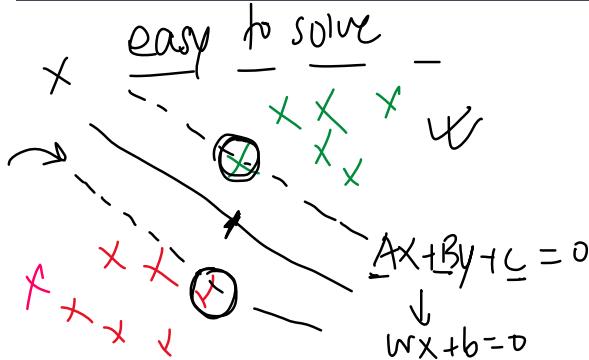
$$\underline{\alpha_i \alpha_j}$$

x y
↑
training data

$$\begin{aligned} & \text{maximize}_{\mathbf{a}} \quad \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j a_i a_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ & \text{subject to} \quad \begin{cases} a_i \geq 0, & i = 1, \dots, n \\ \sum_{i=1}^n a_i y_i = 0 \end{cases} \end{aligned}$$

← kernel trick

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$



$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

W → support vectors

$$\begin{cases} \alpha > 0 \\ \alpha = 0 \\ \alpha < 0 \end{cases}$$

for support vector
 $\alpha > 0$
 Non support vector
 $\alpha = 0$

$$\begin{aligned} & \text{maximize}_{\mathbf{a}} \quad \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j a_i a_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ & \text{subject to} \quad \begin{cases} a_i \geq 0, & i = 1, \dots, n \\ \sum_{i=1}^n a_i y_i = 0 \end{cases} \end{aligned}$$

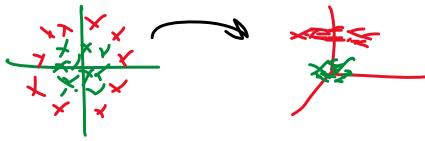
multiply
 2 support
 vectors

$$\begin{aligned} \text{row 1} & \rightarrow [x_1 \ y_1 \ \alpha_1] \\ \text{row 2} & \rightarrow [x_2 \ y_2 \ \alpha_2] \\ \text{row n} & \rightarrow [x_n \ y_n \ \alpha_n] \end{aligned}$$

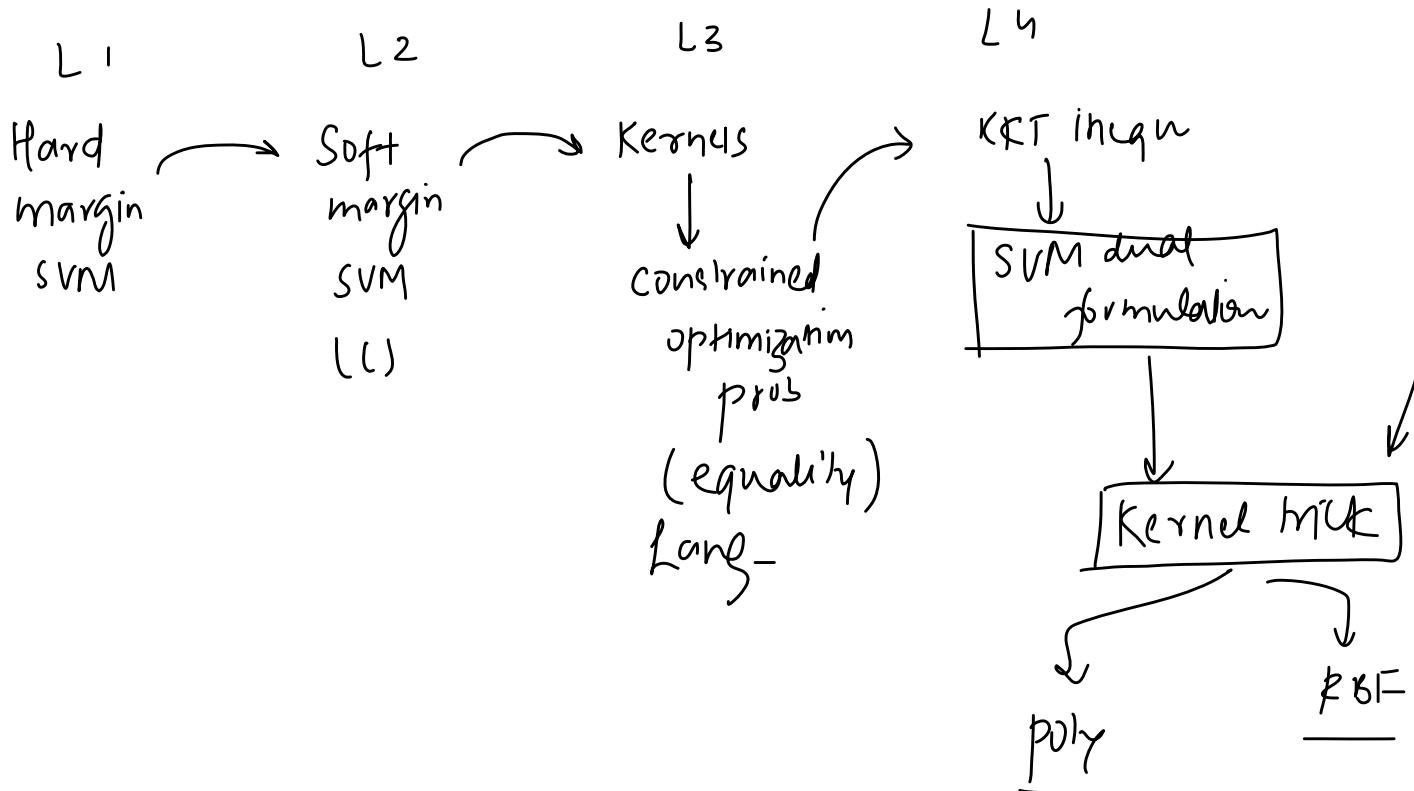
$$\begin{aligned} & \downarrow \text{kernel} \\ & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \rightarrow 5, 10, 15 \\ & \downarrow \text{next class} \end{aligned}$$

Recap

19 July 2023 17:04



SVM



SVM Dual Formulation

19 July 2023 19:03

$$\max_{\alpha_i} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$x_i \cdot x_j$



$$x_i \cdot x_j$$

$$x_1 \cdot x_1 \rightarrow x_{11} x_{11} + x_{12} x_{12}$$

(25) terms
 $\alpha_i > 0$

$$\sum_{i=1}^n \alpha_i y_i = 0$$

$$SV$$

$$x_{11} x_{21} + x_{12} x_{22}$$

$$x_1 \cdot x_1$$

$$y_i y_j = 1 \times 1$$

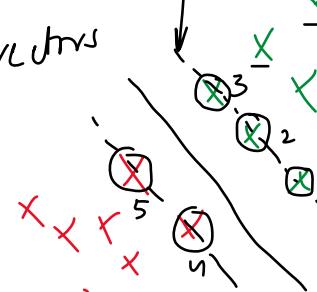
$$x_j \rightarrow \alpha_j$$

$$\alpha_i > 0$$

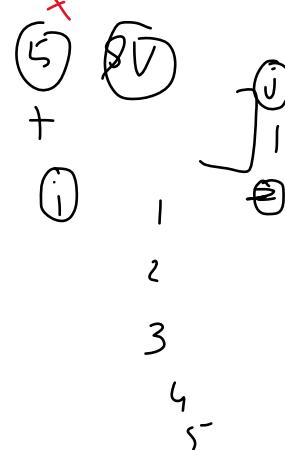
$$\alpha_i = 0$$

$$\left\{ \begin{array}{l} \alpha_i = 0 \text{ for all non support vectors} \\ \alpha_i > 0 \text{ for all SV} \end{array} \right.$$

$$x_1 \cdot x_2$$



$$\max_{\alpha_i} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 - \frac{1}{2} \left[+ - + \right]$$



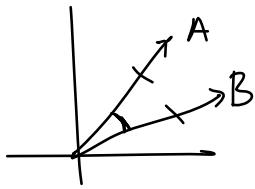
The Similarity Perspective

18 July 2023 08:52

$$\text{cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

$$\|\mathbf{A}\|=1$$

$$\|\mathbf{B}\|=1$$



Kernel trick

\rightarrow $\boxed{\mathbf{A} \cdot \mathbf{B}}$ \rightarrow dot product of $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$

$\underline{x_i}$ and $\underline{x_j}$

$$\max_{\alpha_i} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\underline{x_i} \cdot \underline{x_j})$$

similarity

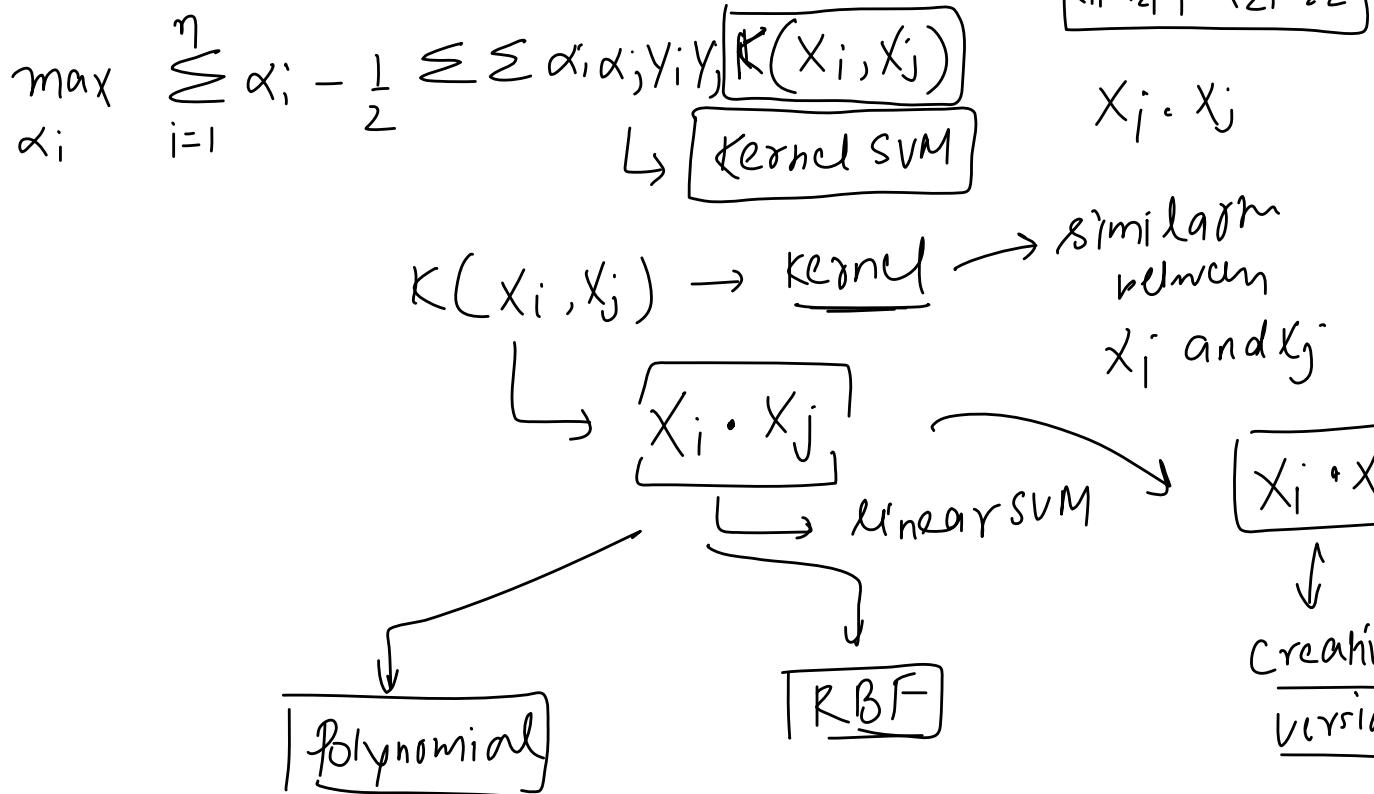
\downarrow \downarrow maximizing the similarity of $\underline{s_v}$ based on margin

$$(\underline{x_i} \cdot \underline{x_j}) \rightarrow \boxed{\text{sim}(\underline{x_i}, \underline{x_j})}$$

\leftarrow Kernel

Kernel SVM

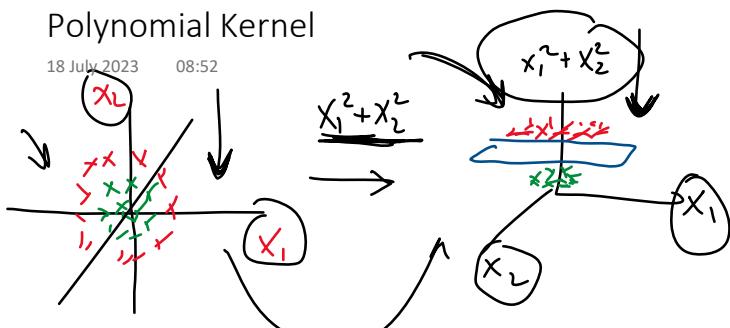
19 July 2023 07:41



Polynomial Kernel

18 July 2023

08:52



$$x_i \rightarrow \frac{x_{i1}}{x_{i2}} \quad | \quad x_2 \\ x_j \rightarrow \frac{x_{j1}}{x_{j2}}$$

$$x_{11}x_{21} + x_{12}x_{22}$$

$$K(x_i, x_j) = \frac{(r + x_i \cdot x_j)^d}{(1 + x_i \cdot x_j)^k}$$

$r=1, d=2$

$$= (1 + x_i \cdot x_j)^2 = (1 + x_{11}x_{21} + x_{12}x_{22})^2$$

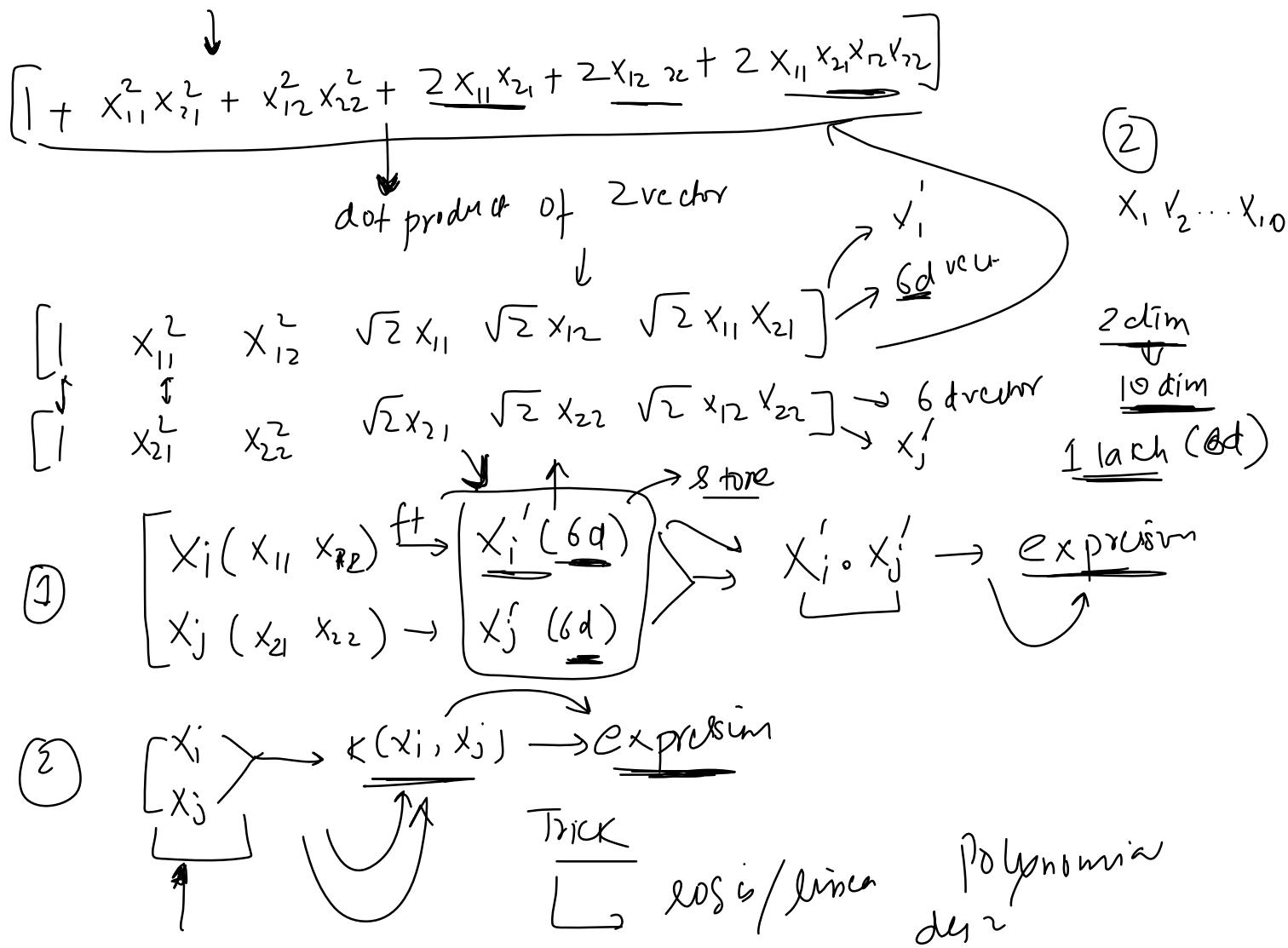
$$= 1 + x_{11}^2x_{21}^2 + x_{12}^2x_{22}^2 + 2\underline{x_{11}x_{21}} + 2\underline{x_{12}x_{22}} + 2\underline{\underline{x_{11}x_{21}x_{12}x_{22}}}$$

Kernel trick

↳ polynomial term

The Trick

19 July 2023 07:30



$$x_1 \quad x_2$$

$$x_1 \left| \begin{array}{c|c} x_1^2 & x_2 \end{array} \right| x_2^2 \left| \begin{array}{c} x_1 v_2 \end{array} \right|$$

What about the other Polynomial terms

19 July 2023 10:15

$$1 + \underbrace{x_{11}^2 x_{21}^2}_{\text{circular}} + \underbrace{x_{12}^2 x_{22}^2}_{\text{circular}} + \underbrace{2 \overline{x_{11} x_{21}} + 2 \overline{x_{12} x_{22}} + 2 \overline{x_{11} x_{21} x_{12} x_{22}}}_{\text{other shapes}}$$

non sech

(3)

RBF Kernel

18 July 2023 08:52

Radial basis function { Normal dist)

→ popular

→ best out of the box kernel

→ powerful

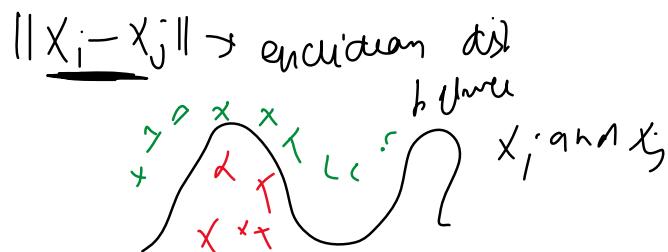
$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

polynomial

x x
x x
xx

$$e^{-\gamma \|x_i - x_j\|^2}$$

hyperparameter



$$K(x_i, x_j) = e^{-\frac{\text{dist}^2}{2\sigma^2}}$$

neural networks

$$\gamma = \frac{1}{2\sigma^2}$$

$K \propto \frac{1}{\text{dist}}$

- Non-linear Transformations: The RBF kernel enables the use of non-linear transformations, which can map the original feature space to a higher-dimensional space where the data becomes linearly separable. This is particularly useful for problems where the decision boundary is not linear.

- Local Decisions: Unlike some other kernels, the RBF kernel makes "local" decisions. That is, the effect of each data point is limited to a certain region around that point. This can make the model more robust to outliers and create complex decision boundaries.

- Flexibility: The RBF kernel has a parameter γ (related to the standard deviation of the Gaussian distribution) that determines the complexity of the decision boundary. By tuning this parameter, we can adjust the trade-off between bias and variance, allowing for a flexible range of decision boundaries.

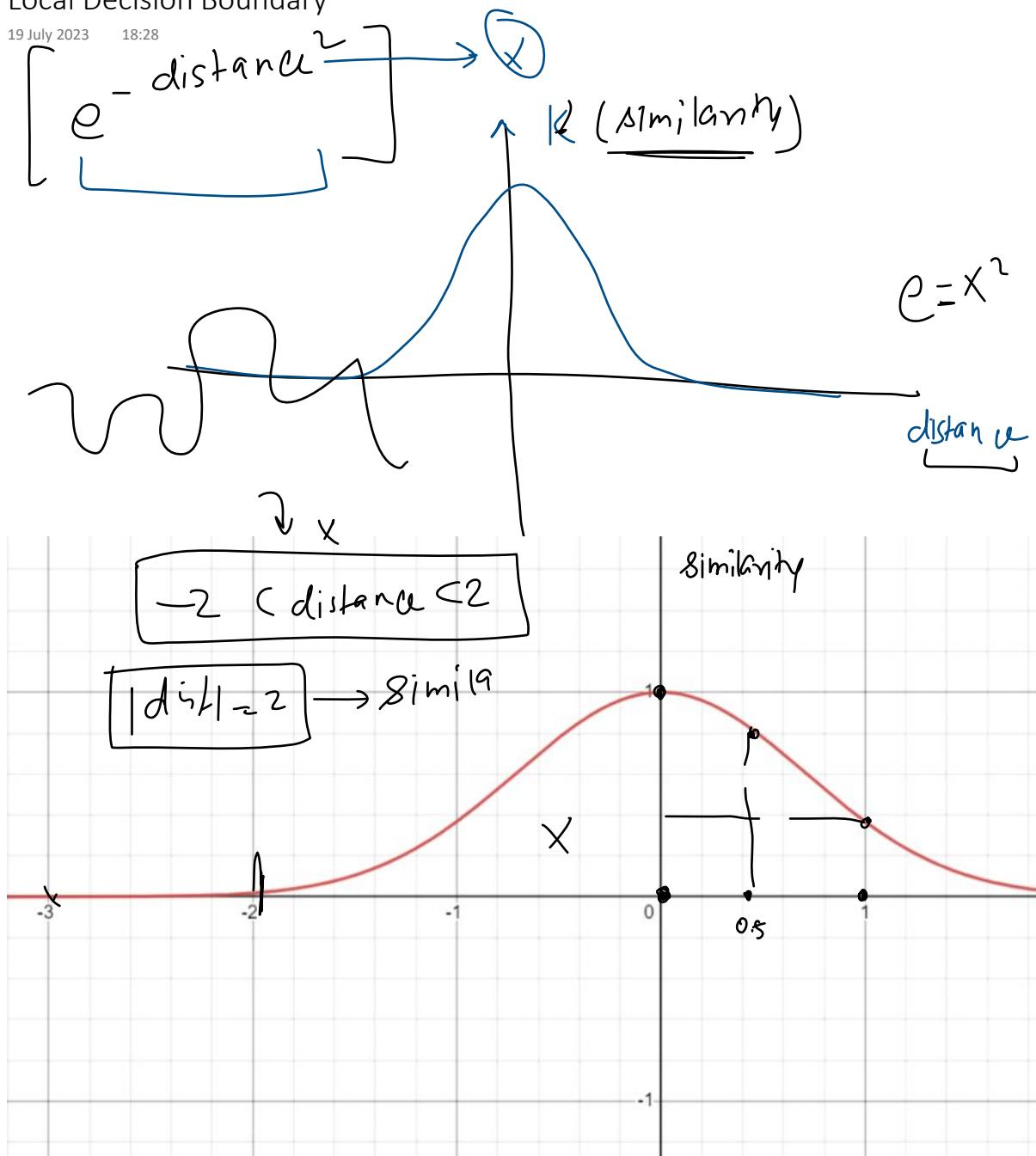
- Universal Approximation Property: The RBF kernel has a property known as the "universal approximation" property, meaning it can approximate any continuous function to a certain degree of accuracy given enough data points. This makes it highly versatile and capable of modelling a wide variety of relationships in data.

- General-Purpose: The RBF kernel does not make any strong assumptions about the data and can therefore be a good choice in many different situations, making it a versatile, general-purpose kernel.

distance

Local Decision Boundary

19 July 2023 18:28



Effect of Gamma

19 July 2023 11:11



$$e^{-\frac{\|x_i - y_j\|^2}{2\sigma^2}}$$

$$\frac{1}{2\sigma^2} = \gamma$$

$$\sigma^2 = 1$$

$$\sigma = 10 \quad \sigma^2 = 100$$

$$\sigma = 0.1 \quad \sigma^2 = 0.01$$

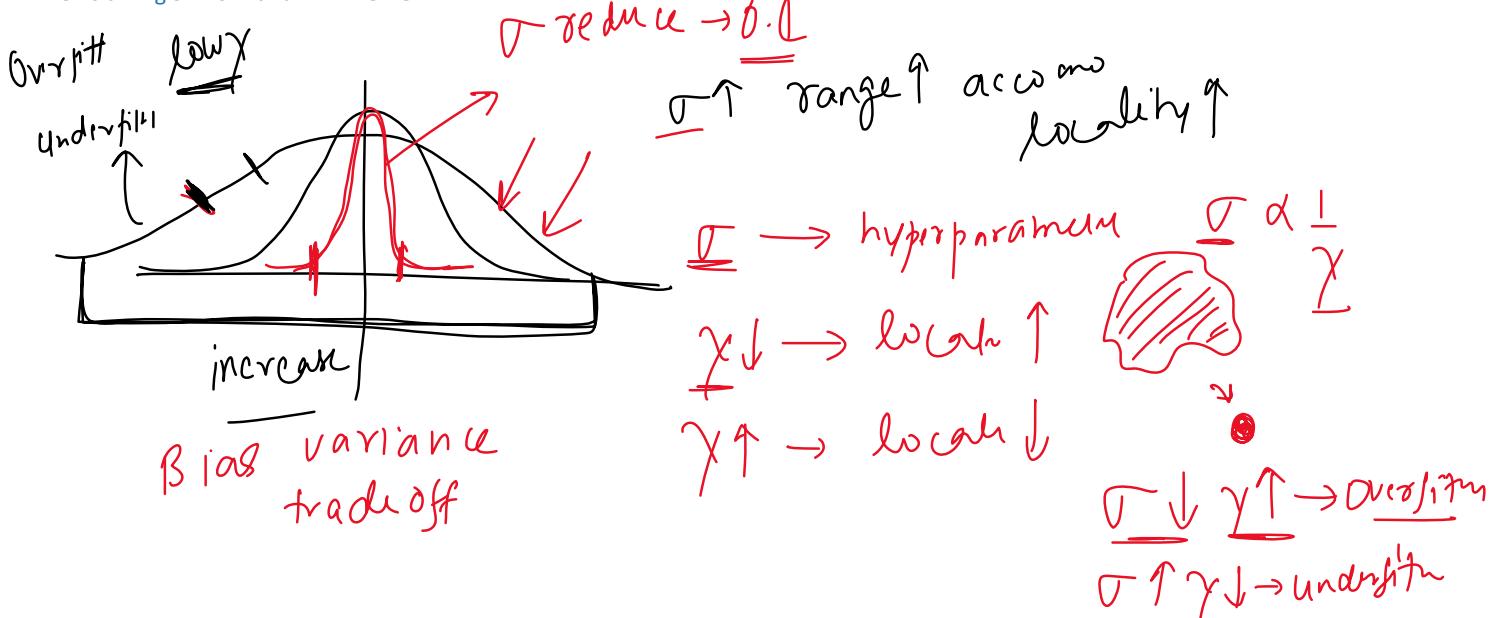
The parameter γ in the Radial Basis Function (RBF) kernel of a Support Vector Machine (SVM) is a hyperparameter that determines the spread of the kernel and therefore the decision region.

The effect of γ can be summarized as follows:

- If γ is too large, the exponential will decay very quickly, which means that each data point will only have an influence in its immediate vicinity. The result is a more complex decision boundary, which might overfit the training data.
- If γ is too small, the exponential will decay slowly, which means that each data point will have a wide range of influence. The decision boundary will therefore be smoother and more simplistic, which might underfit the training data.

In a sense, γ in the RBF kernel plays a role similar to that of the inverse of the regularization parameter: it controls the trade-off between bias (underfitting) and variance (overfitting). High γ values can lead to high variance (overfitting) due to more flexibility in shaping the decision boundary, while low γ values can lead to high bias (underfitting) due to a more rigid, simplistic decision boundary.

Tuning the γ parameter using cross-validation or a similar technique is typically a crucial step when training SVMs with an RBF kernel.



Relationship Between RBF and Polynomial Kernel

19 July 2023 14:36

Infinite Dimensional Mapping: The RBF kernel implicitly maps input data to an infinite-dimensional feature space, which allows for even greater flexibility in forming decision boundaries

Input $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ $\xrightarrow{\text{deg} \rightarrow 2}$ $x_1 \ x_2 \ x_1^2 \ x_2^2 \ x_1 x_2$

Input $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ $\xrightarrow{\text{dim} \rightarrow \infty}$ $\begin{bmatrix} x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \\ x_1^3 & x_2^3 & x_1^2 x_2 & x_2^2 x_1 & x_1^3 x_2 \\ x_1^4 & x_2^4 & x_1^3 x_2^1 & x_2^3 x_1^1 & x_1^4 x_2^2 \end{bmatrix} \xrightarrow{d=2}$

Input $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ $\xrightarrow{\text{dim} \rightarrow \infty}$ $\begin{bmatrix} x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \\ x_1^3 & x_2^3 & x_1^2 x_2 & x_2^2 x_1 & x_1^3 x_2 \\ x_1^4 & x_2^4 & x_1^3 x_2^1 & x_2^3 x_1^1 & x_1^4 x_2^2 \end{bmatrix} \xrightarrow{d=3}$

Input $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ $\xrightarrow{\text{dim} \rightarrow \infty}$ $\begin{bmatrix} x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \\ x_1^3 & x_2^3 & x_1^2 x_2 & x_2^2 x_1 & x_1^3 x_2 \\ x_1^4 & x_2^4 & x_1^3 x_2^1 & x_2^3 x_1^1 & x_1^4 x_2^2 \end{bmatrix} \xrightarrow{d=4}$

Input $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ $\xrightarrow{\text{dim} \rightarrow \infty}$ $\begin{bmatrix} x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \\ x_1^3 & x_2^3 & x_1^2 x_2 & x_2^2 x_1 & x_1^3 x_2 \\ x_1^4 & x_2^4 & x_1^3 x_2^1 & x_2^3 x_1^1 & x_1^4 x_2^2 \end{bmatrix} \xrightarrow{d=5}$

Input $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ $\xrightarrow{\text{dim} \rightarrow \infty}$ $\begin{bmatrix} x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \\ x_1^3 & x_2^3 & x_1^2 x_2 & x_2^2 x_1 & x_1^3 x_2 \\ x_1^4 & x_2^4 & x_1^3 x_2^1 & x_2^3 x_1^1 & x_1^4 x_2^2 \end{bmatrix} \xrightarrow{d=6}$

Input $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$ $\xrightarrow{\text{dim} \rightarrow \infty}$ $\begin{bmatrix} x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \\ x_1^3 & x_2^3 & x_1^2 x_2 & x_2^2 x_1 & x_1^3 x_2 \\ x_1^4 & x_2^4 & x_1^3 x_2^1 & x_2^3 x_1^1 & x_1^4 x_2^2 \end{bmatrix} \xrightarrow{d=\infty}$

$\rightarrow e^{-\gamma \|x_i - x_j\|^2}$ $\xrightarrow{\text{rbf}} \text{poly}$

$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$ $\sigma = 1 \quad \sigma^2 = 1$

$= e^{-\frac{\|x_i - x_j\|^2}{2}}$

$= e^{-\frac{(x_i^\top - x_j^\top)(x_i - x_j)}{2}}$

$= e^{-\frac{(x_i - x_j)^\top (x_i - x_j)}{2}}$

$= e^{-\frac{x_{11}x_{21} + x_{12}x_{22}}{2}}$

$x_i = \begin{bmatrix} x_{11} & x_{12} \end{bmatrix}$

$x_j = \begin{bmatrix} x_{21} & x_{22} \end{bmatrix}$

$x_i^\top x_j =$

$x_j^\top x_i =$

$= e^{-\frac{x_{11}^\top x_i + x_{12}^\top x_j - x_{11}^\top x_j - x_{12}^\top x_i}{2}}$

$= e^{-\frac{x_{11}^\top x_i + x_{12}^\top x_j - 2x_{11}^\top x_j}{2}}$

$= e^{-\frac{1}{2} [x_{11}^\top x_i + x_{12}^\top x_j]} e^{x_{11}^\top x_j}$

$\sim \boxed{1 + x_i^\top x_j - 1}$

$$\begin{aligned}
 &= C e^{\overbrace{1+x_i^T x_j}^{\text{Term}}} \\
 &= C e^{\overbrace{1+x_i^T x_j}^{\text{Term}}} e^{-1} \\
 &= C' e^{\overbrace{1+x_i^T x_j}^{\text{Term}}} \rightarrow = C' \sum_{k=0}^{\infty} \frac{(1+x_i^T x_j)^k}{k!} \\
 C^X &= \sum_{k=0}^{\infty} \frac{x^k}{k!} \\
 C' &\sum_{k=0}^{\infty} \frac{k \text{poly}(x_i, x_j)^k}{k!} \\
 C' \left[1 + \frac{(1+x_i^T x_j)}{1!} + \frac{(1+x_i^T x_j)^2}{2!} + \frac{(1+x_i^T x_j)^3}{3!} \right] \\
 \partial_b f &= \sum_{i=0}^{\infty} \frac{k(x_i, x_j)}{k!} \dots + \frac{(1+x_i^T x_j)^0}{0!} \quad]
 \end{aligned}$$

Custom Kernels

19 July 2023 16:27

Custom kernels

1. **String kernels:** These are used for classifying text or sequences, where the input data is not numerical. String kernels measure the similarity between two strings. For example, a simple string kernel might count the number of common substrings between two strings.
2. **Chi-square kernel:** This kernel is often used in computer vision problems, especially for histogram comparison. It's defined as $K(x, y) = \exp(-\gamma\chi^2(x, y))$, where $\chi^2(x, y)$ is the chi-square distance between the histograms x and y .
3. **Intersection kernel:** This is another kernel commonly used in computer vision, which computes the intersection between two histograms (or generally non-negative feature vectors).
4. **Hellinger's kernel:** Hellinger's kernel, or Bhattacharyya kernel, is used for comparing probability distributions and is popular in image recognition tasks.
5. **Radial basis function network (RBFN) kernels:** These are similar to the standard RBF kernel, but the centers and widths of the RBFs are learned from the data, rather than being fixed a priori.
6. **Spectral kernels:** These kernels use spectral analysis techniques to compare data points. They can be particularly useful for dealing with cyclic or periodic data.

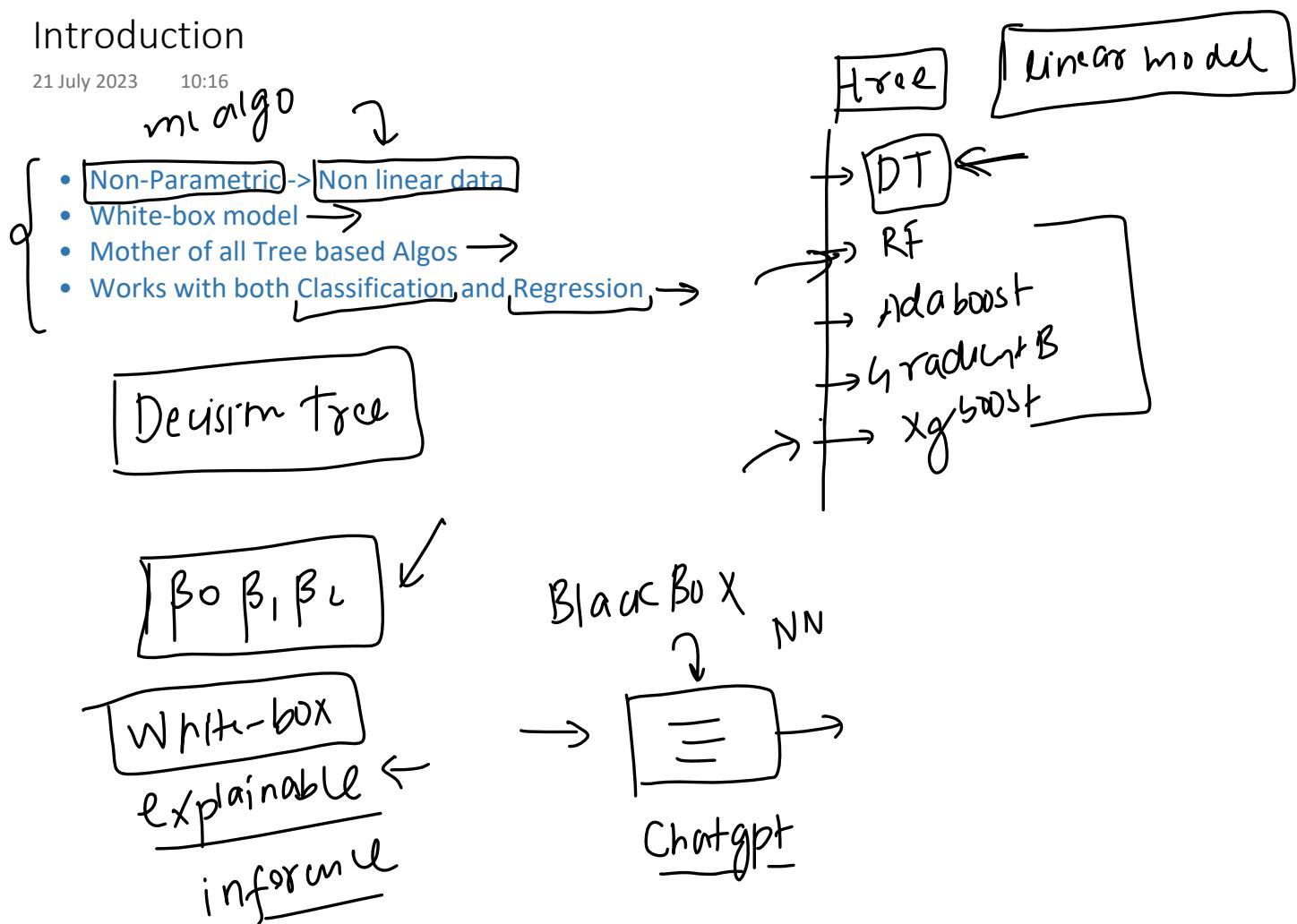
Doubts

20 July 2023 15:42

1. What is complementary Slackness
2. Why did min change to max?
3. Observations
 - a. Effect on high dimensional data
4. Prediction is done?

Introduction

21 July 2023 10:16



Intuition behind DT

20 July 2023 16:21

Gender	Occupation	Suggestion
F	<u>Student</u>	PUBG
<u>F</u>	Programmer	Github
M	Programmer	<u>Whatsapp</u>
<u>F</u>	Programmer	Github
M	<u>Student</u>	PUBG
M	Student	PUBG

$\{m, p\} \text{ of } F, \underline{\text{student}}$

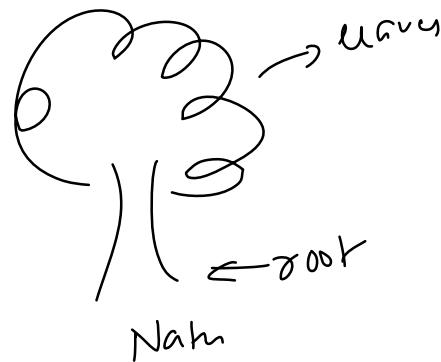
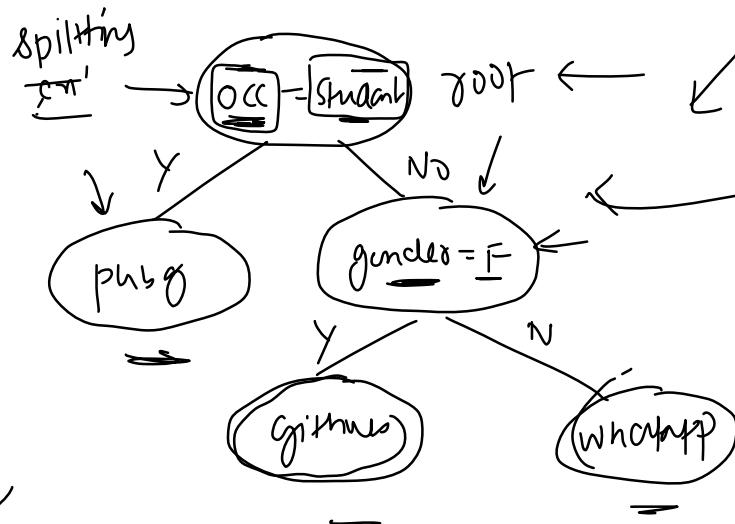
Nested
giant if-else structure

```
if occupation = student
    print pubg
```

```
else
    if gender = F
        print github
```

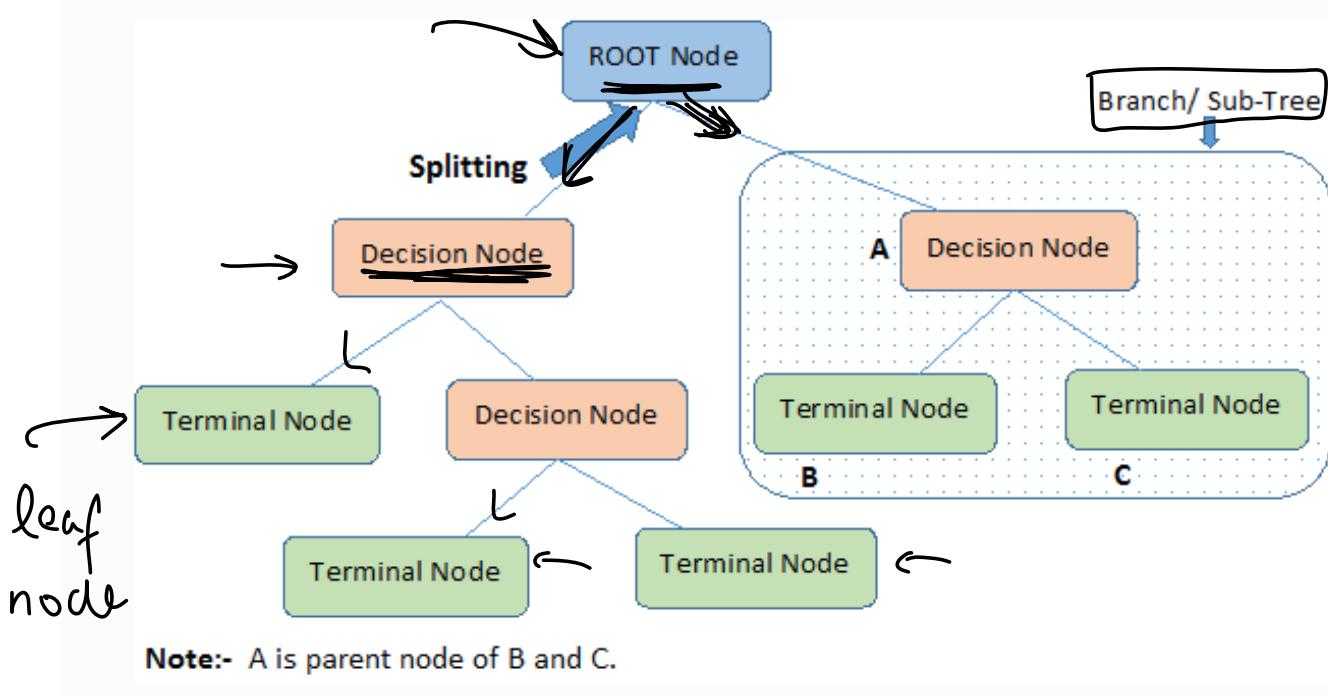
```
else
    print whatsapp
```

basic
dt



Vocab

20 July 2023 16:22



Example 2

20 July 2023 16:25

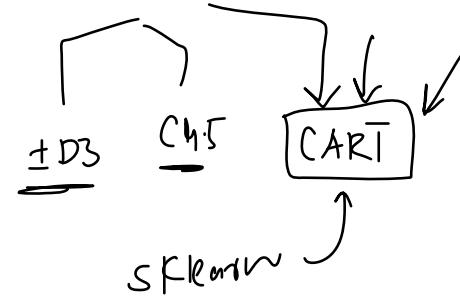
	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed
7	Postgraduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed
10	Undergraduate	Arts	91	Unemployed
11	Postgraduate	Arts	96	Employed
12	PhD	Science	87	Employed
13	Undergraduate	Science	90	Unemployed
14	Postgraduate	Science	95	Employed

Input Output

{ UG, Sunu, 90 }

→ 15 rows

↳ 1.5 min



skewness

The CART Algorithm - Classification

21 July 2023 10:26

sklearn

Given training vectors $x_i \in R^n$, $i=1, \dots, l$ and a label vector $y \in R^l$, a decision tree recursively partitions the feature space such that the samples with the same labels or similar target values are grouped together.

Let the data at node m be represented by Q_m with n_m samples. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m , partition the data into $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ subsets

$$Q_m^{left}(\theta) = \{(x, y) | x_j \leq t_m\}$$

$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta)$$

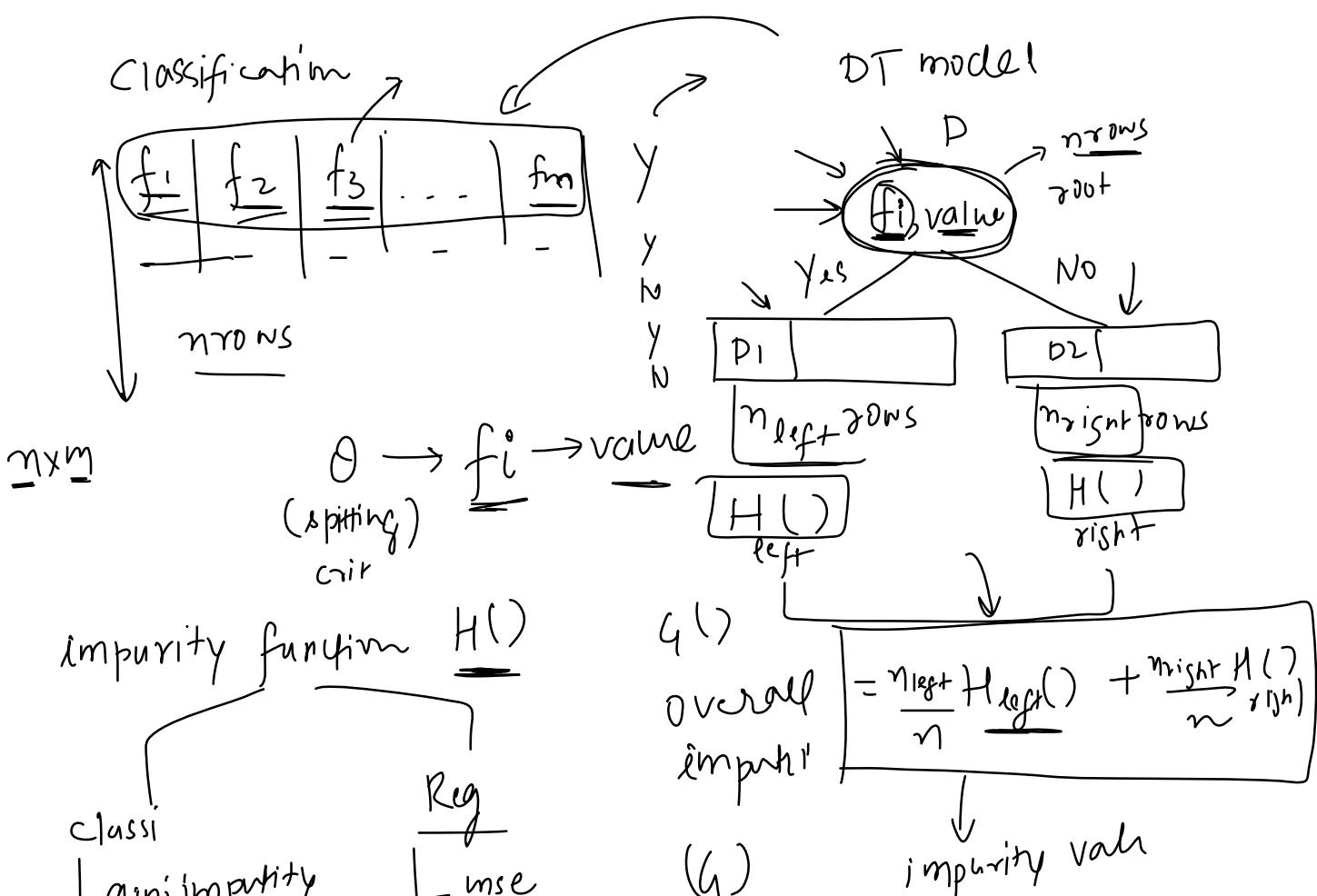
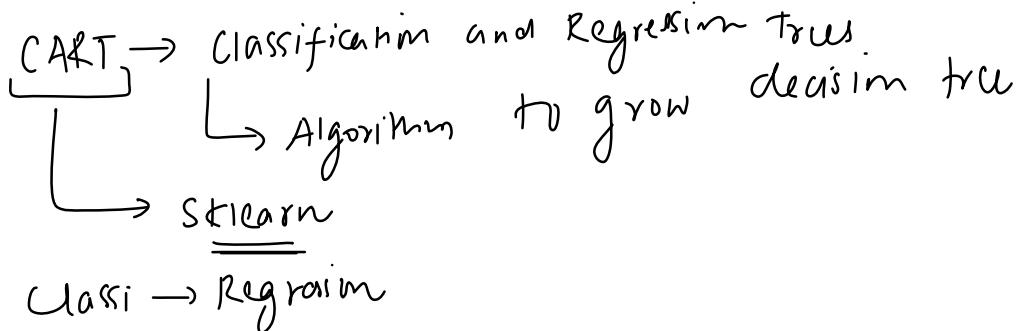
The quality of a candidate split of node m is then computed using an impurity function or loss function $H()$, the choice of which depends on the task being solved (classification or regression)

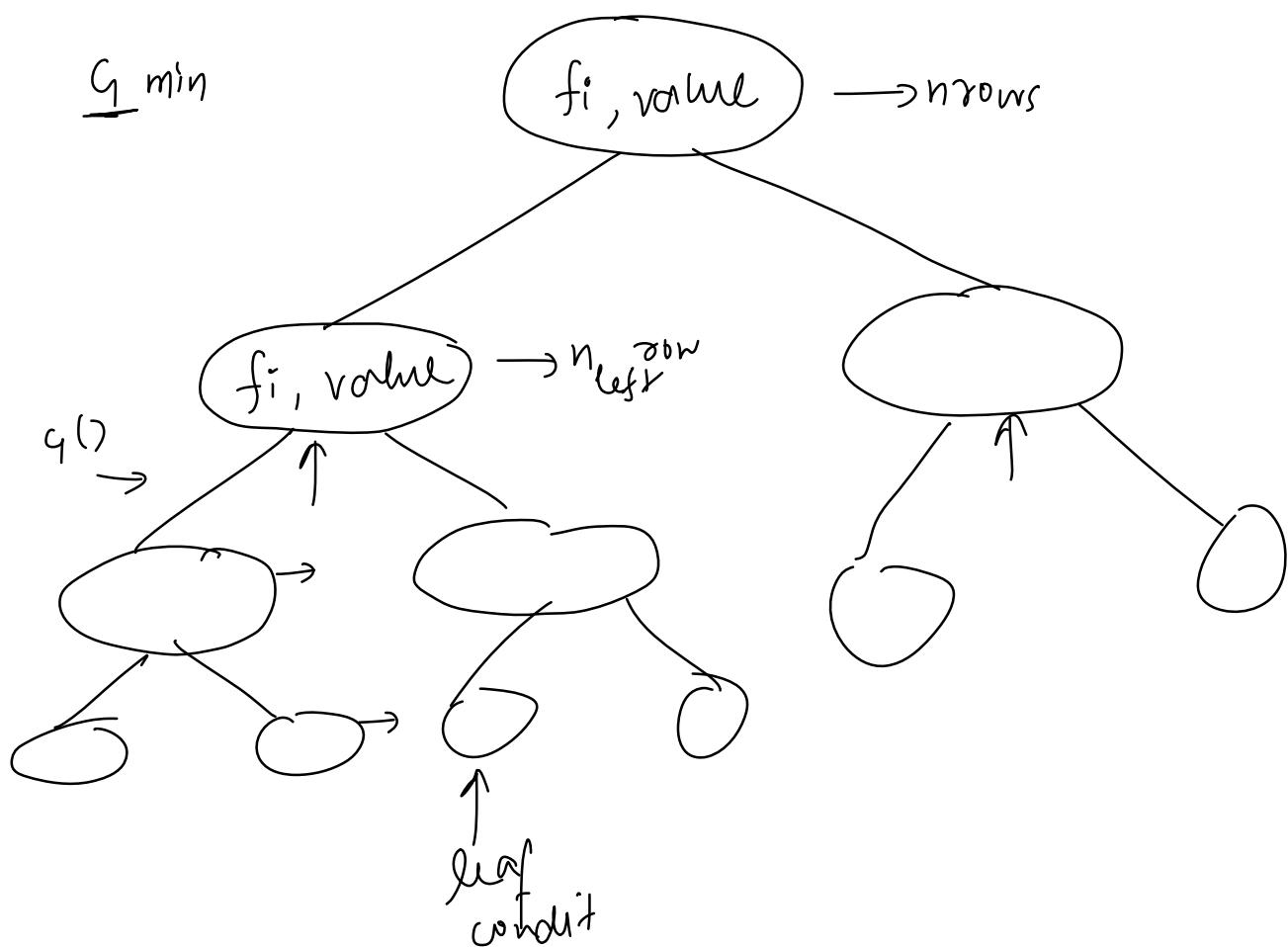
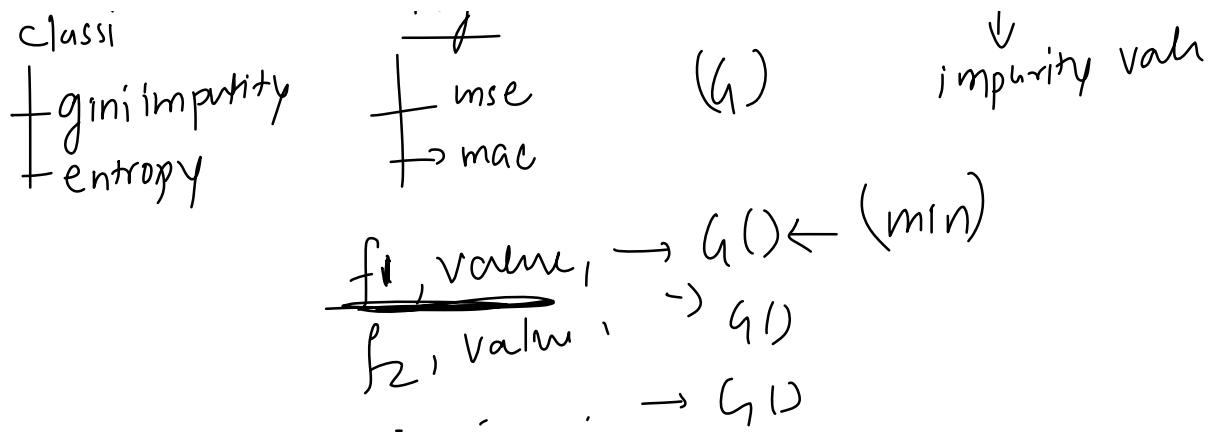
$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta))$$

Select the parameters that minimises the impurity

$$\theta^* = \operatorname{argmin}_{\theta} G(Q_m, \theta)$$

Recurse for subsets $Q_m^{left}(\theta^*)$ and $Q_m^{right}(\theta^*)$ until the maximum allowable depth is reached, $n_m < \min_{samples}$ or $n_m = 1$.





Splitting Categorical Features

21 July 2023

17:45

	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed
7	Postgraduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed

multi class

binary class

Numerical

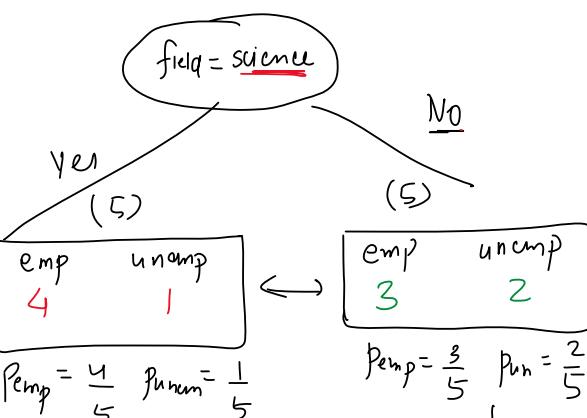
$$\text{gini impurity} = 1 - \sum_{k=1}^K P_k^2$$

$$P_{\text{Science}} = \left(\frac{5}{10}\right) \quad P_{\text{Arts}} = \left(\frac{5}{10}\right)$$

$$P_k = \text{prob of each class}$$

$$1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2$$

	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed
7	Postgraduate	Arts	93	Employed
8	Undergraduate	Arts	94	Unemployed
9	PhD	Science	86	Employed



$$\frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{12}{25}$$

$$0.16 + 0.20 \longrightarrow$$

$G()$ value

0.40

$G \rightarrow (\text{field, science})$

degree, ug
degree, phd
 $G() = 0.5$

$G()$

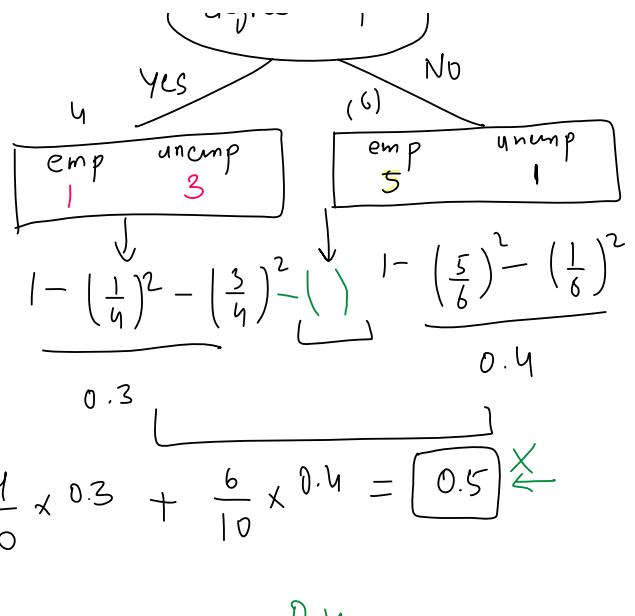
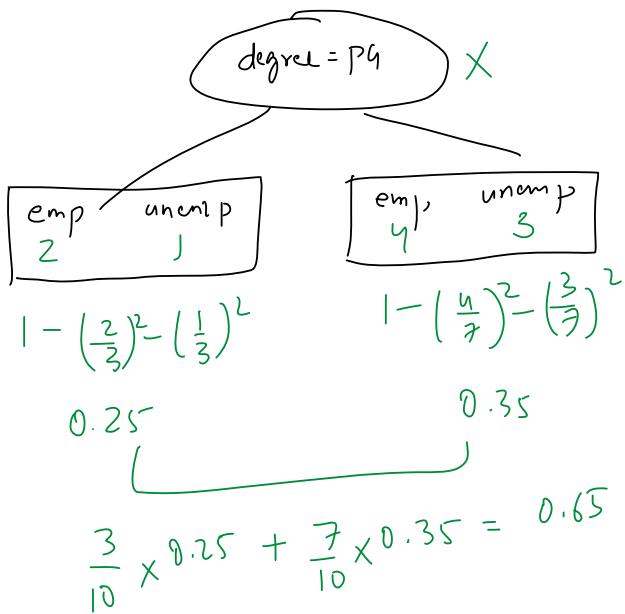
min

degree = ug

YES NO

	Degree_Type	Field	Average_Grade	Job_Outcome
0	Undergraduate	Science	89	Employed
1	Undergraduate	Arts	92	Unemployed
2	Postgraduate	Science	95	Employed
3	PhD	Science	85	Employed
4	Postgraduate	Arts	98	Unemployed
5	PhD	Arts	90	Employed
6	Undergraduate	Science	88	Unemployed

4	Postgraduate	-	Arts	98	Unemployed
5	PhD	-	Arts	90	Employed
6	Undergraduate	Science	88	-	Unemployed
7	Postgraduate	-	Arts	93	Employed
8	Undergraduate	Arts	94	-	Unemployed
9	PhD	Science	86	-	Employed

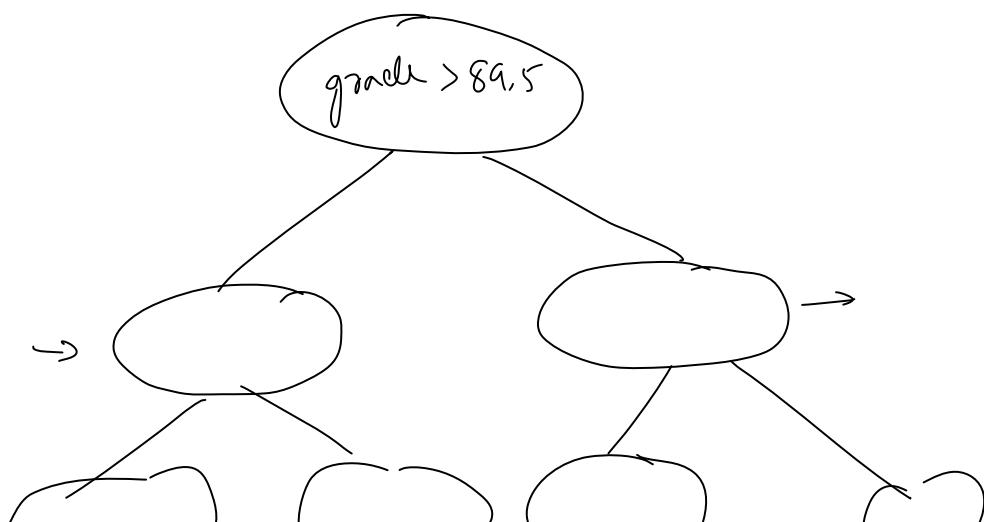
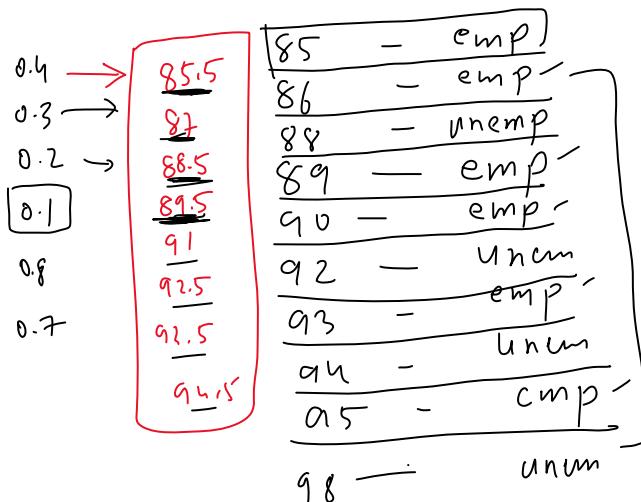


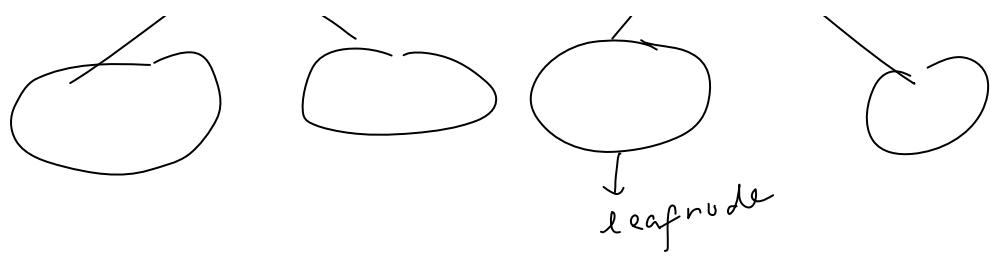
Splitting Numerical Features

21 July 2023 17:45

c

Average_Grade	Job_Outcome
89	Employed
92	Unemployed
95	Employed
85	Employed
98	Unemployed
90	Employed
88	Unemployed
93	Employed
94	Unemployed
86	Employed





Understanding Gini Impurity?

21 July 2023 16:40

The Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

$$H() = 1 - \sum_{i=1}^k p_i^2$$

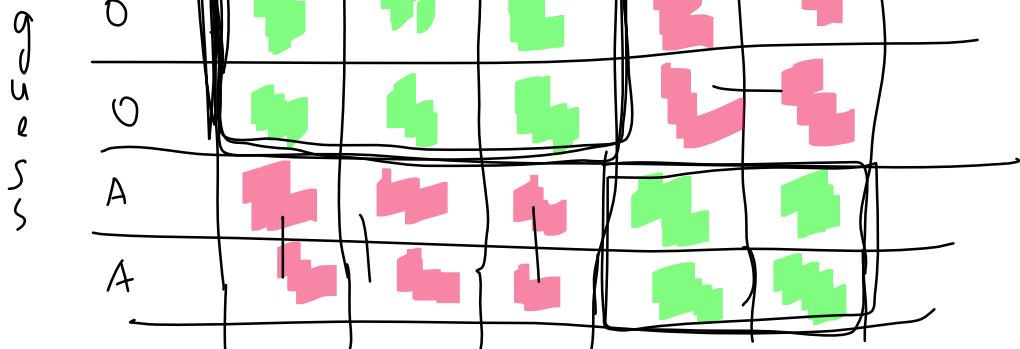
O_1, O_2, O_3
 A_1, A_2

O_1, O_2, O_3
 O_4, O_5

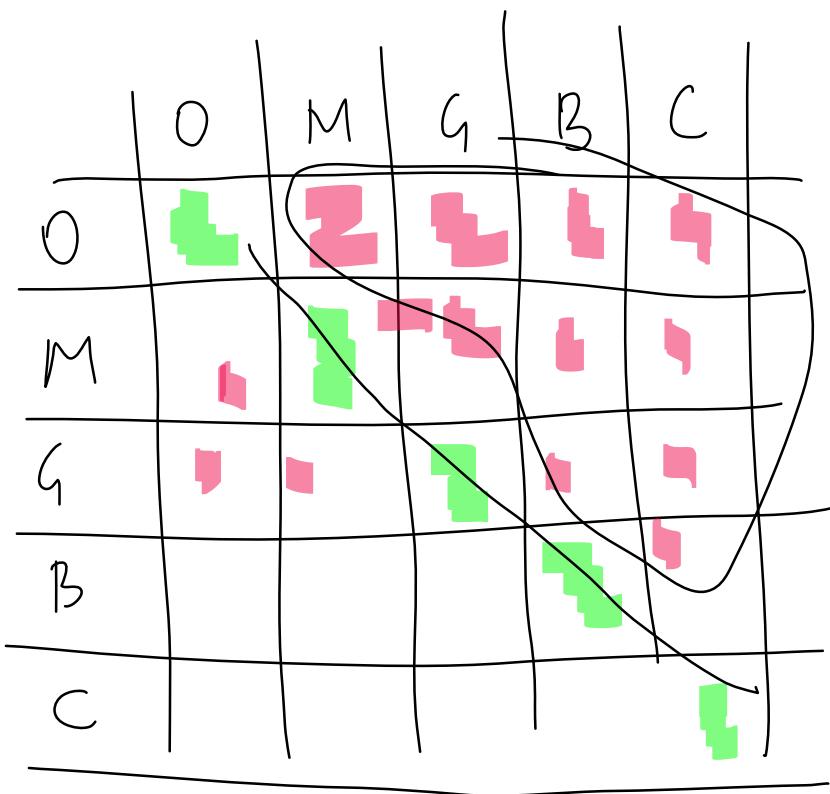
3 oranges 2 apples
 $\frac{1}{5}$ 0 $\frac{1}{5}$ 0 $\frac{1}{5}$ 0 actual label
 $\frac{1}{5}$ A $\frac{1}{5}$ A

$$\frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5}$$

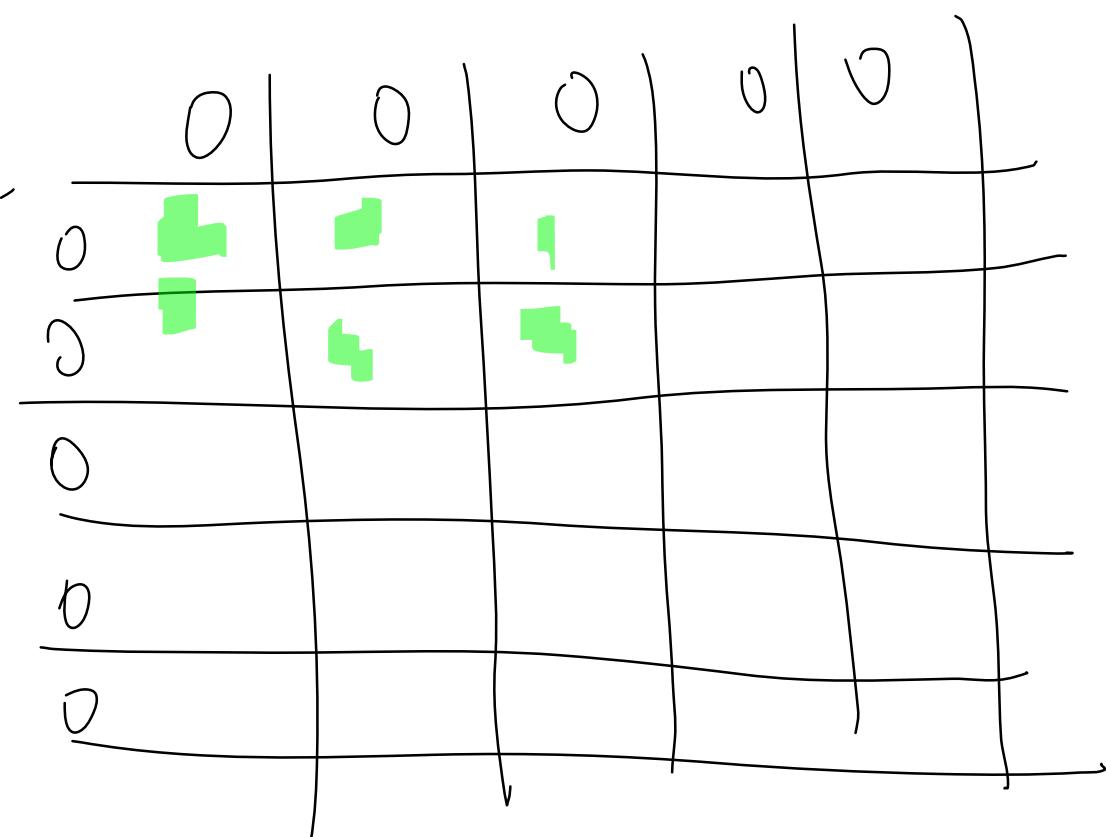
$$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$



= gini impur'n'hi



$$1 - \left(\frac{1}{5}\right)^5 = \left(\frac{4}{5}\right)^5 = \left(\frac{1}{5}\right)^5 - 1$$

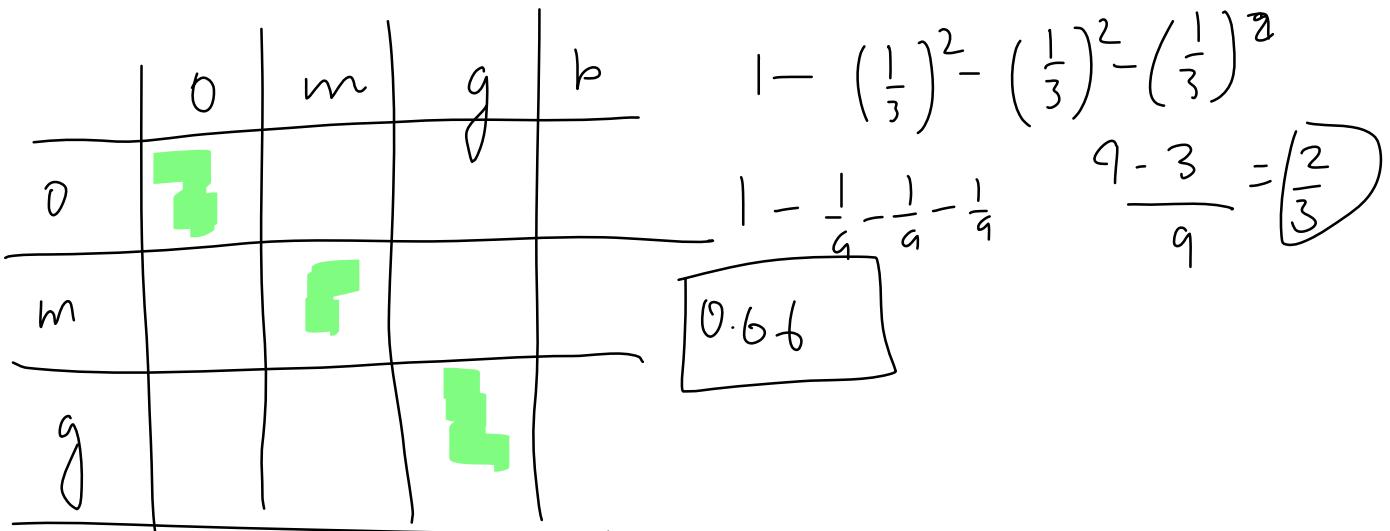


$$1 - \left(\frac{5}{5}\right)^2$$

$$1 - 1 = 0$$

$|I| > g_{in} \geq 0$

$O \rightarrow \underline{\text{preferred}}$



$$1 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 \quad (1)$$

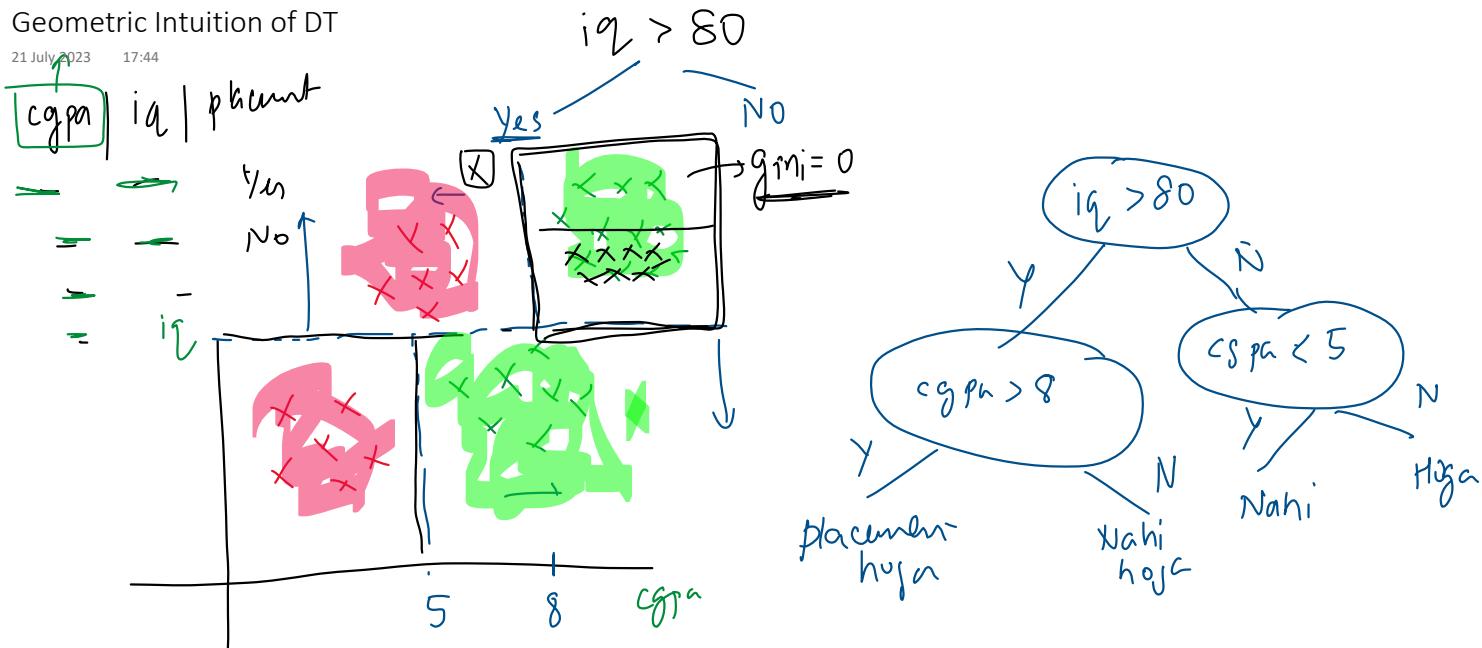
$$1 - \frac{4}{16} = \frac{12}{16}$$

0.75

Geometric Intuition of DT

21 July 2023

17:44

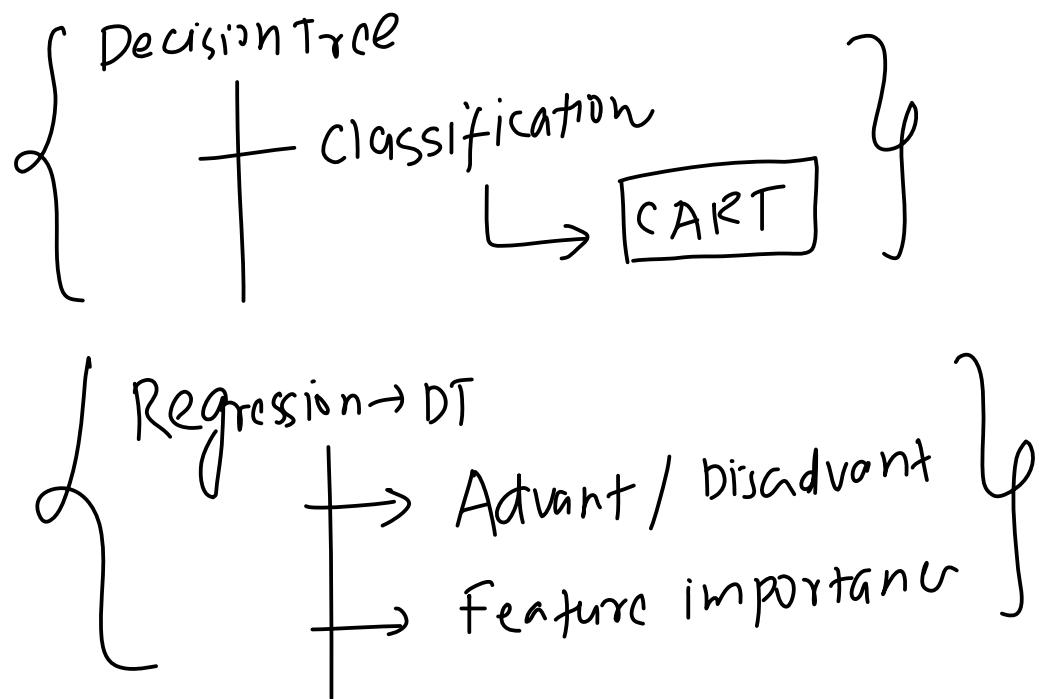


Code

21 July 2023 17:49

Recap

24 July 2023 13:31



CART for Regression

24 July 2023

13:31

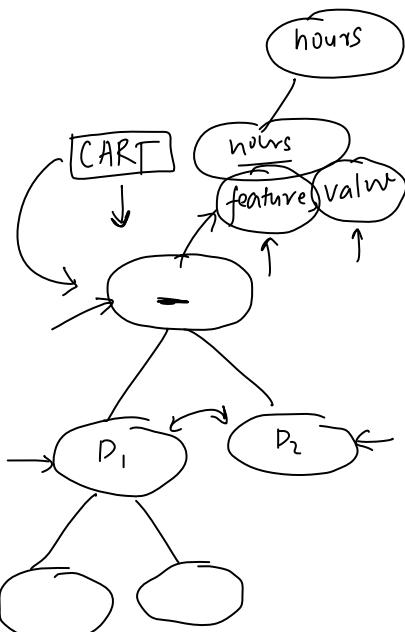
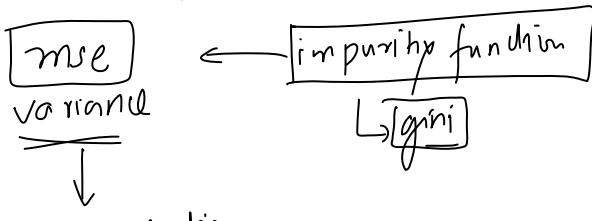
$\downarrow + \text{variable} \downarrow$

$\downarrow \text{dataset}$

$\downarrow \text{ML}$

Subject	Grade_Level	Hours_Studied	Test_Score
0 Math	Freshman	4	59
1 Physics	Freshman	1	82
2 Physics	Freshman	4	81
3 Math	Junior	6	60
4 Physics	Sophomore	1	73
5 Physics	Junior	3	85
6 Physics	Junior	4	61
7 Physics	Freshman	9	78

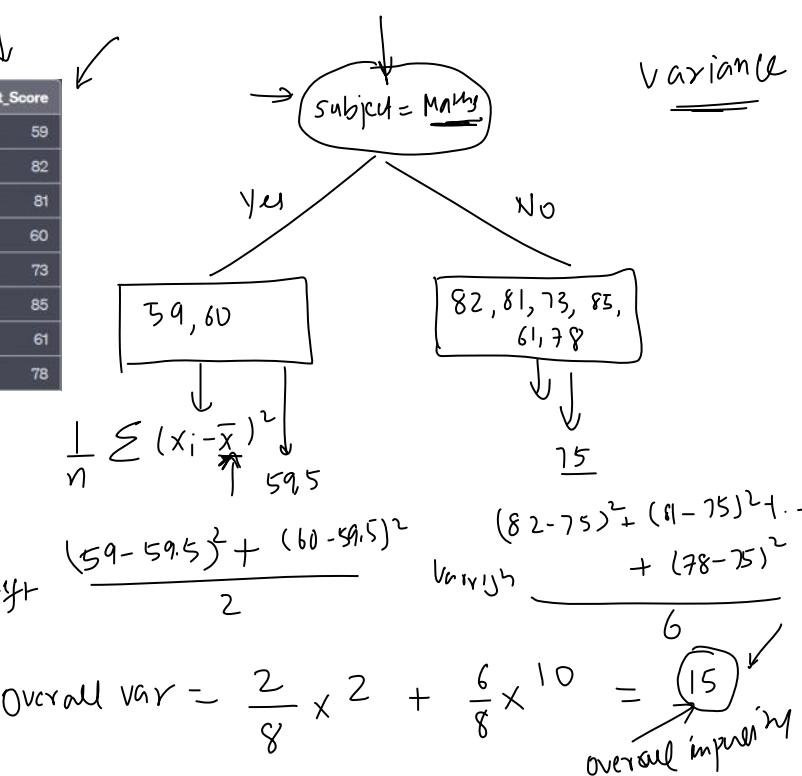
$\uparrow \text{binary}$ $\uparrow \text{multi-class}$ $\uparrow \text{numerical}$



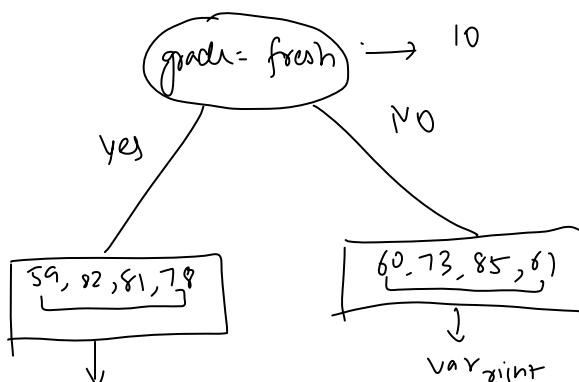
Subject	Grade_Level	Hours_Studied	Test_Score
0 Math	Freshman	4	59
1 Physics	Freshman	1	82
2 Physics	Freshman	4	81
3 Math	Junior	6	60
4 Physics	Sophomore	1	73
5 Physics	Junior	3	85
6 Physics	Junior	4	61
7 Physics	Freshman	9	78

$\cancel{15}$ $\cancel{\text{fresh} = 7}$

$\cancel{\text{hours} > 2}$



Grade_Level	Hours_Studied	Test_Score
Freshman	4	59
Freshman	1	82
Freshman	4	81
Junior	6	60
Sophomore	1	73
Junior	3	85
Junior	4	61
Freshman	9	78



Junior	4	61
Freshman	9	78

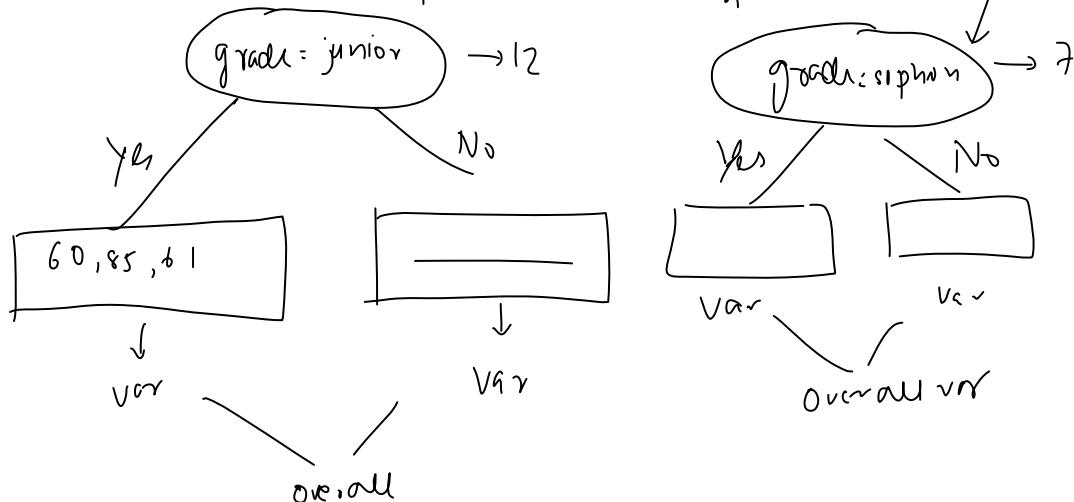
3 categories:
 { freshman, (junior, so sophomore) }
 { junior, (freshman, sophomore) }
 { sophomore, (freshman, junior) }



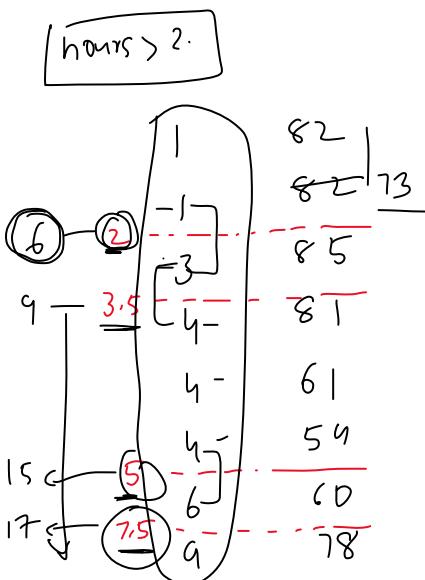
var_left

var_right

$$V_{overall} = \frac{n_{left}}{n} V_{left} + \frac{n_{right}}{n} V_{right}$$



Hours_Studied	Test_Score
4	59
1	82
4	81
6	60
1	73
3	85
4	61
9	78



17

15

9

1

-1

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

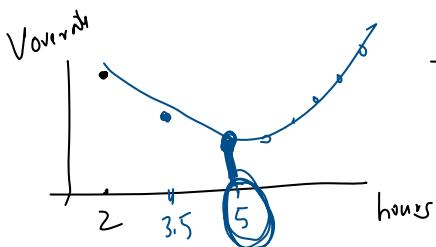
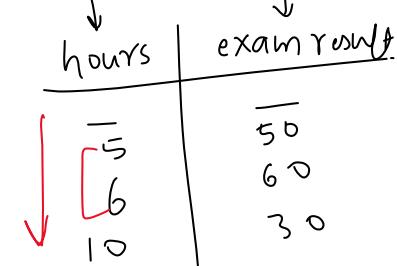
268

269

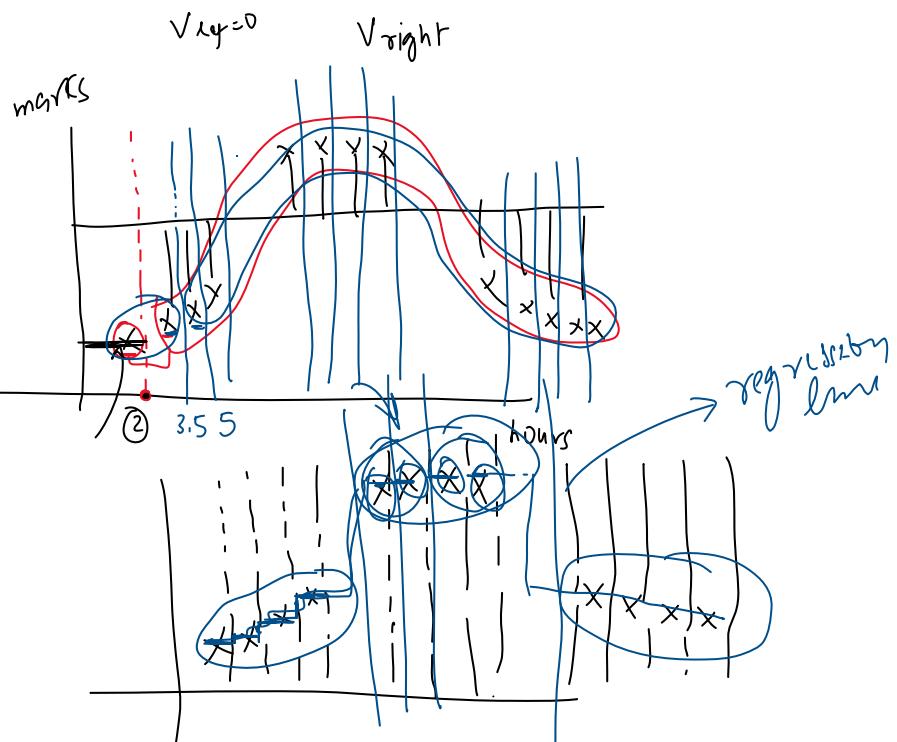
270

Geometric Intuition

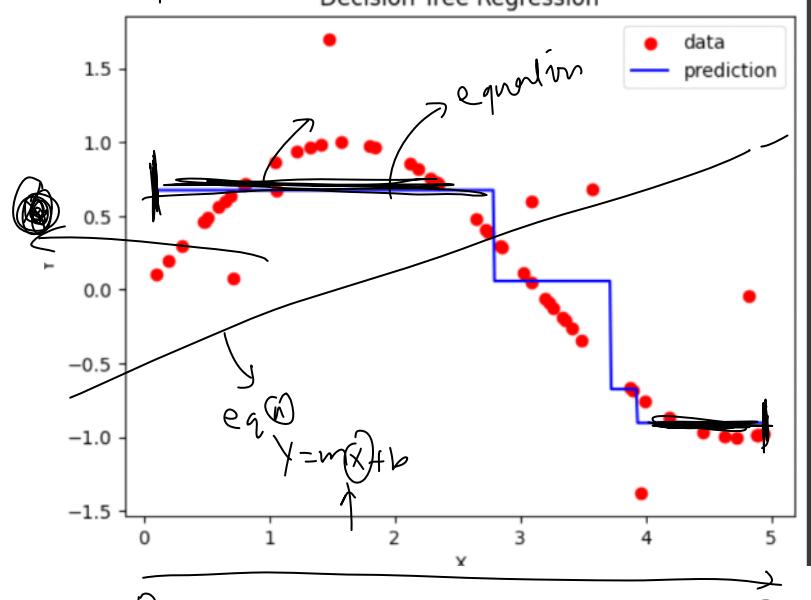
24 July 2023 13:31



piecewise constant approximation

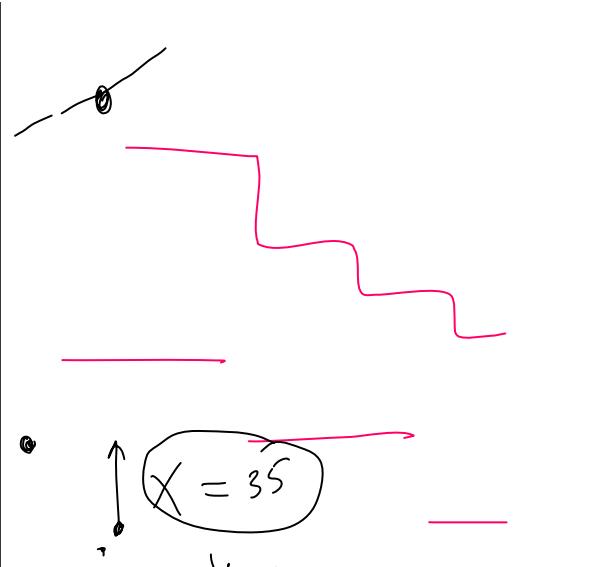


Decision Tree Regression



marks
(0-100)

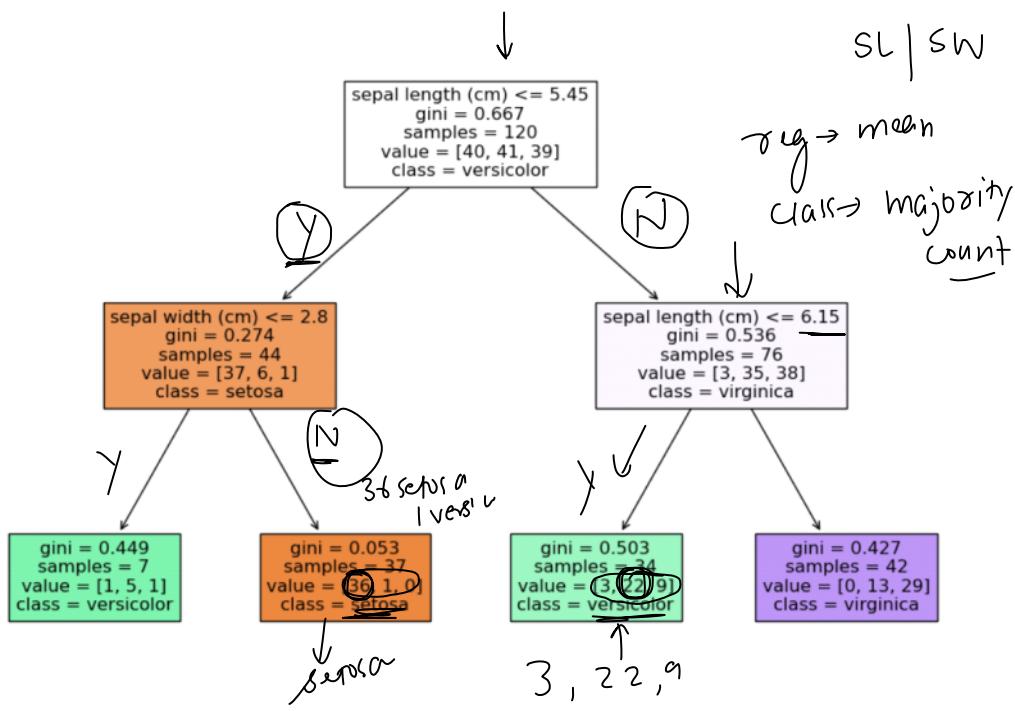
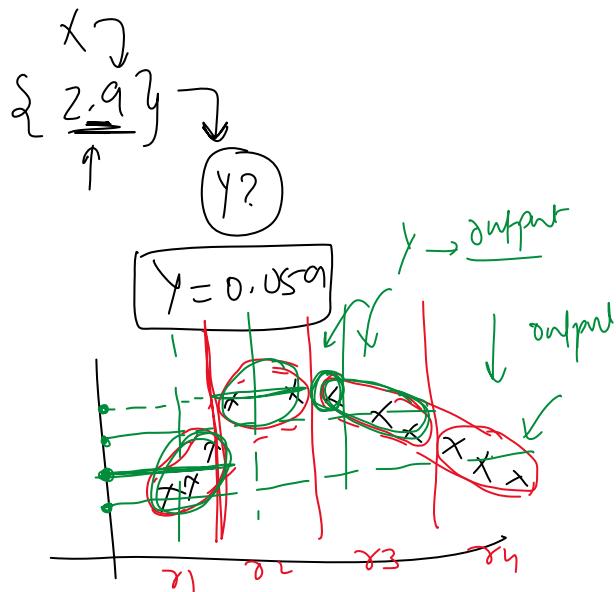
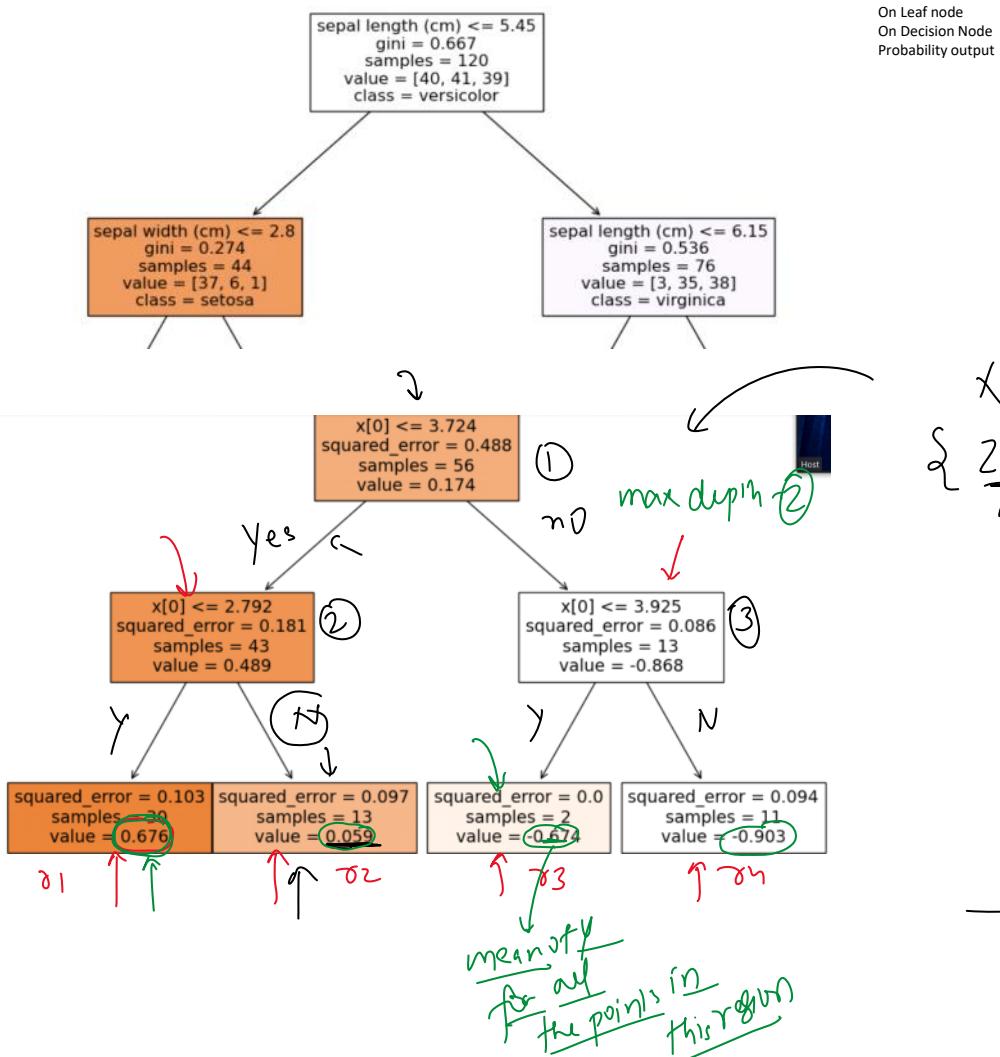
profit
dis-



How Prediction is Done

24 July 2023 13:36

Classification
Regression



$$\frac{\text{setosa}}{\text{virginica}} = \frac{3}{34}$$

$$\left\{ \frac{4.9}{\text{SL}}, \frac{3}{\text{SW}} \right\}$$

$$\left\{ \frac{5.2}{\text{SL}}, \frac{1}{\text{SW}} \right\}$$

$$\left\{ \frac{6}{\text{SL}}, \frac{2}{\text{SW}} \right\} \rightarrow \text{setosa} = \frac{3}{34}$$

\downarrow
benzene

\uparrow
3, 22, 9

$$\text{benzene} = \frac{3}{3n}$$

$$\text{verci} = \frac{22}{3n}$$

$$\text{virgin} = \frac{9}{3n}$$

Code

24 July 2023 13:32

Advantages & Disadvantages

24 July 2023 13:32

KNN

Advantages

- Simple to understand and to interpret. Trees can be visualized.
- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data.
- Can work on non-linear datasets
- Can give you feature importance.

Disadvantages

- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- Predictions of decision trees are neither smooth nor continuous, but piecewise constant approximations as seen in the above figure. Therefore, they are not good at extrapolation.
This limitation is inherent to the structure of decision tree models. They are very useful for interpretability and for handling non-linear relationships within the range of the training data, but they aren't designed for extrapolation. If extrapolation is important for your task, you might need to consider other types of models.

distance ($\log n$)

$$\begin{array}{c} 500 \\ \downarrow \\ \underline{\log(500)} \end{array}$$

pruning

linear

Feature Importance

25 July 2023 17:40

The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance.

$$f_{ik} = \frac{\sum_{j \in \text{node split on feature } k} n_i}{\sum_{j \in \text{all nodes}} n_i}$$

$$ni = \frac{N-t}{N} \left[\text{impurity} - \left(\frac{N-t-\gamma}{N-t} \times \text{right-impurity} \right) - \left(\frac{N-t-\gamma}{N-t} \times \text{left-impurity} \right) \right]$$

Regression

-> flat tentative price system(prediction)

Recommender System

-> Suggest more flats like this

-> Society suggestion

Analysis

-> City Level

-> Sector Level

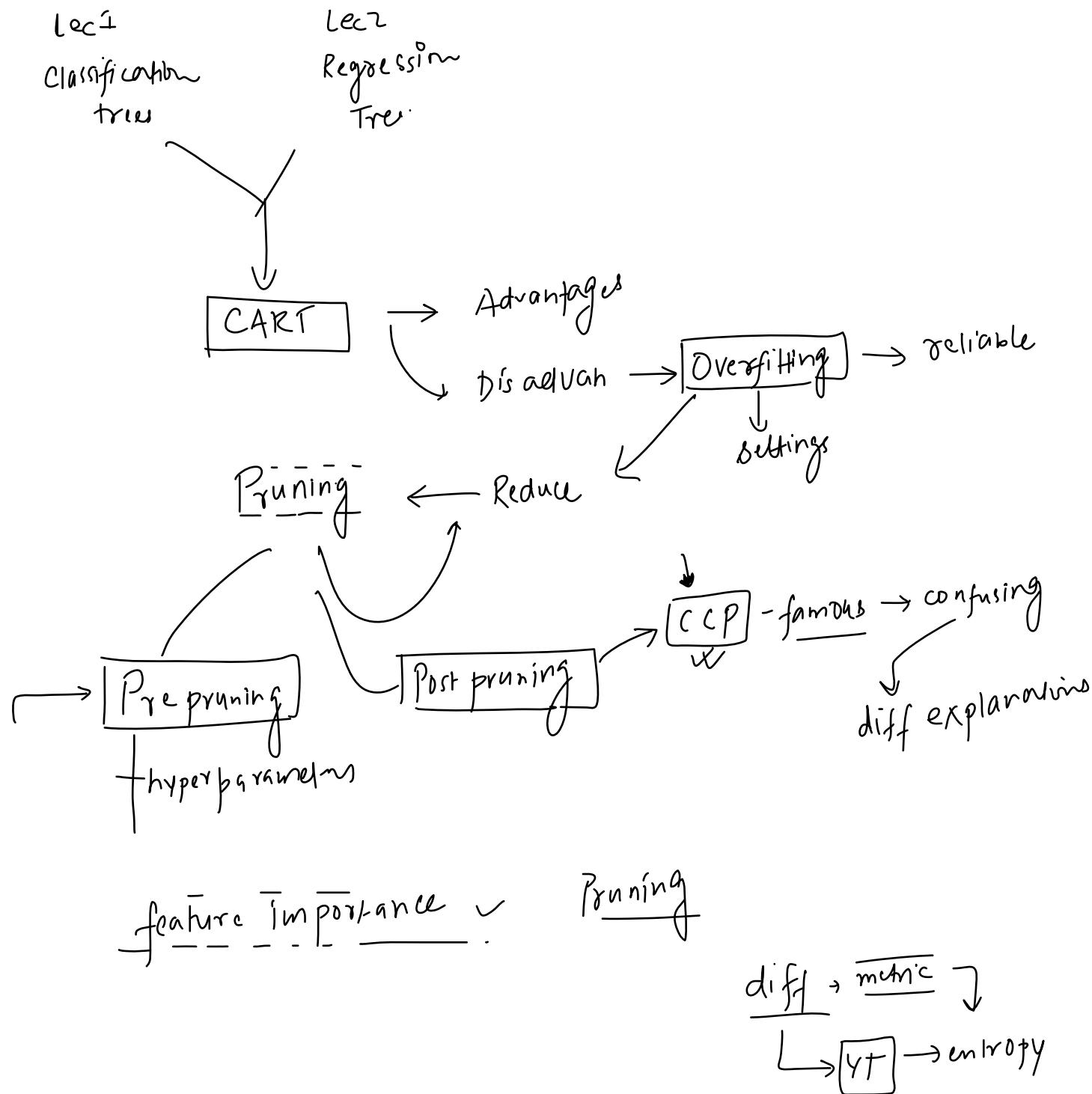
-> Insight System(Factors) -> inference(ml model)

Deploy on AWS

CI/CD pipelines
efficient

Recap

26 July 2023 15:14



Feature Importance

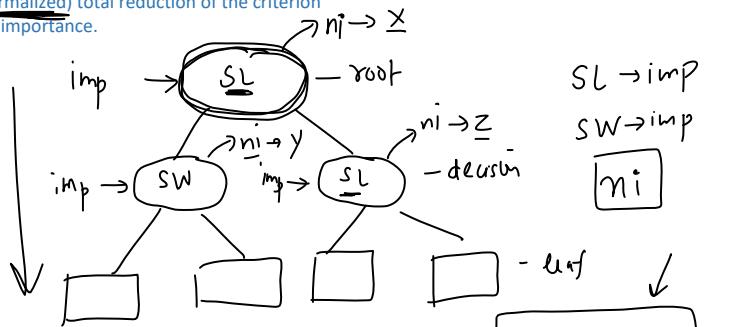
27 July 2023 07:43

$$SL \mid SW \boxed{ni}$$

The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance.

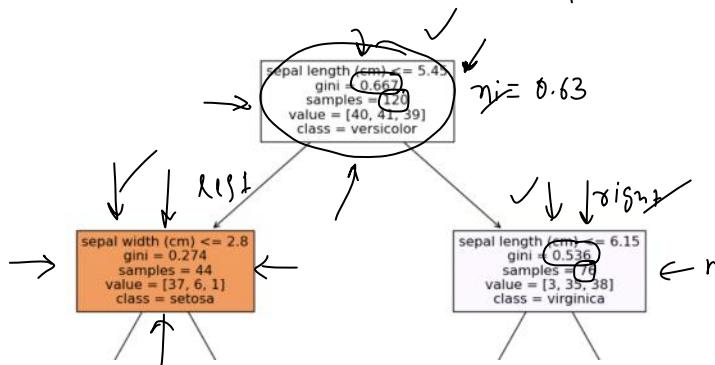
$$fi_K = \frac{\sum_{j \in \text{node split on feature } K} ni}{\sum_{j \in \text{all nodes}} ni}$$

$$ni = \frac{N-t}{N} \left[\text{impurity} - \left(\frac{N-t-x}{N-t} \times \text{right_impurity} \right) - \left(\frac{N-t-y}{N-t} \times \text{left_impurity} \right) \right]$$



$$SL = \frac{X+Z}{X+Y+Z}$$

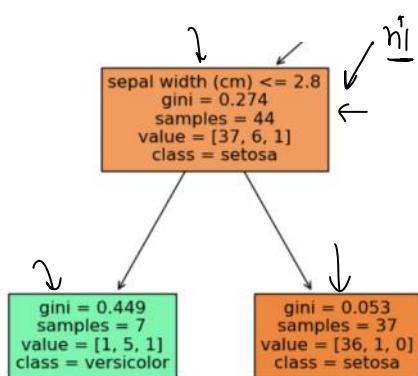
$$SW = \frac{Y}{X+Y+Z}$$



$$\frac{120}{120} \left[0.667 - \left(\frac{76}{120} \times 0.536 \right) - \left(\frac{44}{120} \times 0.27 \right) \right]$$

$$\text{sepal length} = \frac{X+Y}{-}$$

$$SN = \frac{Z}{X+Y+Z}$$

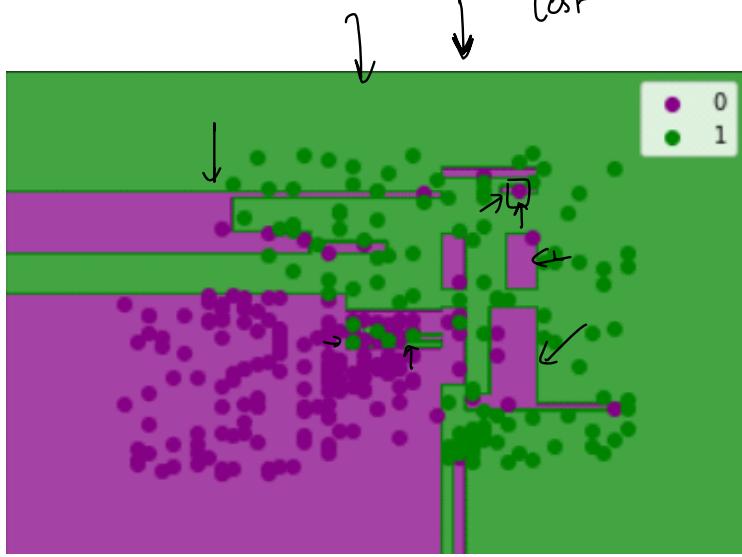


$$\frac{44}{120} \left[0.274 - \left(\frac{7}{44} \times 0.66 \right) - \left(\frac{37}{44} \times 0.05 \right) \right]$$

$$ni = \frac{N-t}{N} \left[\text{impurity} - \left(\frac{N-t-x}{N-t} \times \text{right_impurity} \right) - \left(\frac{N-t-y}{N-t} \times \text{left_impurity} \right) \right]$$

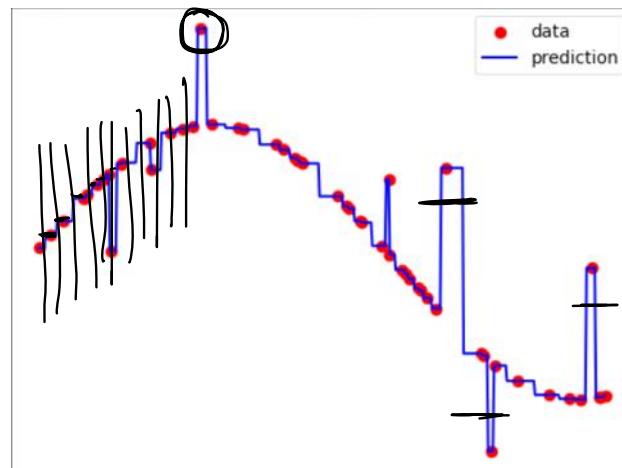
The Problem of Overfitting

26 July 2023 15:15



Classification

Pruning

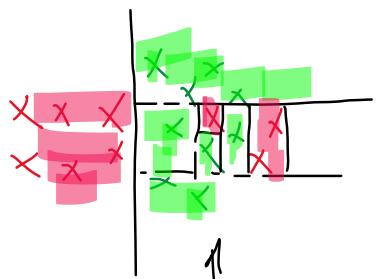
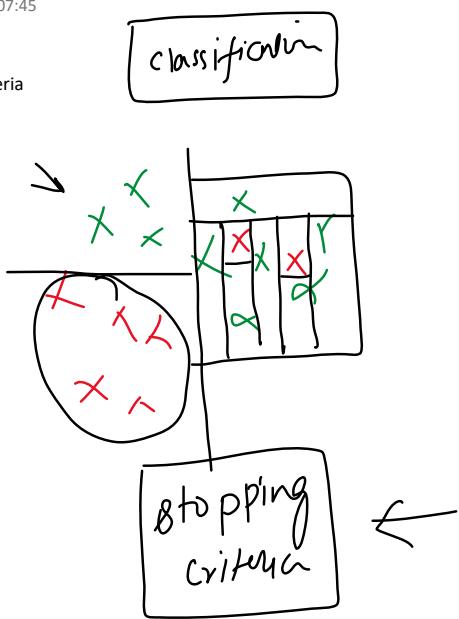


Regression

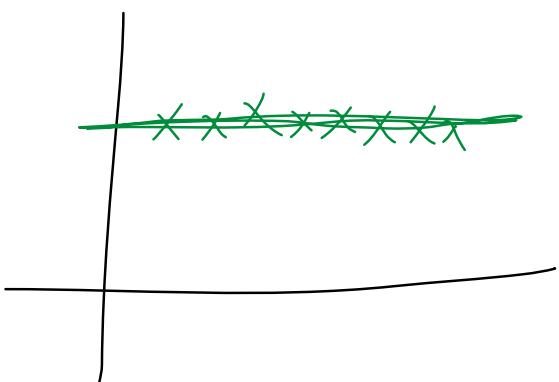
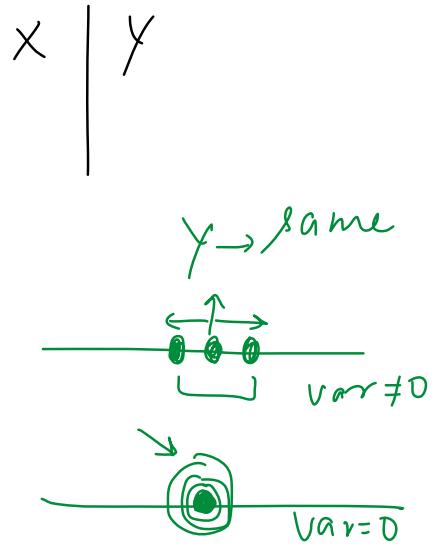
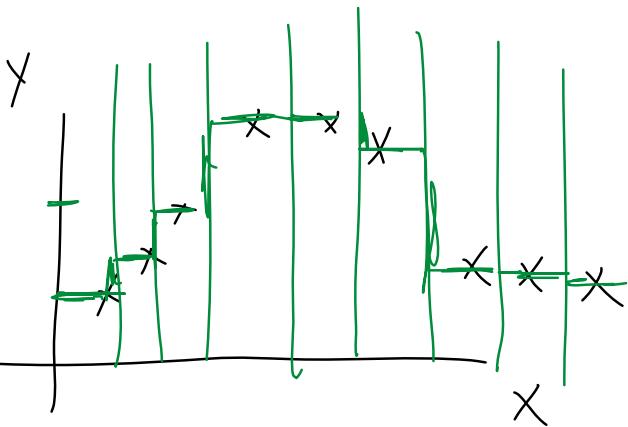
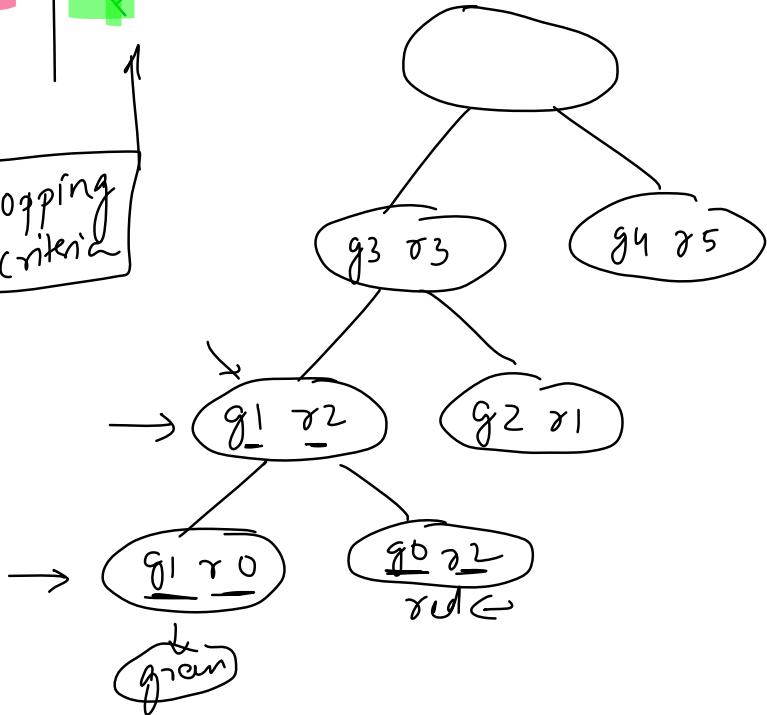
Why Overfitting happens

27 July 2023 07:45

-> Stopping criteria

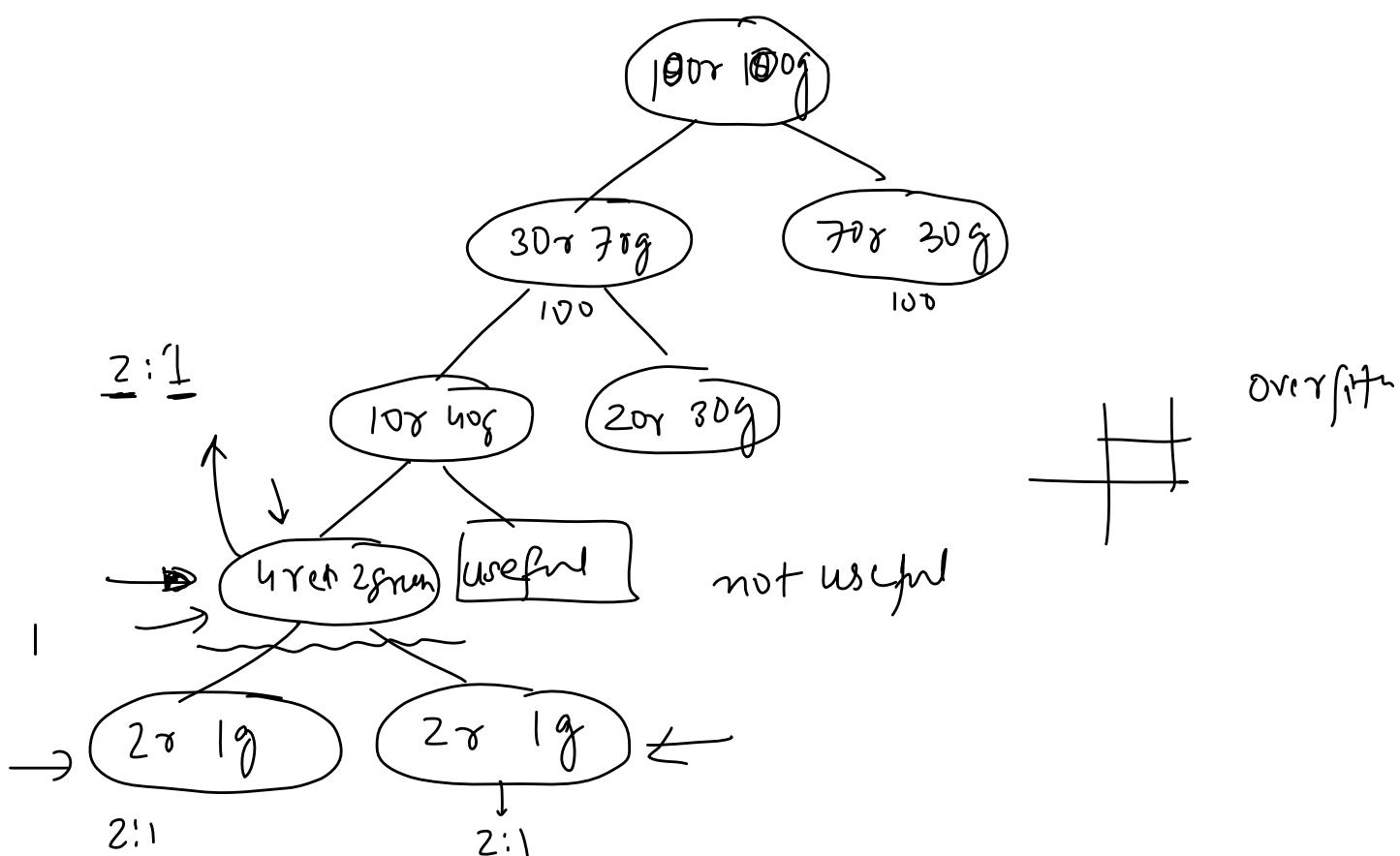
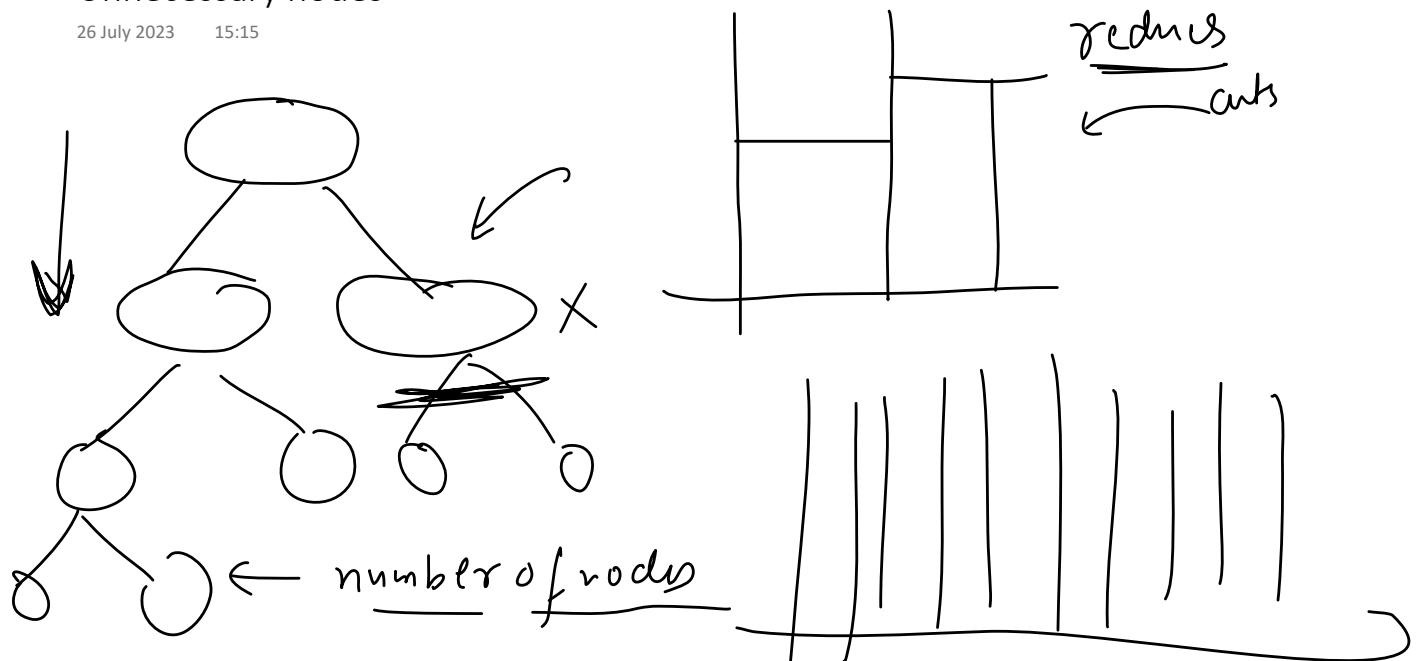


stopping
criteria



Unnecessary nodes

26 July 2023 15:15



$dt \rightarrow$ overfitting \rightarrow stopping rule

\downarrow
depth rule (a lot of nodes)

\downarrow
(not useful) \rightarrow cut these nodes
||

Pruning

Pruning

(^v not useful) → cut more
↓
reduce overfitting ← tree size reduced

Pruning & it's types

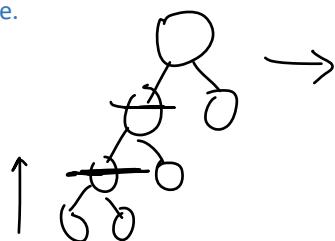
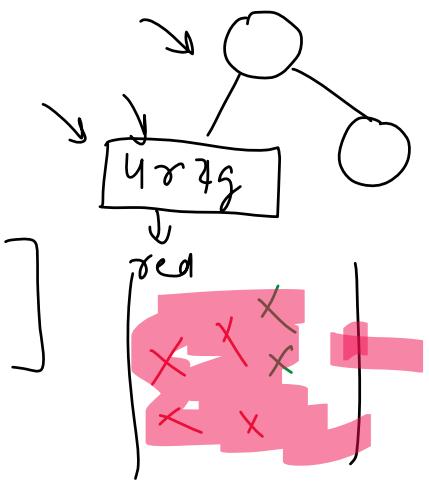
26 July 2023 15:15

Pruning is a technique used in machine learning to reduce the size of decision trees and to avoid overfitting. Overfitting happens when a model learns the training data too well, including its noise and outliers, which results in poor performance on unseen or test data.

Decision trees are susceptible to overfitting because they can potentially create very complex trees that perfectly classify the training data but fail to generalize to new data. Pruning helps to solve this issue by reducing the complexity of the decision tree, thereby improving its predictive power on unseen data.

There are two main types of pruning: pre-pruning and post-pruning.

1. Pre-pruning (Early stopping): This method halts the tree construction early. It can be done in various ways: by setting a limit on the maximum depth of the tree, setting a limit on the minimum number of instances that must be in a node to allow a split, or stopping when a split results in the improvement of the model's accuracy below a certain threshold.
2. Post-pruning (Cost Complexity Pruning): This method allows the tree to grow to its full size, then prunes it. Nodes are removed from the tree based on the error complexity trade-off. The basic idea is to replace a whole subtree by a leaf node, and assign the most common class in that subtree to the leaf node.



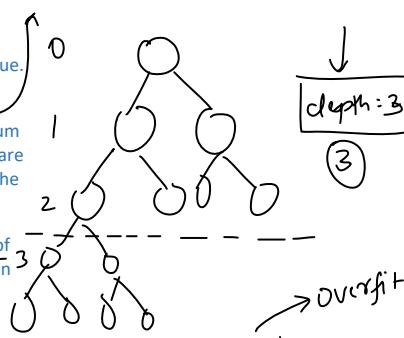
Pre-pruning

26 July 2023 15:27

Pre-pruning, also known as early stopping, is a technique where the decision tree is pruned during the learning process as soon as it's clear that further splits will not add significant value. There are several strategies for pre-pruning:

1. **Maximum Depth:** One of the simplest forms of pre-pruning is to set a limit on the maximum depth of the tree. Once the tree reaches the specified depth during training, no new nodes are created. This strategy is simple to implement and can effectively prevent overfitting, but if the maximum depth is set too low, the tree might be overly simplified and underfit the data.

hyperparam



2. **Minimum Samples Split:** This is a condition where a node will only be split if the number of samples in that node is above a certain threshold. If the number of samples is too small, then the node is not split and becomes a leaf node instead. This can prevent overfitting by not allowing the model to learn noise in the data. (parent)

overfit

3. **Minimum Samples Leaf:** This condition requires that a split at a node must leave at least a minimum number of training examples in each of the leaf nodes. Like the minimum samples split, this strategy can prevent overfitting by not allowing the model to learn from noise in the data. (child)

max_depth = None

4. **Maximum Leaf Nodes:** This strategy limits the total number of leaf nodes in the tree. The tree stops growing when the number of leaf nodes equals the maximum number.

max_depth = 1 underfit

5. **Minimum Impurity Decrease:** This strategy allows a node to be split if the impurity decrease of the split is above a certain threshold. Impurity measures how mixed the classes within a node are. If the decrease is too small, the node becomes a leaf node.

max_depth = 9 overfit

6. **Maximum Features:** This strategy considers only a subset of features for deciding a split at each node. The number of features to consider can be defined and this helps in reducing overfitting.

underfit

Advantages of Pre-Pruning:

1. **Simplicity:** Pre-pruning criteria such as maximum depth or minimum number of samples per leaf are easy to understand and implement.

2. **Computational Efficiency:** By limiting the size of the tree, pre-pruning can substantially reduce the computational cost of training and prediction.

3. **Reduced Overfitting:** By preventing the tree from becoming overly complex, pre-pruning can help avoid overfitting the training data and thereby improve the model's generalization performance.

4. **Improved Interpretability:** Simpler trees (with fewer nodes) are often easier for humans to interpret.

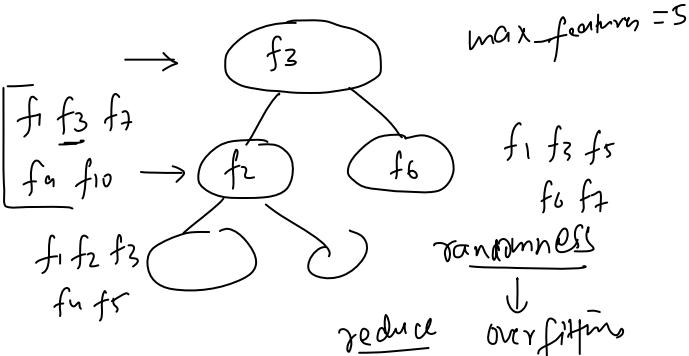
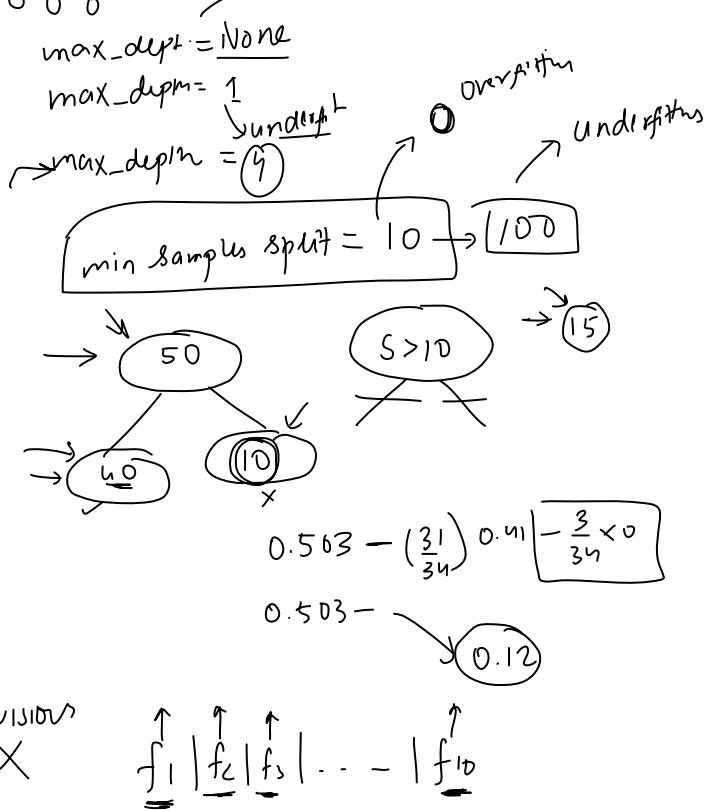
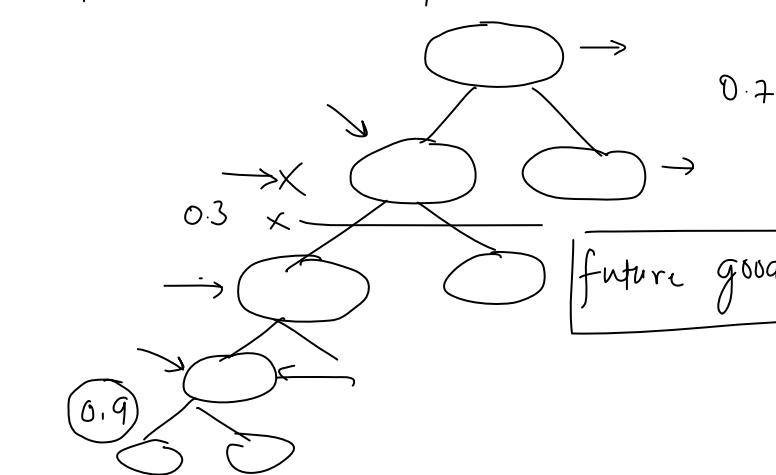
Disadvantages of Pre-Pruning:

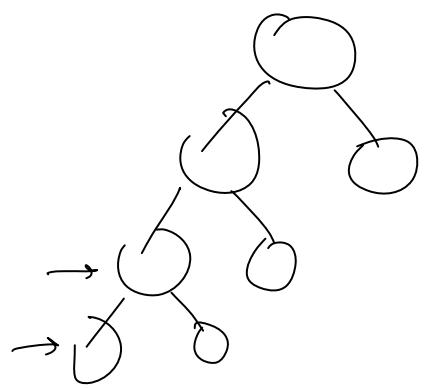
1. **Risk of Underfitting:** If the stopping criteria are too strict, pre-pruning can halt the growth of the tree too early, leading to underfitting. The model may become overly simplified and fail to capture important patterns in the data.

2. **Requires Fine-Tuning:** The pre-pruning parameters (like maximum depth or minimum samples per leaf) often require careful tuning to find the right balance between underfitting and overfitting.

3. **Short Sightedness:** Can prune good nodes if they come after a bad node.

min impurity decr = 0.5





Post Pruning

26 July 2023 15:31

Post-pruning, also known as backward pruning, is a technique used to prune the decision tree after it has been built. There are several strategies for post-pruning:

1. **Cost Complexity Pruning (CCP):** Also known as Weakness Pruning, this technique introduces a tuning parameter (α) that trades off between tree complexity and its fit to the training data. For each value of α , there is an optimal subtree that minimizes the cost complexity criterion. The subtree that minimizes the cost complexity criterion over all values of α is chosen as the pruned tree.
2. **Reduced Error Pruning:** In this method, starting at the leaves, each node is replaced with its most popular class. If the accuracy is not affected in the validation set, the change is kept.

Advantages of Post-Pruning:

1. **Reduced Overfitting:** Post-pruning methods can help to avoid overfitting the training data, which can lead to better model generalization and thus better performance on unseen data.
2. **Preserving Complexity:** Unlike pre-pruning, post-pruning allows the tree to grow to its full complexity first, which means it can capture complex patterns in the data before any pruning is done.
3. **Better Performance:** Post-pruned trees often outperform pre-pruned trees, as they are able to better balance the bias-variance trade-off.

Disadvantages of Post-Pruning:

1. **Increased Computational Cost:** Post-pruning can be more computationally intensive than pre-pruning, as the full tree must be grown first before it can be pruned.
2. **Requires Validation Set:** Many post-pruning methods require a validation set to assess the impact of pruning. This reduces the amount of data available for training the model.
3. **Complexity of Implementation:** Post-pruning methods, especially those involving tuning parameters (like cost complexity pruning), can be more

complex to implement and understand than pre-pruning methods.

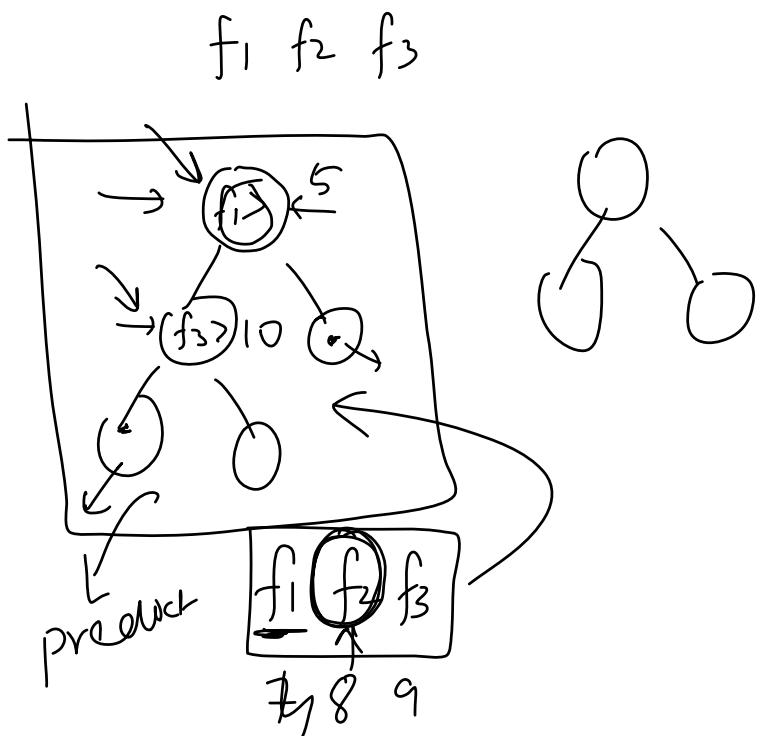
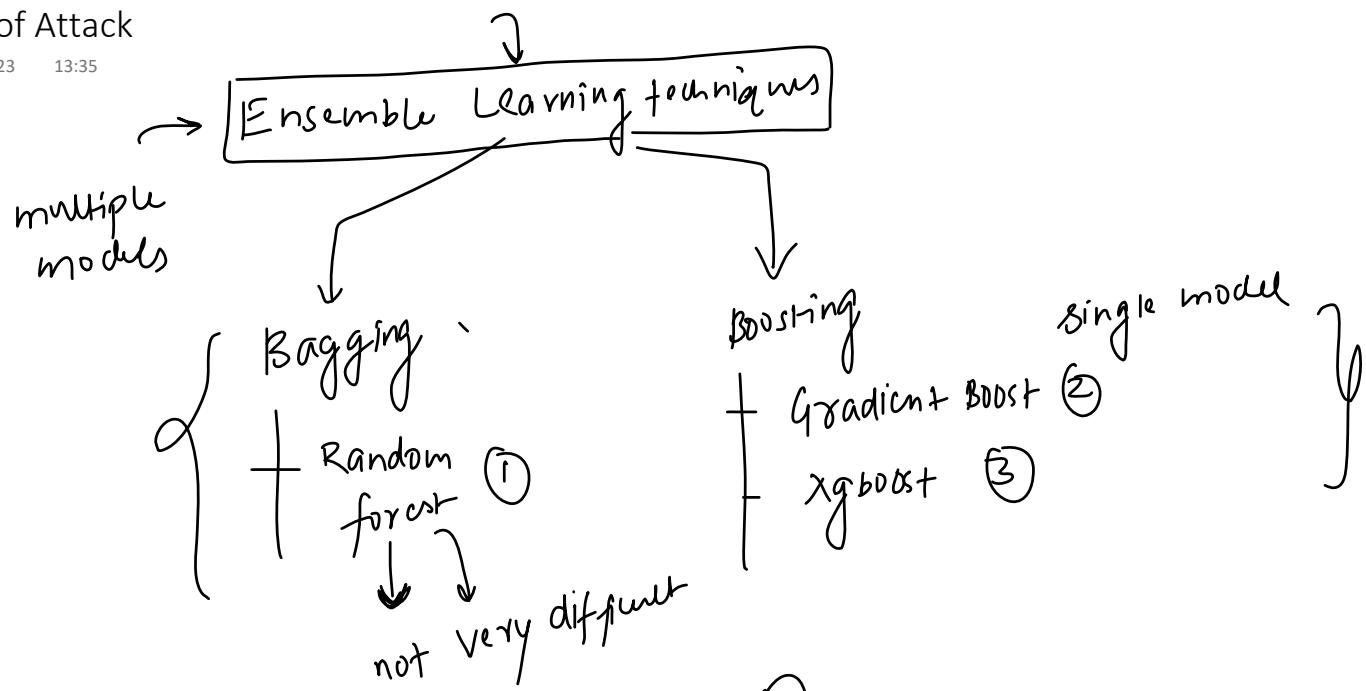
4. Risk of Underfitting: Similar to pre-pruning, if too much pruning is done, it can lead to underfitting where the model becomes overly simplified and fails to capture important patterns in the data.

Cost Complexity Pruning

26 July 2023 15:34

Plan of Attack

29 July 2023 13:35

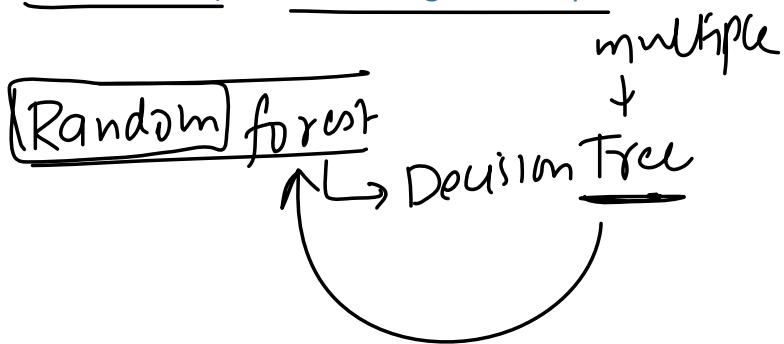


Introduction to Random Forest

29 July 2023 15:52

Random Forest is a versatile and widely used machine learning algorithm that belongs to the class of ensemble methods. Specifically, it is a type of bagging technique, which involves training many individual models (in this case, decision trees) and combining their outputs to make a final prediction.

Bagging

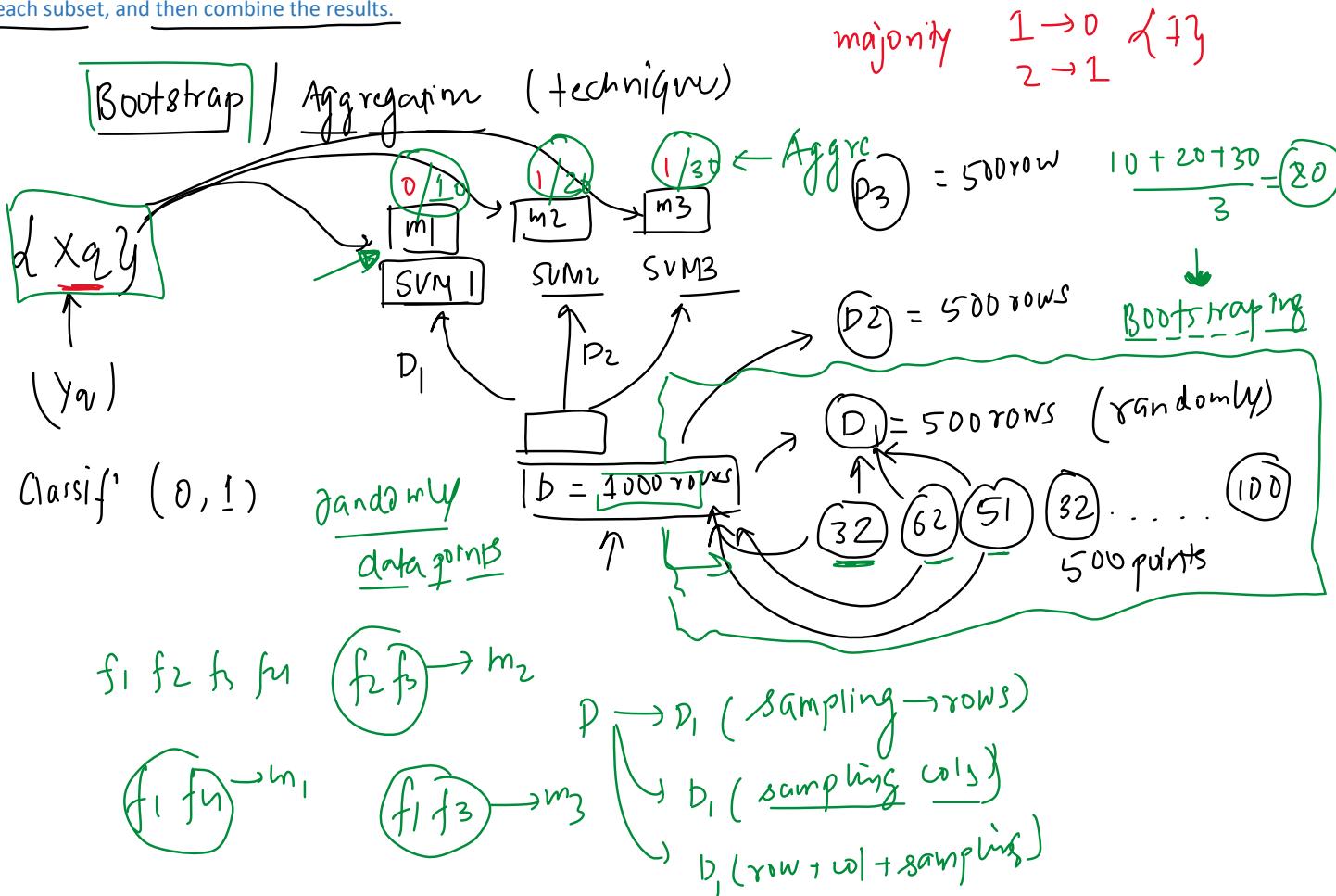


Bagging

29 July 2023 15:52

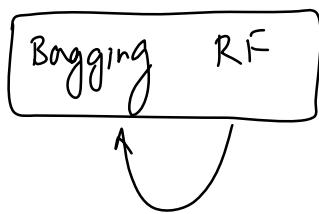
Bagging

Bagging, short for bootstrap aggregating, is a machine learning ensemble method designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also helps to avoid overfitting. The key principle of bagging is to generate multiple subsets of the original data (with replacement), train a separate model for each subset, and then combine the results.



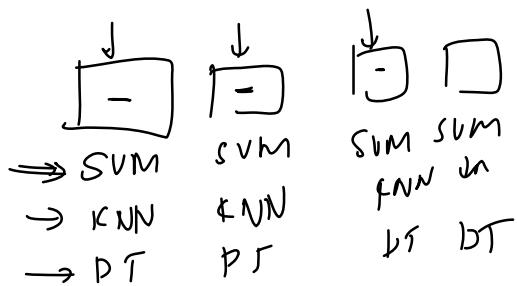
Random Forest Intuition [Code]

29 July 2023 15:52



(2) major diff

↳ Bagging → any ml algorithm



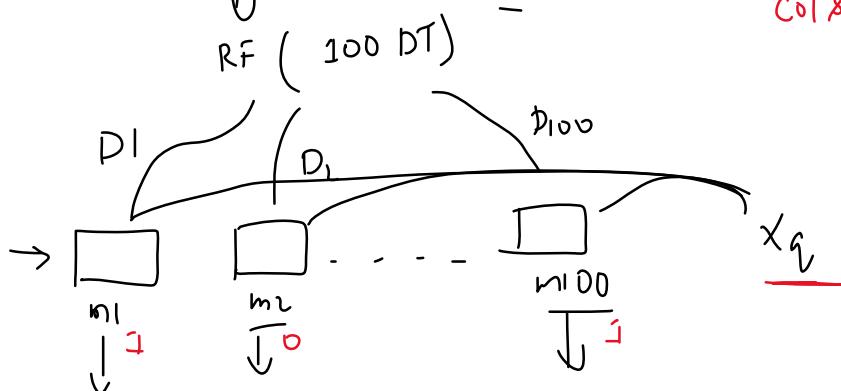
RF → base model → \hat{P}^T

$D \rightarrow 1000 \text{ rows} / 5 \text{ cols}$ (classification)

rows = 500 row with replacement

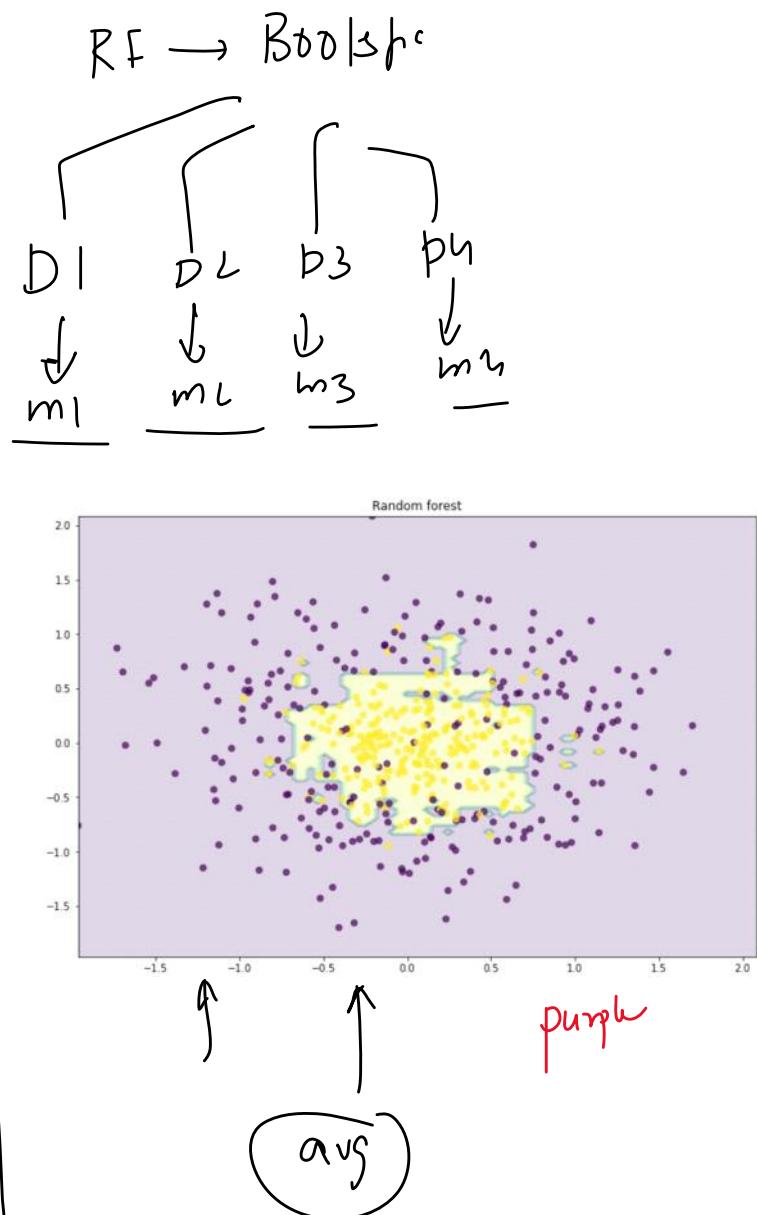
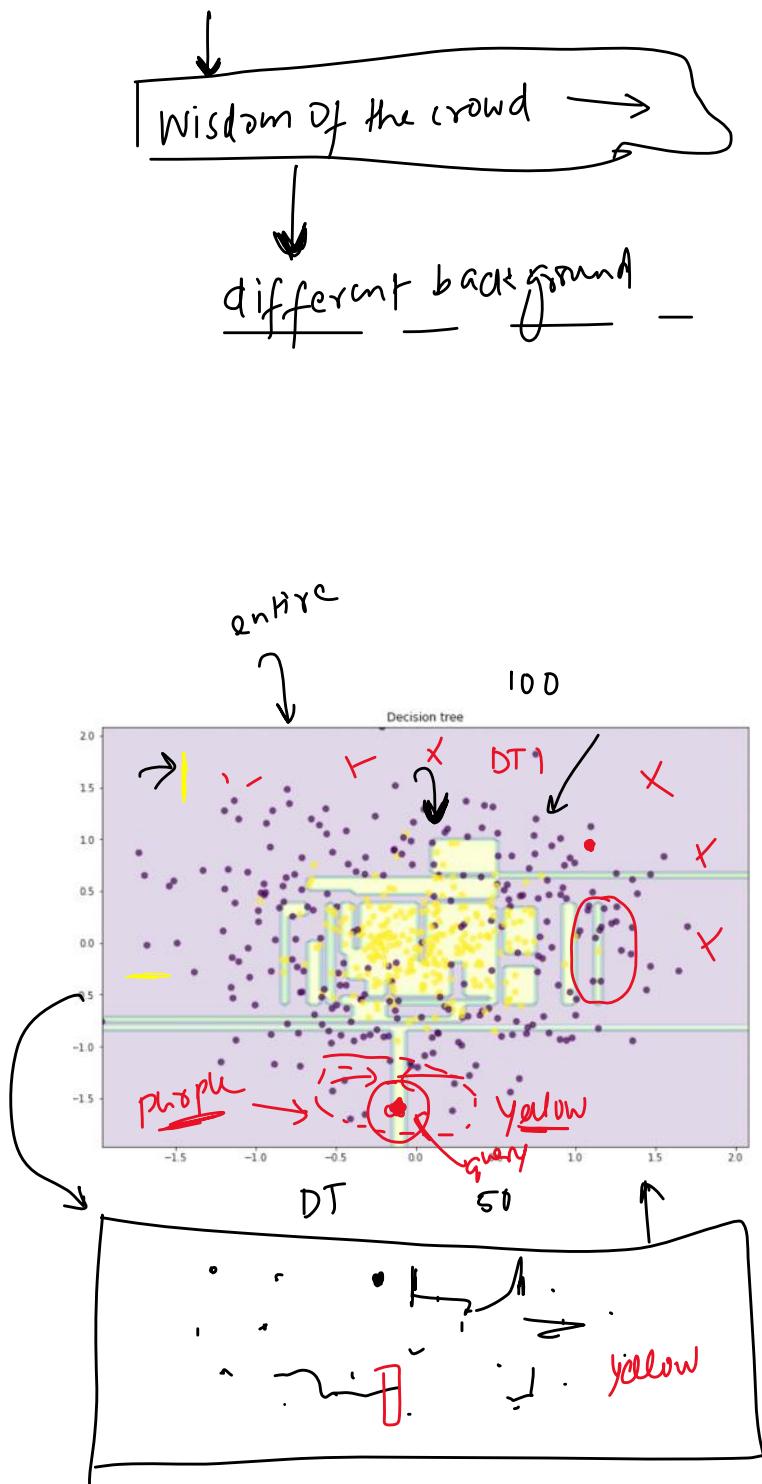
col sampling 200 cols sampled

60 → 6
40 → 1 → 0



Why Random Forest works? [Code]

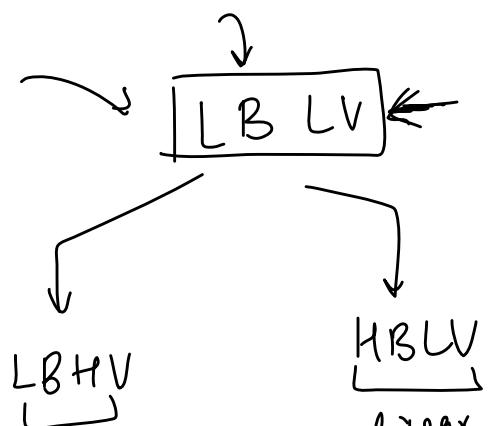
29 July 2023 15:56

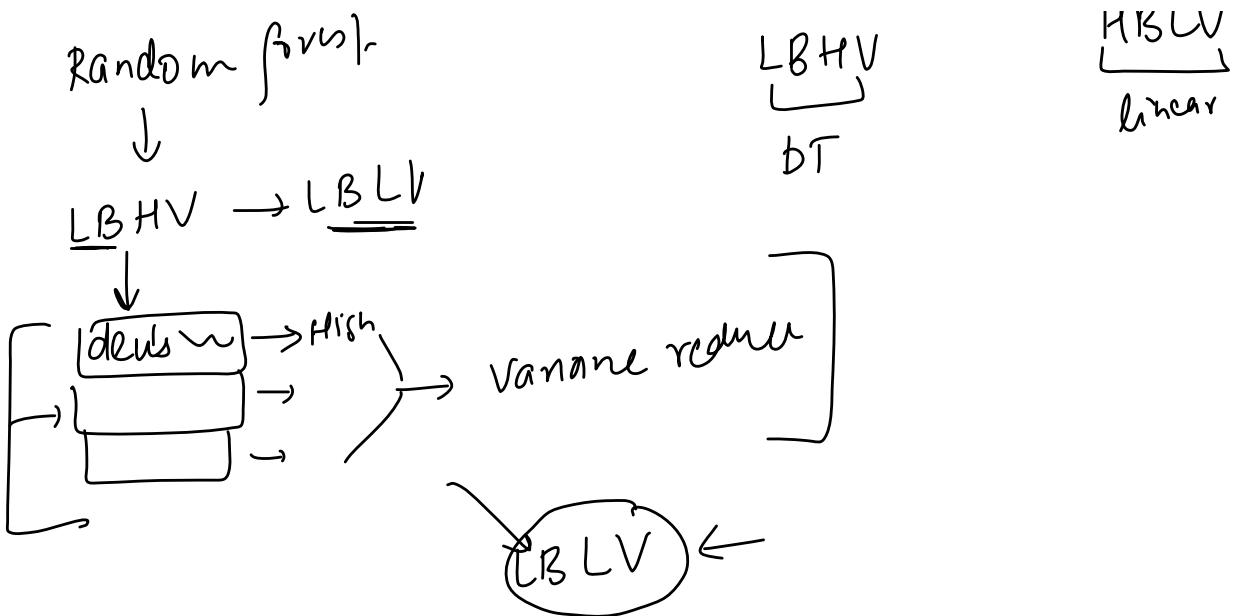


Bias - variance Tradeoff

$$\text{Bias} \propto \frac{1}{\text{Variance}}$$

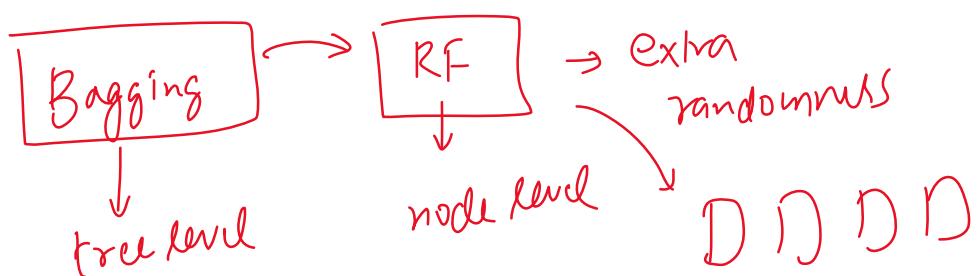
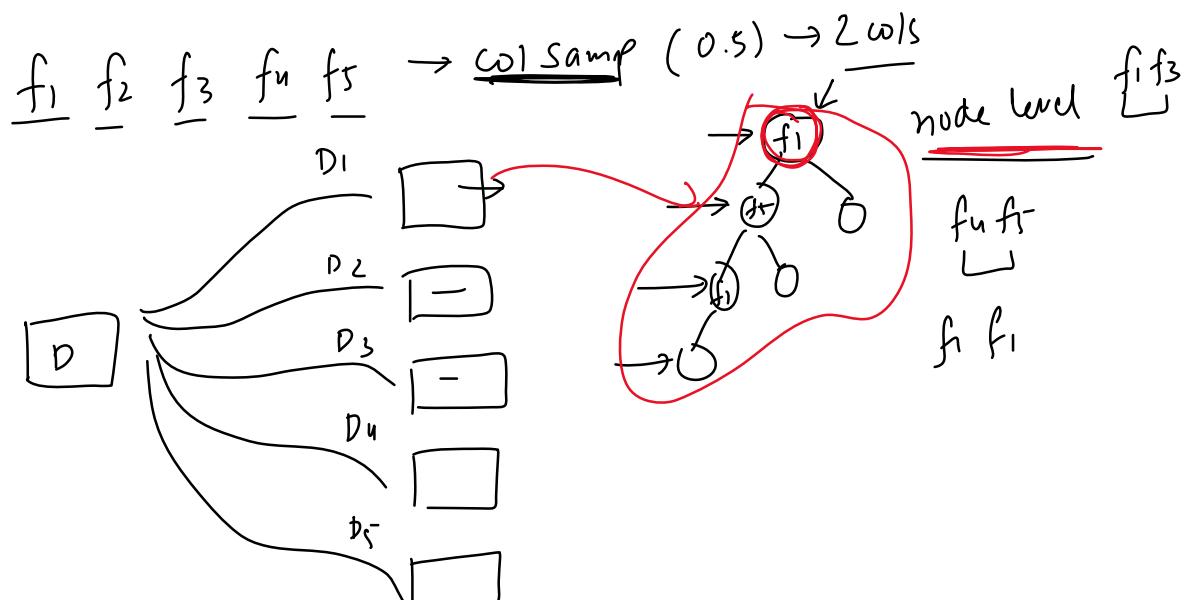
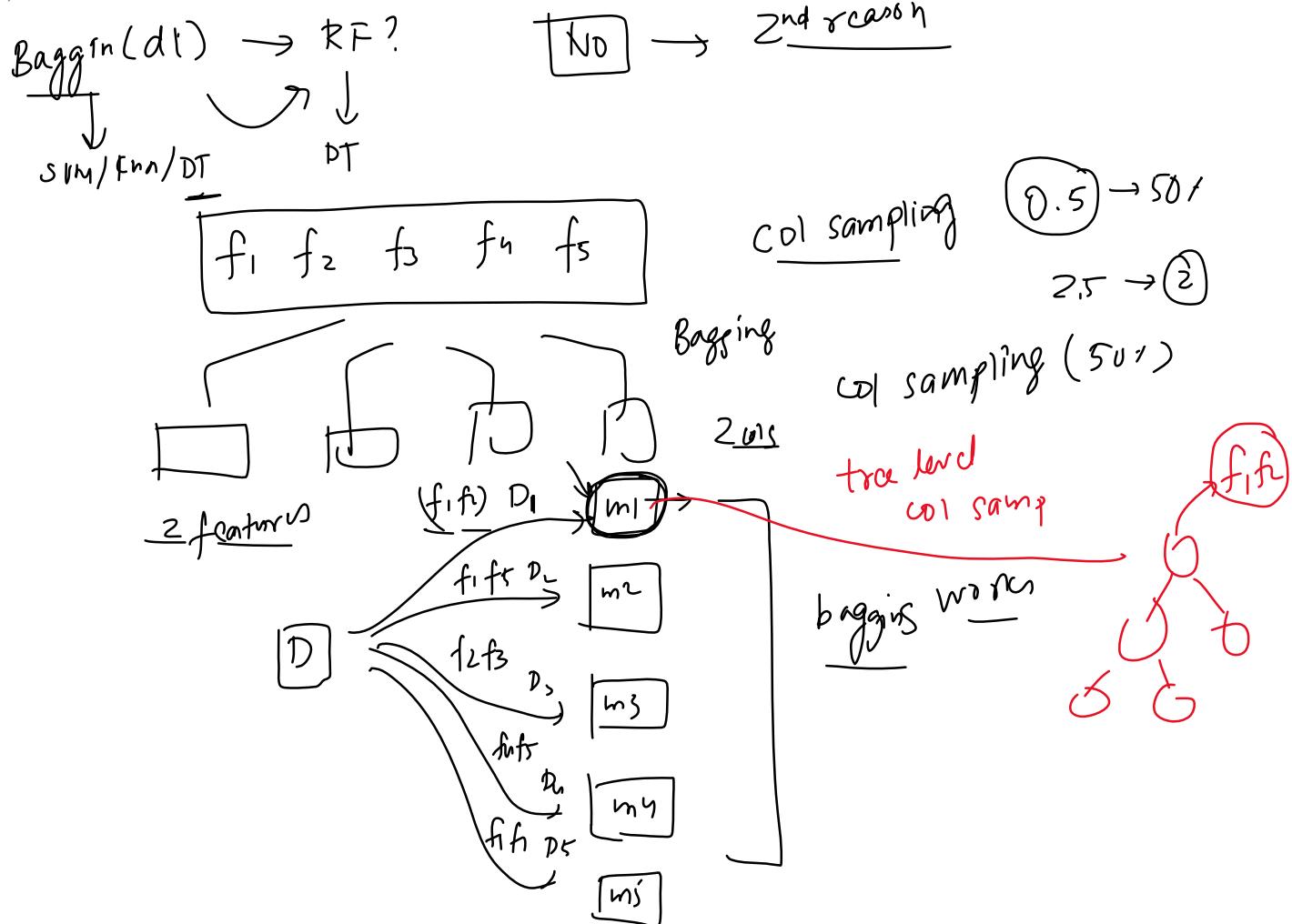
Random forest





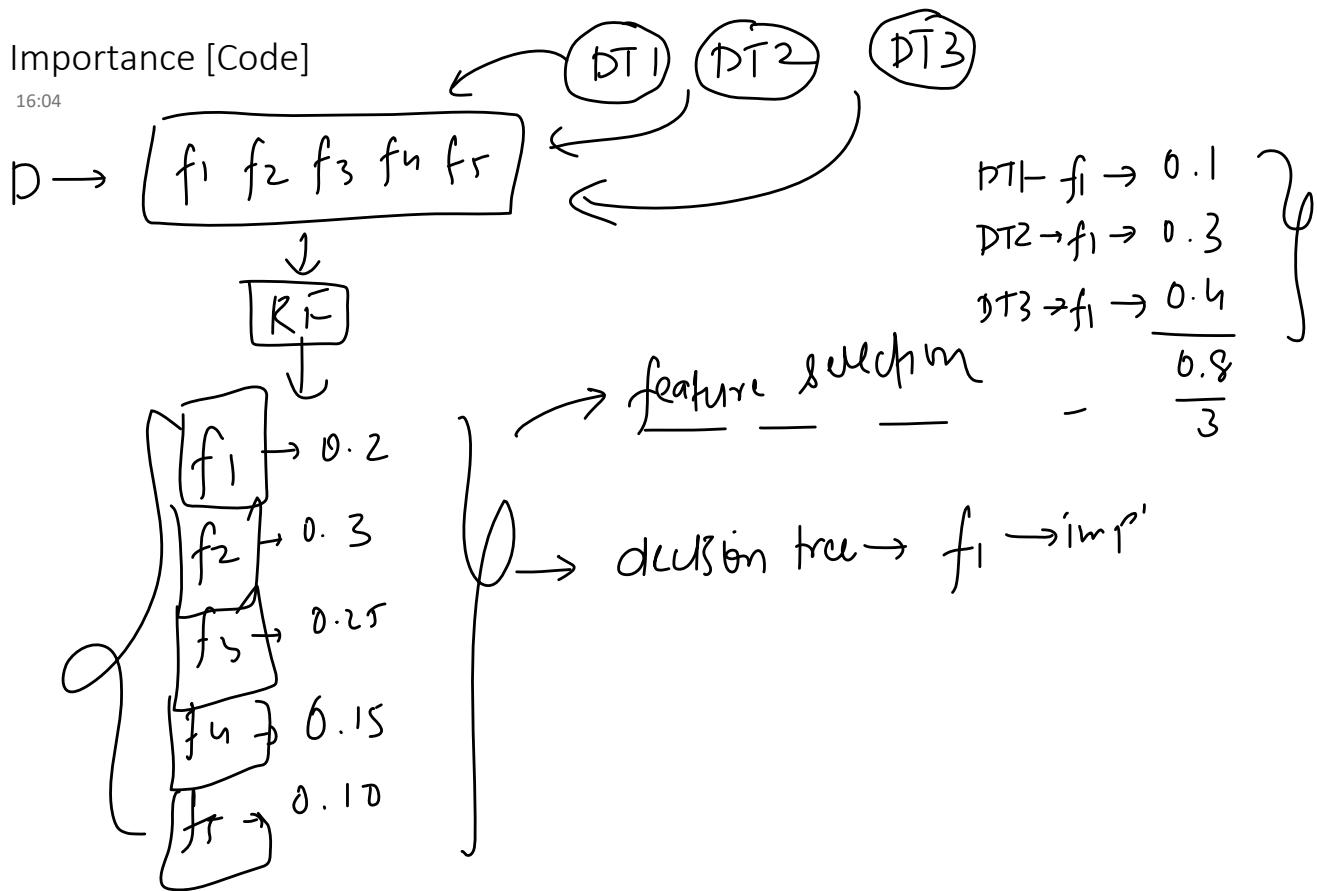
Bagging Vs Random Forest [Code]

29 July 2023 15:57



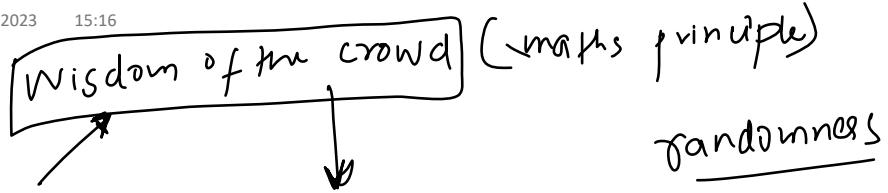
Feature Importance [Code]

29 July 2023 16:04



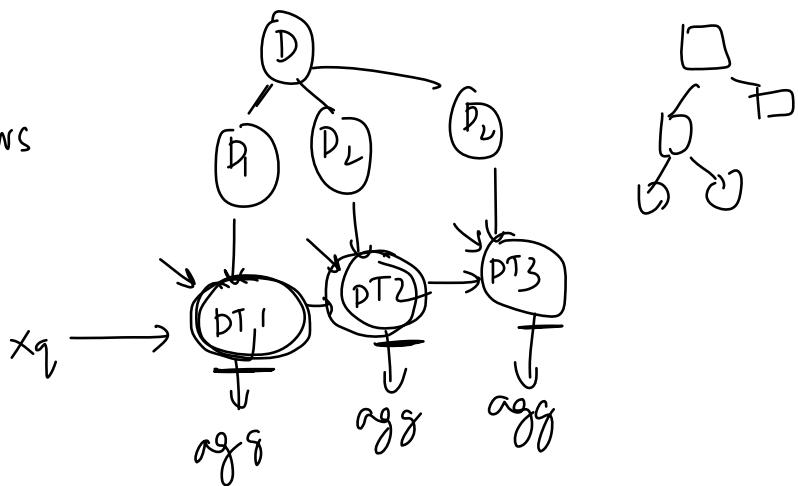
Recap

31 July 2023 15:16



randomness

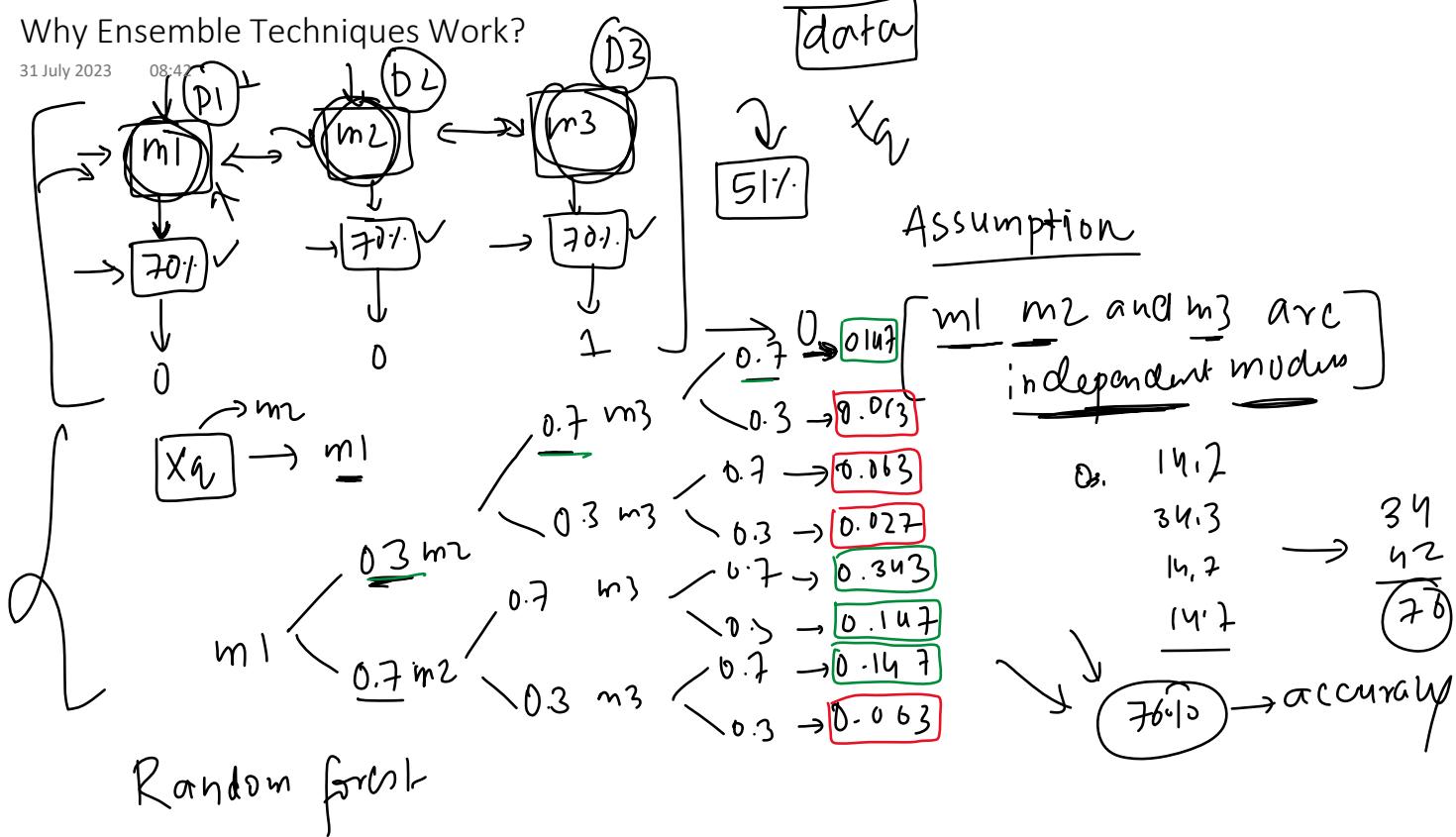
1) Bootstrapping \rightarrow yours
decorrelated



Why Ensemble Techniques Work?

31 July 2023

08:42



Random Forest Hyperparameters

31 July 2023 08:44

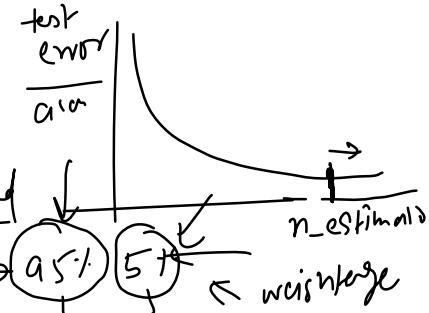
\sqrt{p}

fully grown (5)

$\sqrt{r(5)}$

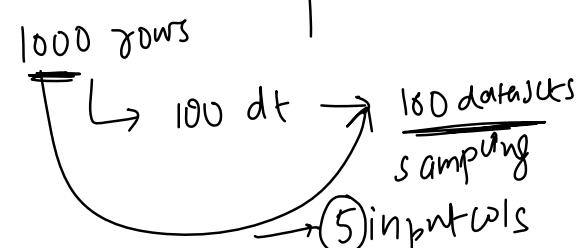
tree

Forest Level HP	<u>Tree Level HP</u>	Miscellaneous HP
N_estimators	Criterion	Oob_score
Max_features	Max_depth	N_jobs
Bootstrap	Min_Samples_split	Random_state
Max_samples	Min_samples_leaf	verbose
Bootstrap_feature	Min_weight_fraction_leaf	Warm_start
	Max_leaf_nodes	Class_weight
	Min_impurity_decrease	imbalanced
	Ccp_alpha	



row sampling how many dt 100 → deep trees → low 5, 10, overfitting 100, 500, 1000

with replacement with

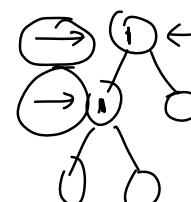


col sampling

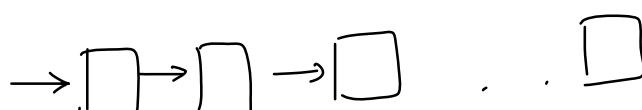
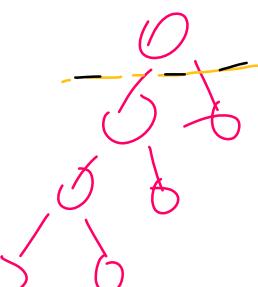
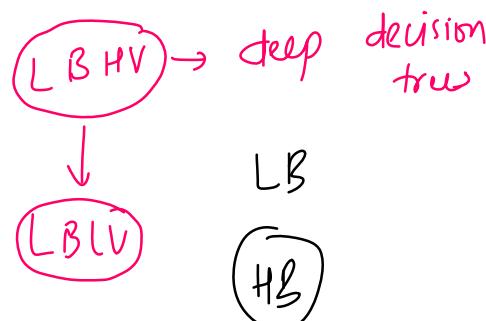
row samp

max_features = 2

1000
500
100



node level
col samp

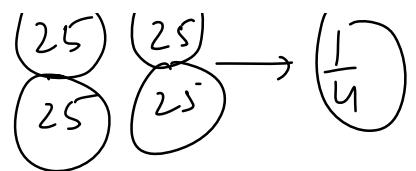


multicore

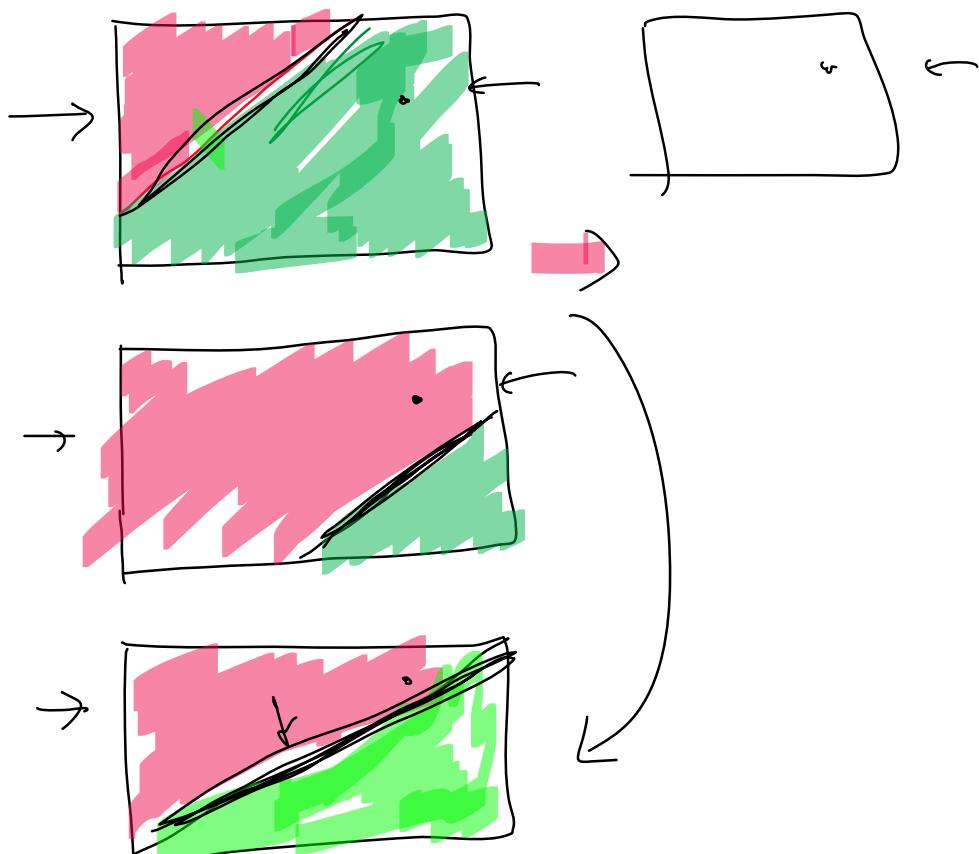
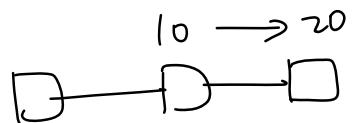
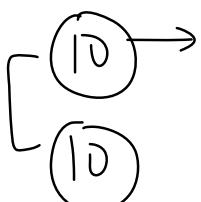
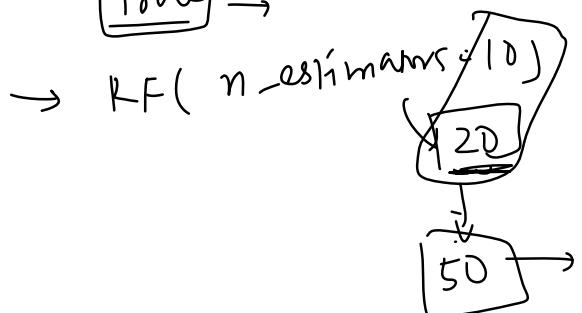
quad process

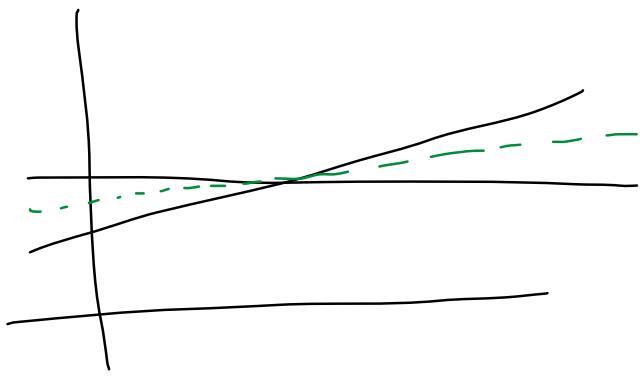
25 25 → $\left(\frac{1}{4}\right)$

↓
multi-core → quad process



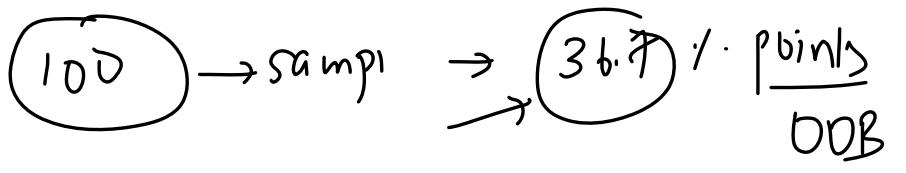
warm_start → training time
is high
false





[OOB Score] →

31 July 2023 08:47

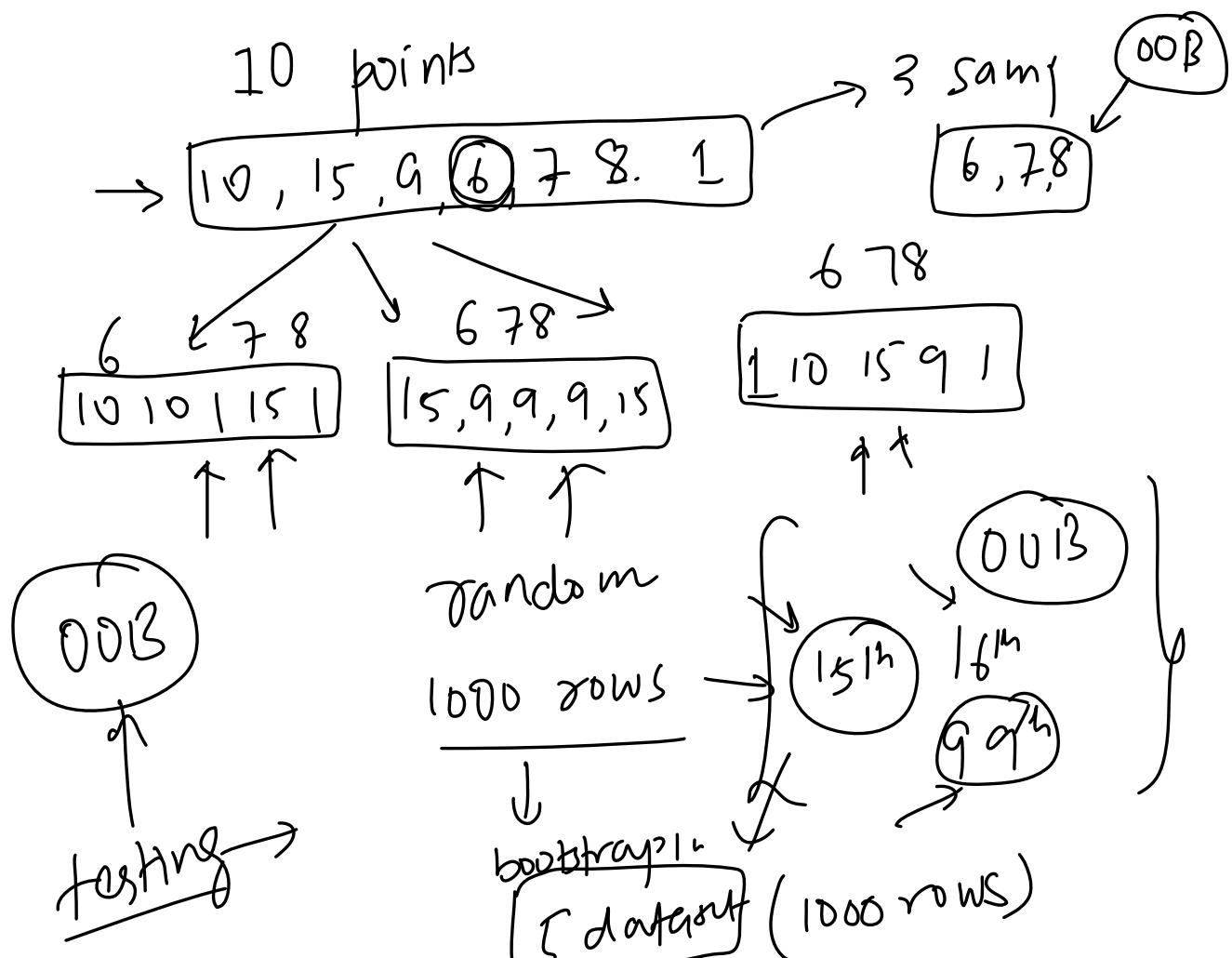


"OOB" stands for "out-of-bag". In the context of machine learning, an out-of-bag score is a method of measuring the prediction error of random forests, bagging classifiers, and other ensemble methods that use bootstrap aggregation (bagging) when sub-samples of the training dataset are used to train individual models.

Here's how it works:

1. Each tree in the ensemble is trained on a distinct bootstrap sample of the data. By the nature of bootstrap sampling, some samples from the dataset will be left out during the training of each tree. These samples are called "out-of-bag" samples.
2. The out-of-bag samples can then be used as a validation set. We can pass them through the tree that didn't see them during training and obtain predictions.
3. These predictions are then compared to the actual values to compute an "out-of-bag score", which can be thought of as an estimate of the prediction error on unseen data.

One of the advantages of the out-of-bag score is that it allows us to estimate the prediction error without needing a separate validation set. This can be particularly useful when the dataset is small and partitioning it into training and validation sets might leave too few samples for effective learning.



Test

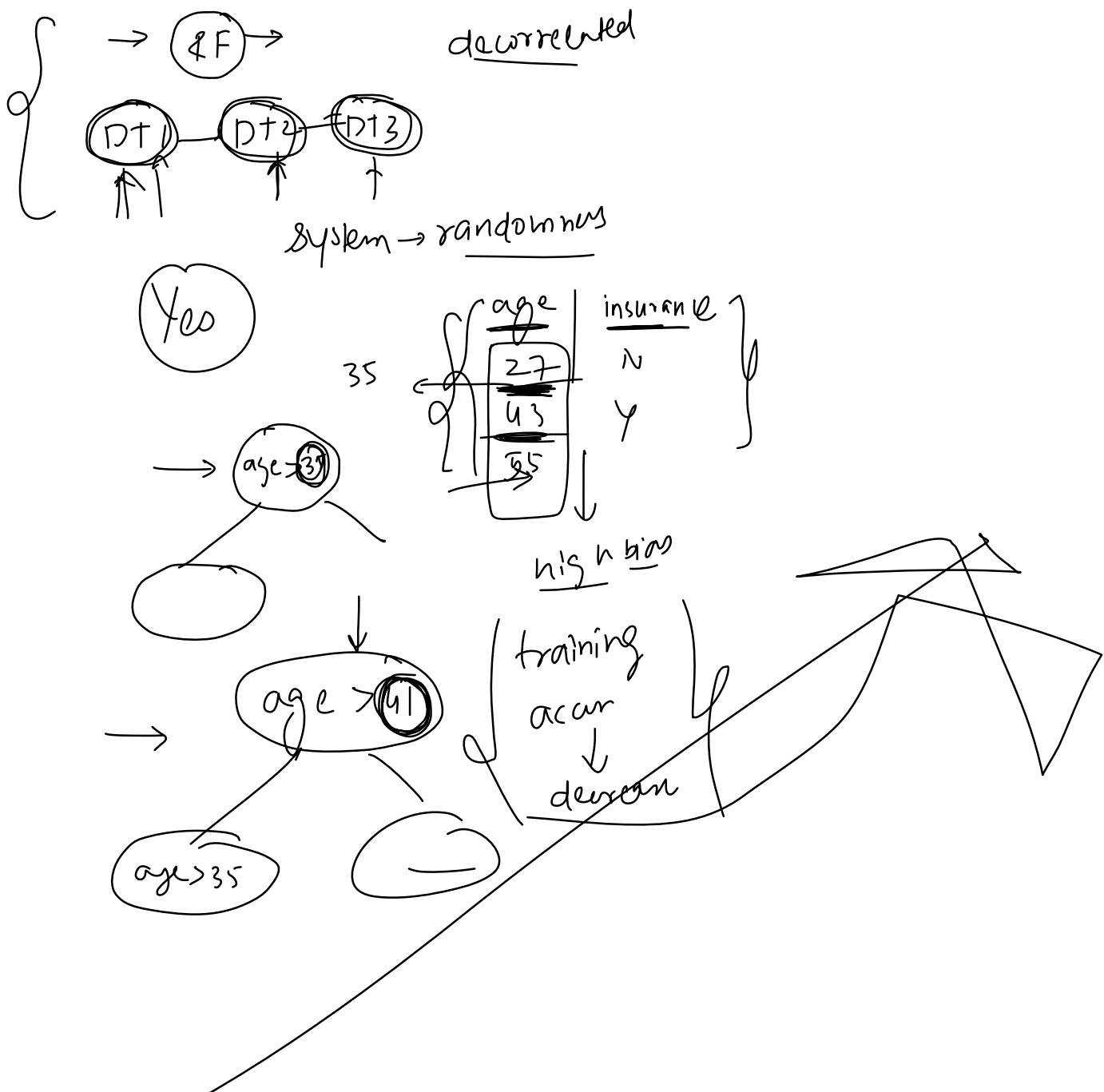
[5 datasets] (1000 rows)

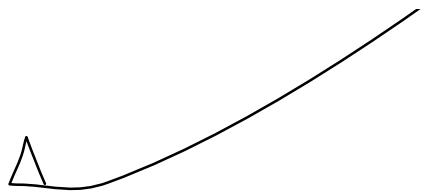
they are not a
part of training

Extra Trees is short for "Extremely Randomized Trees". It's a modification of the Random Forest algorithm that changes the way the splitting points for decision tree branches are chosen.

In traditional decision tree algorithms (and therefore in Random Forests), the optimal split point for each feature is calculated, which involves a degree of computation. For a given node, the feature and the corresponding optimal split point that provide the best split are chosen. On the other hand, in the Extra Trees algorithm, for each feature under consideration, a split point is chosen completely at random. The best-performing feature and its associated random split are then used to split the node. This adds an extra layer of randomness to the model, hence the name "Extremely Randomized Trees".

Because of this difference, Extra Trees tend to have more branches (be deeper) than Random Forests, and the splits are made more arbitrarily. This can sometimes lead to models that perform better, especially on tasks where the data may not have clear optimal split points. However, like all models, whether Extra Trees will outperform Random Forests (or any other algorithm) depends on the specific dataset and task.





Advantages and Disadvantages

31 July 2023 08:53

Advantages

- Robustness to Overfitting: Random Forests are less prone to overfitting compared to individual decision trees, because they average the results from many different trees, each of which might overfit the data in a different way.
- Handling Large Datasets: They can handle large datasets with high dimensionality effectively.
- Less Pre-processing: Random Forests can handle both categorical and numerical variables without the need for scaling or normalization. They can also handle missing values.
- Variable Importance: They provide insights into which features are most important in the prediction.
- Parallelizable: The training of individual trees can be parallelized, as they are independent of each other. This speeds up the training process.
- Non-Parametric: Random Forests are non-parametric, meaning they make no assumptions about the functional form of the transformation from inputs to output. This makes them very flexible and able to model complex, non-linear relationships.

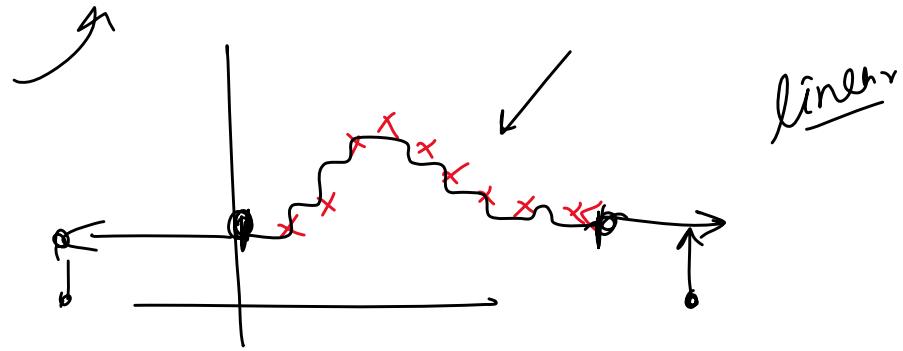
Linear
non-linear

Disadvantages

- Model Interpretability: One of the biggest drawbacks of Random Forests is that they lack the interpretability of simpler models like linear regression or decision trees. While you can rank features by their importance, the model as a whole is essentially a black box.
- Performance with Unbalanced Data: Random Forests can be biased towards the majority class when dealing with unbalanced datasets. This can sometimes be mitigated by balancing the dataset prior to training.
- Predictive Performance: Although Random Forests generally perform well, they may not always provide the best predictive performance. Gradient boosting machines, for instance, often outperform Random Forests. If the relationships within the data are linear, a linear model will likely perform better than a Random Forest.
- Inefficiency with Sparse Data: Random Forests might not be the best choice for sparse data or text data where linear models or other algorithms might be more suitable.
- Parameters Tuning: Although Random Forests require less tuning than some other models, there are still several parameters (like the number of trees, tree depth, etc.) that can affect model performance and need to be optimized.
- Difficulty with High Cardinality Features: Random Forests can struggle with high cardinality categorical features (features with a large number of distinct values). These types of features can lead to trees that are biased towards the variables with more levels, and may cause overfitting.
- Can't Extrapolate - This is because they do not predict beyond the range of the training data, and that they may not predict as accurately as other regression models.

OHE

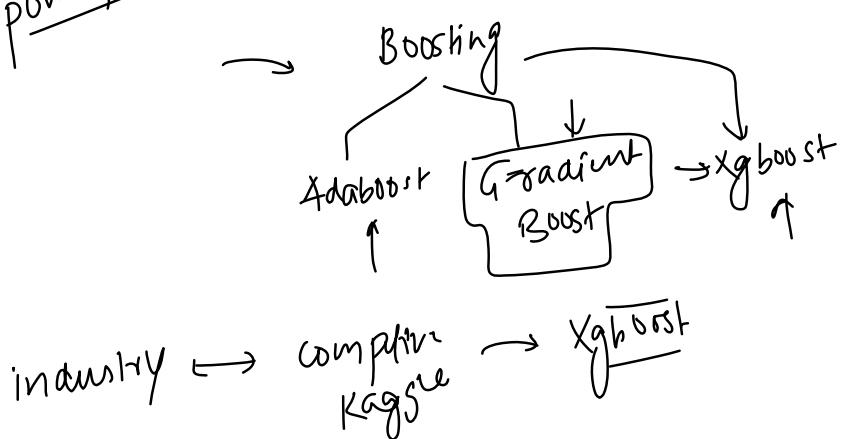
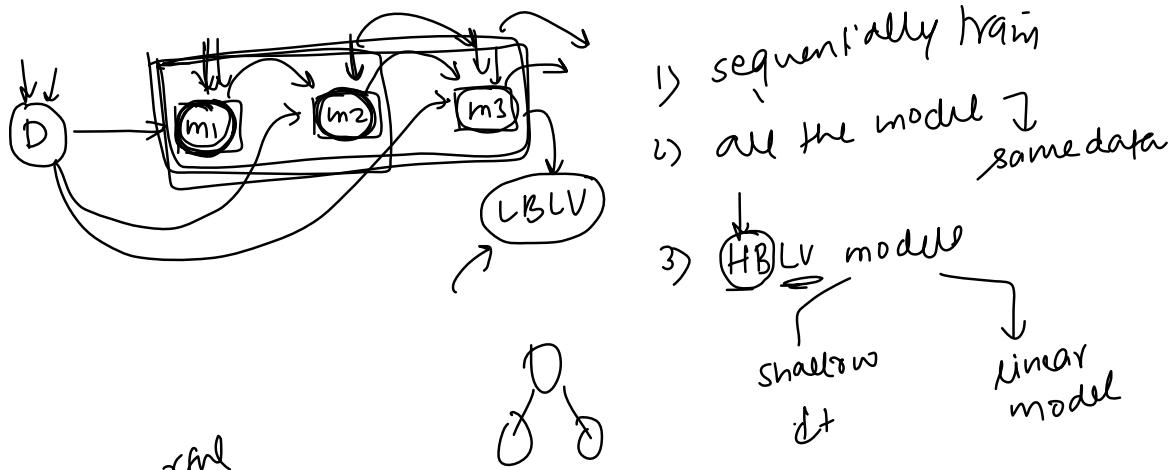
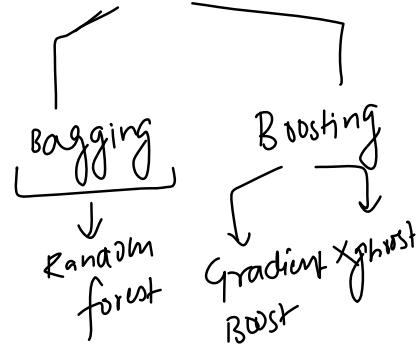
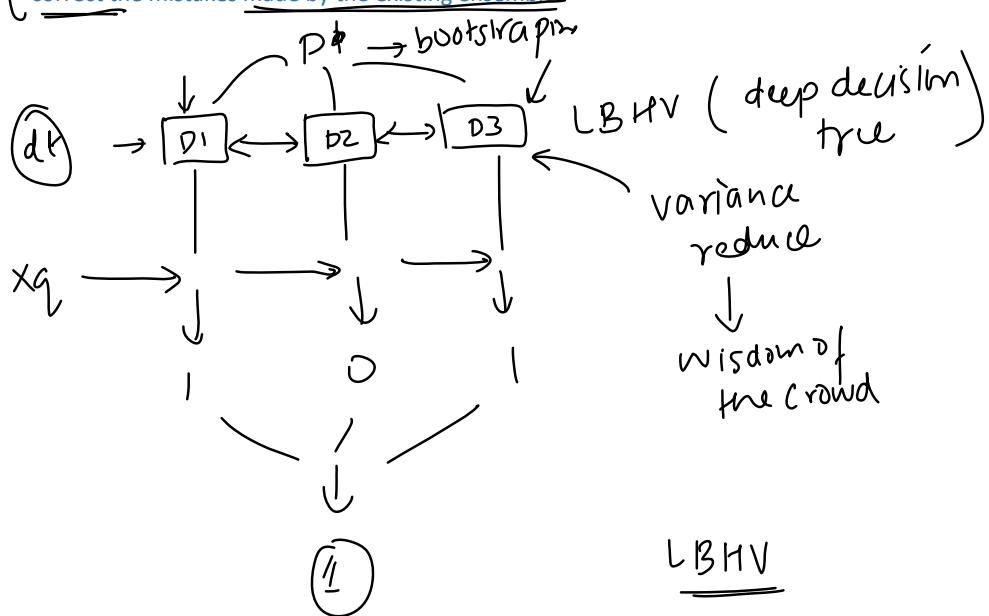
L data, and that they may not predict as accurately as other regression models.



Boosting

02 August 2023 15:55

Boosting is a general ensemble method in machine learning that aims to create a strong classifier or regressor by combining the predictions of several weaker models. The idea is to build the strong model incrementally, by sequentially adding weak models that are trained to correct the mistakes made by the existing ensemble.



What is Gradient Boosting

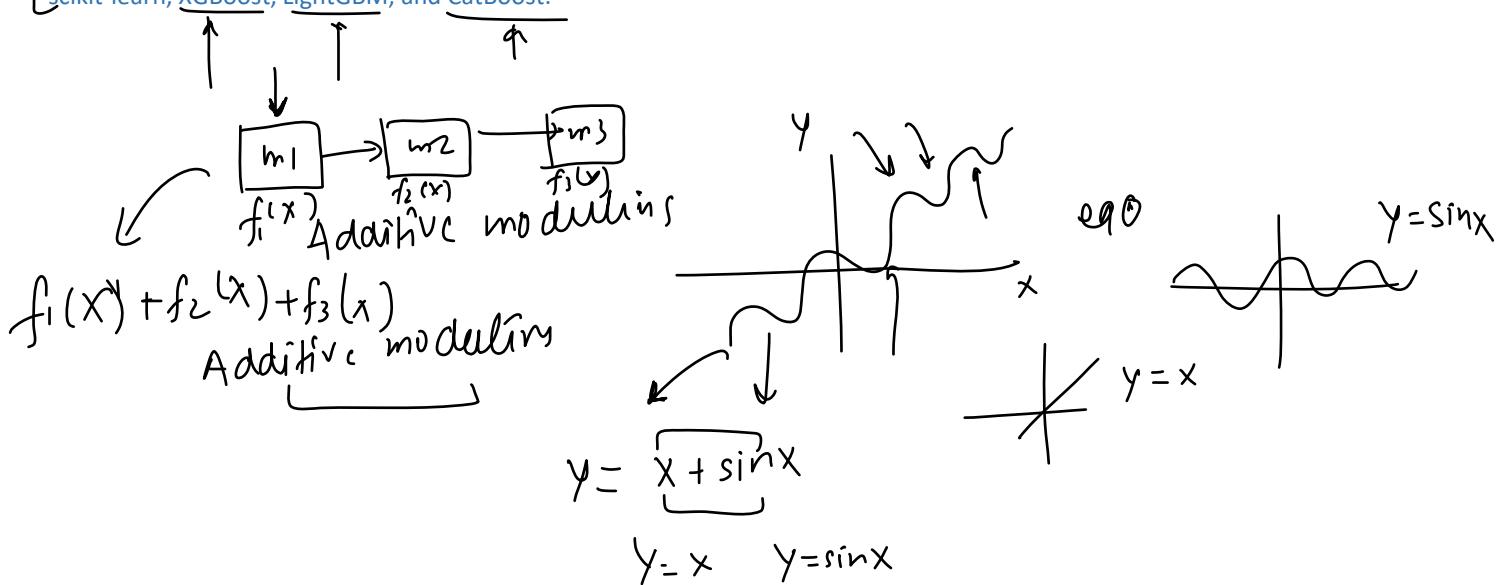
02 August 2023 07:38

Gradient Boosting is a machine learning ensemble technique that aims to build a strong predictive model by combining the predictions of several weaker models using the concept of Additive Modelling, typically decision trees. The method works by iteratively adding models to the ensemble, with each new model trained to correct the mistakes made by the combined ensemble of existing models.

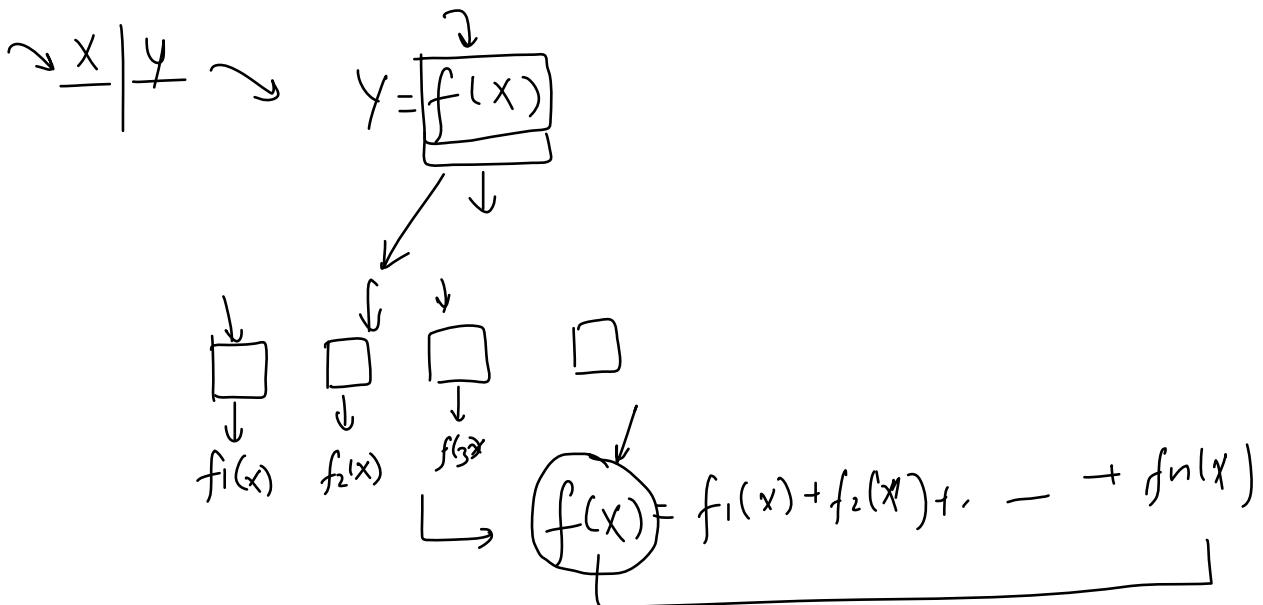
Gradient Boosting is a powerful and flexible method that can be used for both regression and classification tasks. → handwriting systems

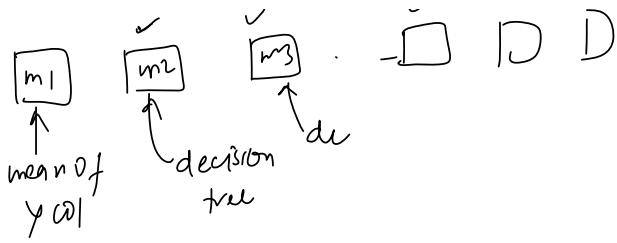
It is particularly effective when the data has complex, non-linear relationships, and it often performs well even with little hyperparameter tuning.

Its popularity in real-world applications and machine learning competitions is testament to its effectiveness, and it has implementations in most major machine learning libraries, such as scikit-learn, XGBoost, LightGBM, and CatBoost.



Additive modelling



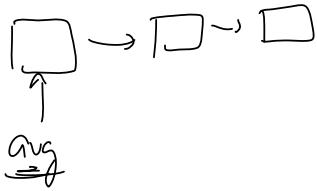


$m=1$

$$\tau_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}$$

$$\tau_{i1} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_0} = \left[\frac{\partial L(y_i, f_0(x_i))}{\partial f_0(x_i)} \right]$$

$$L(y, \hat{y}) = \frac{1}{2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \tau_{i1} = \frac{1}{2} \sum_{i=1}^n (y_i - f_0(x_i))^2$$

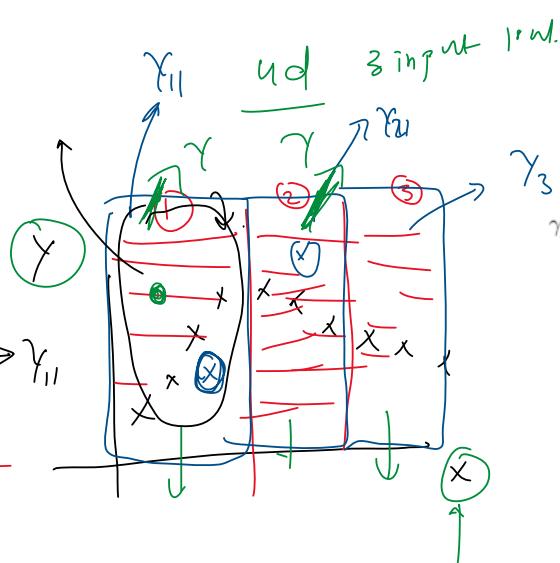
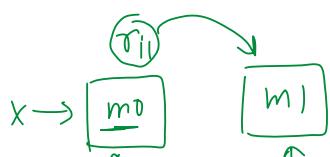
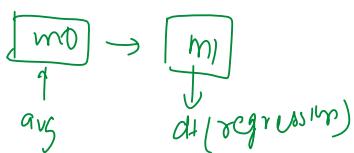


$$= - \sum (y_i - f_0(x_i)) \Rightarrow \tau_{i1} = f_0(x_i) - y_i$$

$$\tau_{11} = f_0(x_1) - y_1 = 142 - 192$$

$$\tau_{21} = f_0(x_2) - y_2 = 142 - 144$$

$$\tau_{31} = f_0(x_3) - y_3 = 144 - 91$$



$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

γ_1

m=1 x γ_{21}

$x_i \in R_{j1} \rightarrow$ terminal region

$$\underline{\gamma}_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

$$\begin{aligned} \underline{\gamma}_{j1} &= \arg \min_{\gamma} \sum_{x_i \in R_{j1}} L(y_i, f_0(x_i) + \gamma) \\ \underline{\gamma}_{11} &= \arg \min_{\gamma} \sum_{x_3 \in R_{11}} L(y_3, f_0(x_3) + \gamma) \end{aligned}$$

$$\begin{aligned} L &= (y_i - \hat{y}_i)^2 \\ L &= (y_3 - f_0(x_3) + \gamma)^2 \end{aligned}$$

$$\underline{\gamma}_{11} = \arg \min_{\gamma} (y_3 - f_0(x_3) - \gamma)^2$$

$$= \frac{\partial}{\partial \gamma} (y_3 - f_0(x_3) - \gamma)^2 \Rightarrow -2(y_3 - f_0(x_3) - \gamma) = 0$$

$$y_3 - f_0(x_3) - \gamma = 0$$

$$\begin{aligned} \gamma &= y_3 - f_0(x_3) \\ &= \underline{q_1} - \underline{142} \end{aligned}$$

$$\boxed{\gamma_{11} = -51}$$

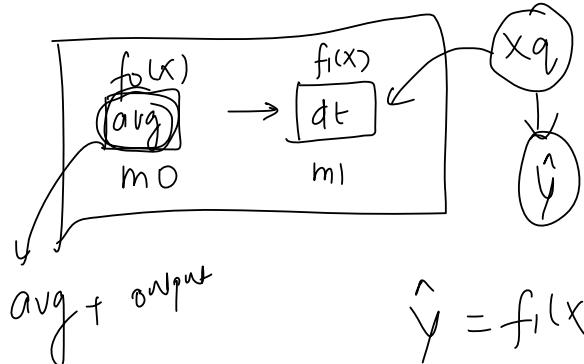
$$\boxed{\gamma = -51}$$

m=1

$$f_m(x) = f_{m-1}(x) + \left[\sum_{j=1}^{J_m} \underline{\gamma}_{jm} I(x \in R_{jm}) \right]$$

$$\begin{aligned} f_1(x) &= f_0(x) + \text{output of decision tree} \\ 142 + q_1 &= 142 + 1.66 = 143.6 \end{aligned}$$

141



$$\begin{aligned} \boxed{m_0} + \boxed{m_1} dt + \boxed{m_2} &\rightarrow \boxed{m_0 + m_1} + m_2 \\ &\downarrow \text{avg} \end{aligned}$$

$$f_m(x) = f_{m-1}(x) + \underbrace{\sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})}_{d+5}.$$

$$\underline{f_5(x)} = \underline{f_4(x)} + \underline{d+5}$$

$$\underline{f_4(x)} = \underline{f_3(x)} + \underline{d+4}$$

$$\underline{f_3(x)} + \underline{d+3}$$

$$\underline{f_2(x)} + \underline{d+2}$$

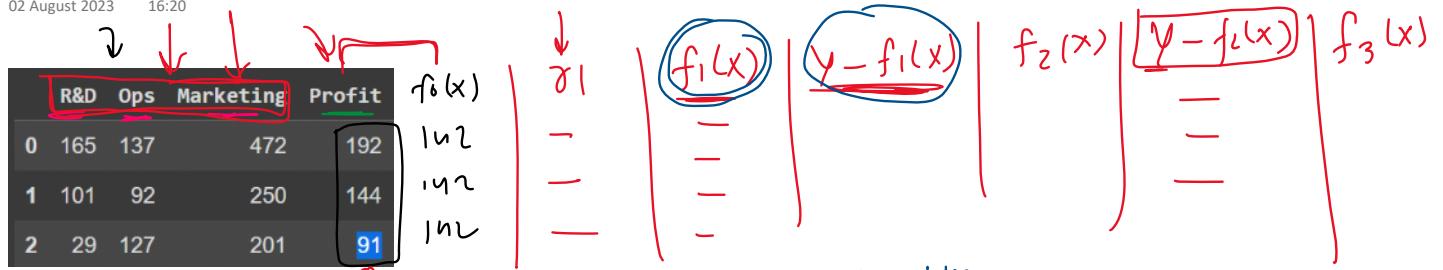
$$\boxed{\underline{f_1(x)} + \underline{d+1}}$$

$$f_5(x) = \underline{f_0(x)} + \underline{f_1(x)} + \underline{f_2(x)} + \dots$$

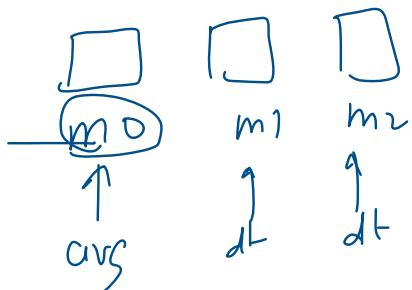
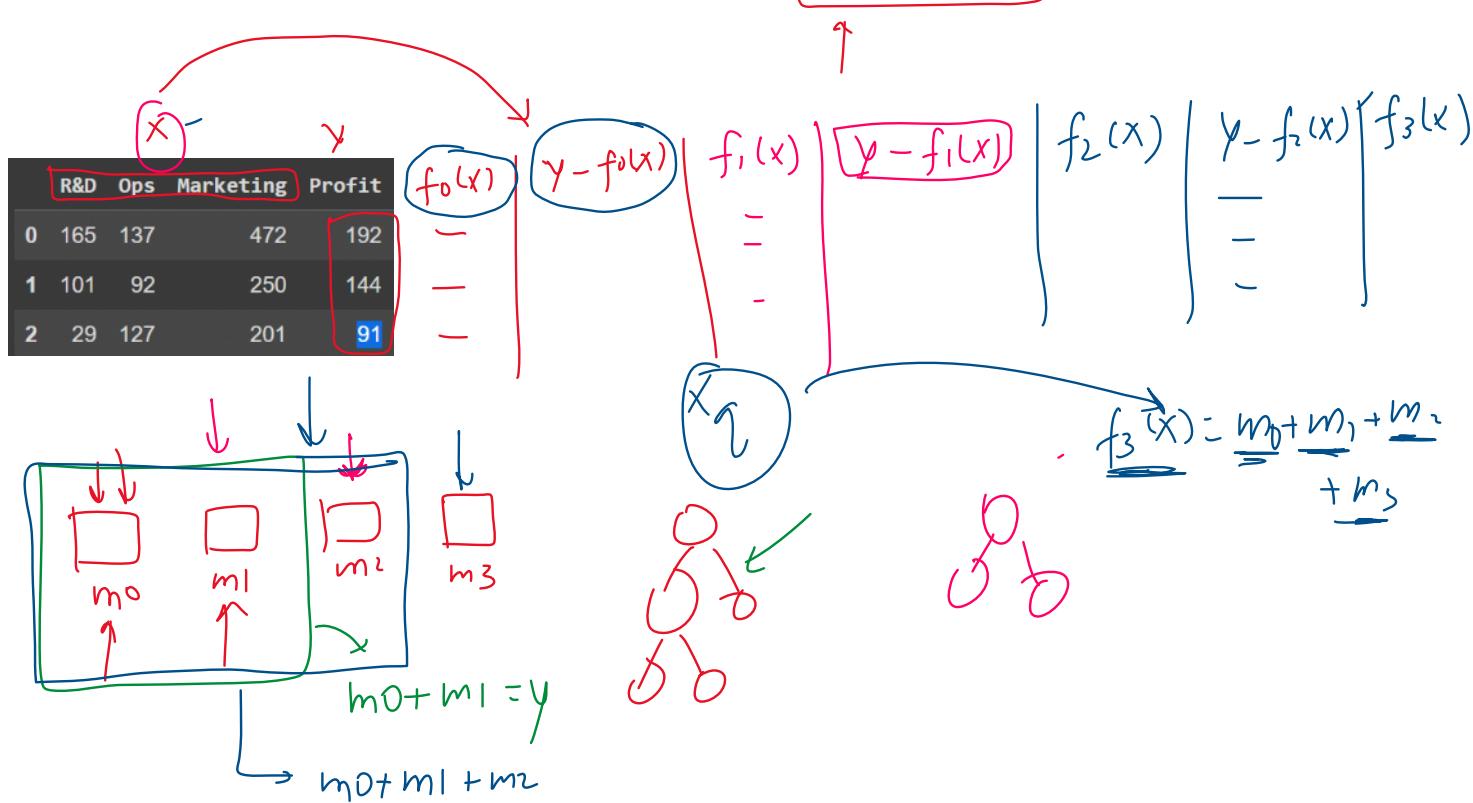
$$= f_0(x) + d+1 + d+2 + d+3 + \dots + d+4$$

The What?

02 August 2023 16:20



gradient
boosting
regression



The Why?

02 August 2023 16:20

1. Initialize $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$.

2. For $m = 1$ to M :

(a) For $i = 1, 2, \dots, N$ compute

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

(b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm} , $j = 1, 2, \dots, J_m$.

(c) For $j = 1, 2, \dots, J_m$ compute

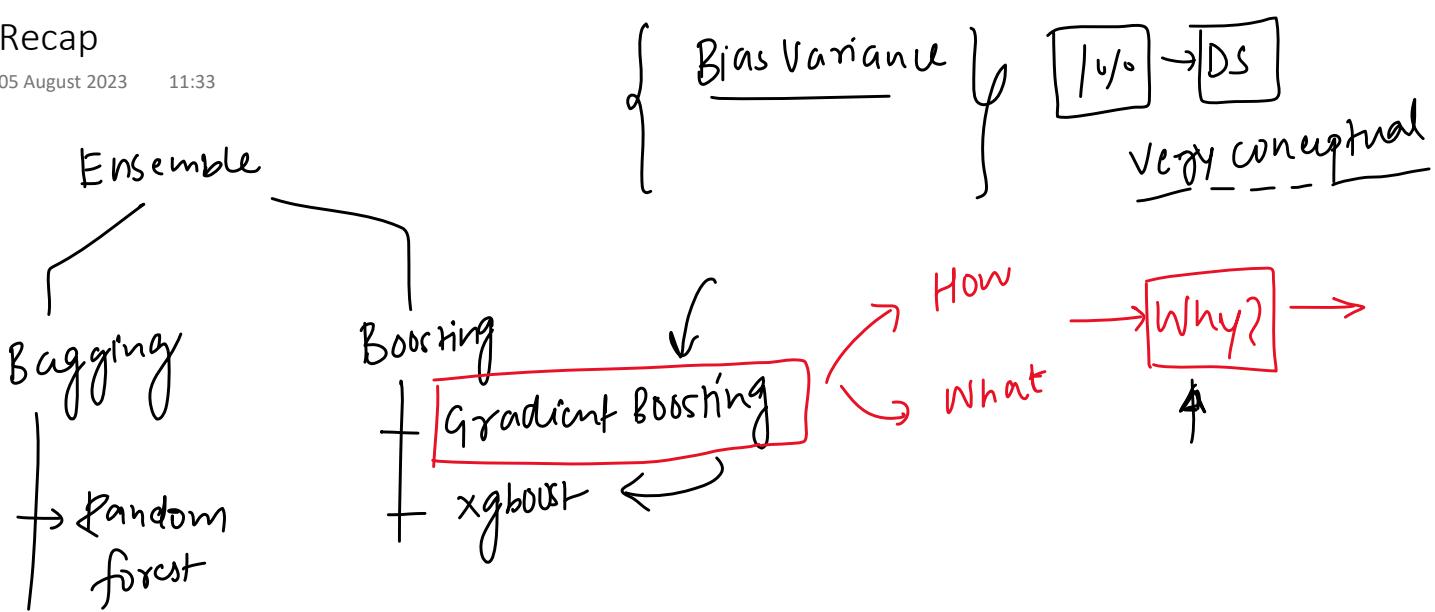
$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

3. Output $\hat{f}(x) = f_M(x)$.

Recap

05 August 2023 11:33



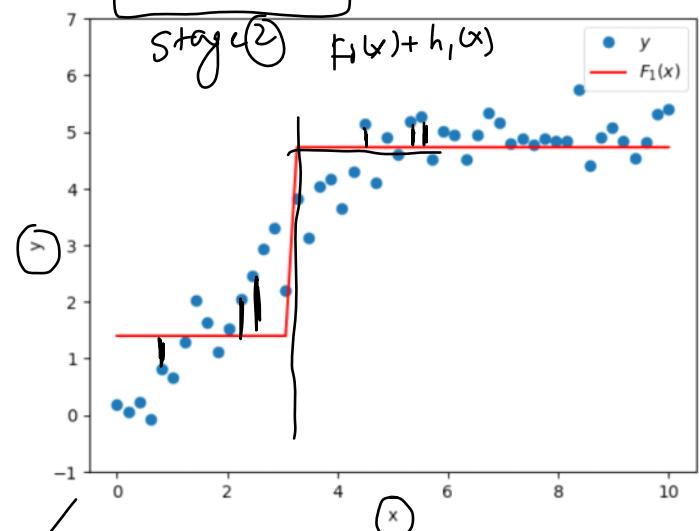
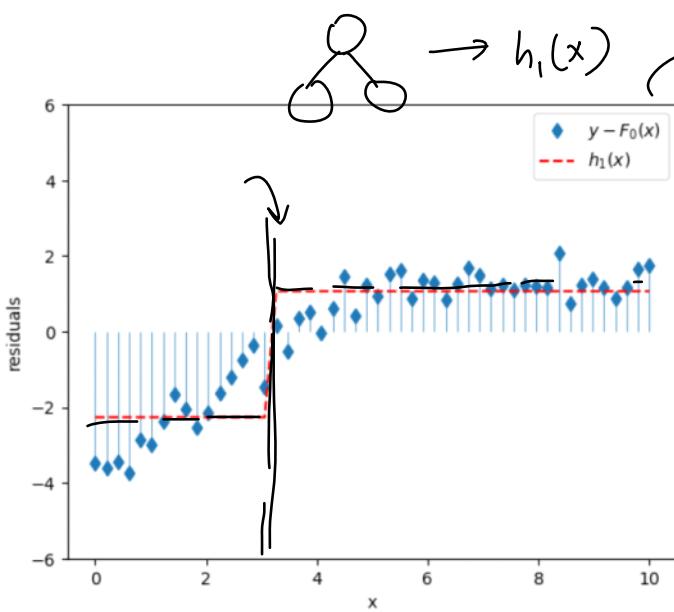
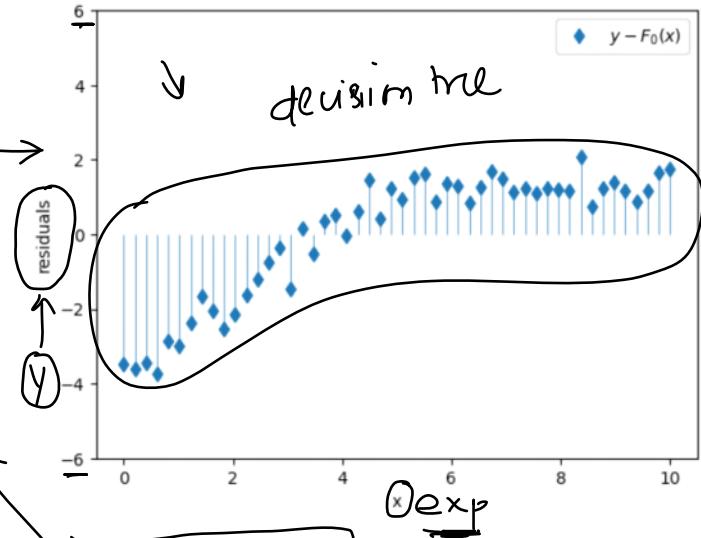
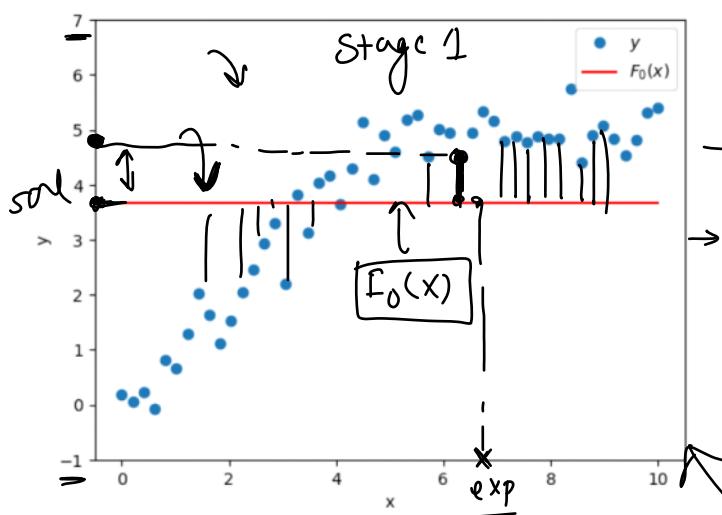
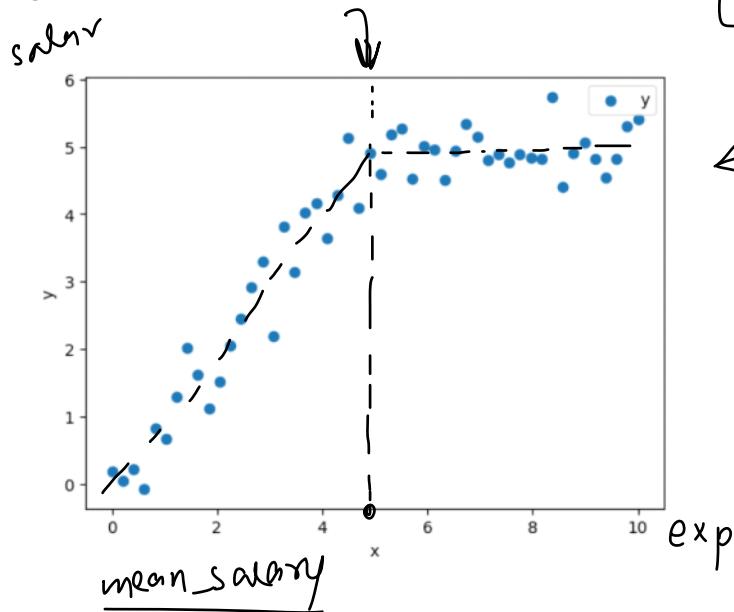
Today's class in one line

05 August 2023 11:34

Gradient
Boosting is
performing
Gradient
Descent in
Function Space

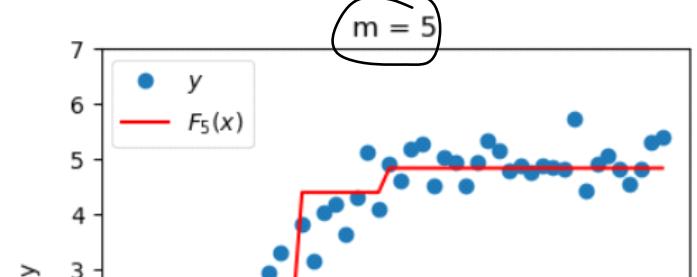
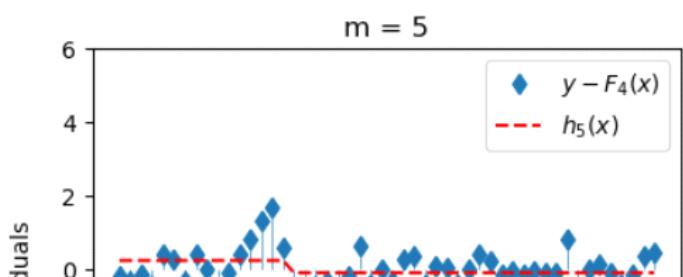
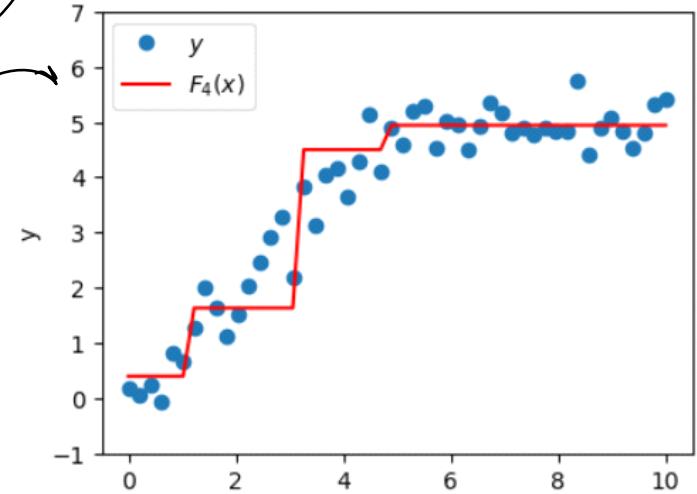
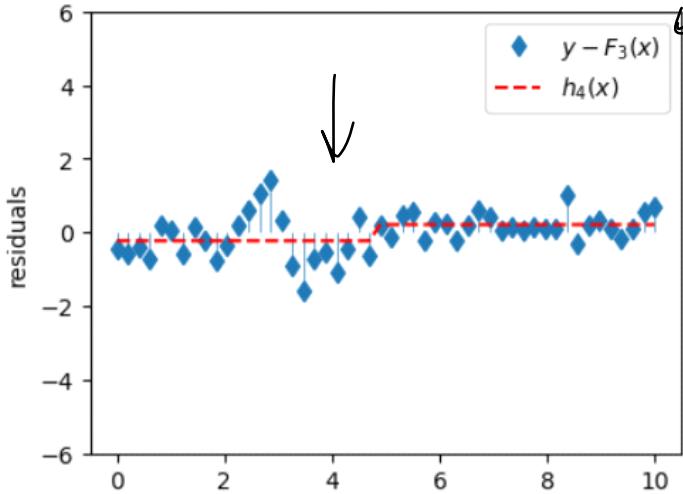
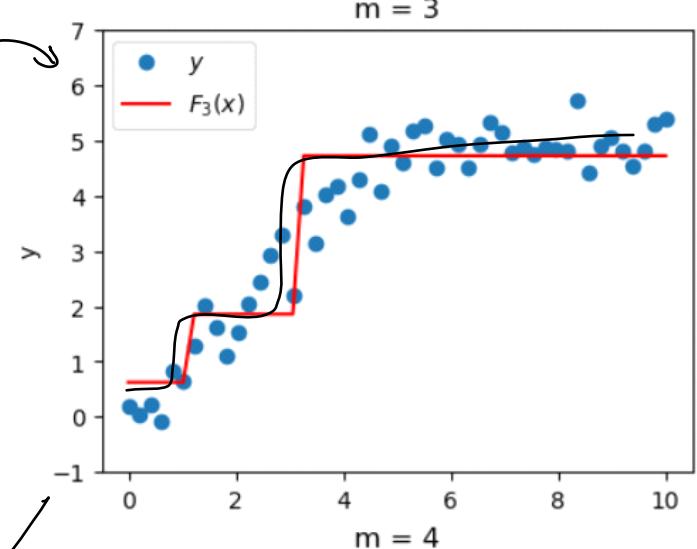
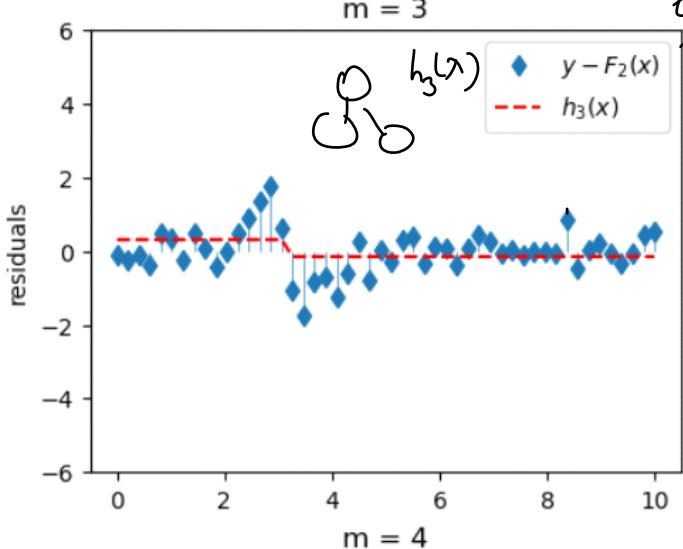
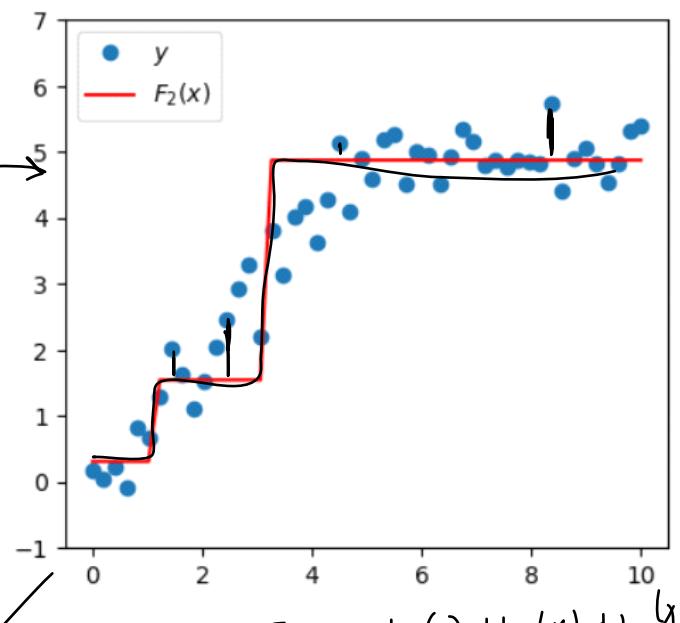
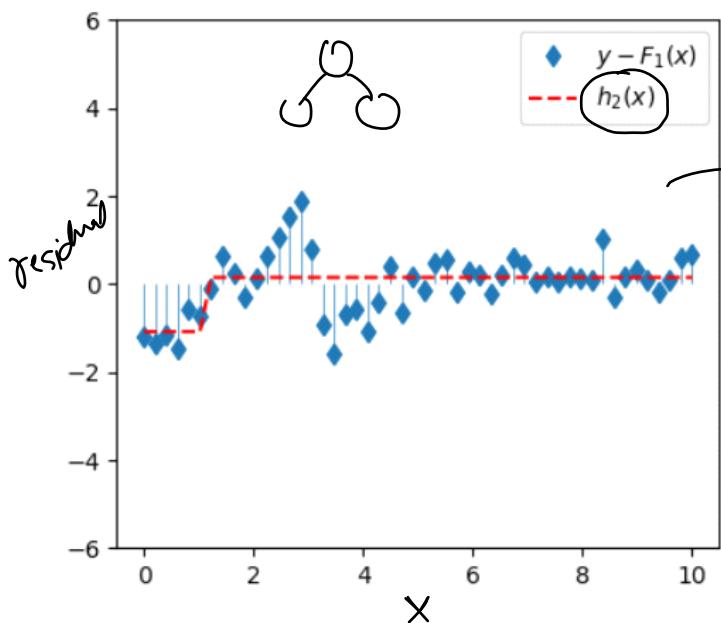
How Gradient Boosting Works?

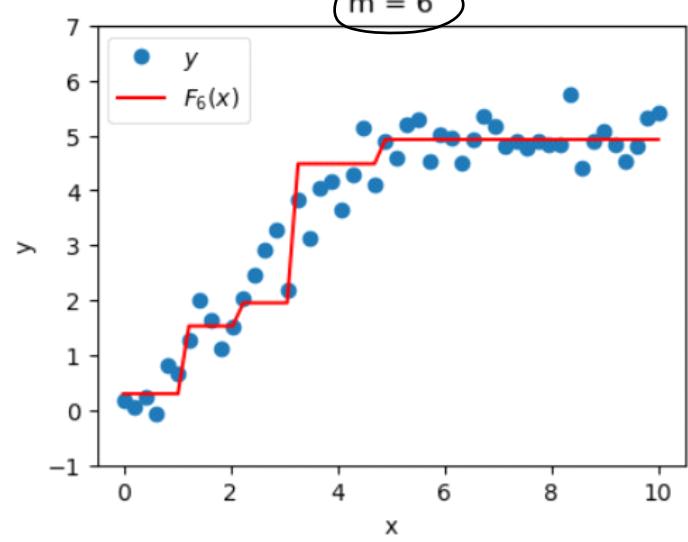
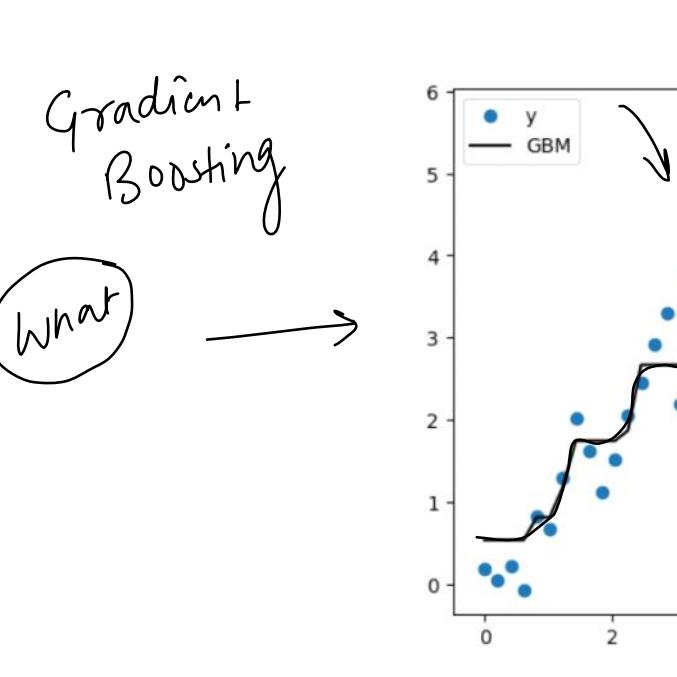
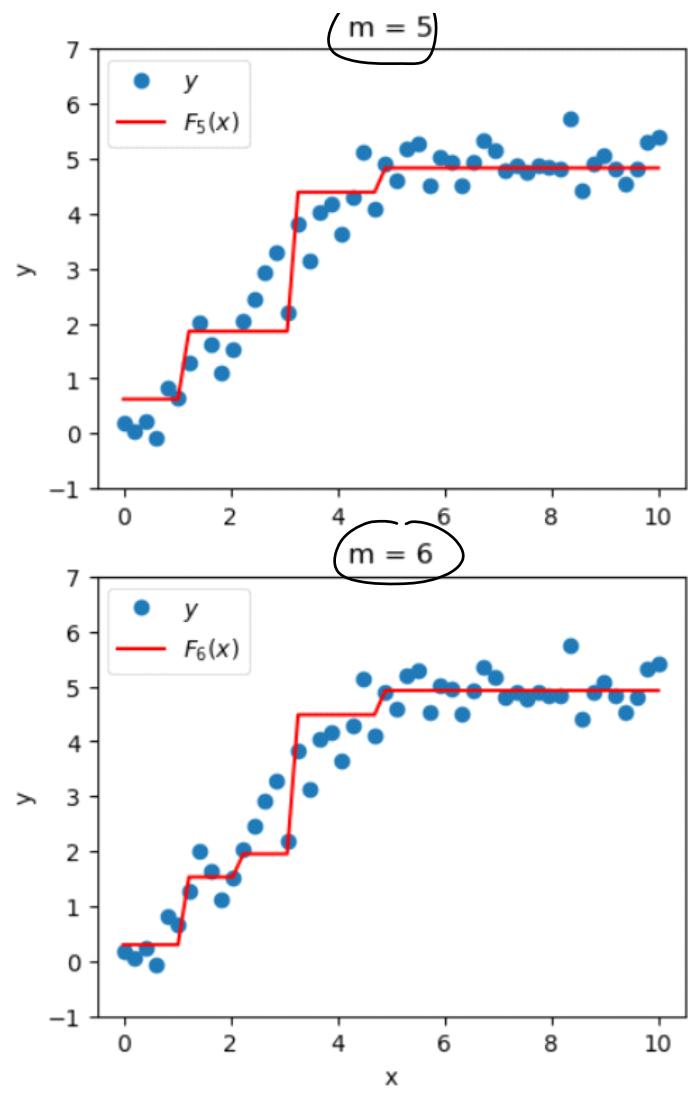
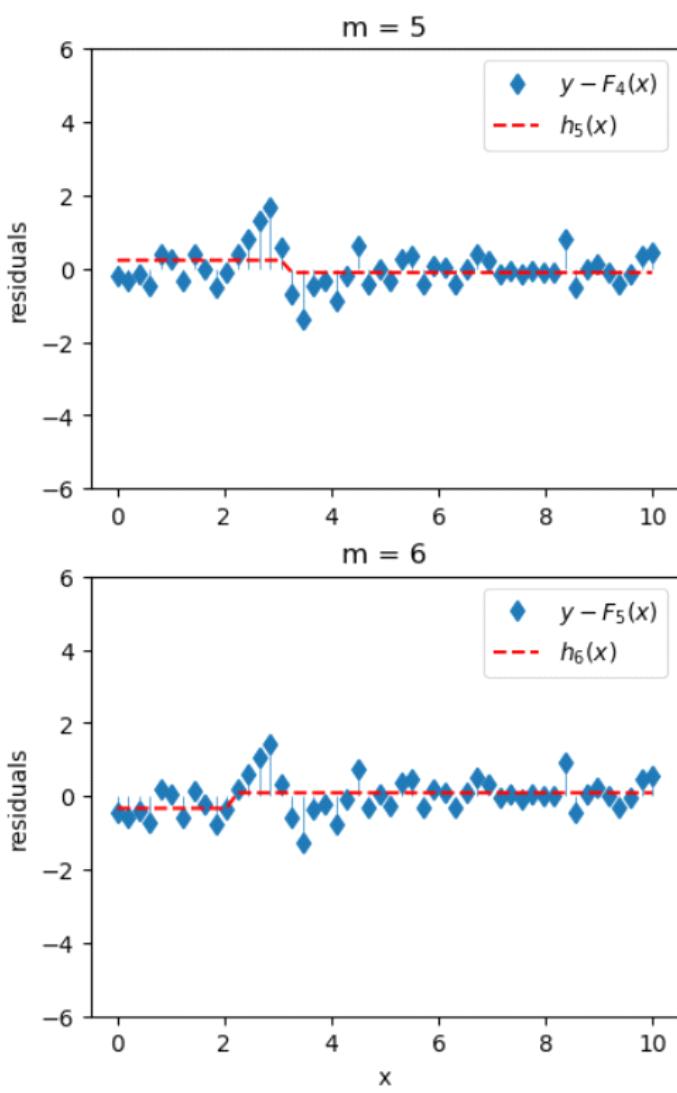
04 August 2023 14:58



$$F_2(x) = F_0(x) + h_1(x) + h_2(x)$$

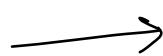






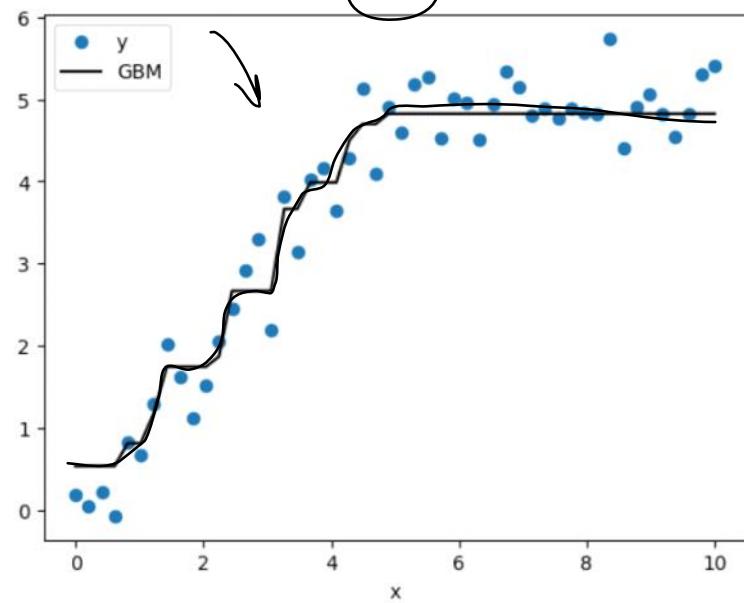
Gradient
Boosting

What



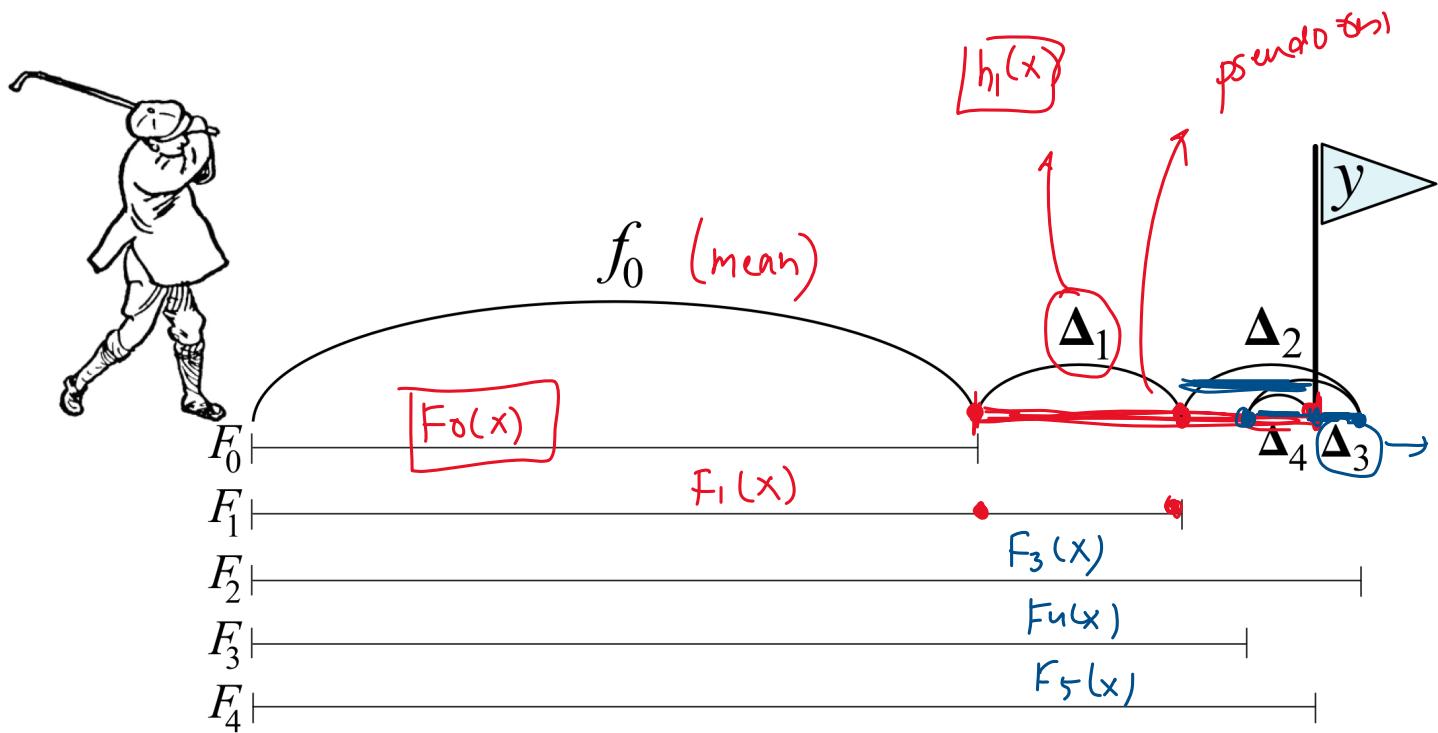
Q1

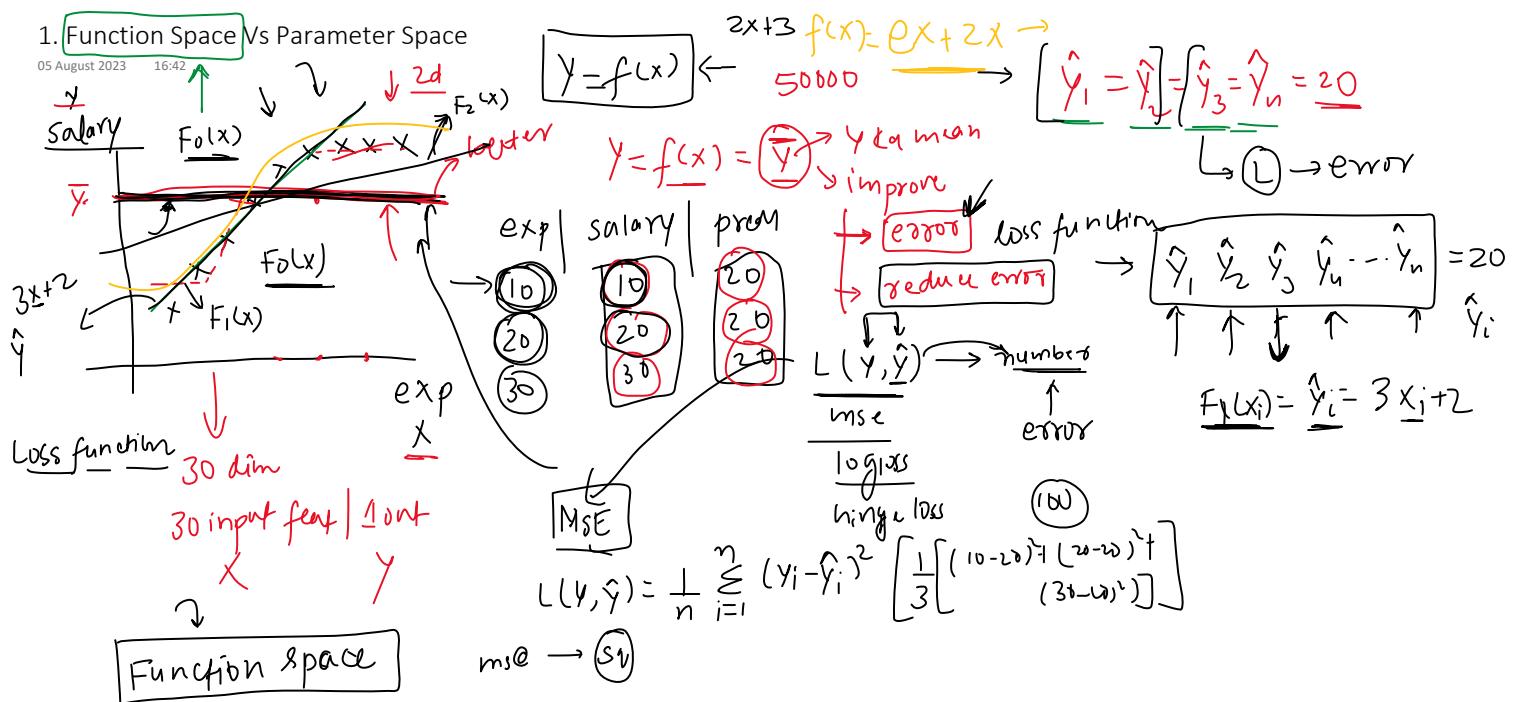
Revision



Intuition

04 August 2023 14:59



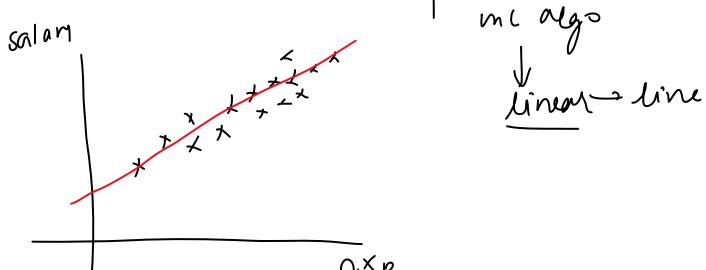
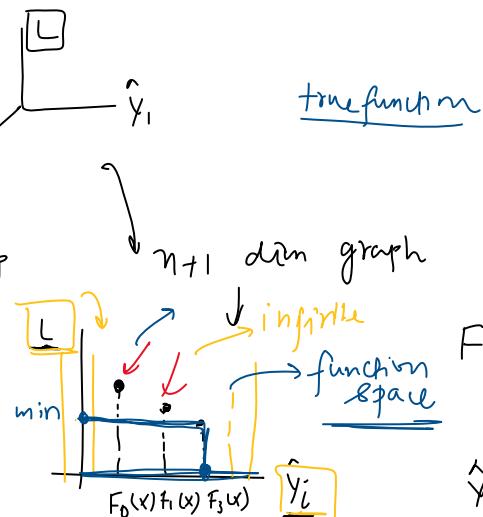
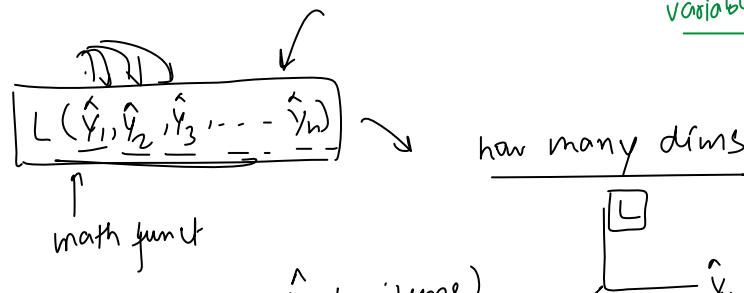


$$L(y, \hat{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\rightarrow L(y, \hat{y}) = (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \dots + (\hat{y}_n - y_n)^2$$

variables → may vary (model)
function

Constant (data)



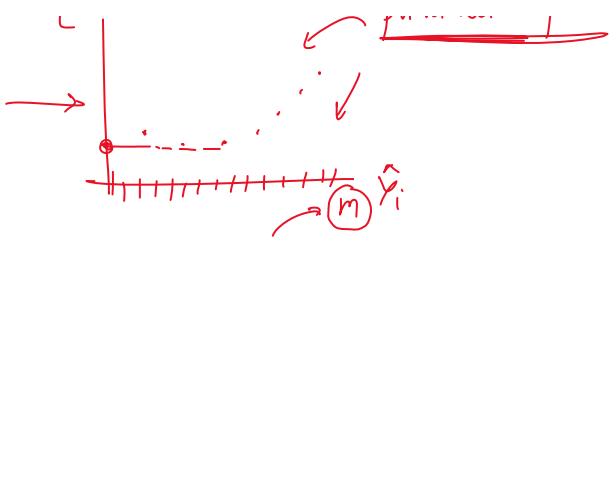
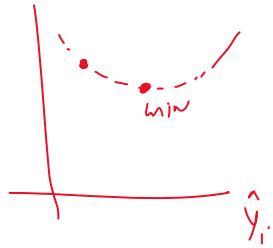
$$y = f(x) = mx + b = e^x \log x \sin x$$

parametric space

$$y = f(x) \quad \text{with } y = \log x \quad \sin x$$

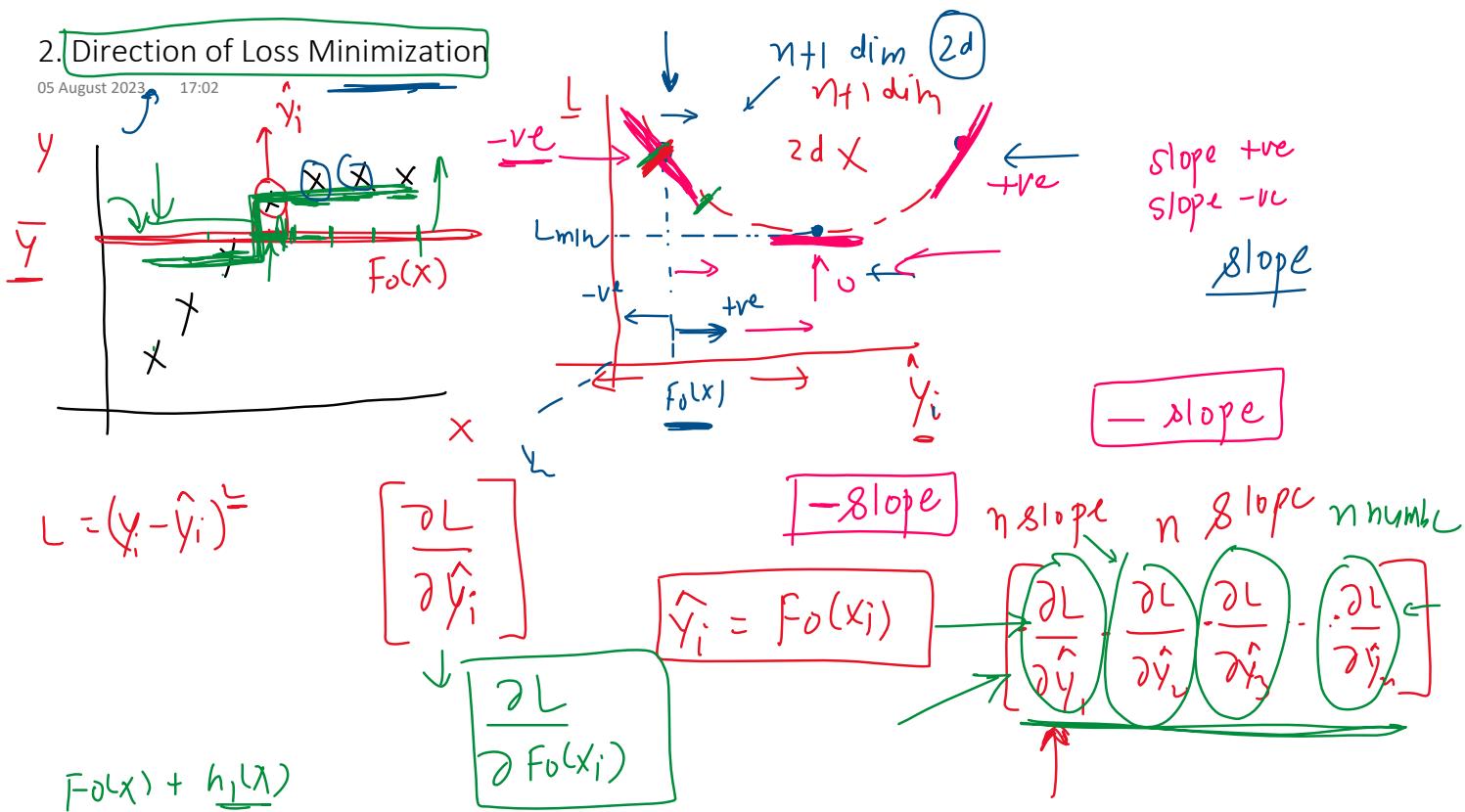
m, b

$L = (y - \hat{y})^2$



2. Direction of Loss Minimization

05 August 2023 17:02



$$\sigma_i = \frac{1}{2} \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_0}$$

$$= \frac{\partial}{\partial f(x_i)} \sum_{i=1}^n (y_i - f(x_i))^2 \Big|_{f=f_0}$$

$$[y_i - f(x_i)] \Big|_{f=f_0}$$

$$y_i - f_0(x_i) \rightarrow \text{slope}$$

$$y_1 - f_0(x_1)$$

$$y_2 - f_0(x_2)$$

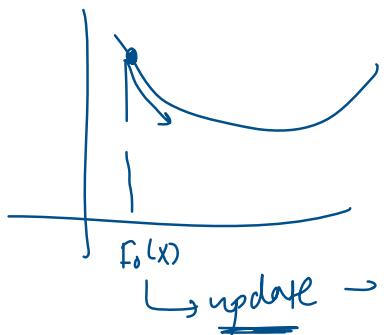
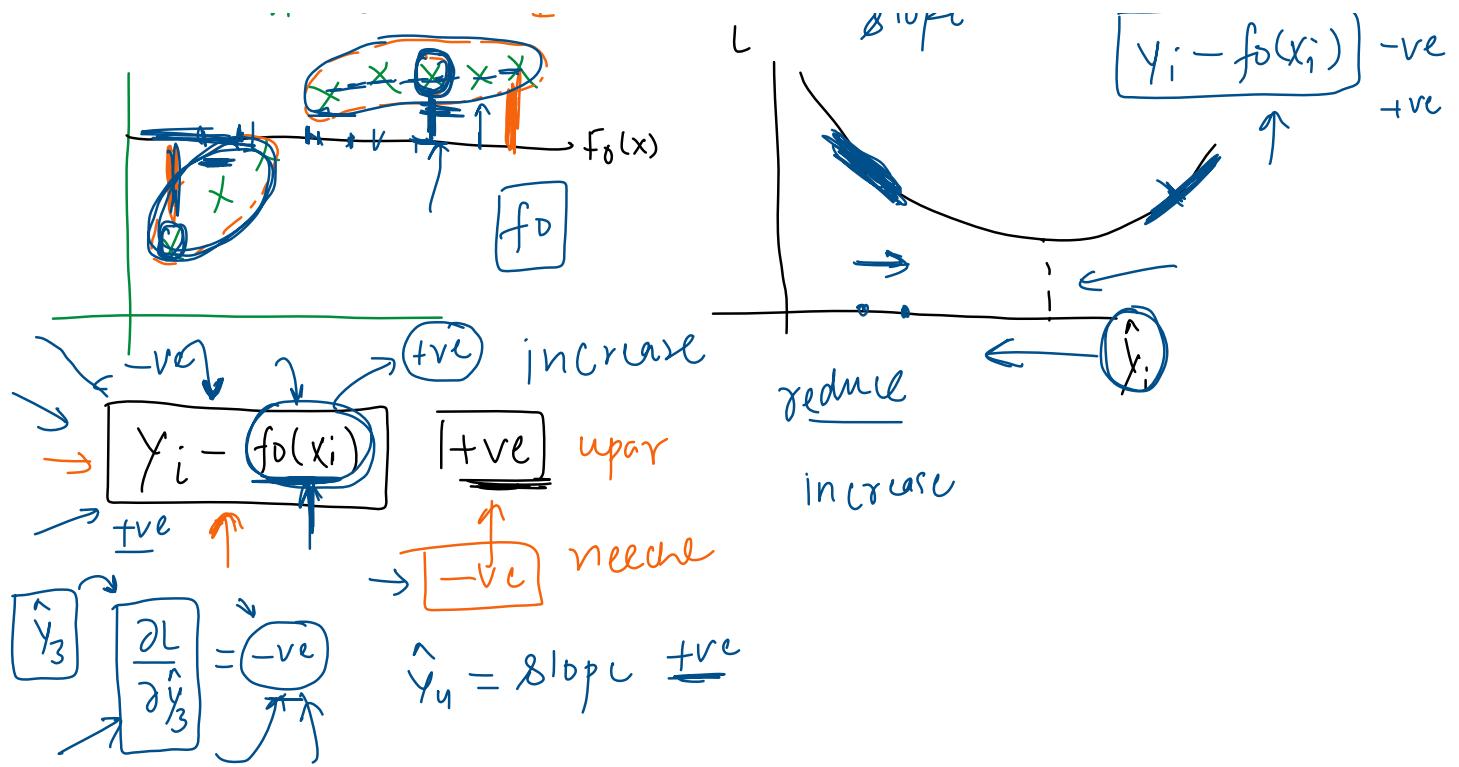
$$y_n - f_0(x_n)$$



$$L \downarrow \text{-slope}$$

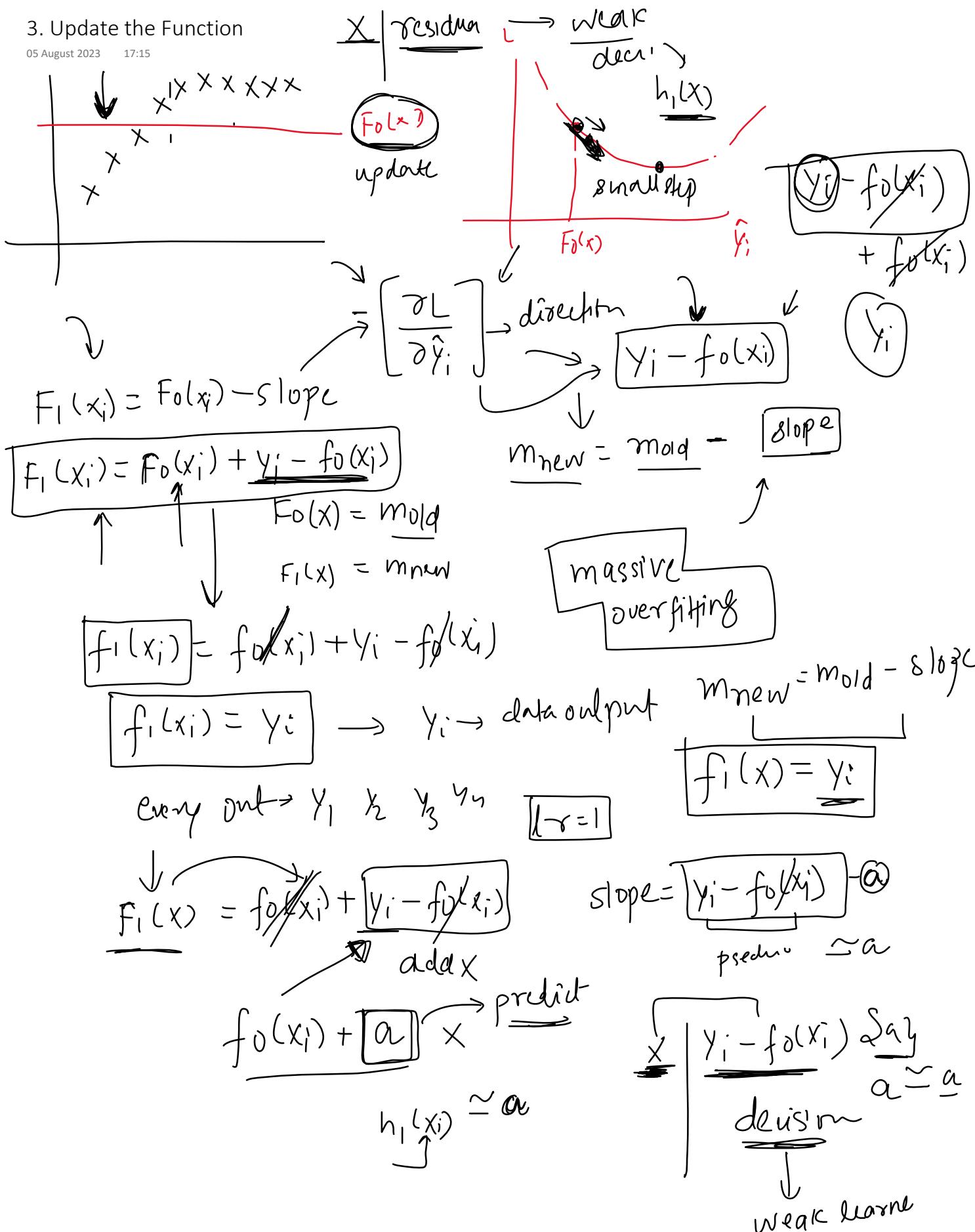
$$y_i - f_0(x_i)$$

-ve
+ve



3. Update the Function

05 August 2023 17:15



$$f_1(x) = f_0(x) + h_1(x)$$

right dir

$$f_i(x) = f_0(x) + h_i(x)$$

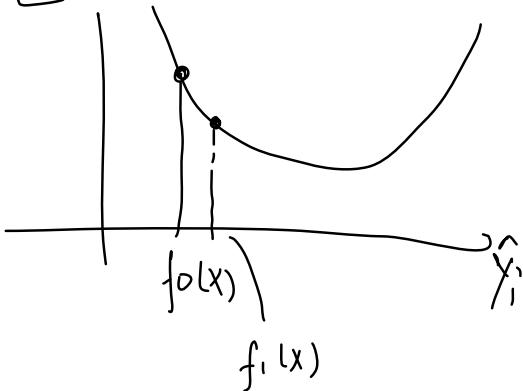
→ weak fit

$$\underline{x_i} \mid \overrightarrow{x_i - f_0(x)}$$

resid

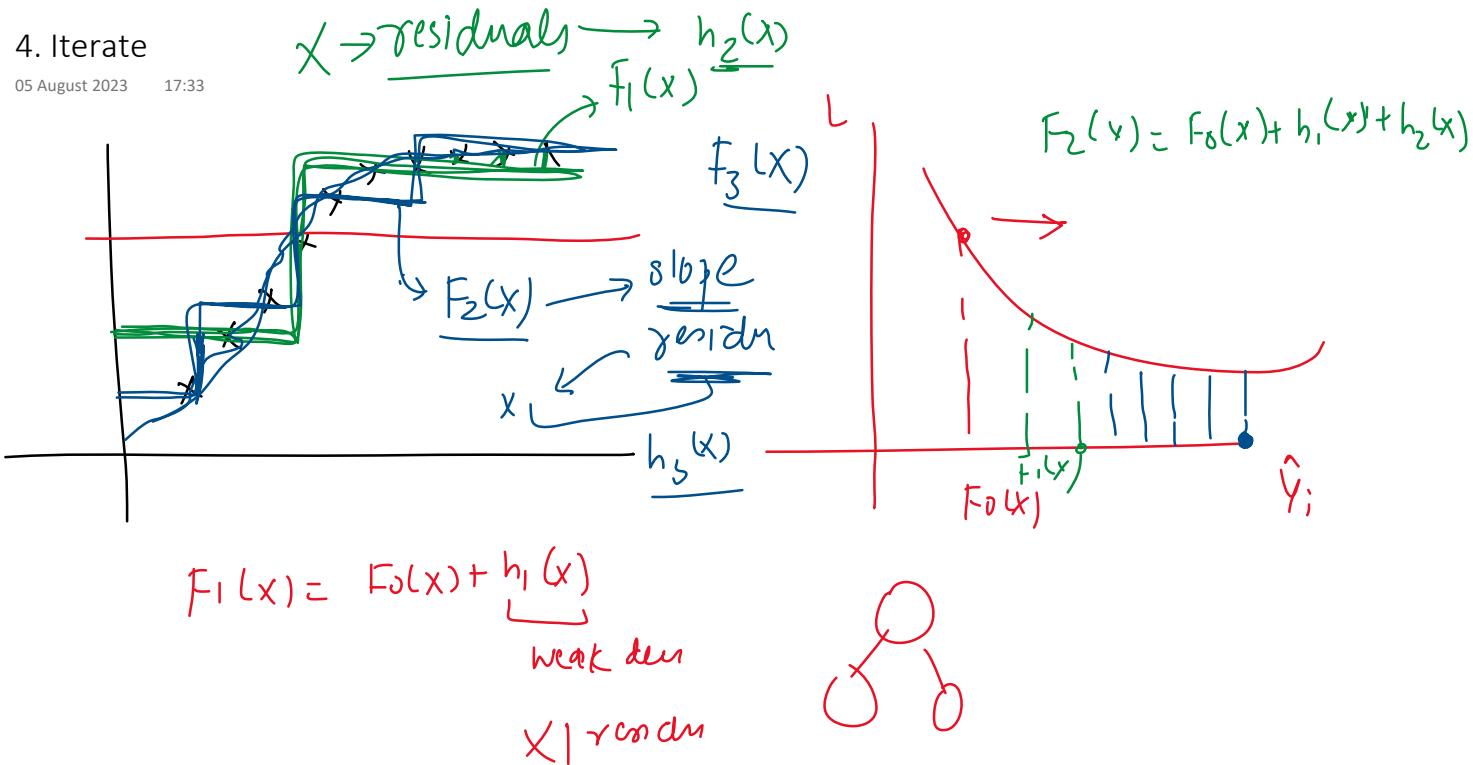
$$f_i(x) = f_0(x) + h_i(x)$$

L



4. Iterate

05 August 2023 17:33

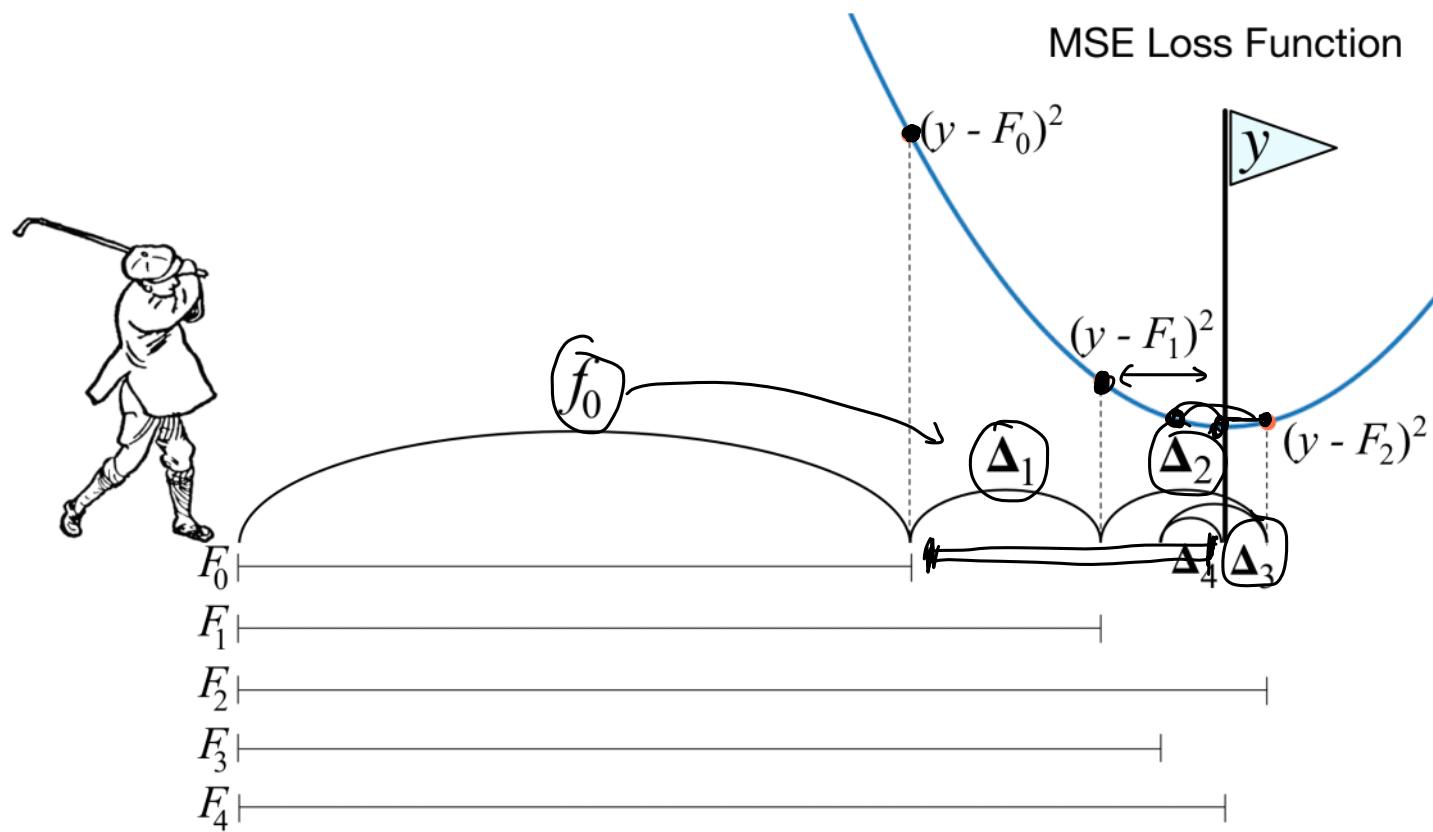


Detour - Gradient Descent

05 August 2023 14:10

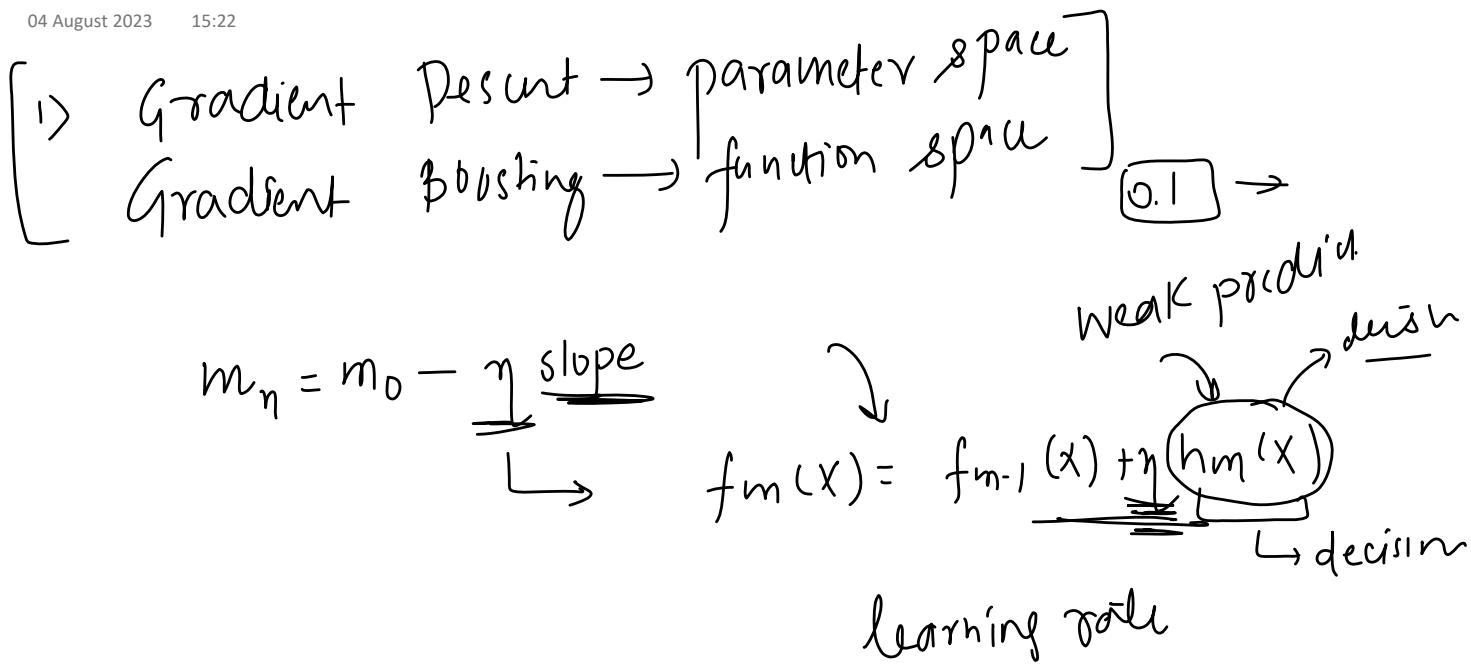
Another Perspective

05 August 2023 07:58



Difference Between Gradient Boosting and Gradient Descent

04 August 2023 15:22



Advantages of Gradient Boosting

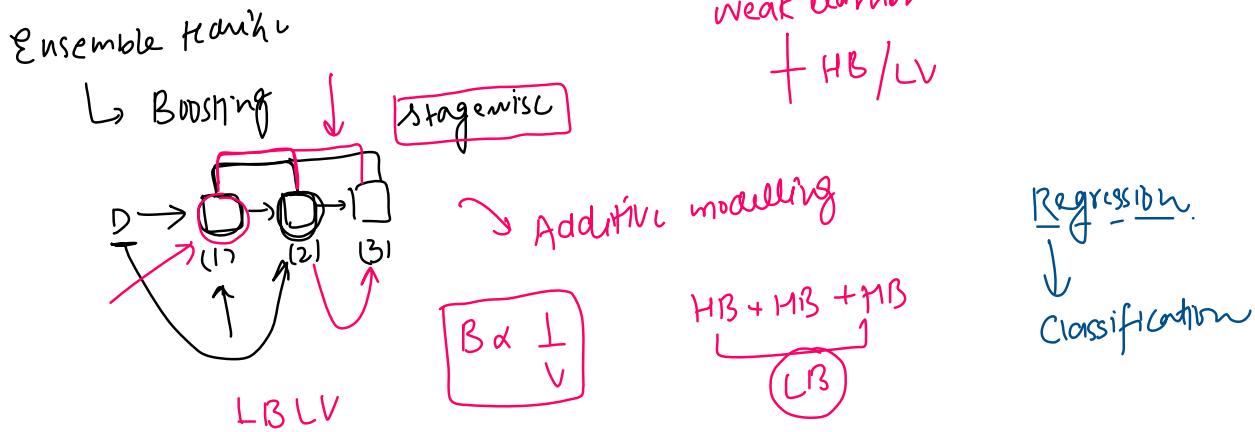
05 August 2023 07:52

Let's Appreciate

05 August 2023 18:31

Gradient Boosting

04 August 2023 20:07



Classification Vs Regression

04 August 2023 20:14

Gradient Boosting

mse / log loss

Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M .

$$1. \text{ Initialize } f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma).$$

2. For $m = 1$ to M :

(a) For $i = 1, 2, \dots, N$ compute

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

(b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm} , $j = 1, 2, \dots, J_m$.

(c) For $j = 1, 2, \dots, J_m$ compute

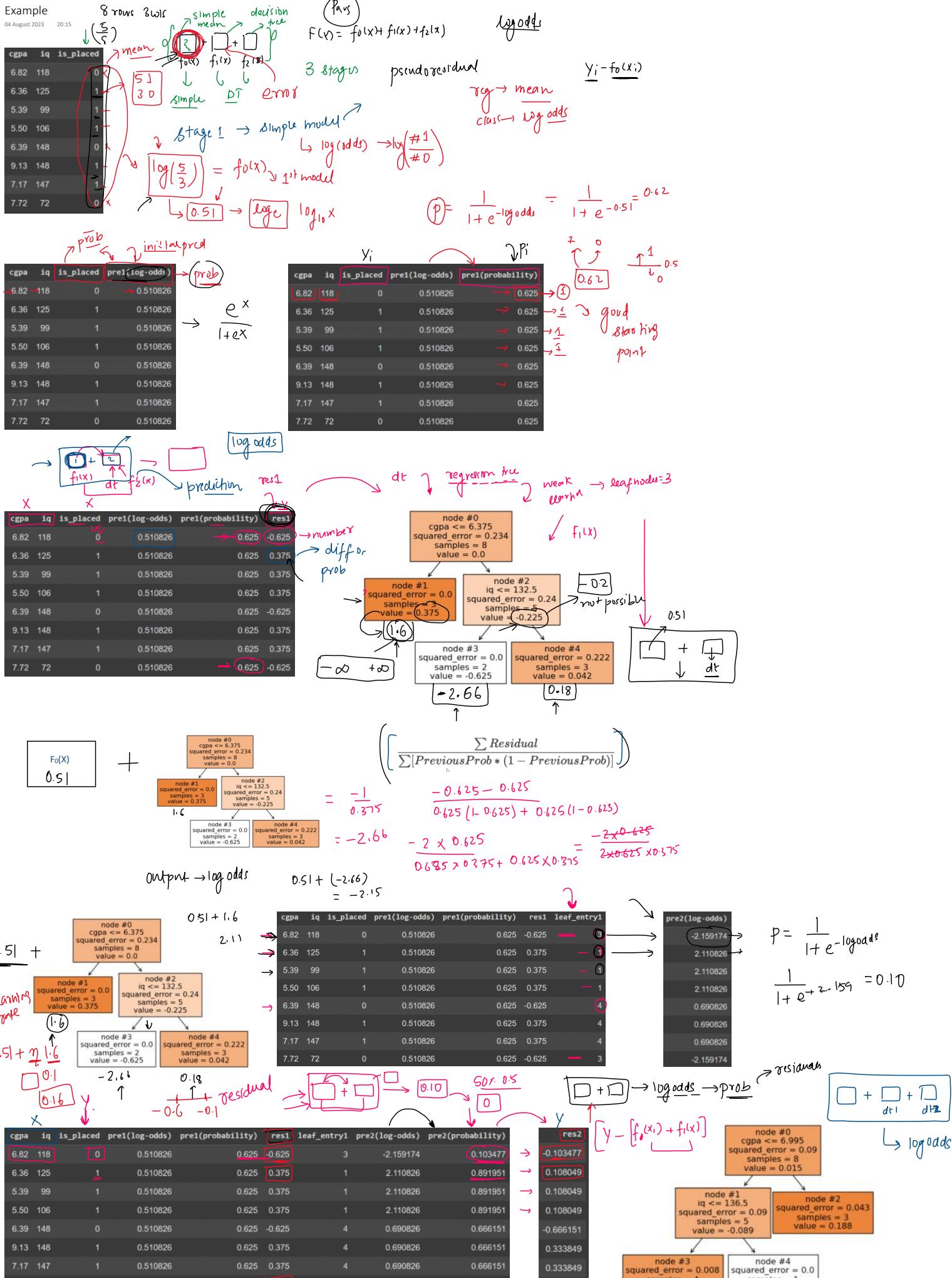
$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

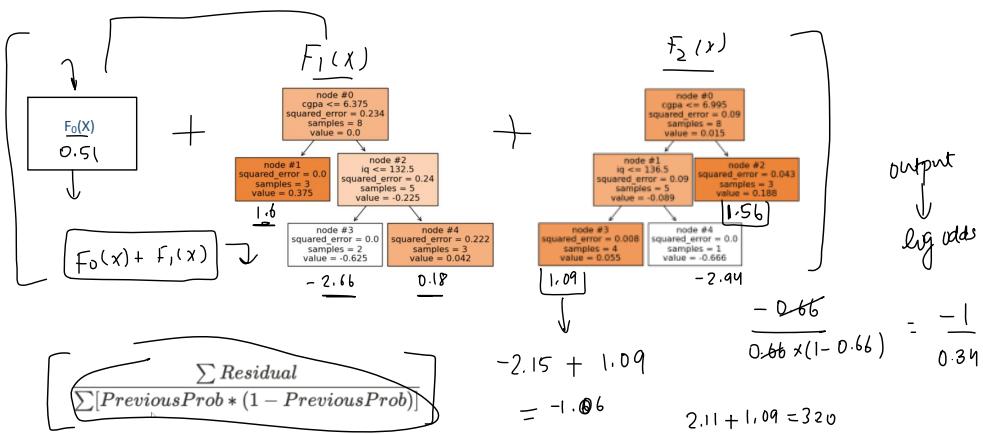
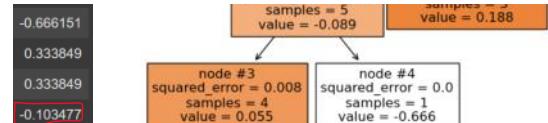
3. Output $\hat{f}(x) = f_M(x)$.

Example

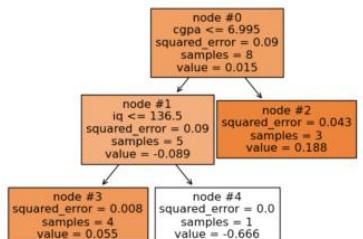
04 August 2023 20:15



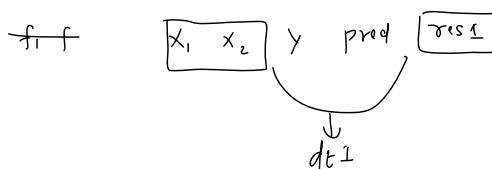
6.39	148	0	0.510826	0.625	-0.625	4	0.690826	0.666151
9.13	148	1	0.510826	0.625	0.375	4	0.690826	0.666151
7.17	147	1	0.510826	0.625	0.375	4	0.690826	0.666151
7.72	72	0	0.510826	0.625	-0.625	3	-2.159174	0.103477



cgpa	iq	is_placed	pre1(log-odds)	pre1(probability)	res1	leaf_entry1	pre2(log-odds)	pre2(probability)	res2	leaf_entry2	pre3(log-odds)	pre3(probability)
6.82	118	0	0.510826	0.625	-0.625	3	-2.159174	0.103477	-0.103477	3	-1.068349	0.255717
6.36	125	1	0.510826	0.625	0.375	1	2.110826	0.891951	0.108049	3	3.201651	0.960896
5.39	99	1	0.510826	0.625	0.375	1	2.110826	0.891951	0.108049	3	3.201651	0.960896
5.50	106	1	0.510826	0.625	0.375	1	2.110826	0.891951	0.108049	3	3.201651	0.960896
6.39	148	0	0.510826	0.625	-0.625	4	0.690826	0.666151	-0.666151	4	-1.798349	0.142052
9.13	148	1	0.510826	0.625	0.375	4	0.690826	0.666151	0.333849	2	2.251651	0.904793
7.17	147	1	0.510826	0.625	0.375	4	0.690826	0.666151	0.333849	2	2.251651	0.904793
7.72	72	0	0.510826	0.625	-0.625	3	-2.159174	0.103477	-0.103477	2	-0.598349	0.354722



cgpa	iq	is_placed	pre1(log-odds)	pre1(probability)	res1	leaf_entry1	pre2(log-odds)	pre2(probability)	res2	leaf_entry2	pre3(log-odds)	pre3(probability)
6.82	118	0	0.510826	0.625	-0.625	3	-2.159174	0.103477	-0.103477	3	-1.068349	0.255717
6.36	125	1	0.510826	0.625	0.375	1	2.110826	0.891951	0.108049	3	3.201651	0.960896
5.39	99	1	0.510826	0.625	0.375	1	2.110826	0.891951	0.108049	3	3.201651	0.960896
5.50	106	1	0.510826	0.625	0.375	1	2.110826	0.891951	0.108049	3	3.201651	0.960896
6.39	148	0	0.510826	0.625	-0.625	4	0.690826	0.666151	-0.666151	4	-1.798349	0.142052
9.13	148	1	0.510826	0.625	0.375	4	0.690826	0.666151	0.333849	2	2.251651	0.904793
7.17	147	1	0.510826	0.625	0.375	4	0.690826	0.666151	0.333849	2	2.251651	0.904793
7.72	72	0	0.510826	0.625	-0.625	3	-2.159174	0.103477	-0.103477	2	-0.598349	0.354722

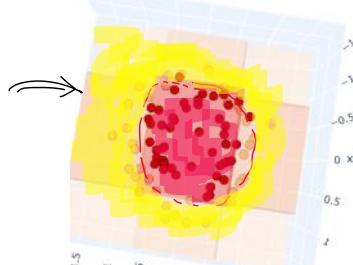


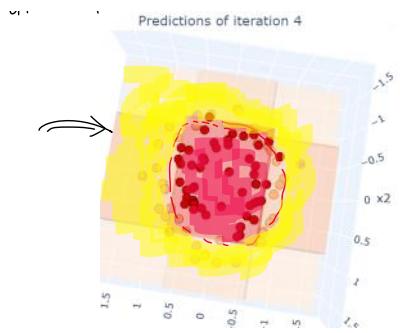
Purple $\rightarrow F_0(x)$
Yellow + Yellow $\rightarrow [F_0(x) + F_1(x)] \downarrow dt^1$ \rightarrow Purple

$$F_0(x) + F_1(x) + F_2(x)$$

$$\downarrow dt^1 \quad \downarrow dt^2$$

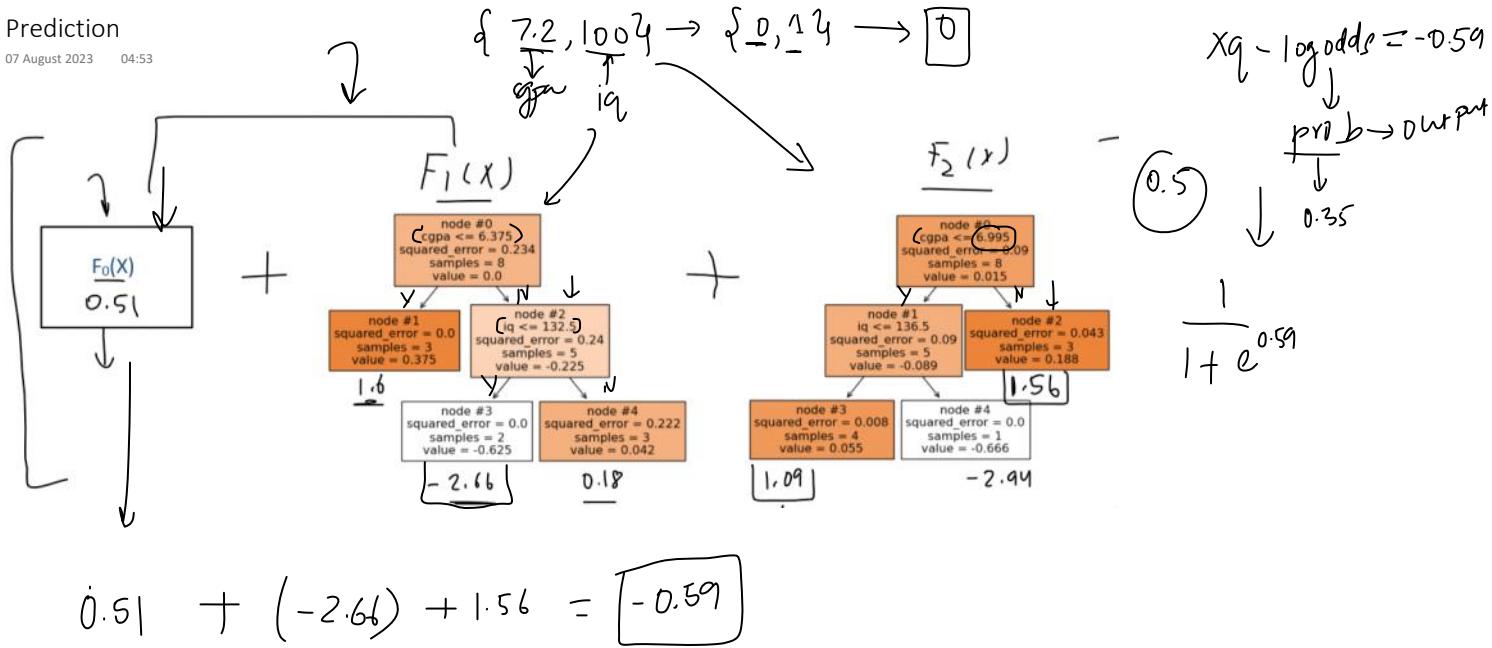
Predictions of iteration 4





Prediction

07 August 2023 04:53

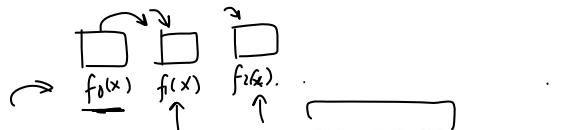


Geometric Intuition

04 August 2023 20:15

Maths behind Classification

08 August 2023 22:54



Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M .

- 1. Initialize $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$. $\min \left[\log \left(\frac{P_{\text{true}}}{1 - P_{\text{true}}} \right) \right]$
- 2. For $m = 1$ to M :

- (a) For $i = 1, 2, \dots, N$ compute

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}$$

- (b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1, 2, \dots, J_m$.

- (c) For $j = 1, 2, \dots, J_m$ compute

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

$$\frac{\sum \text{Residual}}{\sum [\text{PreviousProb} * (1 - \text{PreviousProb})]}$$

- (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

- 3. Output $\hat{f}(x) = f_M(x)$.

Step 0 -> The Loss Function

10 August 2023 09:43

$$\text{Log loss} \rightarrow L = -\frac{1}{n} \sum_{i=1}^n y_i \log p_i + (1-y_i) \log(1-p_i) \rightarrow \underline{\underline{L(p_i)}} \rightarrow L(\log \text{odds})$$

$\hat{y}_i \rightarrow \text{output prob}$

$\log \text{odds} \rightarrow \underline{\underline{\log \text{odds}}}$

$\log \left(\frac{p_i}{1-p_i} \right)$

category	y_i	placement	$\hat{y}(p_i)$
8	80	1	0.62
7	70	0	0.32
6	60	1	0.51

$$L = -\frac{1}{n} \sum_{i=1}^n y_i \log p_i + (1-y_i) \log(1-p_i)$$

$$L = -\frac{1}{n} \left[\sum_{i=1}^n y_i \log p_i + \log(1-p_i) - y_i \log(1-p_i) \right]$$

$$= -\frac{1}{n} \left[\sum_{i=1}^n y_i [\log p_i - \log(1-p_i)] + \log(1-p_i) \right]$$

$$= -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(\text{odd}_i) + \log(1-p_i) \right]$$

$$p_i = \frac{e^{\log(\text{odd}_i)}}{1+e^{\log(\text{odd}_i)}}$$

$$= -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(\text{odd}_i) + \log \left(\frac{1}{1+e^{\log(\text{odd}_i)}} \right) \right]$$

$$= -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(\text{odd}_i) + \log 1 - \log(1+e^{\log(\text{odd}_i)}) \right]$$

$$L = -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(\text{odd}_i) - \log(1+e^{\log(\text{odd}_i)}) \right] \leftarrow \underline{\underline{\text{data}}}$$

$$L = -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(\text{odds}_i) - \underbrace{\log(1 + e^{\log(\text{odds})})}_{\text{link function}} \right]$$

Step 1 -> Finding $F_0(x)$

10 August 2023 09:44

Loss function



$$1. \text{ Initialize } f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma).$$

$$\rightarrow L = -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(\text{odds}_i) - \log(1 + e^{\frac{\log(\text{odds}_i)}{\gamma}}) \right]$$

$$\begin{matrix} L(y_i, \log(\text{odds}_i)) \\ \hookrightarrow L(y_i, \gamma) \end{matrix}$$

$$L = \underbrace{-\frac{1}{n} \left[\sum_{i=1}^n y_i \gamma - \log(1 + e^\gamma) \right]}_{\text{argmin } \gamma}$$

cg pali placm
1
0
F
0

$$\frac{\partial L}{\partial \gamma} = -\frac{1}{n} \left[\sum_{i=1}^n y_i - \frac{e^\gamma}{1 + e^\gamma} \right] = 0$$

$$= -\frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{n} \sum_{i=1}^n \frac{e^\gamma}{1 + e^\gamma} = 0$$

$$= -p_{avg} + \frac{n}{n} \frac{e^\gamma}{1 + e^\gamma} = 0$$

$$\frac{e^\gamma}{1 + e^\gamma} = p_{avg}$$

$$e^\gamma = p_{avg} + p_{avg} e^\gamma \Rightarrow e^\gamma - p_{avg} e^\gamma = p_{avg}$$

$$e^\gamma (1 - p_{avg}) = p_{avg}$$

$$e^\gamma = \frac{p_{avg}}{1 - p_{avg}}$$

$$\boxed{\gamma = \log\left(\frac{p_{avg}}{1 - p_{avg}}\right)}$$

$$f_0(x) = \gamma = \log\left(\frac{p_{avg}}{1 - p_{avg}}\right)$$

$$\log\left(\frac{5/8}{1 - 5/8}\right) = \log\left(\frac{5/8}{3/8}\right) = \log\left(\frac{5}{3}\right)$$

Step 2.a -> Pseudo Residuals

10 August 2023 11:23

n rows

2. For $m = 1$ to M :

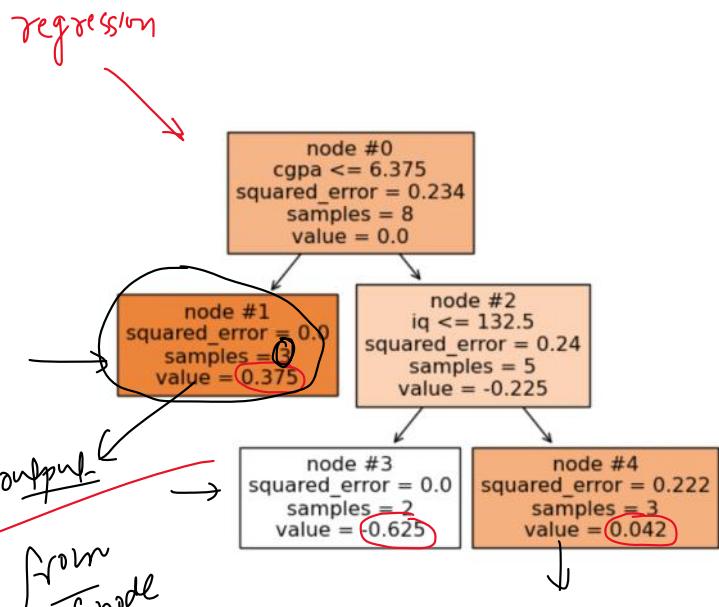
(a) For $i = 1, 2, \dots, N$ compute

$$\begin{aligned}
 r_{im} &= - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}} \\
 L &= -y \log(\text{odds}) + \log(1 + e^{\log(\text{odds})}) \\
 \frac{\partial L}{\partial \log(\text{odds})} &= -y + \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}} \rightarrow \text{prob output} \\
 &= -y + p = y - p \rightarrow \log \text{odds}
 \end{aligned}$$

Step 2.b -> Train Regression Tree

10 August 2023 11:41

cgpa	iq	is_placed	pre1(log-odds)	pre1(probability)	res1
6.82	118	0	0.510826	0.625	-0.625
6.36	125	1	0.510826	0.625	0.375
5.39	99	1	0.510826	0.625	0.375
5.50	106	1	0.510826	0.625	0.375
6.39	148	0	0.510826	0.625	-0.625
9.13	148	1	0.510826	0.625	0.375
7.17	147	1	0.510826	0.625	0.375
7.72	72	0	0.510826	0.625	-0.625



$$\frac{\sum \text{Residual}}{\sum [\text{PreviousProb} * (1 - \text{PreviousProb})]}$$

Step 2.c -> Compute Lambda for all leaf nodes

10 August 2023 12:02

$$m = 1$$

$$\gamma_j = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

$$\gamma_1$$

$$\gamma_2$$

$$\gamma_3$$

$$\gamma_4$$

$$\gamma_5$$

$$\gamma_6$$

$$\gamma_7$$

$$\gamma_8$$

$$\gamma_9$$

$$\gamma_{10}$$

$$\gamma_{11}$$

$$\gamma_{12}$$

$$\gamma_{13}$$

$$\gamma_{14}$$

$$\gamma_{15}$$

$$\gamma_{16}$$

$$\gamma_{17}$$

$$\gamma_{18}$$

$$\gamma_{19}$$

$$\gamma_{20}$$

$$\gamma_{21}$$

$$\gamma_{22}$$

$$\gamma_{23}$$

$$\gamma_{24}$$

$$\gamma_{25}$$

$$\gamma_{26}$$

$$\gamma_{27}$$

$$\gamma_{28}$$

$$\gamma_{29}$$

$$\gamma_{30}$$

$$\gamma_{31}$$

$$\gamma_{32}$$

$$\gamma_{33}$$

$$\gamma_{34}$$

$$\gamma_{35}$$

$$\gamma_{36}$$

$$\gamma_{37}$$

$$\gamma_{38}$$

$$\gamma_{39}$$

$$\gamma_{40}$$

$$\gamma_{41}$$

$$\gamma_{42}$$

$$\gamma_{43}$$

$$\gamma_{44}$$

$$\gamma_{45}$$

$$\gamma_{46}$$

$$\gamma_{47}$$

$$\gamma_{48}$$

$$\gamma_{49}$$

$$\gamma_{50}$$

$$\gamma_{51}$$

$$\gamma_{52}$$

$$\gamma_{53}$$

$$\gamma_{54}$$

$$\gamma_{55}$$

$$\gamma_{56}$$

$$\gamma_{57}$$

$$\gamma_{58}$$

$$\gamma_{59}$$

$$\gamma_{60}$$

$$\gamma_{61}$$

$$\gamma_{62}$$

$$\gamma_{63}$$

$$\gamma_{64}$$

$$\gamma_{65}$$

$$\gamma_{66}$$

$$\gamma_{67}$$

$$\gamma_{68}$$

$$\gamma_{69}$$

$$\gamma_{70}$$

$$\gamma_{71}$$

$$\gamma_{72}$$

$$\gamma_{73}$$

$$\gamma_{74}$$

$$\gamma_{75}$$

$$\gamma_{76}$$

$$\gamma_{77}$$

$$\gamma_{78}$$

$$\gamma_{79}$$

$$\gamma_{80}$$

$$\gamma_{81}$$

$$\gamma_{82}$$

$$\gamma_{83}$$

$$\gamma_{84}$$

$$\gamma_{85}$$

$$\gamma_{86}$$

$$\gamma_{87}$$

$$\gamma_{88}$$

$$\gamma_{89}$$

$$\gamma_{90}$$

$$\gamma_{91}$$

$$\gamma_{92}$$

$$\gamma_{93}$$

$$\gamma_{94}$$

$$\gamma_{95}$$

$$\gamma_{96}$$

$$\gamma_{97}$$

$$\gamma_{98}$$

$$\gamma_{99}$$

$$\gamma_{100}$$

$$\gamma_{101}$$

$$\gamma_{102}$$

$$\gamma_{103}$$

$$\gamma_{104}$$

$$\gamma_{105}$$

$$\gamma_{106}$$

$$\gamma_{107}$$

$$\gamma_{108}$$

$$\gamma_{109}$$

$$\gamma_{110}$$

$$\gamma_{111}$$

$$\gamma_{112}$$

$$\gamma_{113}$$

$$\gamma_{114}$$

$$\gamma_{115}$$

$$\gamma_{116}$$

$$\gamma_{117}$$

$$\gamma_{118}$$

$$\gamma_{119}$$

$$\gamma_{120}$$

$$\gamma_{121}$$

$$\gamma_{122}$$

$$\gamma_{123}$$

$$\gamma_{124}$$

$$\gamma_{125}$$

$$\gamma_{126}$$

$$\gamma_{127}$$

$$\gamma_{128}$$

$$\gamma_{129}$$

$$\gamma_{130}$$

$$\gamma_{131}$$

$$\gamma_{132}$$

$$\gamma_{133}$$

$$\gamma_{134}$$

$$\gamma_{135}$$

$$\gamma_{136}$$

$$\gamma_{137}$$

$$\gamma_{138}$$

$$\gamma_{139}$$

$$\gamma_{140}$$

$$\gamma_{141}$$

$$\gamma_{142}$$

$$\gamma_{143}$$

$$\gamma_{144}$$

$$\gamma_{145}$$

$$\gamma_{146}$$

$$\gamma_{147}$$

$$\gamma_{148}$$

$$\gamma_{149}$$

$$\gamma_{150}$$

$$\gamma_{151}$$

$$\gamma_{152}$$

$$\gamma_{153}$$

$$\gamma_{154}$$

$$\gamma_{155}$$

$$\gamma_{156}$$

$$\gamma_{157}$$

$$\gamma_{158}$$

$$\gamma_{159}$$

$$\gamma_{160}$$

$$\gamma_{161}$$

$$\gamma_{162}$$

$$\gamma_{163}$$

$$\gamma_{164}$$

$$\gamma_{165}$$

$$\gamma_{166}$$

$$\gamma_{167}$$

$$\gamma_{168}$$

$$\gamma_{169}$$

$$\gamma_{170}$$

$$\gamma_{171}$$

$$\gamma_{172}$$

$$\gamma_{173}$$

$$\gamma_{174}$$

$$\gamma_{175}$$

$$\gamma_{176}$$

$$\gamma_{177}$$

$$\gamma_{178}$$

$$\gamma_{179}$$

$$\gamma_{180}$$

$$\gamma_{181}$$

$$\gamma_{182}$$

$$\gamma_{183}$$

$$\gamma_{184}$$

$$\gamma_{185}$$

$$\gamma_{186}$$

$$\gamma_{187}$$

$$\gamma_{188}$$

$$\gamma_{189}$$

$$\gamma_{190}$$

$$\gamma_{191}$$

$$\gamma_{192}$$

$$\gamma_{193}$$

$$\gamma_{194}$$

$$\gamma_{195}$$

$$\gamma_{196}$$

$$\gamma_{197}$$

$$\gamma_{198}$$

$$\gamma_{199}$$

$$\gamma_{200}$$

$$\gamma_{201}$$

$$\gamma_{202}$$

$$\gamma_{203}$$

$$\gamma_{204}$$

$$\gamma_{205}$$

$$\gamma_{206}$$

$$\gamma_{207}$$

$$\gamma_{208}$$

$$\gamma_{209}$$

$$\gamma_{210}$$

$$\gamma_{211}$$

$$\gamma_{212}$$

$$\gamma_{213}$$

$$\gamma_{214}$$

$$\gamma_{215}$$

$$\gamma_{216}$$

$$\gamma_{217}$$

$$\gamma_{218}$$

$$\gamma_{219}$$

$$\gamma_{220}$$

$$\gamma_{221}$$

$$\gamma_{222}$$

$$\gamma_{223}$$

$$\gamma_{224}$$

$$\gamma_{225}$$

$$\gamma_{226}$$

$$\gamma_{227}$$

$$\gamma_{228}$$

$$\gamma_{229}$$

$$\gamma_{230}$$

$$\gamma_{231}$$

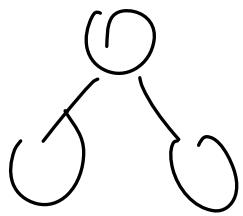
$$\begin{aligned}
& \left[\frac{\partial L}{\partial f_0(x_i)} \right] = \left[-y_i + \left(\frac{e^{\text{log}(\text{odds}_i)}}{1 + e^{\text{log}(\text{odds}_i)}} \right)^{-1} \right] - \left[\frac{1}{1 + e^{\text{log}(\text{odds}_i)}} \right] = \\
& \frac{\partial}{\partial f_0(x_i)} \frac{\partial L}{\partial f_0(x_i)} = \frac{\partial}{\partial \text{log}(\text{odds}_i)} \left(y_i - \frac{e^{\text{log}(\text{odds}_i)}}{1 + e^{\text{log}(\text{odds}_i)}} \right) \\
& \frac{\partial}{\partial \text{log}(\text{odds}_i)} \left[\frac{y_i - e^{\text{log}(\text{odds}_i)} (1 + e^{\text{log}(\text{odds}_i)})^{-1}}{x} \right] \\
& = - \left[e^{\text{log}(\text{odds}_i)} (1 + e^{\text{log}(\text{odds}_i)})^{-1} - \frac{e^{\text{log}(\text{odds}_i)}}{(1 + e^{\text{log}(\text{odds}_i)})^2} (1 + e^{\text{log}(\text{odds}_i)})^{-2} \right] \\
& = \frac{e^{\text{log}(\text{odds}_i)}}{(1 + e^{\text{log}(\text{odds}_i)})^2} - \frac{e^{\text{log}(\text{odds}_i)}}{(1 + e^{\text{log}(\text{odds}_i))})} \\
& = \frac{e^{\text{log}(\text{odds}_i)}}{1 + e^{\text{log}(\text{odds}_i)}} \left[\frac{1}{1 + e^{\text{log}(\text{odds}_i)}} - 1 \right] \rightarrow \frac{1}{1 + e^{\text{log}(\text{odds}_i)}} \\
& = \frac{e^{\text{log}(\text{odds}_i)}}{(1 + e^{\text{log}(\text{odds}_i)})} \quad \frac{1}{(1 + e^{\text{log}(\text{odds}_i)})} \\
& \qquad \qquad \qquad P_i \quad (1 - P_i)
\end{aligned}$$

Step 2d - Update the model

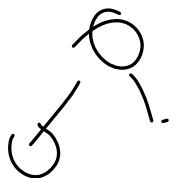
10 August 2023 12:10

$$f_1(x) = f_0(x) + \underbrace{\text{output from } dt}_{\text{log odds}}$$

some leaf node



$$f_2(x)$$



$$f_3(x) \rightarrow M \text{ decisions}$$

Step 3 - Final Model

10 August 2023 12:10

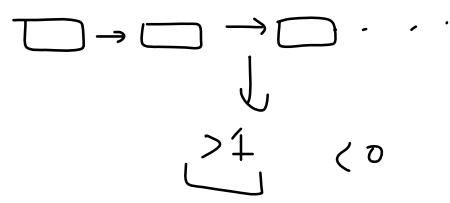
$$f_M(x) = f_{M-1}(x) + \begin{matrix} \text{last} \\ \text{decision} \\ \text{output} \end{matrix}$$

Boosting model

Log(odd) Vs Probability

10 August 2023 13:06

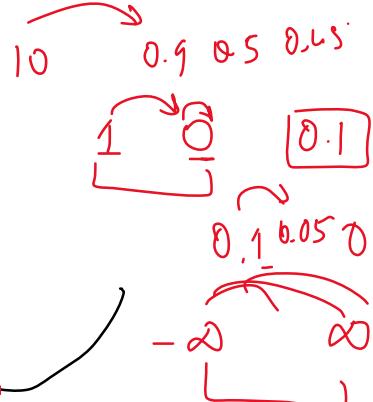
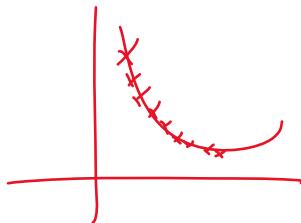
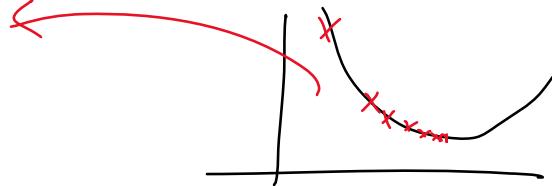
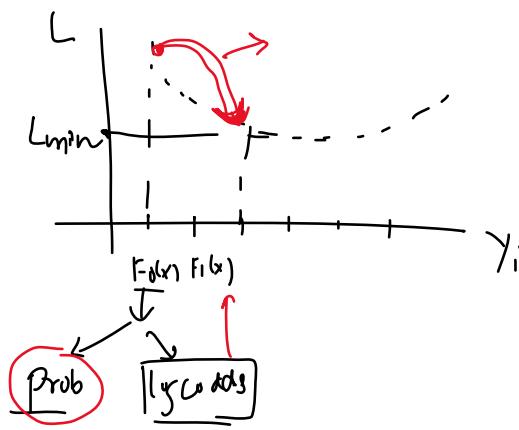
$-\infty$ $+\infty$ $(0 - 1)$

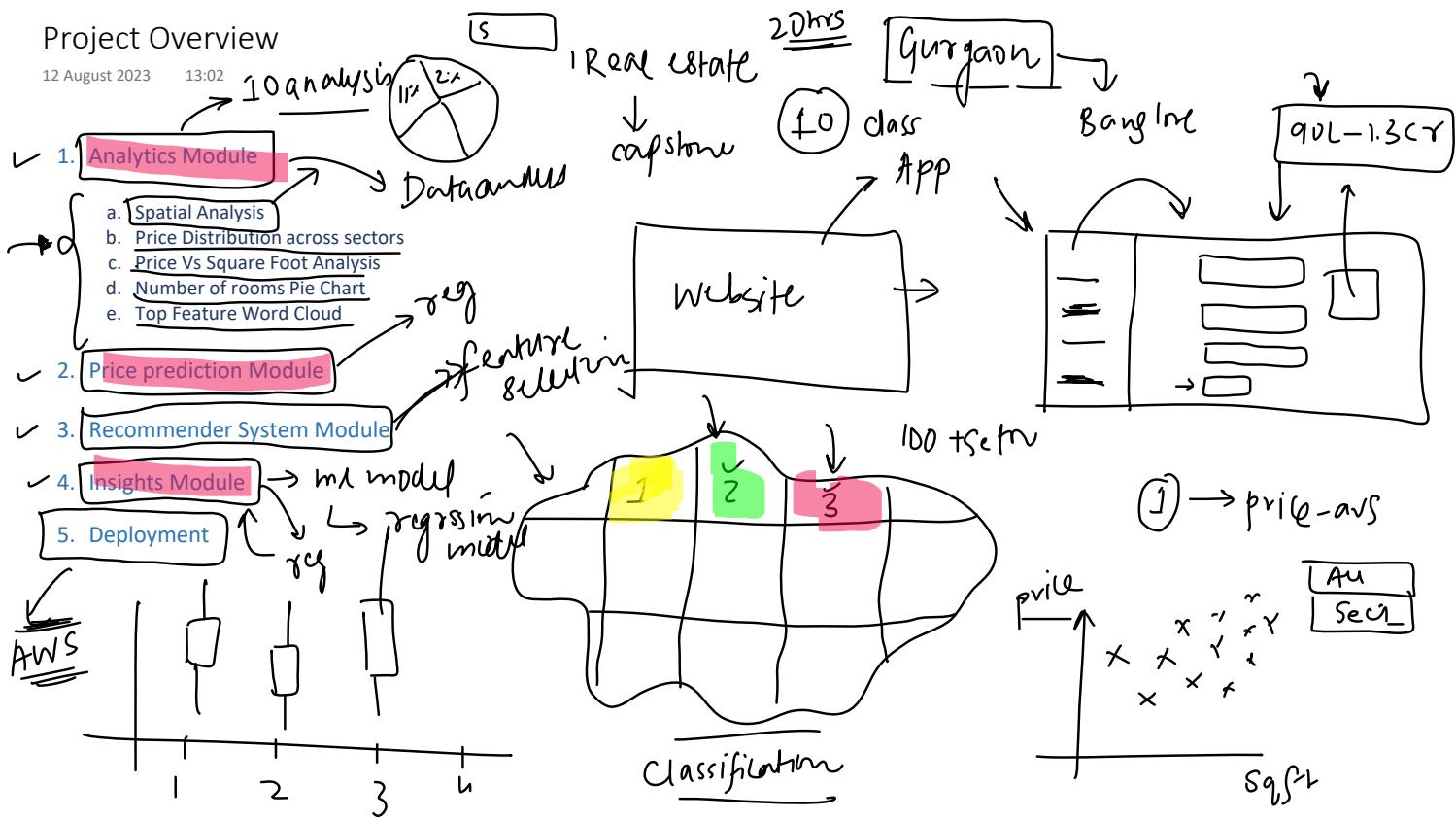


- ① Unconstrained Prediction Space: Log odds can span the entire real line ($-\infty$ to $+\infty$), while probabilities are constrained between 0 and 1. Algorithms like gradient boosting involve adding corrections (via the weak learners) to the predictions iteratively. If you're working in the log odds space, there's no need to worry about your predictions going out of bounds.

$-\infty - +\infty$

- ② **Better Gradients** When computing gradients (which guide the addition of new trees in boosting), the gradients can be more informative and have better magnitudes in the log odds space than in the probability space, especially when probabilities are near 0 or 1.



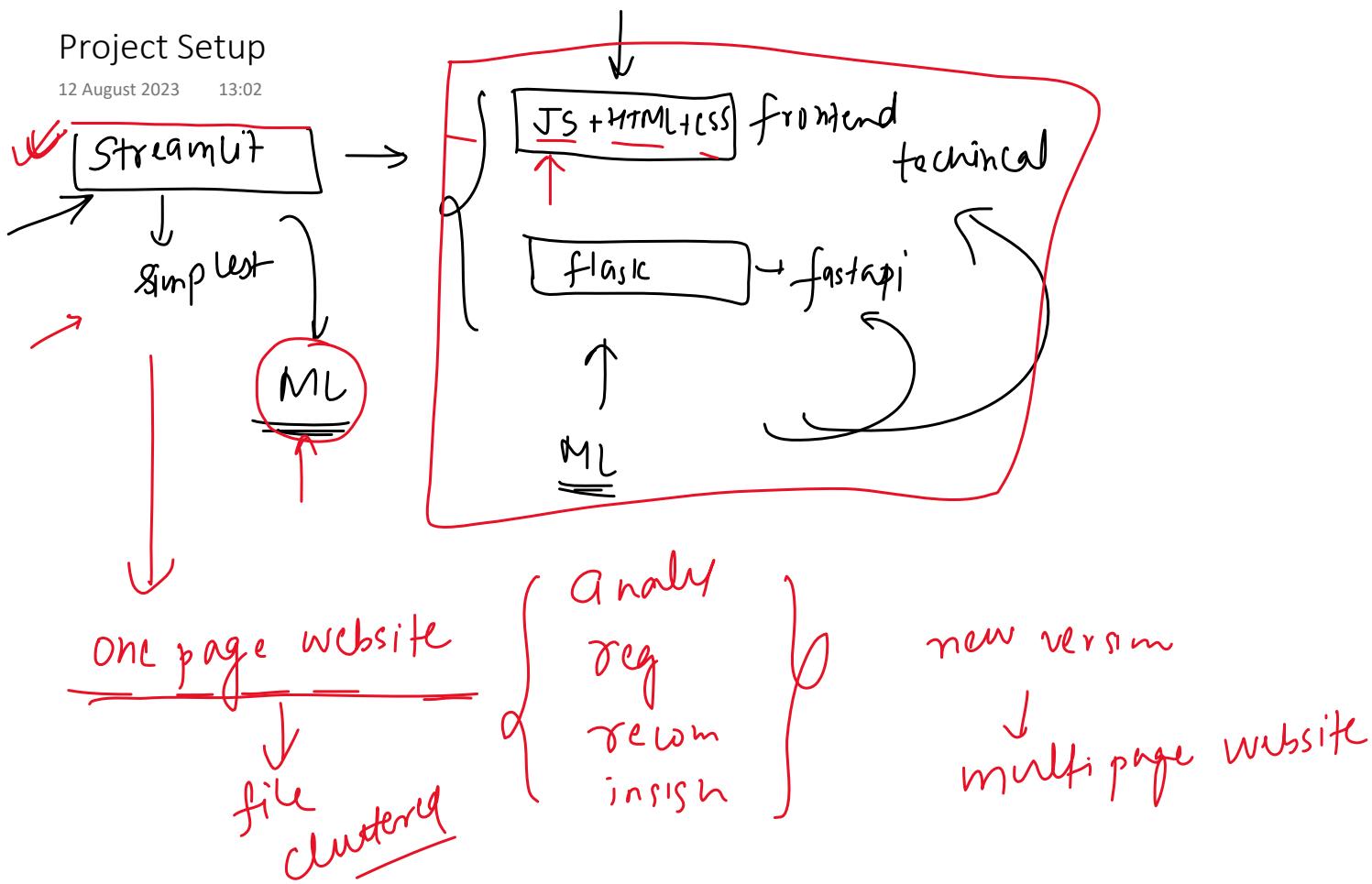


Data Gathering

12 August 2023 13:02

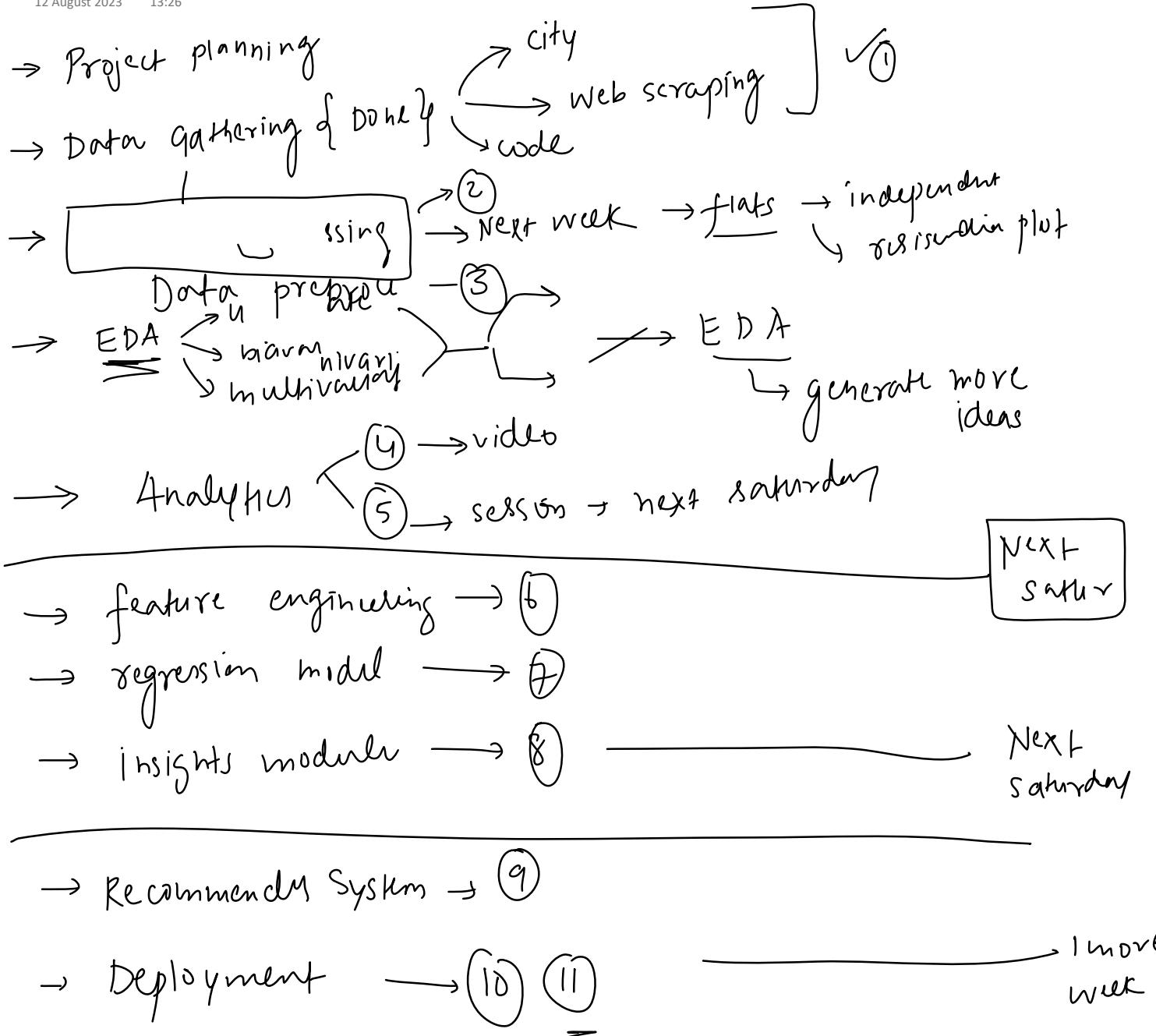
Project Setup

12 August 2023 13:02



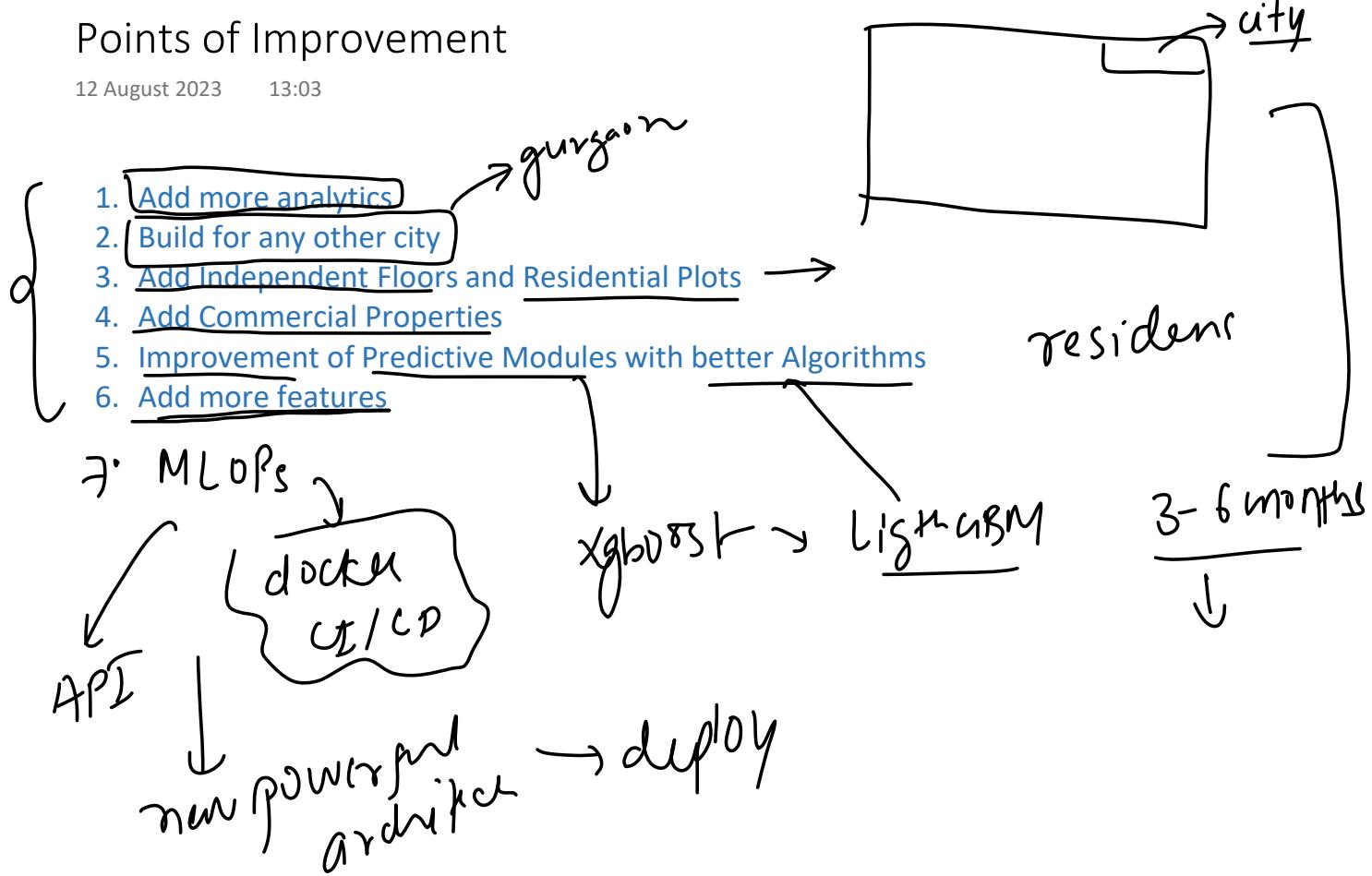
Project Roadmap

12 August 2023 13:26



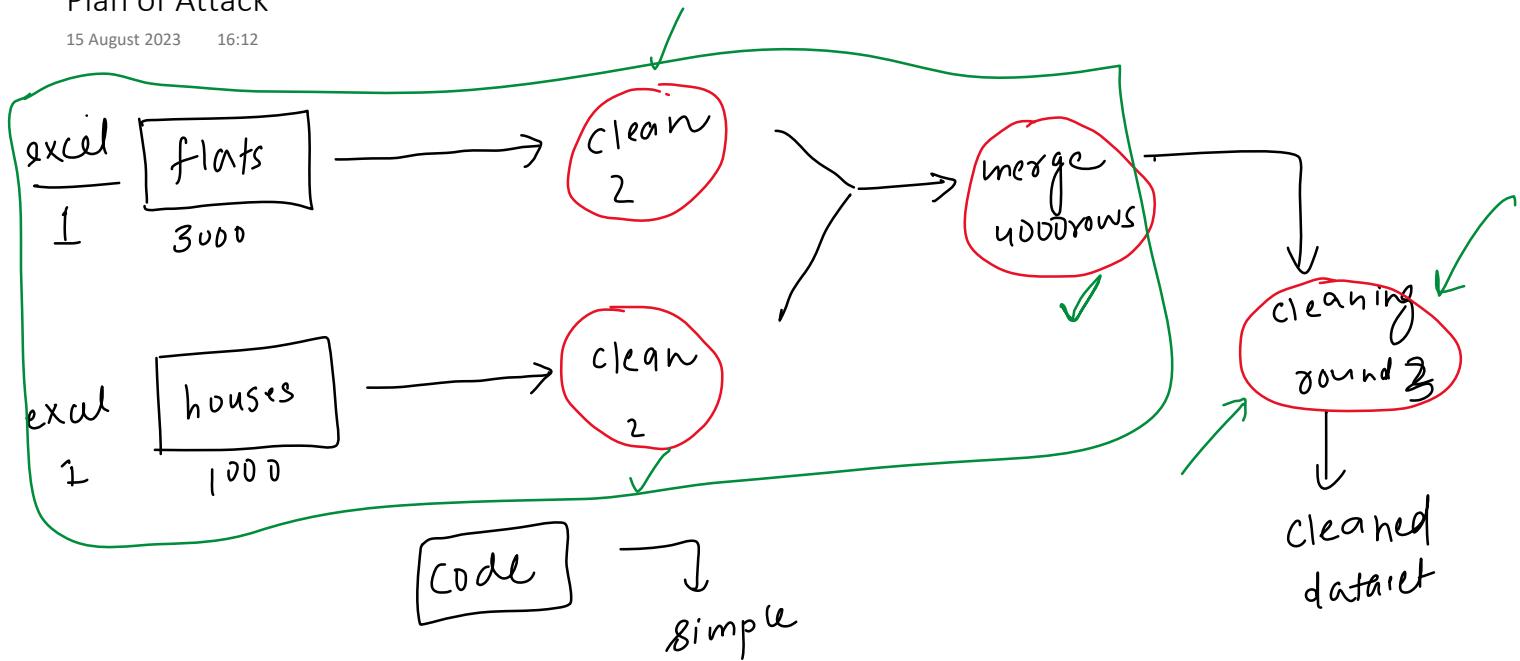
Points of Improvement

12 August 2023 13:03



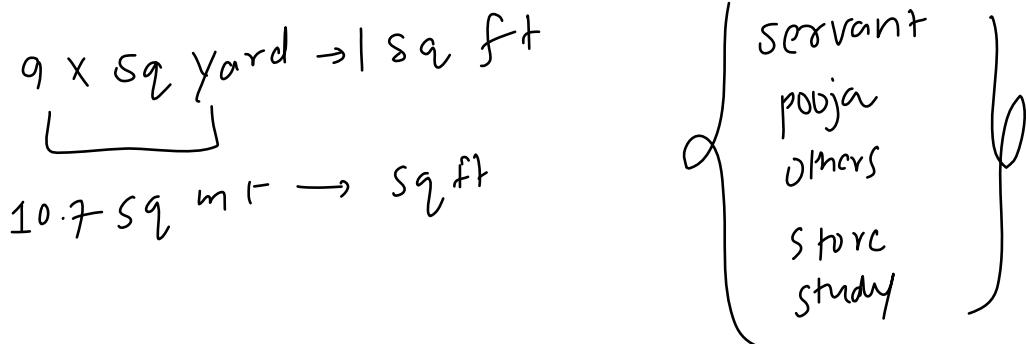
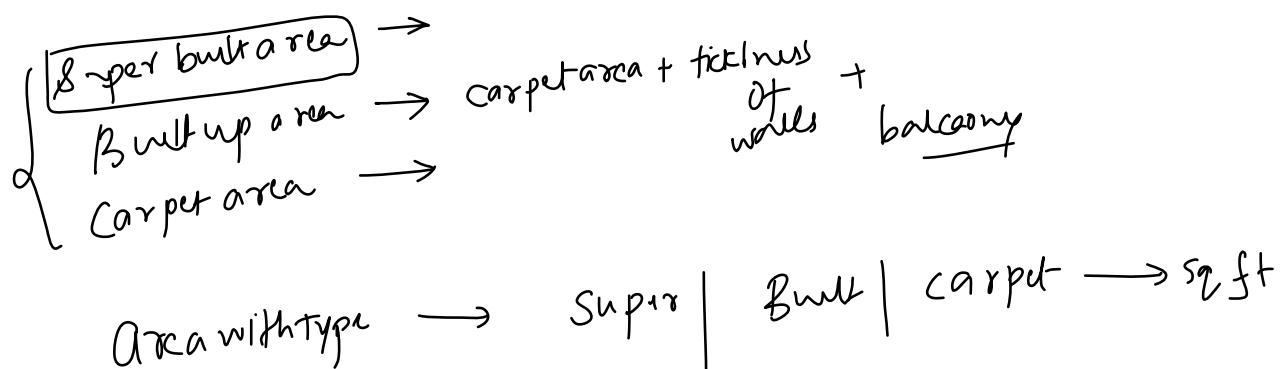
Plan of Attack

15 August 2023 16:12

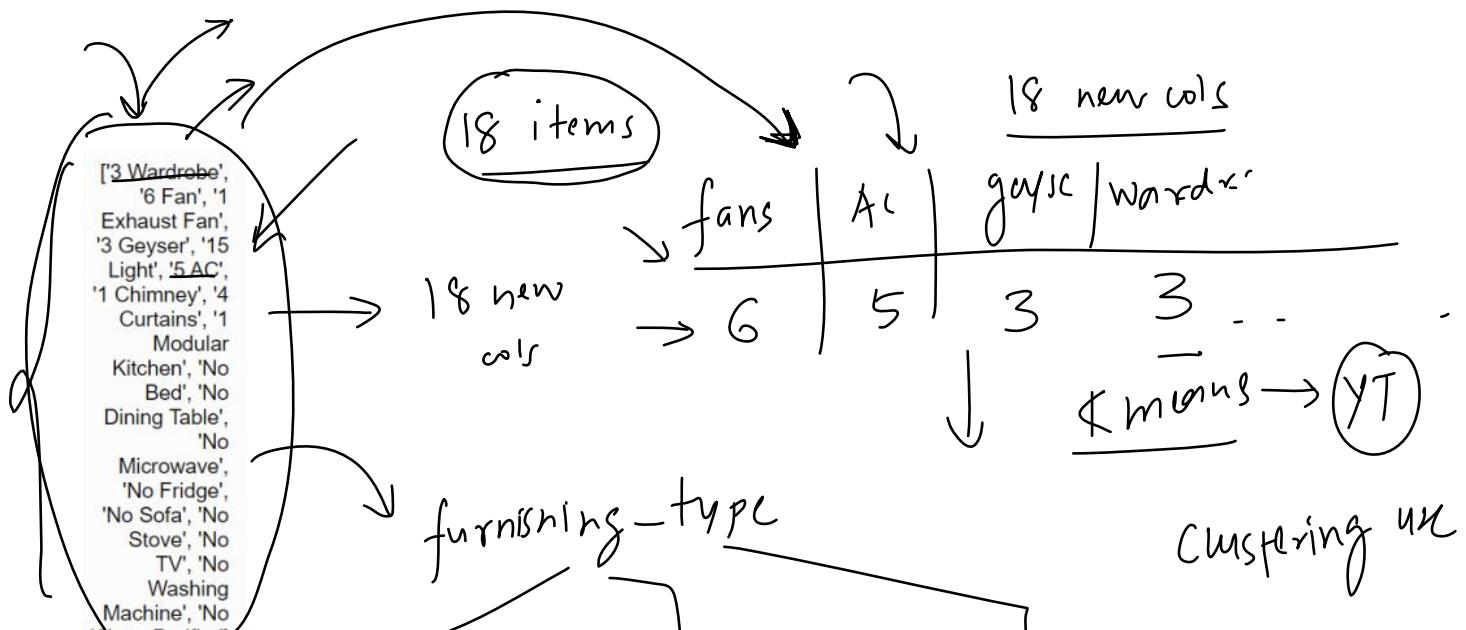


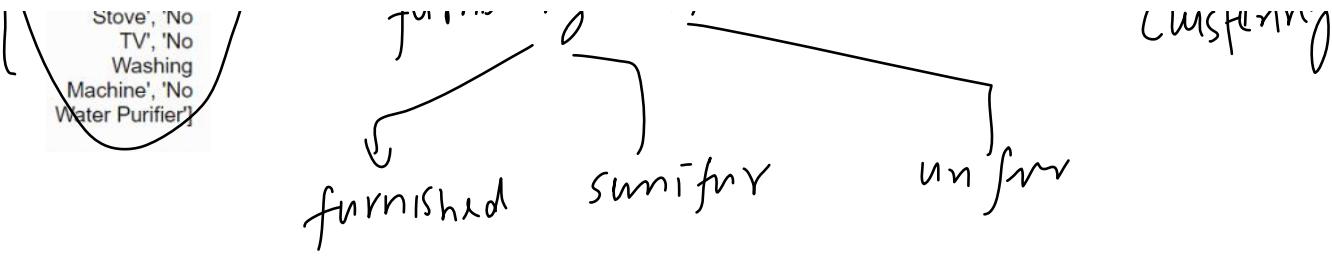
Feature Engineering

17 August 2023 15:57



servant	pooja	study	store	others
0	0	0	0	0
1	0	0	0	0
1	0	1	0	0





security | swimming

13 dim
vector

luxury
semiluxury
budget

NaN
EDA

missing
outlier

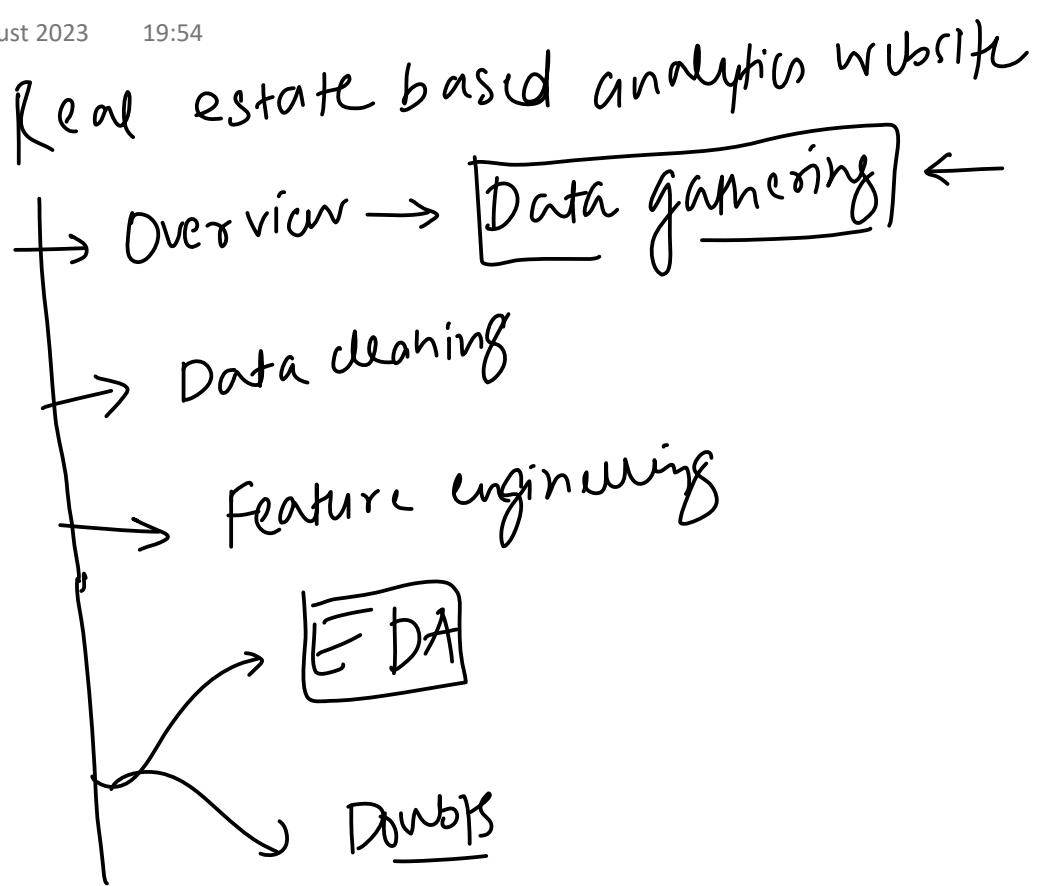
Analytics

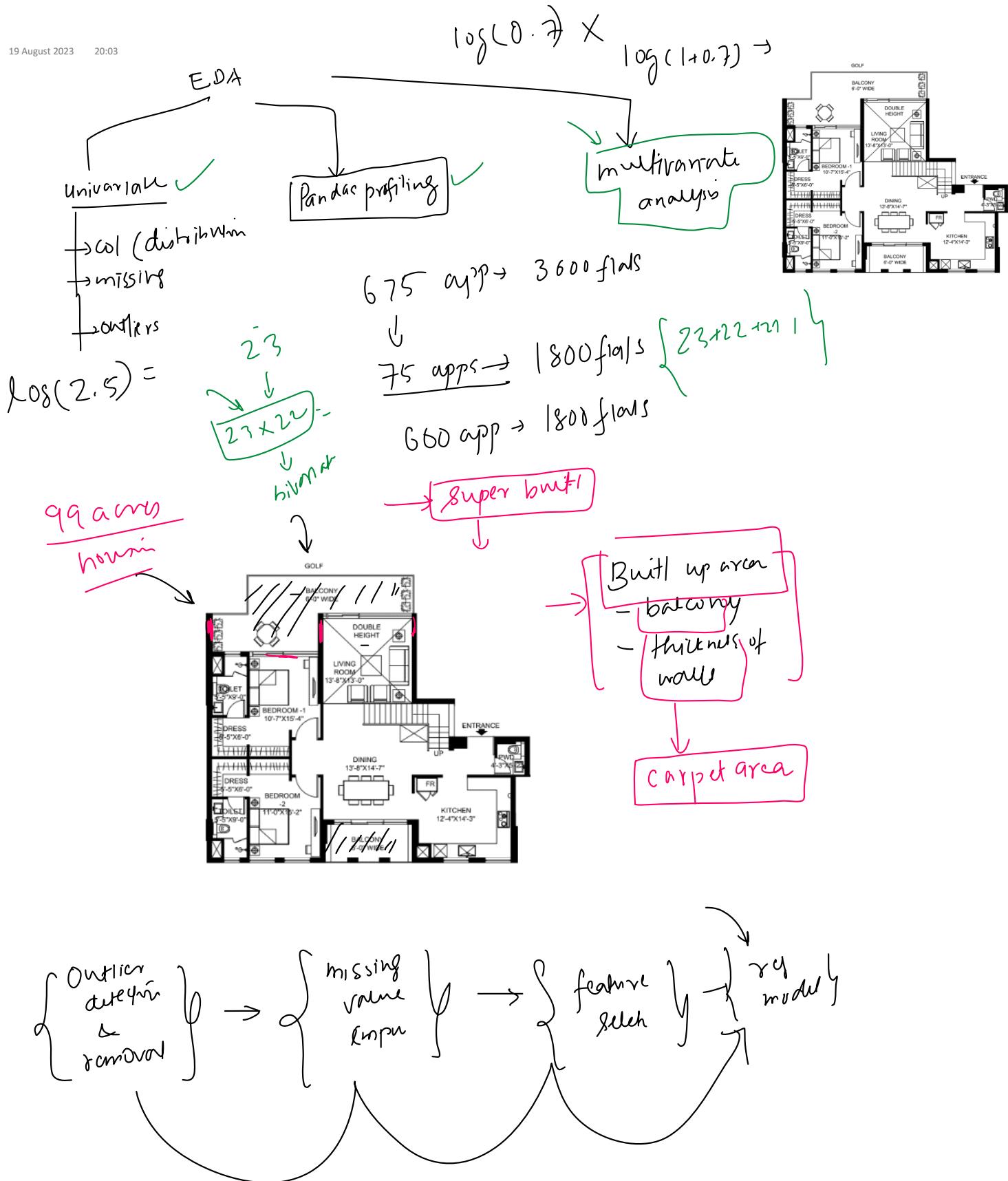
luxury - score

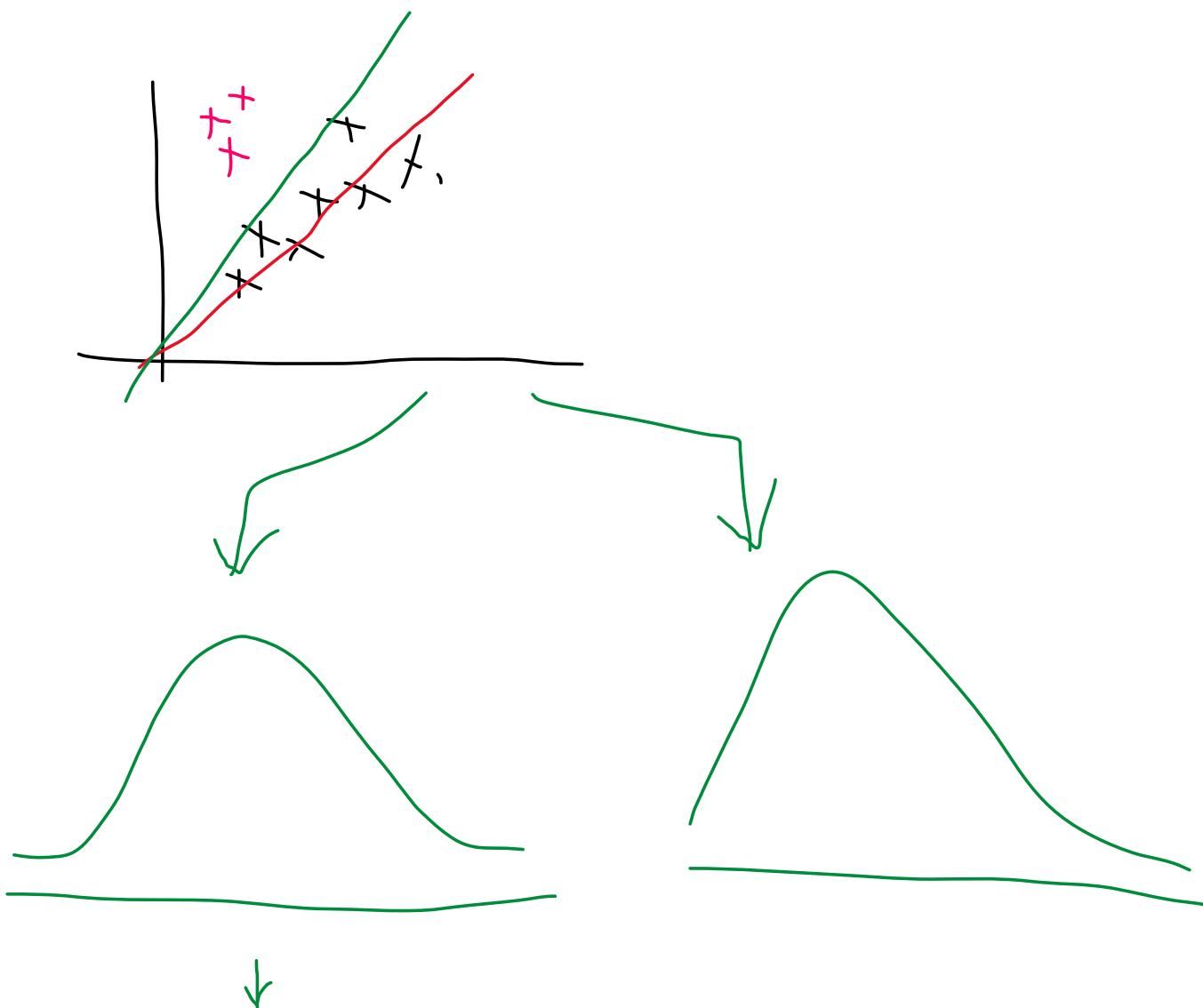
EDA

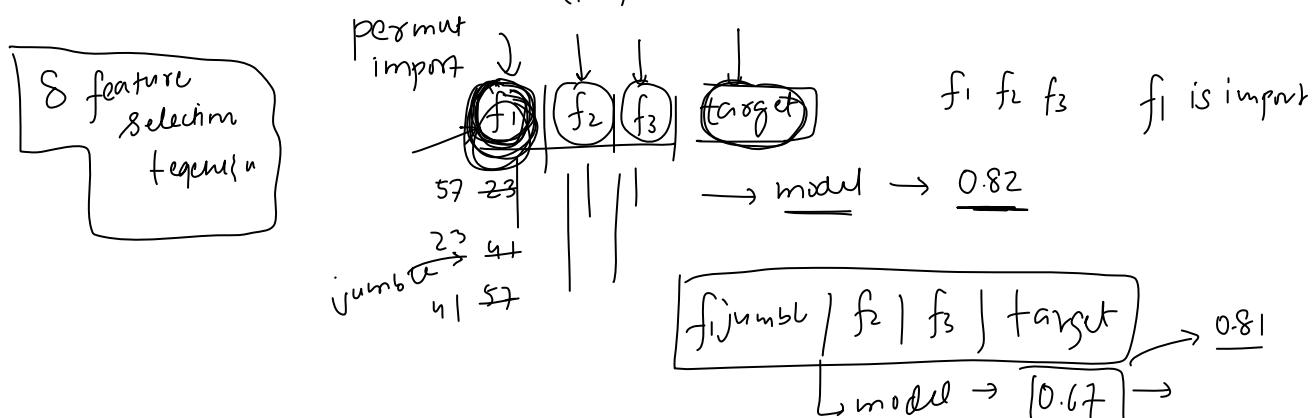
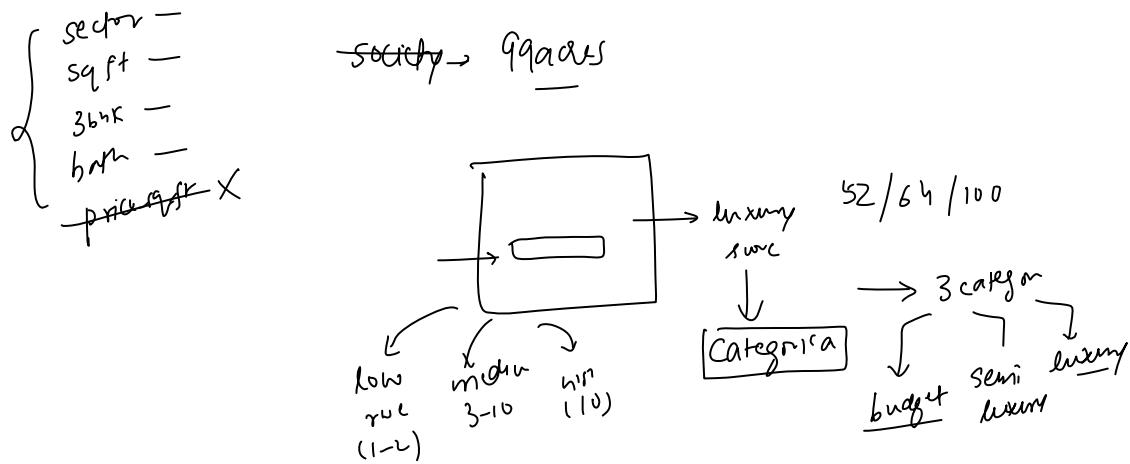
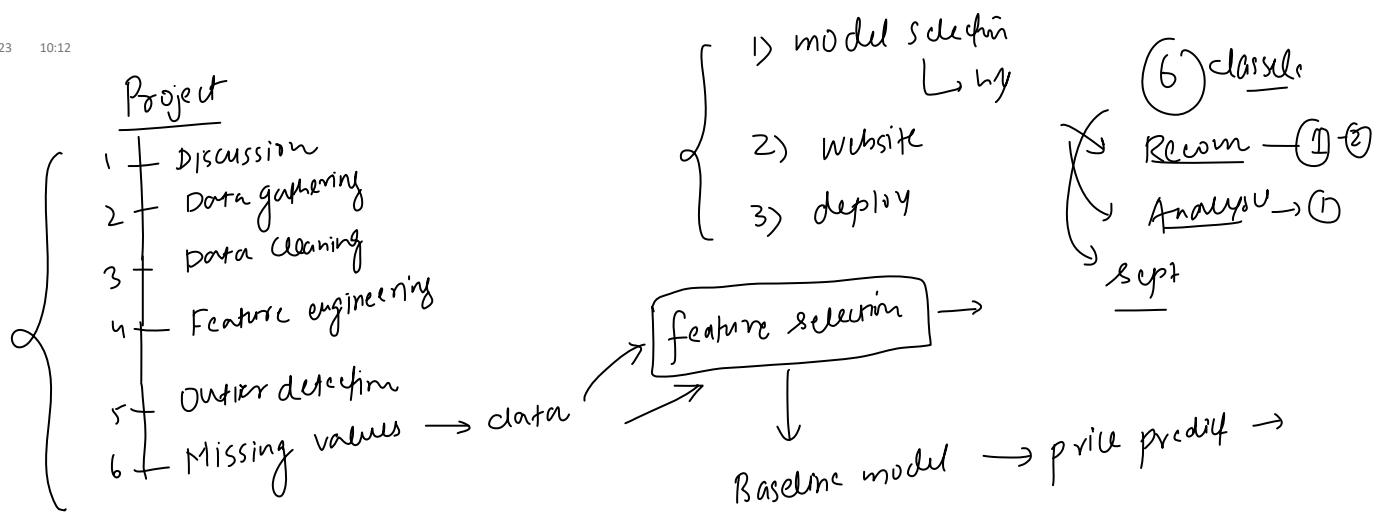
19 August 2023

19:54



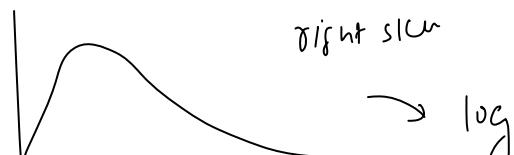
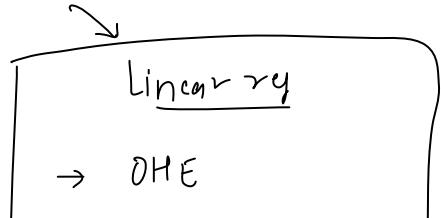






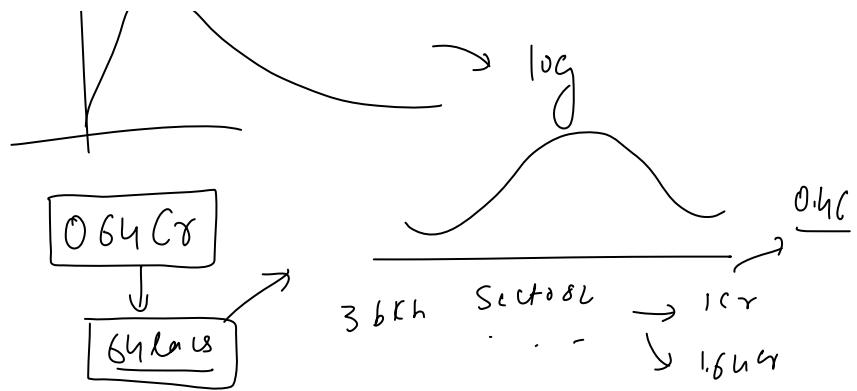
city	city_mum	city_dal	city_hyd
mum	0	1	0
dal	1	0	0
hyd	2	0	1

several → 10h → 10h cols



(53L)

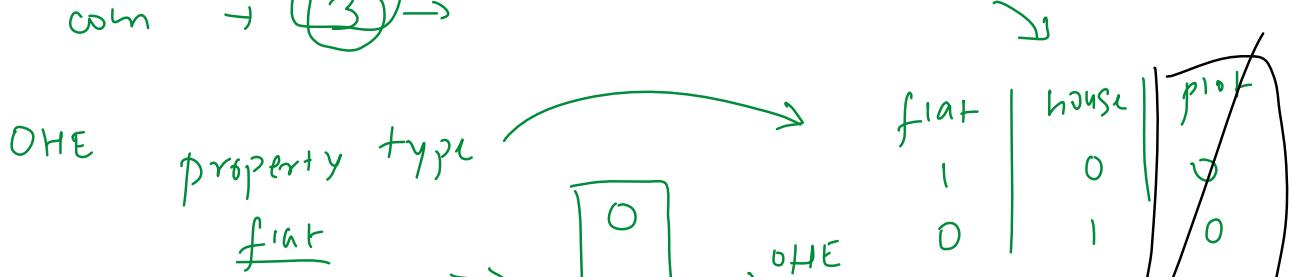
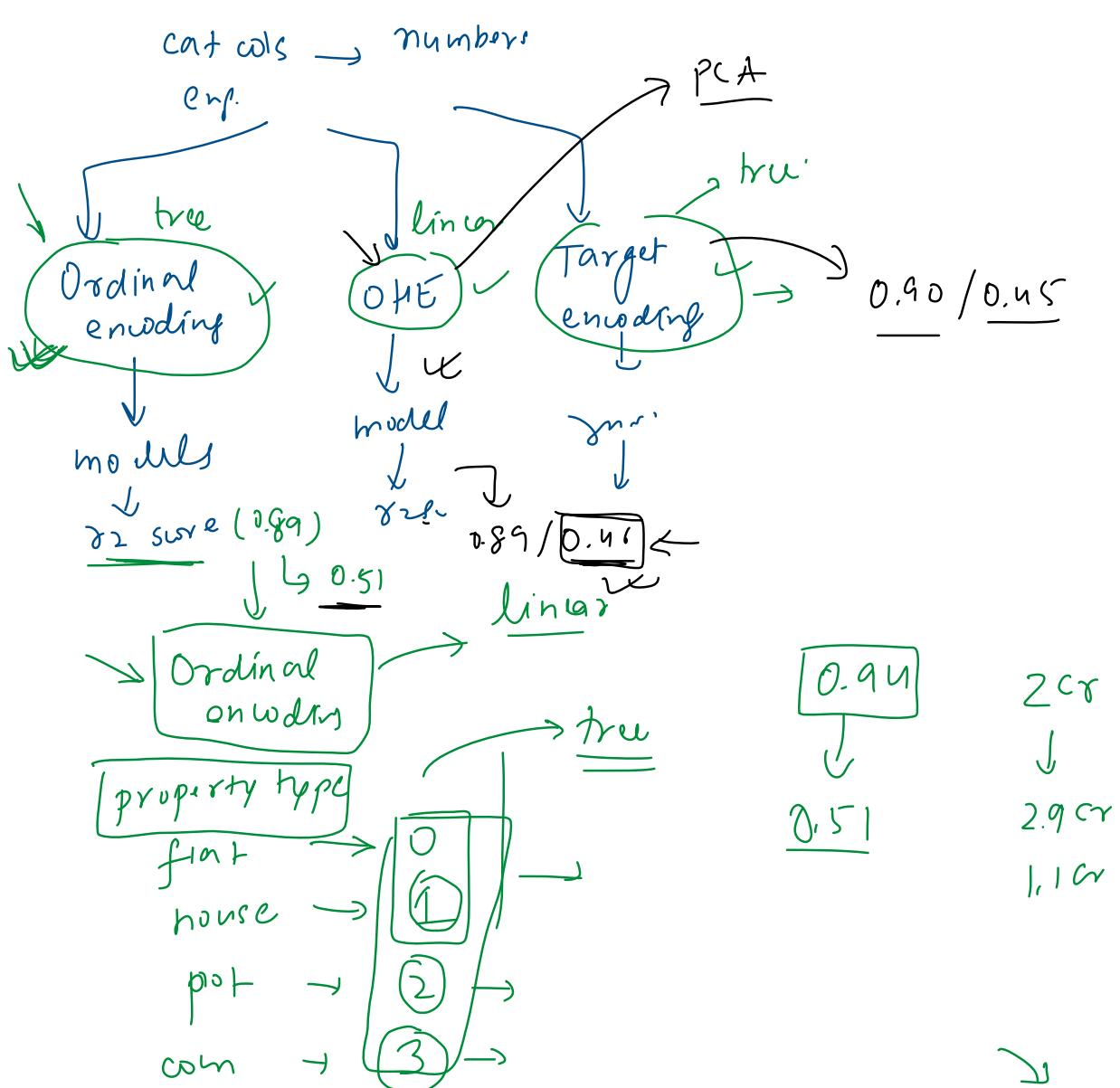
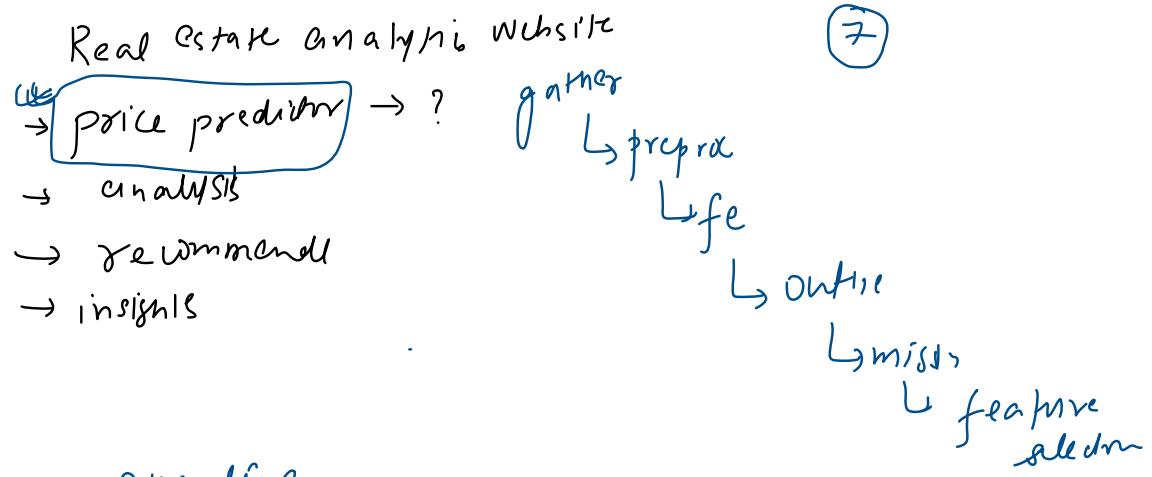
- OHE
- scaling
- log transformation

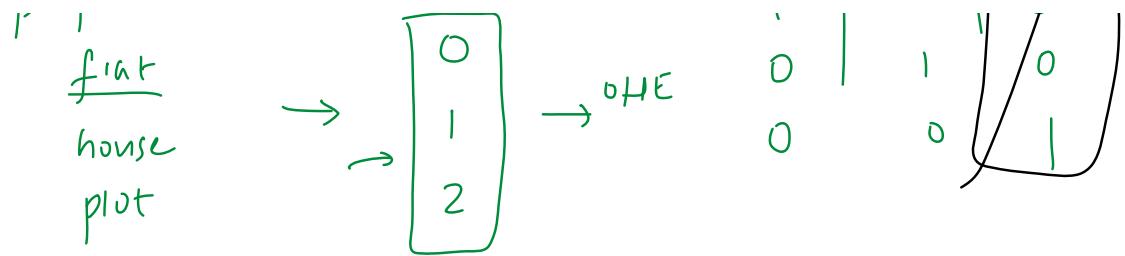


- diff algo
- hyperparameter
- featur engn
- more data

\rightarrow web site \rightarrow dep¹⁰y

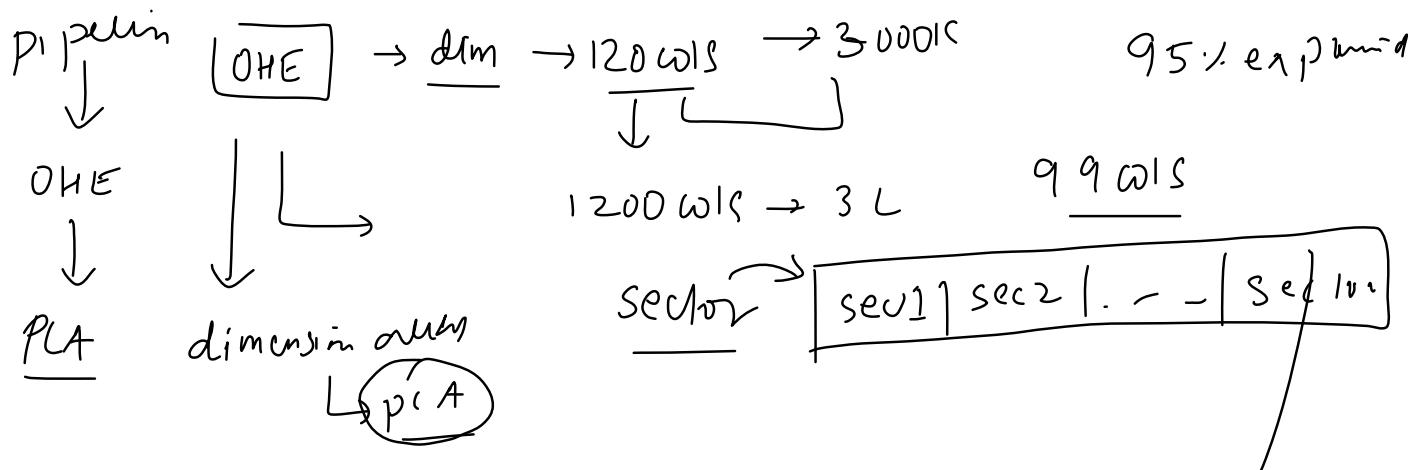
$$\begin{array}{c} y \rightarrow \log(y) \\ \downarrow \\ e^{\log(y)} \\ \downarrow \\ y \end{array}$$





OHE → cat (50) → 100 new cols appear

LE →

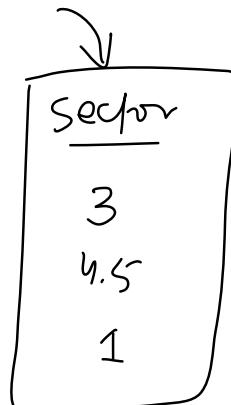


sector (100 cat^e)



target encoding

sector	—	②
Sec 1	—	③
Sec 99	—	④
Sec 1	—	⑤
Sec 100	—	⑥
Sec 49	—	

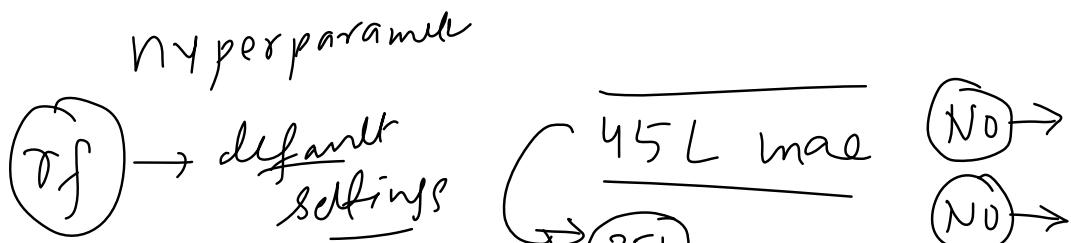
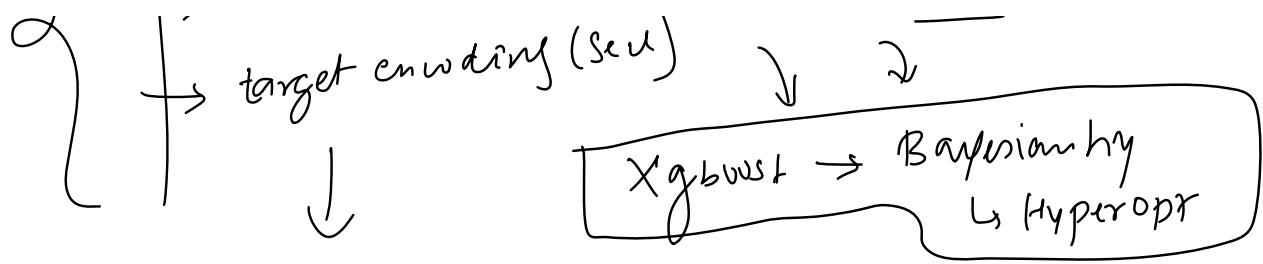


model selection

- tree based → rf | xgbust
- target encoding (Sec)

[0.90] → 0.93

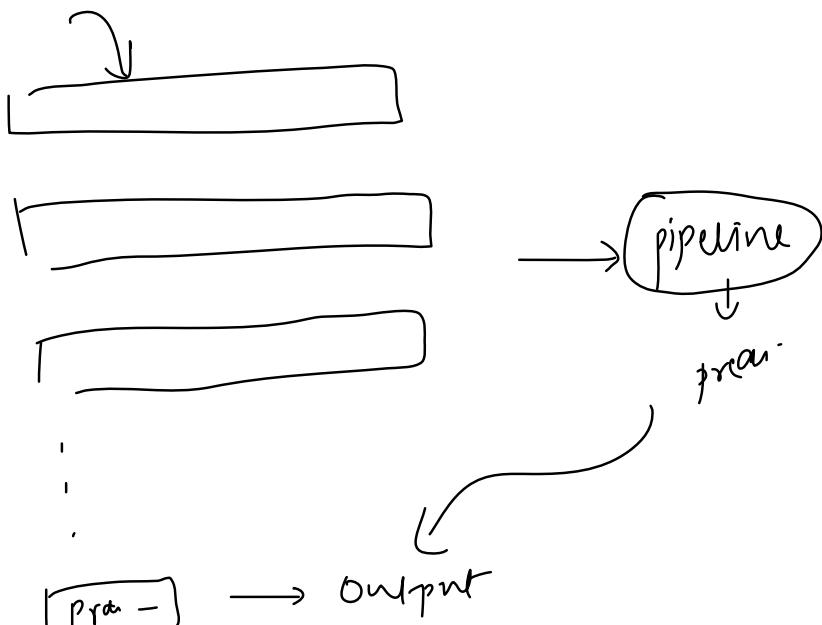
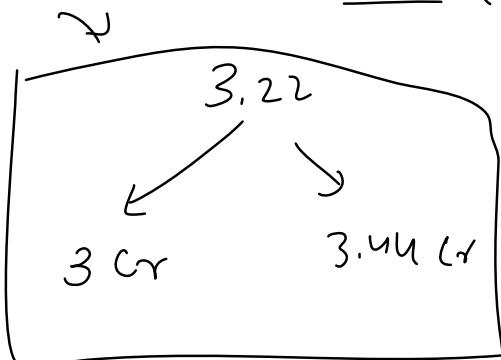
0.65 → 0.35



→ more data (3000 rows)

45L → mae → xgboost → hyperopt
→ revisit the entire process

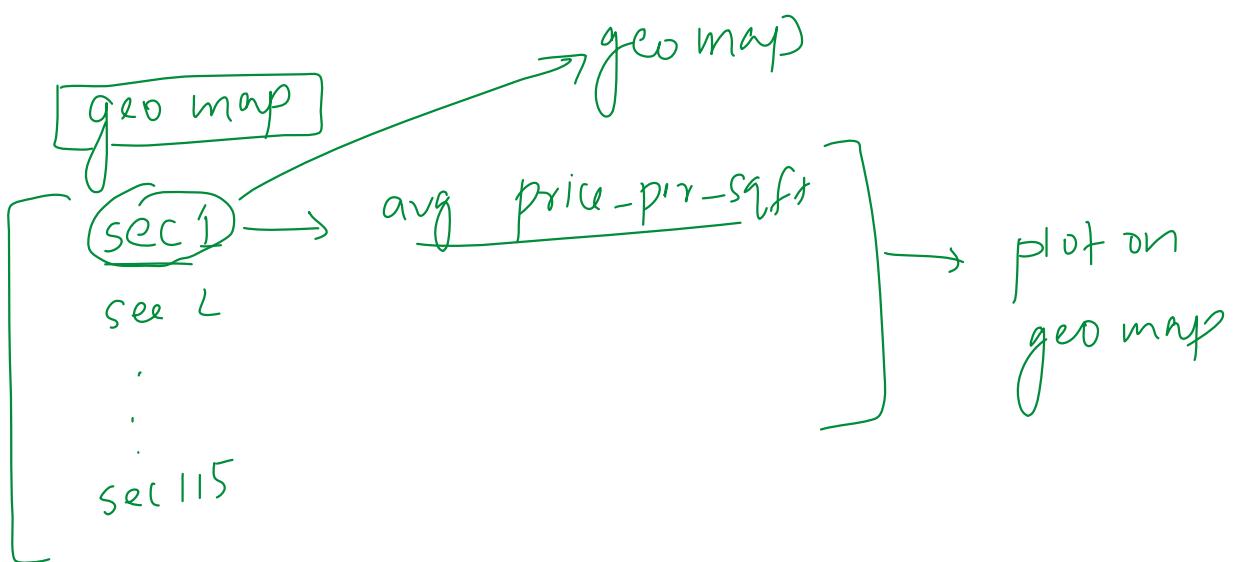
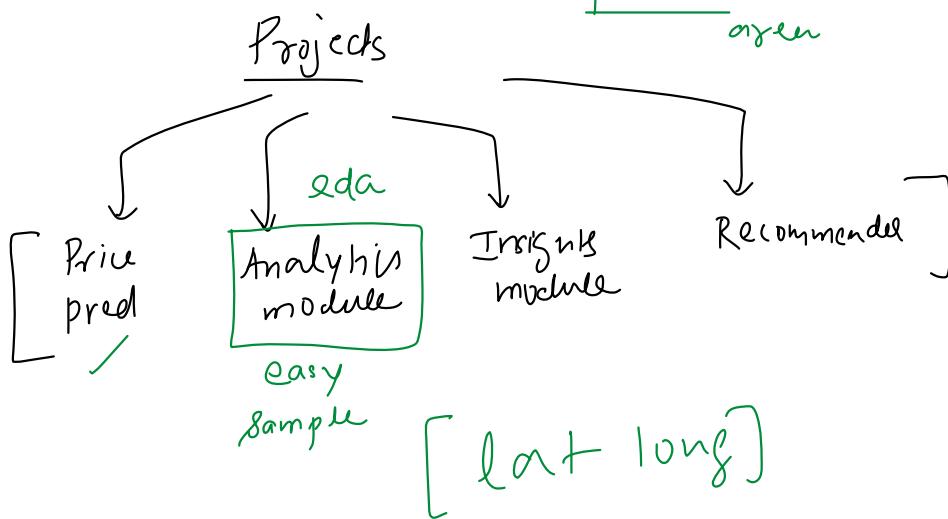
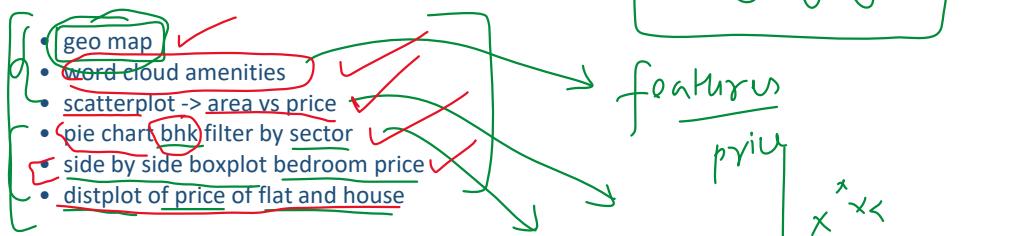
45L (22, 22)

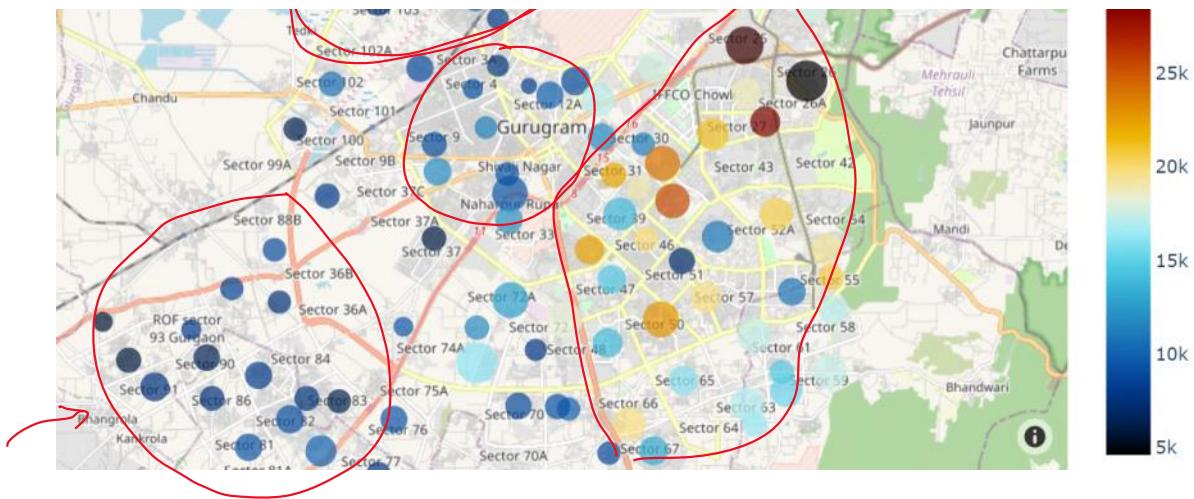


Input
Program → Output

Analytics Module

15 September 2023 15:41





[1, 2, 3]

[4, 1, 2]

[1, 4, 5]

[]

[1, 2, 3, 4, 1, 2, 1, 4, 5, - - -]

A red circle with the word "swing" written inside it in red cursive script. A vertical red line extends downwards from the top center of the circle.

Recommender System

19 September 2023 19:24

Recommender systems are a subclass of information filtering systems that aim to predict the "rating" or "preference" a user would give to an item. There are several types of recommender systems, each with its own strengths and weaknesses:

1. Collaborative Filtering (CF) Recommender Systems

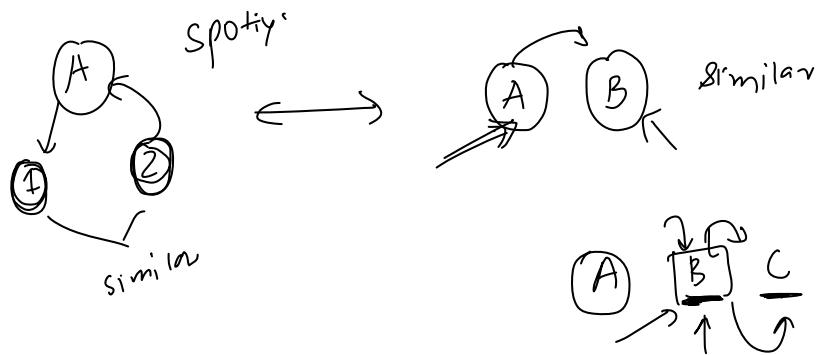
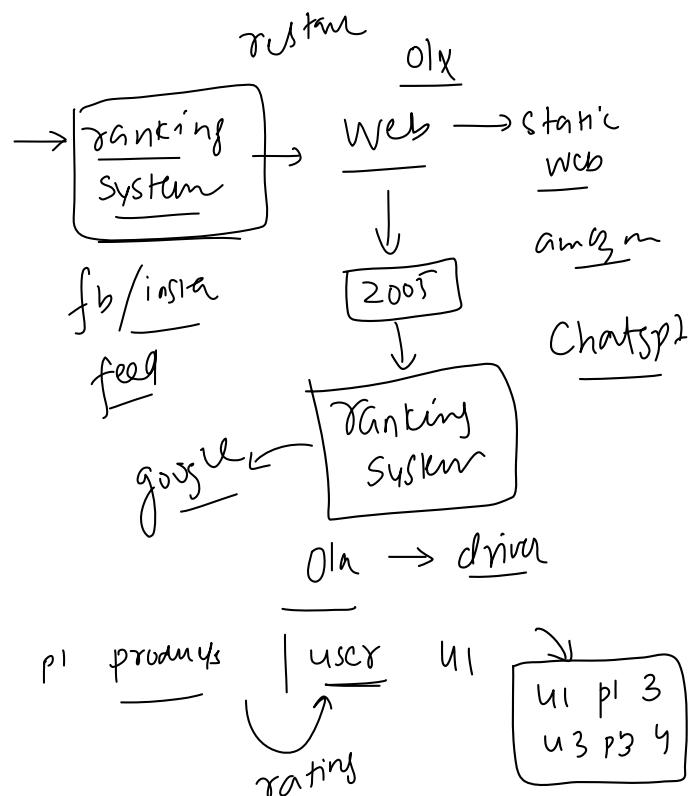
- User-Based Collaborative Filtering (UBCF): This method finds users that are similar to the target user and recommends items that those similar users have liked. It's based on the assumption that users who have agreed in the past tend to agree again in the future.
- Item-Based Collaborative Filtering (IBCF): Instead of finding user similarities, IBCF finds item similarities. If a user likes a particular item, they will likely also like other items that are similar to it.

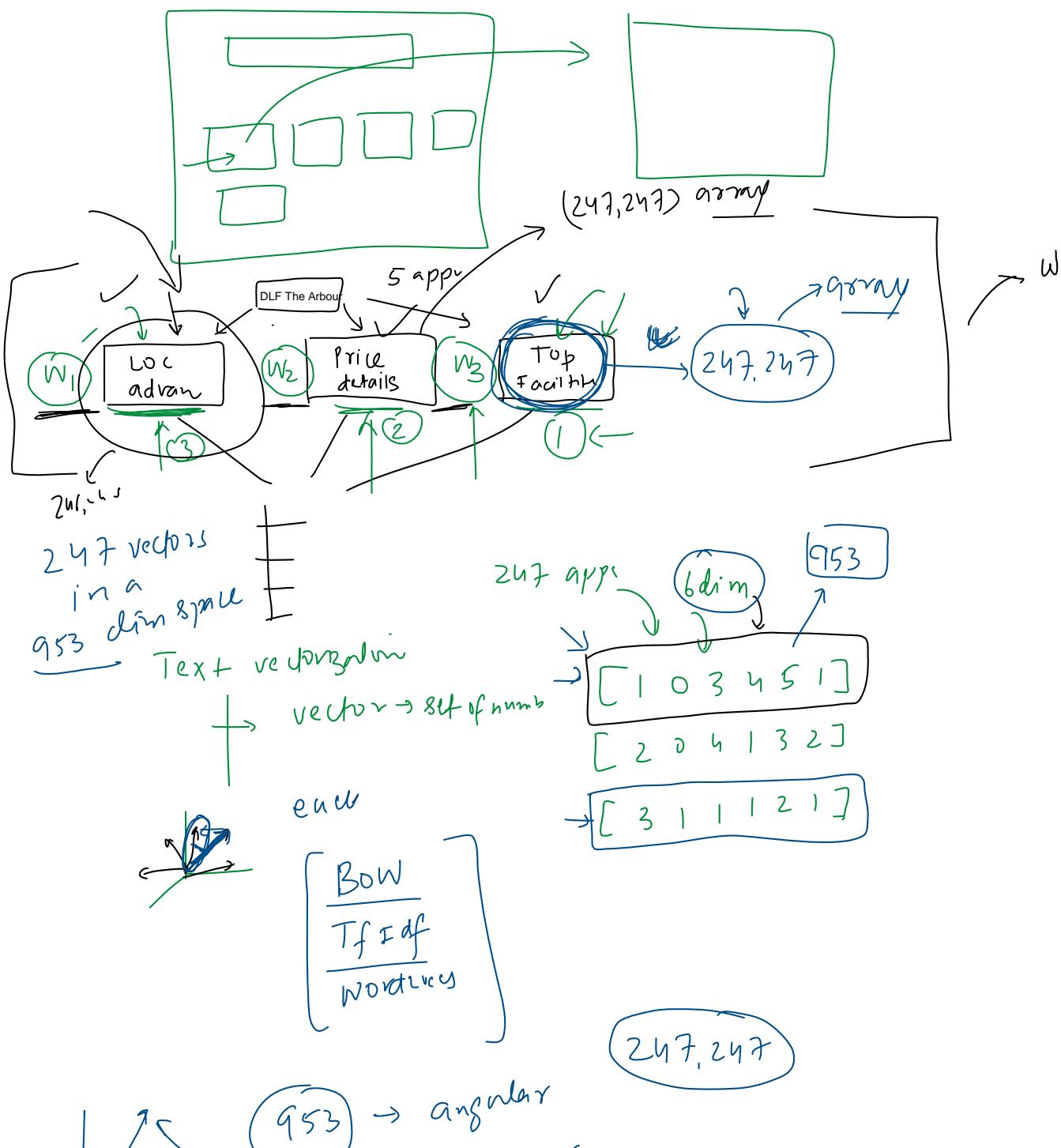
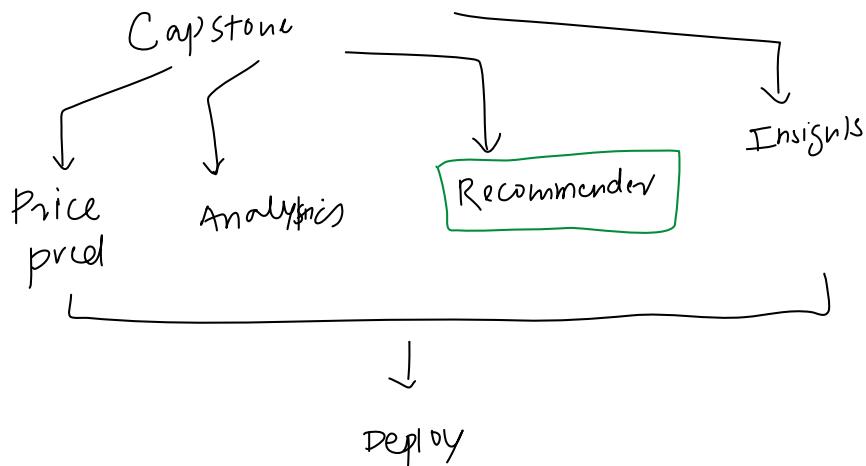
2. Content-Based Recommender Systems:

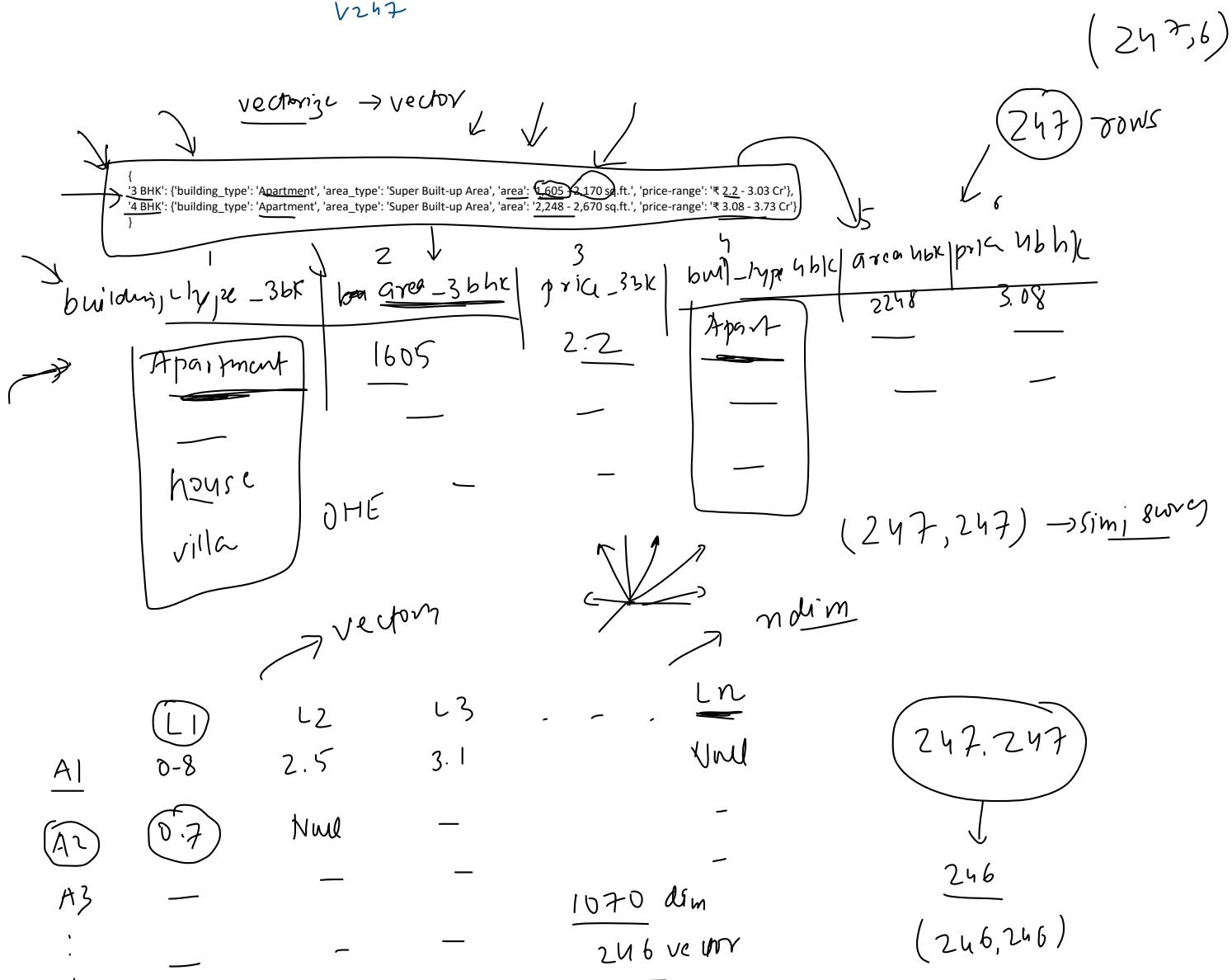
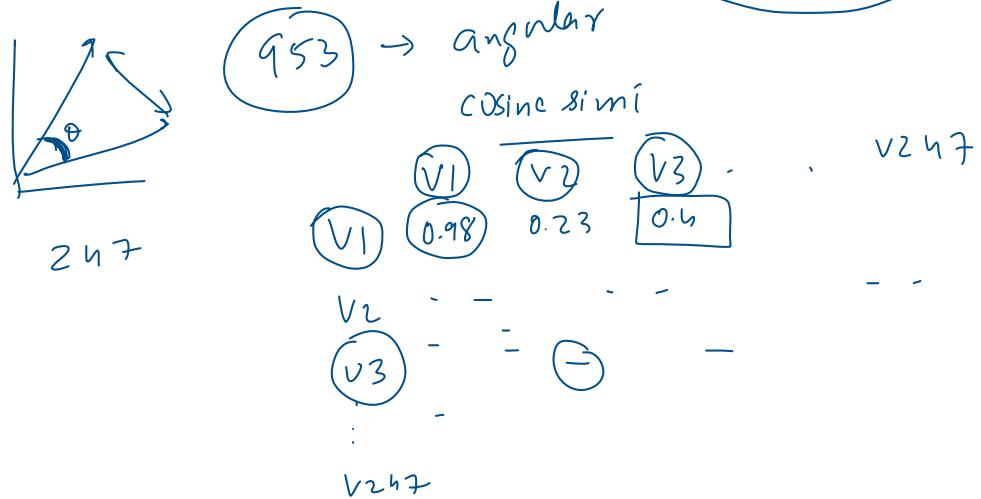
- These systems recommend items by comparing the content of the items and a user profile. Content can be described in terms of several descriptors or terms that are inherent to the item (e.g., a book might be described by its author, its genre, etc.). If a user has interacted positively with certain content attributes in the past, the system will recommend new items with similar attributes.

3. Hybrid Recommender Systems:

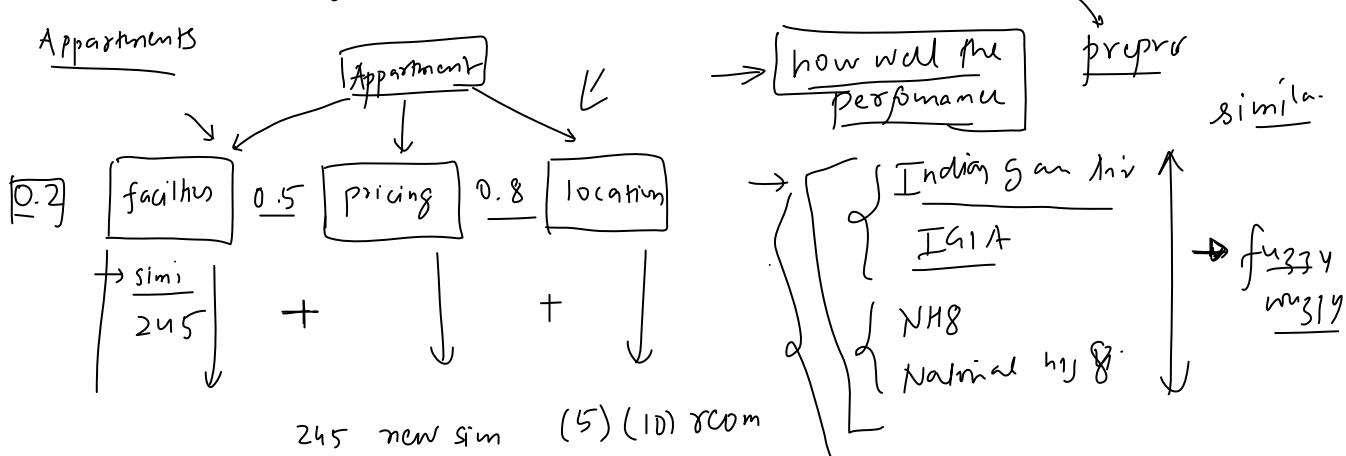
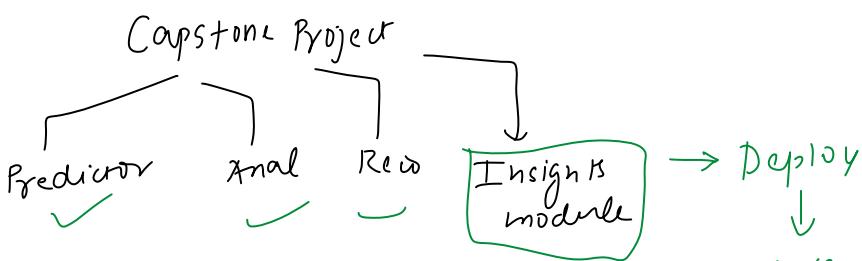
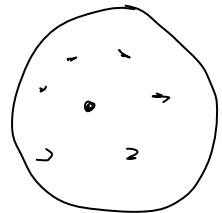
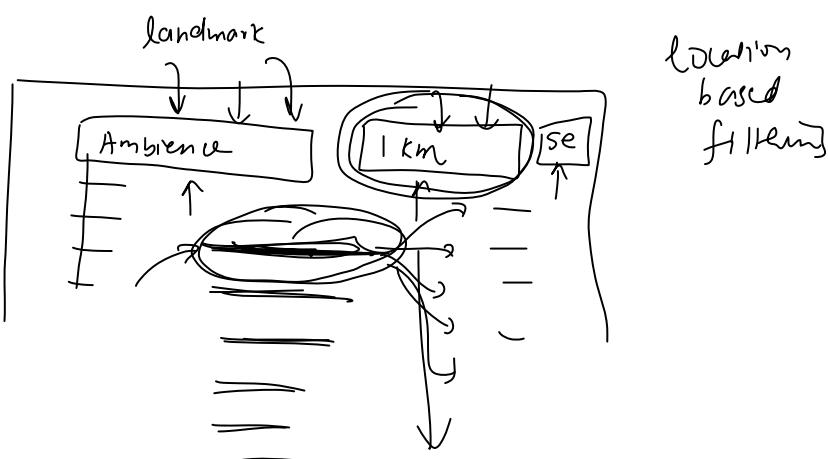
- These systems combine the strengths of both CF and content-based methods. There are several ways to design hybrid systems, such as by making predictions separately with each approach and combining them, adding collaborative and content-based features into a single model, or unifying the models into a single model.







246

Ambience

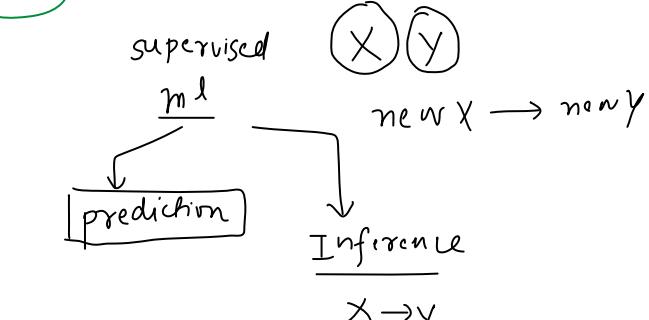
(X) houses
(Y)

understanding
[Insights module]

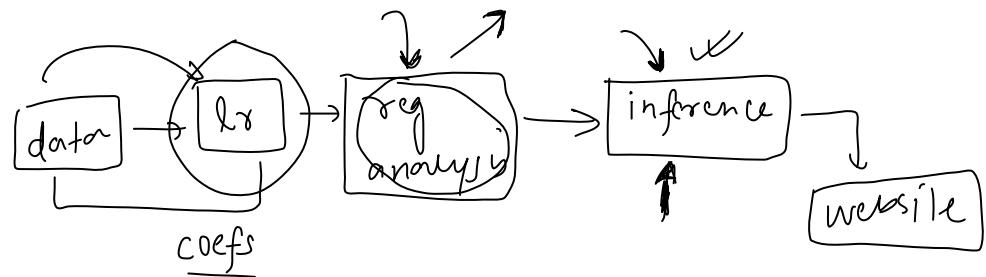
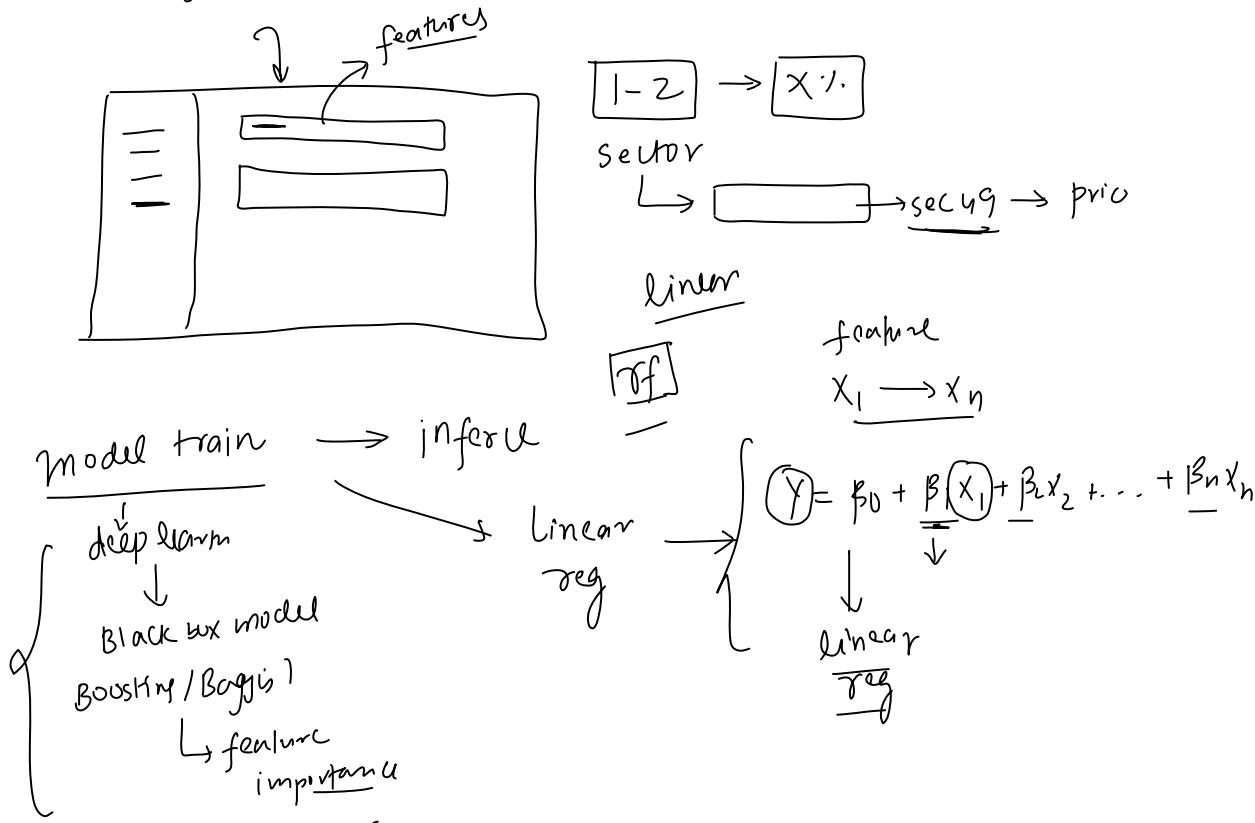
Sector → price ↑ (by percentage)

sec 49 → sec 28 → 16%

area → 100 sqft →
↓ features



T1 → T2 → T3 → T4



bedroom price

1 → 2 $x_1 | x_2 | \dots | x_n | Y$ $\beta_2 = 0.5$

1cr $\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$

Y → log(Y) 1 + 0.5 1 + 0.5 1.5 + 0.5

X → X-scale bedRoom 0.054002 1 Y = 1 $\beta_2 = 0.5$

$\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_n$ 1 → 2 → 3 → 4 2 $y = 1.5$

1cr $1 + 0.5^5$ 3

$= 1cr \ 5^{th}$

1cr 10cr 1cr 1

(a) ctran b) unstem \sqrt{b} $std(b)$

From <http://localhost:8888/notebooks/dsmp-capstone-project/insights-module.ipynb#>

$$\begin{aligned}
 \text{(a) Stan coef (bedroom)} &= \frac{\text{unstndrdn unstem coef (bed)}}{\text{std (log(y))}} \times \frac{\sqrt{b}}{\sqrt{P}} \\
 b = 0.0241 &\rightarrow e^b = 1.0241 \\
 b = \frac{a}{0.054} \times \frac{\sqrt{P}}{\sqrt{b}} &\rightarrow \frac{0.0241}{\sqrt{b}} = 0.557 \\
 y \rightarrow \log(y) &\rightarrow 1.245 \\
 \text{coef} \rightarrow \text{bed} \rightarrow 0.0243 &\rightarrow \text{bedroom} \rightarrow 0.024 \\
 1 - 2 - 3 \rightarrow 1.0241 &\rightarrow 1.0241 \\
 \underline{1 Cr} &\quad \underline{1.0241}
 \end{aligned}$$

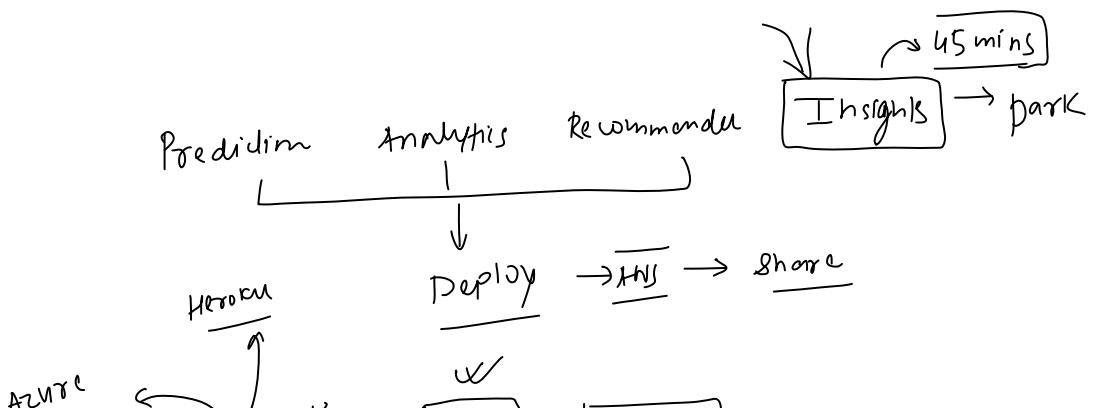
(a) (b)

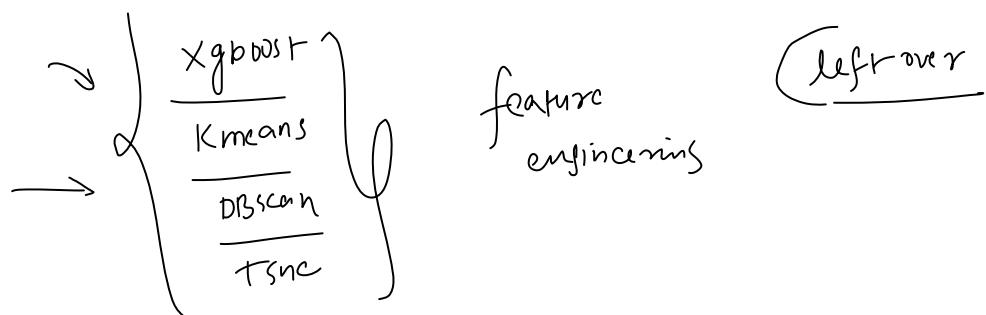
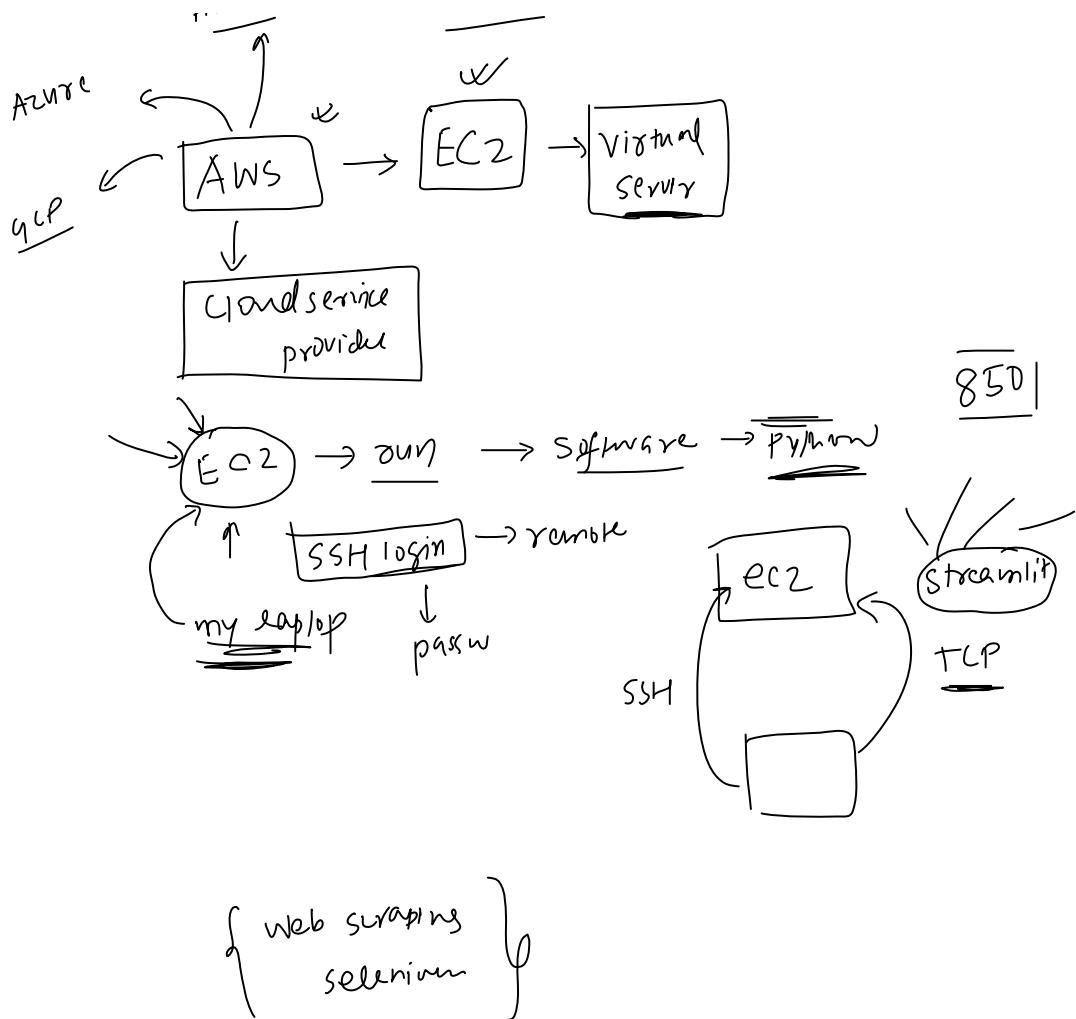
0.210

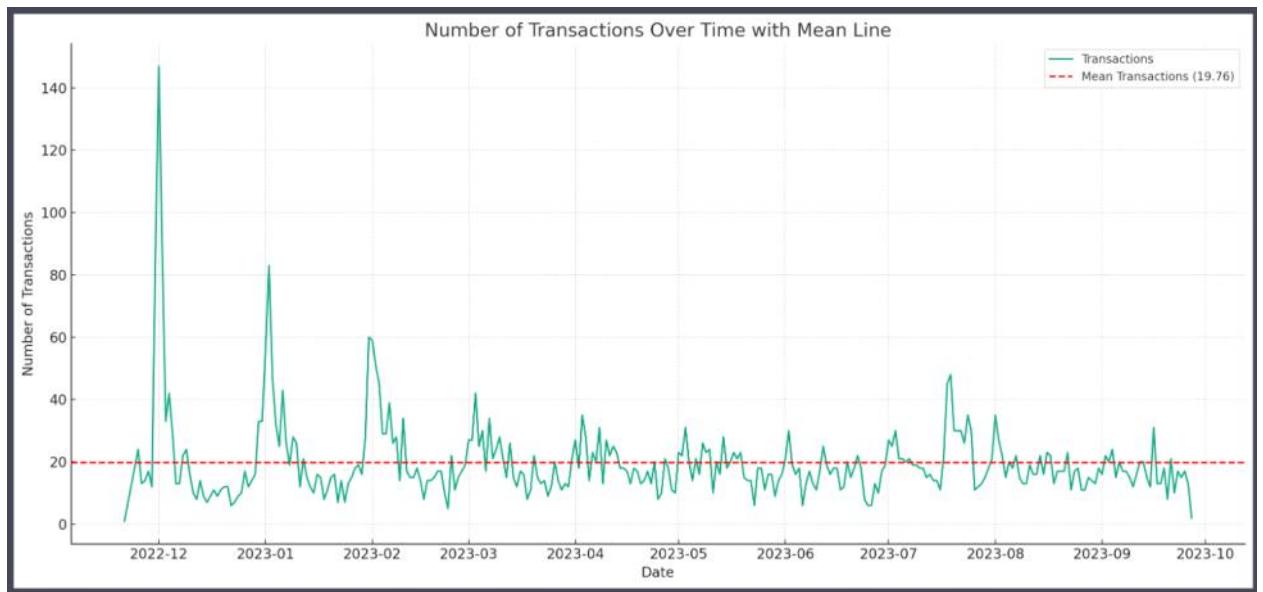
$$b = 0.210 \times \frac{\sqrt{P}}{\sqrt{b_{\text{true}}}} \rightarrow 0.557$$

$$\log(y) \rightarrow e^{\log(y)} \rightarrow y$$

$$SC = \frac{u_C}{\sigma_y} \times \frac{\sqrt{P}}{\sqrt{X}} \rightarrow u_C = SC \frac{\sigma_y}{\sigma_X}$$







Plan of Attack

30 November 2023 10:51