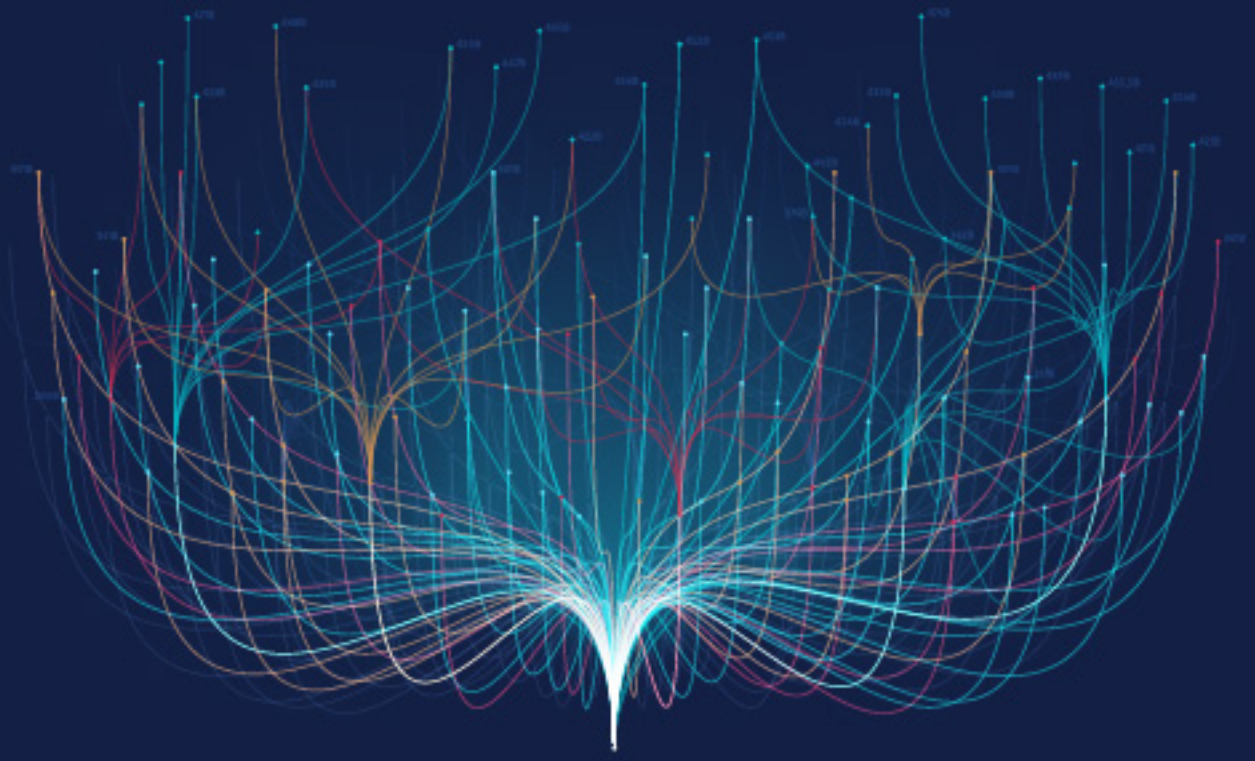
Enough of Theory!

Become a Practical Data Scientist!

100 Days Challenge!



DATA SCIENCE
using Python & R

360 DigiTMG
Digital Transformation | Management | Governance

INDEX

Ingredients of AI

Stages of Analytics

CRISP - DM

- CRISP - DM Data Cleansing / Data Preparation
  - Outlier Treatment
  - Winsorization
  - Alpha Trimmed
  - Missing Values
  - Imputation
  - Transformation
  - Normalization/Standardization
  - Dummy Variables
  - Type Casting
  - Handling Duplicates
  - String Manipulation

- CRISP - DM Exploratory Data Analysis
  - Measure of Central Tendency
  - Measure of Dispersion
  - Measure of Skewness
  - Measure of Kurtosis
  - Graphical Representations
    - Histogram
    - Box Plot
    - Q-Q Plot
    - Bivariate Analysis
      - Scatter Plot
      - Correlation Coefficient
  - Multivariate Analysis
  - Data Quality Analysis
  - Four Errors to Be Avoided During Data Collection
  - Data Integration
  - Feature Engineering
  - Feature Extraction
  - Feature Selection

- CRISP - DM Model Building Using Data Mining
  - Supervised Learning
  - Supervised Learning has Four Broad Problems to Solve:
    - Predict A Categorical Class: Classification
    - Predict A Numerical Value: Prediction
    - Predict User Preference from a Large Pool of Options: Recommendation
    - Predict Relevance of an Entity to a "Query": Retrieval
  - Data Mining Unsupervised
    - A Few of the Algorithms are:
      - Clustering
      - Dimension Reduction
      - Network Analysis
      - Association Rules
      - Online Recommendation Systems

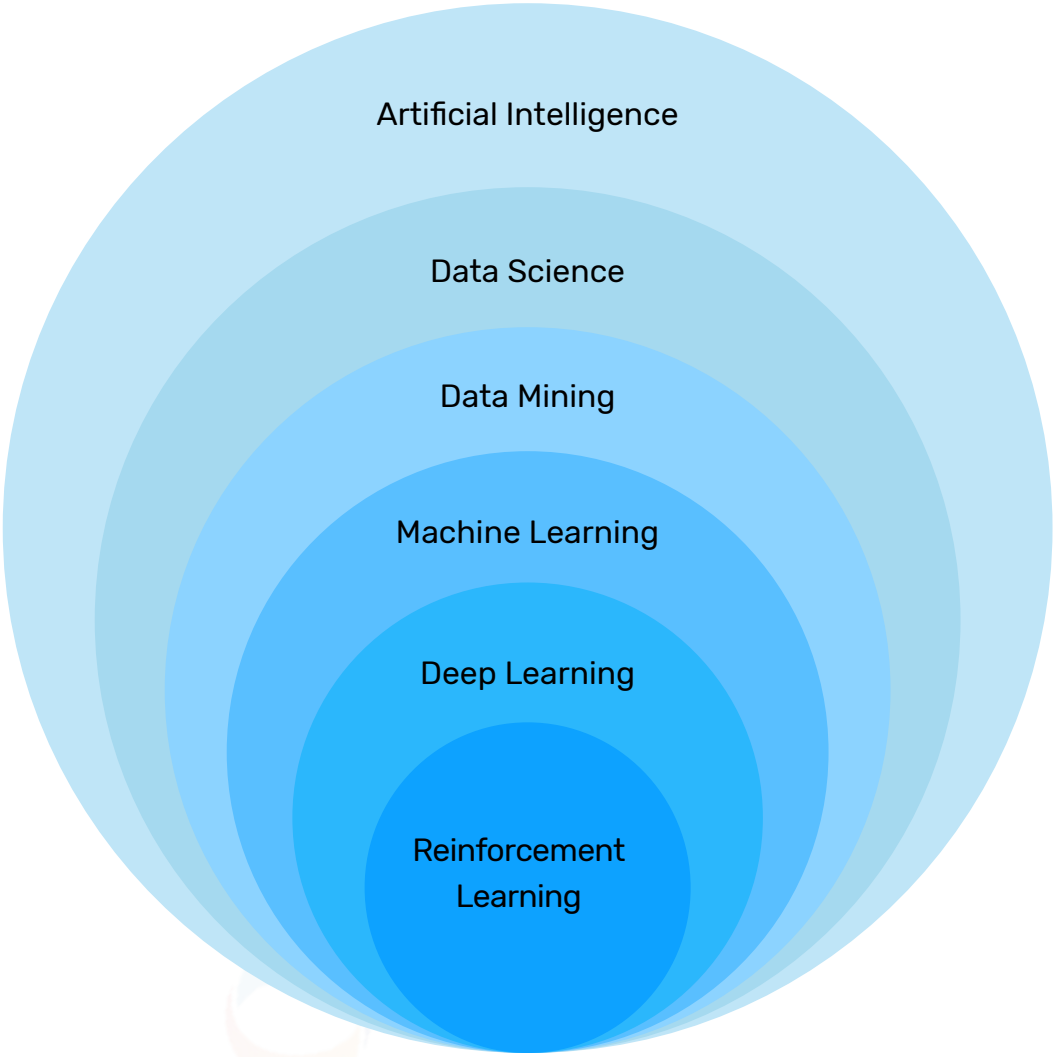
  - Unsupervised Preliminaries
    - Distance Calculation
    - Linkages
  - Clustering / Segmentation
    - K-Means Clustering
    - Disadvantages of K-Means
    - K-Means++ Clustering
    - K-Medians Clustering
    - K-Medoids
    - Partitioning Around Medoids (PAM)
    - CLARA
  - Hierarchical Clustering
    - Disadvantages of Hierarchical Clustering
  - Density Based Clustering: DBSCAN
  - OPTICS
  - Grid-Based Clustering Methods
  - Three Broad Categories of Measurement in Clustering
  - Most Common Measures
  - Clustering Assessment Methods
  - Finding K Value
  - Mathematical Foundations
  - Dimension Reduction
    - PCA
    - SVD
    - LDA

- Association Rules
  - Support
  - Confidence
  - Lift
- Recommender Systems
  - Types of Recommendation Strategies
  - Collaborative Filtering
  - Similarity Measures
  - Disadvantages
  - Alternative Approaches
  - Recommendations vs Association Rules
  - New Users and New Items
- Network Analysis
  - Applications
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality
  - Edge / Link Properties
  - Cluster Coefficient
- Text Mining
  - Examples of Sources
  - Pre-Process the Data
  - Document Term Matrix / Term Document Matrix
  - Word Cloud
  - Natural Language Processing (NLP)
  - Natural Language Understanding (NLU)
  - Natural Language Generation (NLG)
  - Parts of Speech Tagging (Pos)
  - Named Entity Recognition (NER)
  - Topic Modelling
    - LSA / LSI
    - LDA
    - Text Summarization
- Data Mining Supervised Learning
- Machine Learning Primer
  - Key Challenges

- Model Evaluation Techniques
  - Errors
  - Confusion Matrix
  - Cross Table
  - ROC Curve
- K-Nearest Neighbor
  - Choosing K Value
  - Pros and Cons
- Naive Bayes Algorithm
- Decision Tree
  - Three Types of Nodes
  - Greedy Algorithm
  - Information Theory 101
  - Entropy
  - Pros and Cons of Decision Tree
- Scatter Diagram
- Correlation Analysis
- Linear Regression
  - Ordinary Least Squares
  - Model Assumptions
- Logistic Regression
- Support Vector Machine
  - Hyperplane
  - Non-Linear Spaces
  - Kernel Tricks
  - Kernel Functions
- Deep Learning Primer
  - Image Recognition
  - Speech Data
  - Text Data
  - Shallow Machine Learning Models
- Perceptron Algorithm
  - Biological Neuron
  - Simple Neural Network Components
  - Perceptron Algorithm
  - Learning Rate
  - Gradient Primer
  - Gradient Descent Algorithms Variants
  - Empirically Determined Components

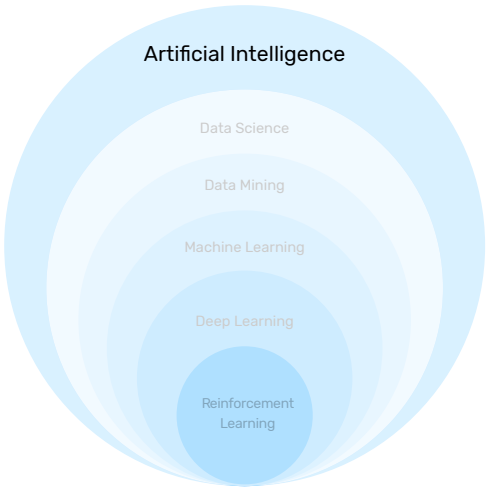
- Multi-Layers Perceptron (MLP) / Artificial Neural Network (ANN)
  - Non-Linear Patterns
    - Integration Function
    - Activation Function
  - Regularization Techniques Used for Overfitting
    - Error-Change Criterion
    - Weight-Change Criterion
  - Dropout
  - Drop Connect
  - Noise
  - Batch Normalization
  - Shuffling Inputs
  - Weight Initialization Techniques
- Forecasting
  - Time Series vs Cross Sectional Data
  - EDA - Components of Time Series
- Systematic Part
  - Level
  - Trend
  - Seasonality
    - Non-Systematic Part
  - Noise/Random
  - Data Partition
  - Forecast Model
    - Model-Driven Techniques
    - Data-Driven Techniques
  - Smoothing Techniques
    - Moving Average
    - Exponential Smoothing
  - De-Trending and De-Seasoning
    - Regression
    - Differencing
    - Moving Average

Ingredients of AI



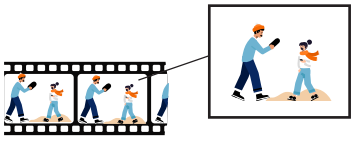
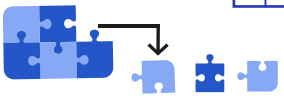
Definition of Artificial Intelligence, Data Science, Data Mining, Machine Learning, Deep Learning, Reinforcement Learning (RL)

Artificial Intelligence

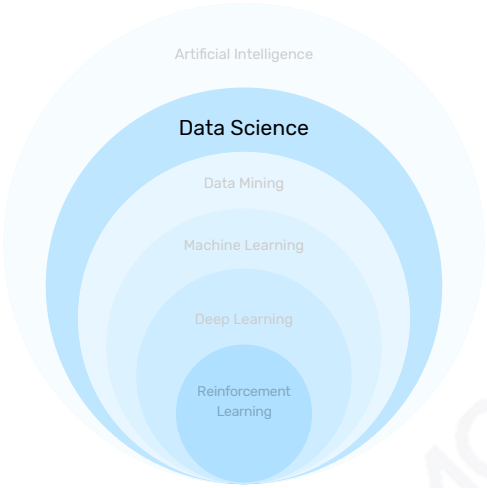


Ability of inanimate objects such as Machines, Robots, Systems, etc., with computing capabilities to perform _____ tasks that are similar to Humans.

Examples of AI

- _____ (Video Analytics & Image Processing) 
- Hearing (Speech to Text Applications) 
- Response to Stimuli (Inputs)  y = f(x)
- _____ 

Data Science

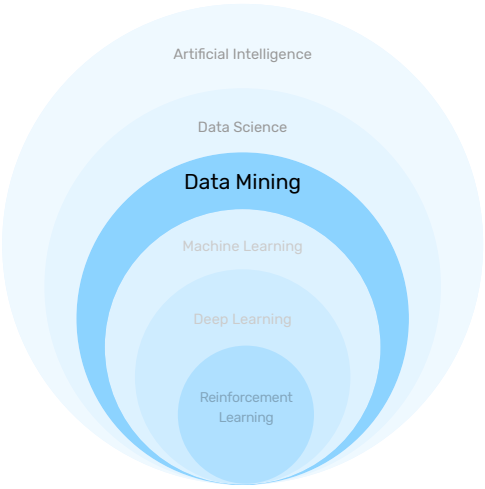


Data Science is a field of study related to data, to bring out meaningful _____ insights for effective _____ making.

Topics of Data Science includes

1. _____ Analysis
2. Hypothesis Testing
3. Data _____
4. Regression Analysis
5. Classification Techniques
6. Black Box Techniques
7. _____ Mining
8. Natural Language Processing
9. _____ Analysis, etc.

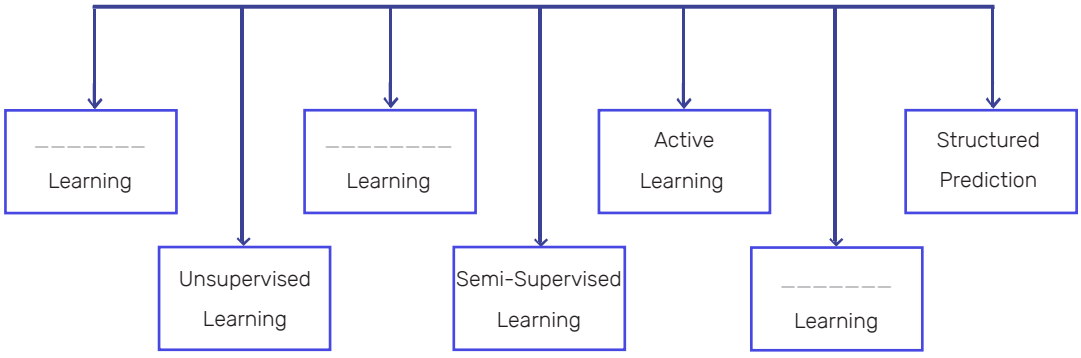
Data Mining



Data Mining is similar to coal mining where we get coal and if lucky one might get precious stones such as diamond.

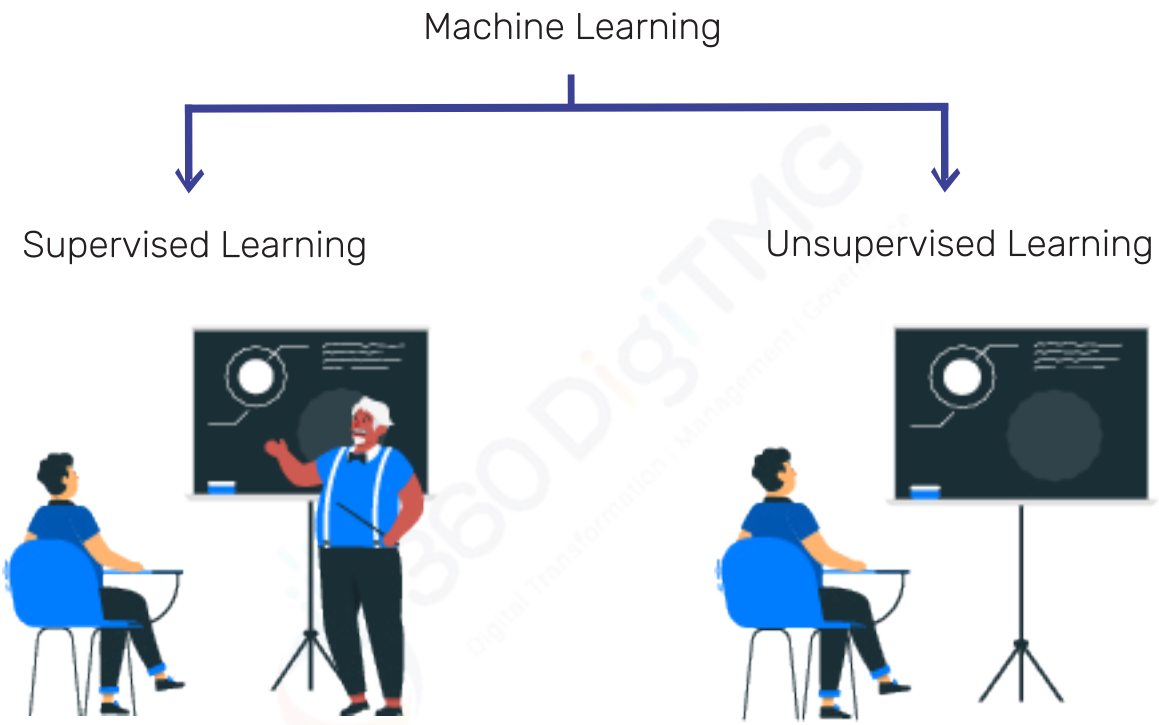
In Data Mining we get _____ from _____ and insights similar to diamond are extremely valuable for _____.

Data Mining (Branches)



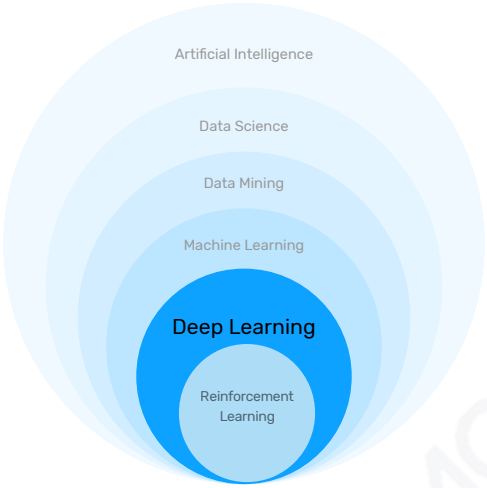
Machine Learning

Machine Learning is learning from the _____ of the historical / past data and then using it on the _____ unseen data to achieve a defined objective.



1. _____ Learning _____ Learning - Both _____and _____ are known in the historical data
2. _____ Learning _____ Learning - Only_____ are known in the historical Data & _____ is not known or assumed as not known

Deep Learning



Deep Learning is a special branch of Machine Learning where the _____ in data are _____ extracted.

Some of the Deep Learning Architecture

- _____ / Multi-Layered Perceptron
- _____ Neural Network
- _____ Neural Network
- Deep Belief Network
- Long Short Term Memory (LSTM)
- Gated Recurrent Units (GRUs)
- Mask R-CNN
- Autoencoders
- Generative Adversarial Network (GAN)
- Boltzmann Machine
- Deep Q-Networks
- Q Learning etc.

Reinforcement Learning

Reinforcement Learning is a special branch of _____ Learning which is heavily used in applications sincluding games, robotics, investment banking, and trading etc.

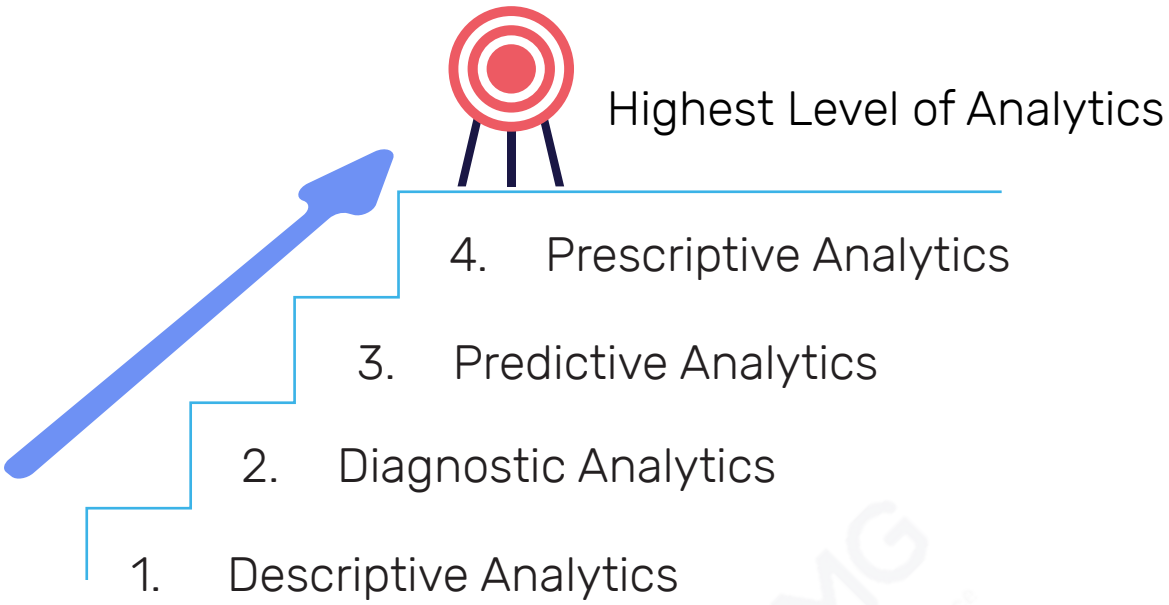
Reinforcement Learning is a _____ based learning, which solves sequential decision problems by _____ with the environment.

The 5 key elements of Reinforcement Learning

_____	_____ is a learning component that makes decision on actions to maximize the reward.
Environment	Environment is the physical world where agents perform actions.
_____	_____ defines behavior of the agent from states to actions.
_____ _____	_____ defines the problem and maps it to a numerical reward.
_____ _____	_____ defines the cumulative future reward.

Model of the Environment is an optional component which predicts the behavior of the environment.

Stages of Analytics



_____ - Answers questions on what happened in the past and present.

Example: Number of Covid-19 cases to date across various countries

Diagnostic Analytics - Answers questions on _____.

Example: Why are the Covid-19 cases increasing?

_____ - Answers questions on _____.

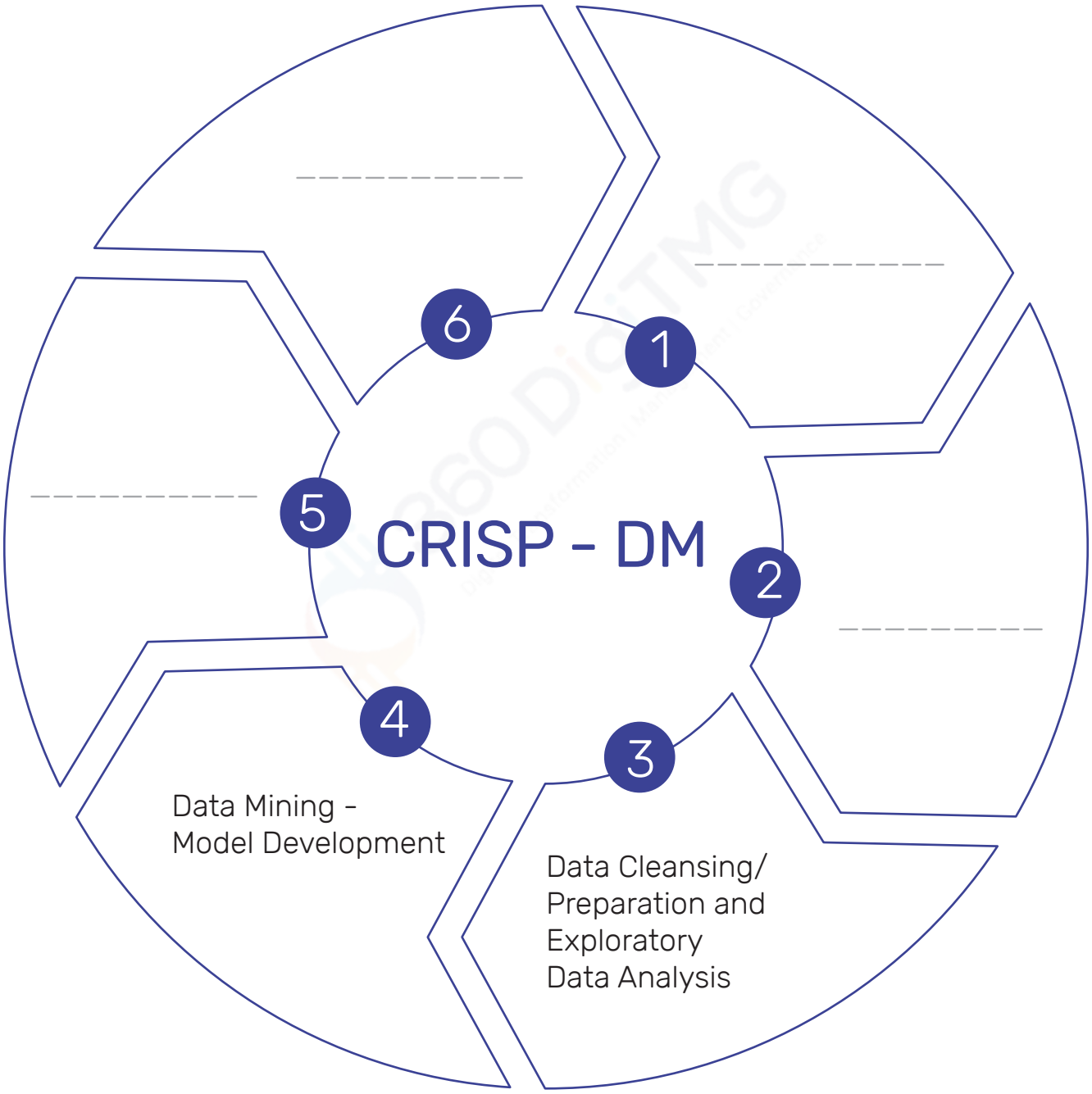
Example: What will be the number of Covid-19 cases for the next month?

_____ - Provides remedies and solutions for what might happen in the future.

Example: What should be done to avoid the spread of Covid-19 cases, which might increase in the next one month?

CRISP - DM

C_____ I_____ S_____ P_____
for Data Mining



CRISP – DM Business Understanding

Articulate the business problem by understanding the client/customer requirements

_____ Objectives	Business _____

A few examples on Business Objective and Business Constraints

Business Problem : Significant proportion of customers who take loan are unable to repay

Business Objective : _____ Loan Defaulters

Business Constraint : Maximize Profits

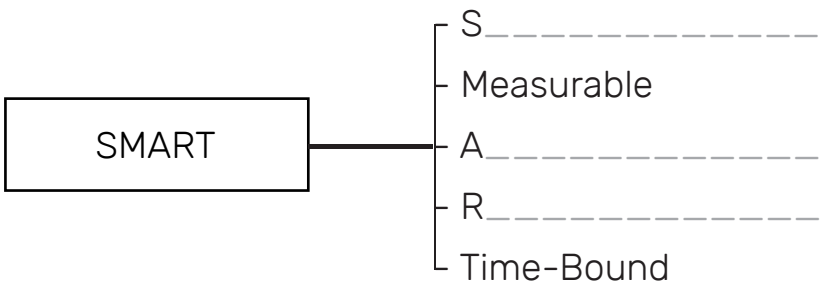
Business Problem : Significant proportion of customers are complaining that they did not do the credit card transaction

Business Objective : Minimize Fraud

Business Constraint : _____ Convenience

Keys points to remember:

Ensure that objectives and constraints are SMART



Key Deliverable: _____

CRISP - DM Data Collection

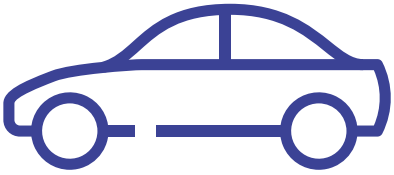
Understanding various _____ is pivotal to proceed further with data collection.

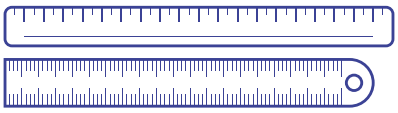
Data Types

_____ _____

Any data, which can be represented in a _____ and makes sense.

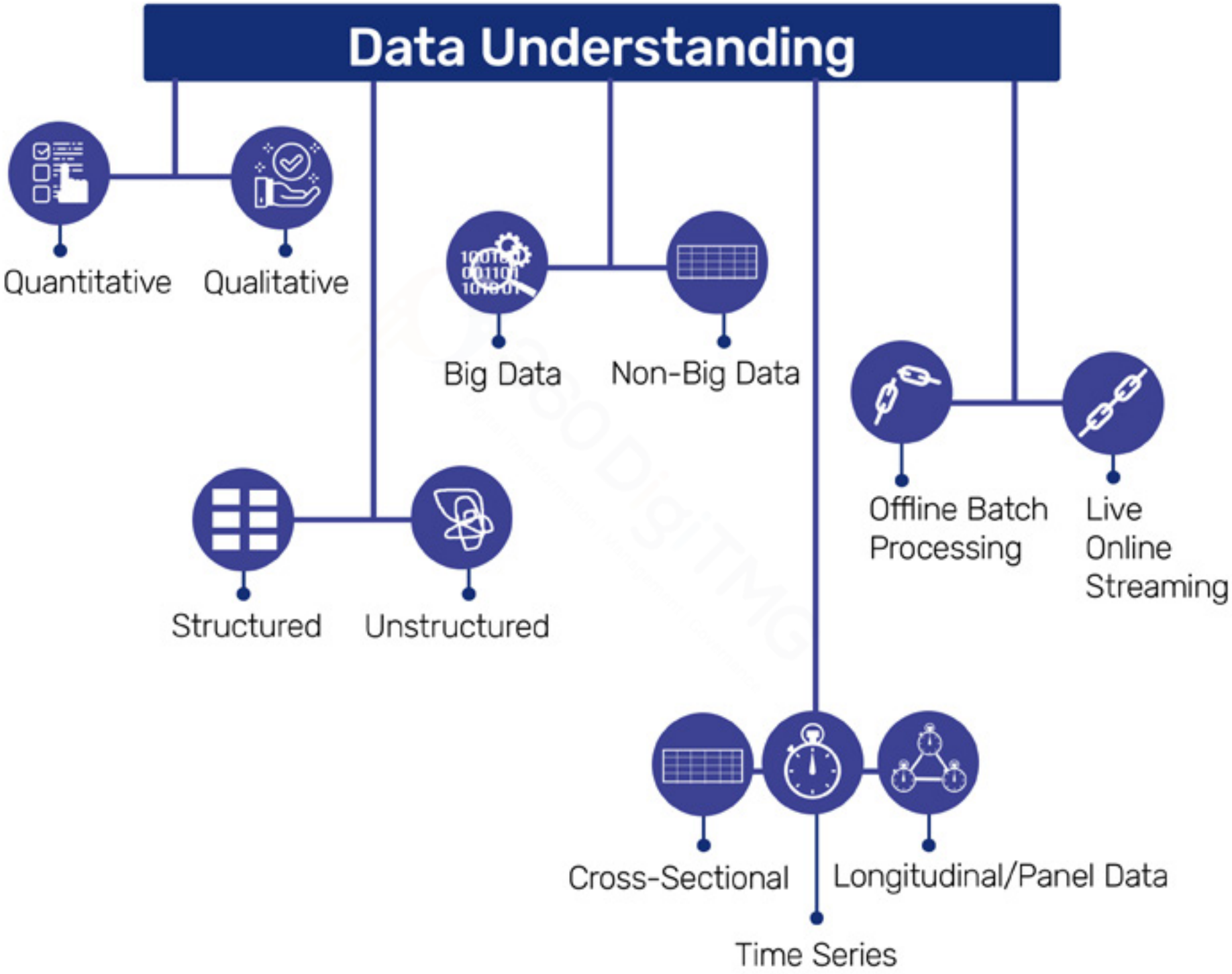
Data when represented in decimal format does not make sense.





CRISP – DM Data Collection

Count Data examples



Data Understanding

_____ VS _____

_____ data is non-numerical data.

Examples

1. This weighs heavy
2. That kitten is small

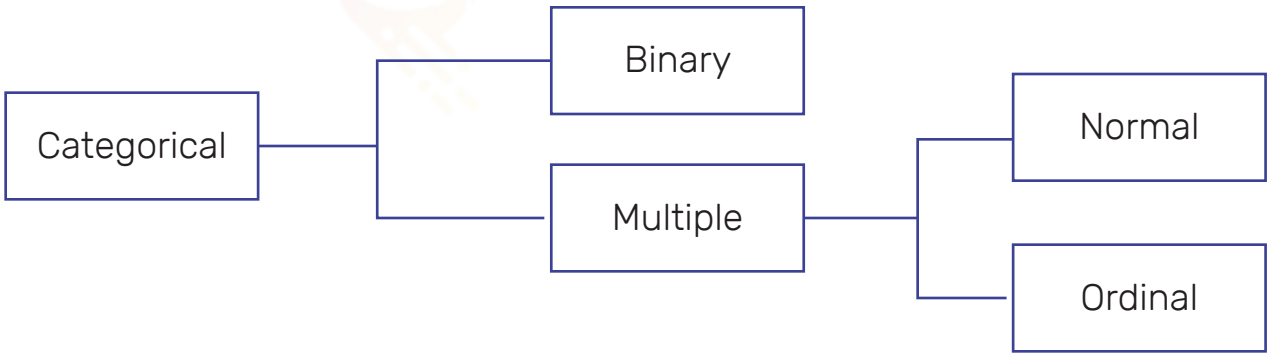
Quantitative data include numbers.

Examples

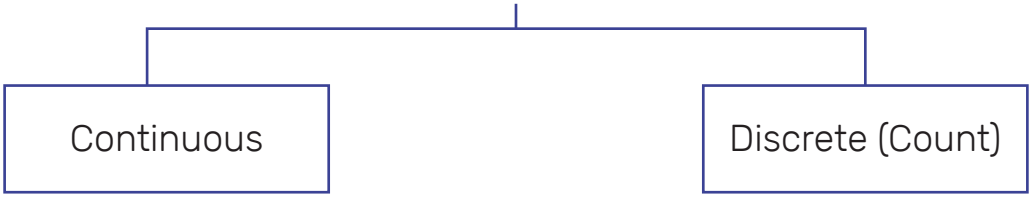
1. Weight 85 kg
2. Height 164.3 cm

_____ Data and _____ Data fall under Quantitative Data.

Qualitative

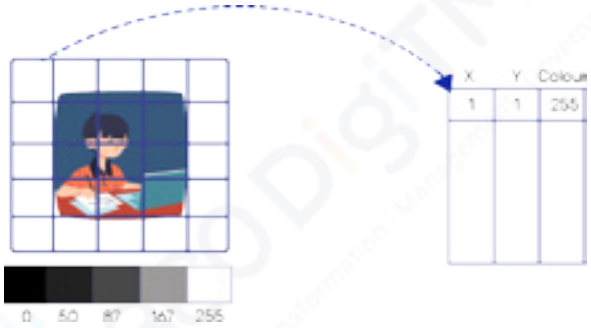


Quantitative



_____ vs Unstructured

Structured data is that data which in raw state can be placed in a _____ format.	Unstructured data is that data which in its raw state cannot be placed in any _____ format.
Video is split into images and images into _____ and each pixel intensity value can be an entry in a column and this becomes structured.	Videos, Images, Audio/Speech, Textual Data are examples of Unstructured data.



Audio Files / Speech data can be converted into features using _____ Frequency _____ Coefficient (MFCC).

Textual data can be converted into _____ of _____ (BoW) as an example to make it Structured.

Example: But I, being poor, have only my dreams; I have spread my dreams under your feet; Tread softly because you tread on my dreams.

Poor	Dream	Spread	Feet	Tread	Soft
1	3	1	1	1	1

Data Understanding

_____ VS _____

Data which is governed by the 5 Vs

5 Vs
- _____
- _____
- _____
- _____
- _____

High Volume, generating at rapid Velocity, from a wide Variety with an element of uncertainty _____, and appropriate Value

_____ is that which cannot be stored in the available hardware and cannot be processed using available software.

_____ is that data which can be stored in the available hardware and can be processed using available software.

Cross-Sectional vs _____ vs _____

1. Cross-sectional data is that data, where date, time, and sequence in which we arrange the data is immaterial
2. Cross-sectional data usually contains more than one variable

Examples:

1. Population survey of demographics
2. Profit & Loss statements of various companies

1 _____ data is that data, where the date, time, and sequence in which we arrange the data is important
2. _____ data usually contains only one variable of interest to be forecasted

Examples:

1. Monitoring patient blood pressure every week
2. Global warming trend

1. _____ is also called _____
2. _____ includes properties of both Cross-Sectional. and Time Series
3. Data as well as _____ , wherein there is more than one variable, which are sorted based on the date and time

Examples:

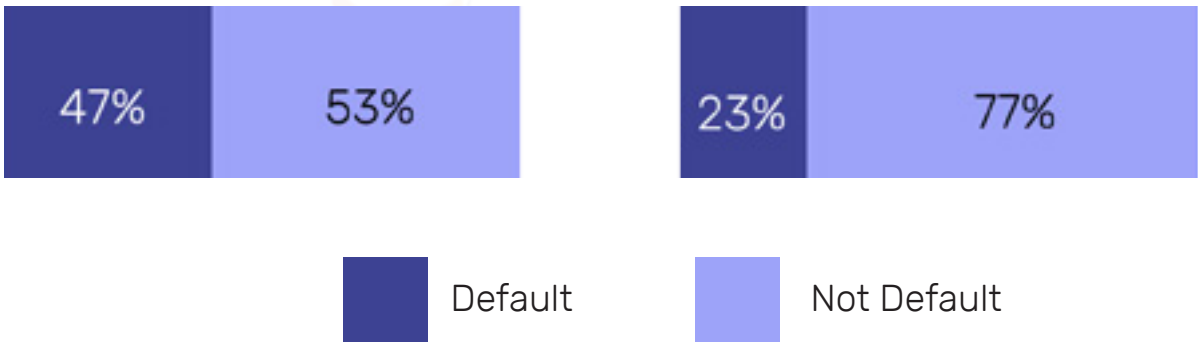
1. Exam scores of all students in a class from sessional to final exams
2. Health scores of all employees recorded every month

_____ VS _____

- Whether a person claims an insurance or not,
- Will a person 'pay on time, 'with a delay' or 'will default', etc,

- _____ is that data where the classes of output variables are more or less in equal proportion
- E.g. 47% of people have defaulted and 53% of data is not defaulted in the loan default variable
- When we have balanced data then we can simply apply random sampling techniques

- _____ is that data where the classes of output variables are in unequal proportion
- E.g. 23% of data is defaulted and 77% of data is not defaulted in the loan default variable



Thumb Rule: if proportion of minority output class is < 30% then data is imbalanced.

Sampling for imbalanced data refer to next page.

When we have imbalanced data then we apply different sampling techniques such as:

- _____ - Undersampling and Oversampling
- Bootstrap Resampling
- K-Fold Cross Validation
- _____ K-Fold Cross Validation
- _____ K-Fold Cross-Validation
- _____ (N-Fold Cross-Validation) LOOCV
- SMOTE (Synthetic Minority Oversampling Technique)
- MSMOTE (Modified SMOTE)
- Cluster-Based Sampling
- Ensemble Techniques

Imbalanced Data

Data Collection Sources

_____: Data Collected at the Source

_____: Data Collected Before Hand

Primary Data

Secondary Data

Primary Data

Examples of _____

Surveys, _____ of _____, _____ Sensors Data, Interviews, Focus Groups, etc.

Survey steps:

1. Understand the business _____ and _____ behind conducting the survey. E.g. Sales are low for a training company
2. Perform _____ analysis - _____ Analysis, 5-_____ Analysis, etc. E.g. Product Pricing is uncompetitive
3. Formulate Decision Problem. E.g. Should product prices be changed
4. Formulate Research _____. E.g. Determine the price elasticity of demand and the impact on sales and profits of various levels of price changes
5. List out Constructs. E.g. Training Enrolment
6. Deduce Aspects based on construct. E.g. Time aspect, Strength aspect, Constraint aspect
7. Devise Survey _____ based on the _____. E.g. I am most likely to enroll for the training program in: In the next one week, In the next one month, In the next one quarter, etc.

_____of _____ examples:

- Coupon marketing with a 10% discount vs 20% discount, to which of these customers are responding well
- Coupon targeting customers within 10 km radius versus 20 km radius
- Combinations of discount & distance to experiment

Secondary Data

Organizational data are stored in _____	_____ databases	_____ (paid) databases

- Oracle DB
- Microsoft DB
- MySQL
- NoSQL - MongoDB
- Big Data, etc.

- Industry reports
- Government reports
- Quasi-government reports, etc.

Meta Data Description: Data about Data

- Obtaining meta data description is mandatory before we proceed further in the project
- Understand the data volume details such as size, number of records, total databases, tables, etc.
- Understand the data attributes/variables - description and values which these variables take

Preliminaries for Data Analysis

Probability can be explained as the extent to which an event is likely to occur, measured by the ratio of the _____ cases to the whole number of cases possible.

$$\text{Probability} = \frac{\#\ \_\_\_\_\_}{\#\ \text{Total events}}$$

Properties of Probability:

- Ranges from 0 to 1
- Summation of probabilities of all values of an event will be equal to 1

Example:

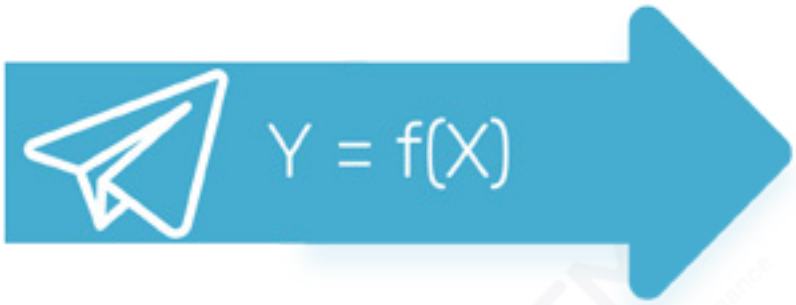
$$P(H) = \frac{H}{H\ \&\ T} = \frac{1}{2} = 0.5$$

$$P(\text{Red}) = \frac{2}{2(R)+2(B)} = \frac{2}{4} = 0.5$$

Base Equation

Random variables can be broadly classified as Output and Input variables.

Mathematically the relation between these is expressed using base equation:



Y is known as:

- _____ variable
- Response
- _____
- Explained variable
- Criterion
- Measures variable
- _____ variable
- _____
- _____ variable

X is known as:

- _____
- Explanatory
- _____
- Covariates
- _____
- Factors
- _____
- Controlled variable
- _____ variable
- Exposure variable

_____

If there is a chance / probability associated with each of the possible output then it is called _____

Any output on any event which _____ is called Variable

_____ are always represented using Upper case.

Values that a random variable takes are represented using _____.

Ex: Roll of a single die

X = {1, 2, 3, 4, 5, 6}

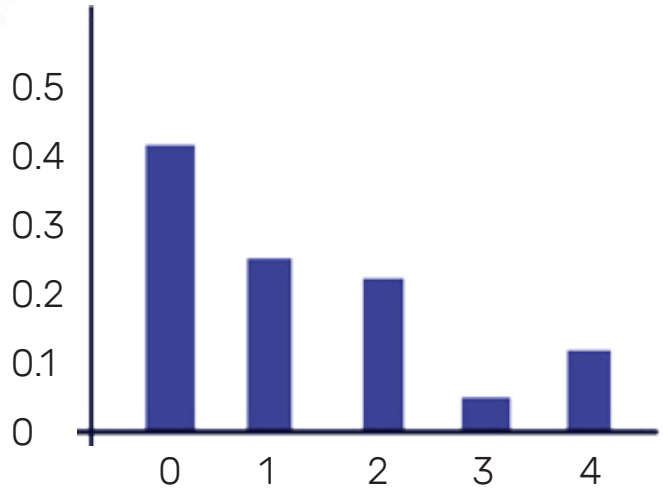
Probability Distribution

Representing the probabilities of all possible outcomes of an event in a tabular format or a graphical representation is called Probability Distribution.

If a random variable is continuous then the underlying probability distribution is called _____.

If a random variable is discrete then the underlying probability distribution is called _____.

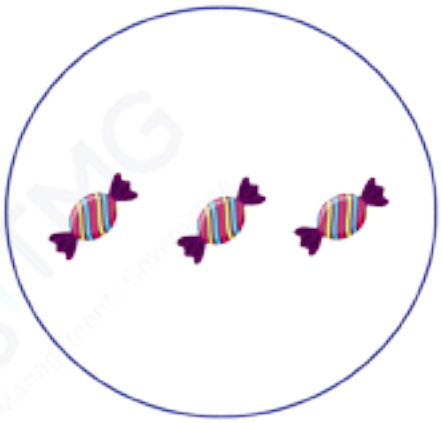
X	P(X=x)
0	0.40
1	0.25
2	0.20
3	0.05
4	0.10



Sampling Techniques

Sampling is a technique to collect the _____ of population data.

These techniques are broadly classified into 2 types.

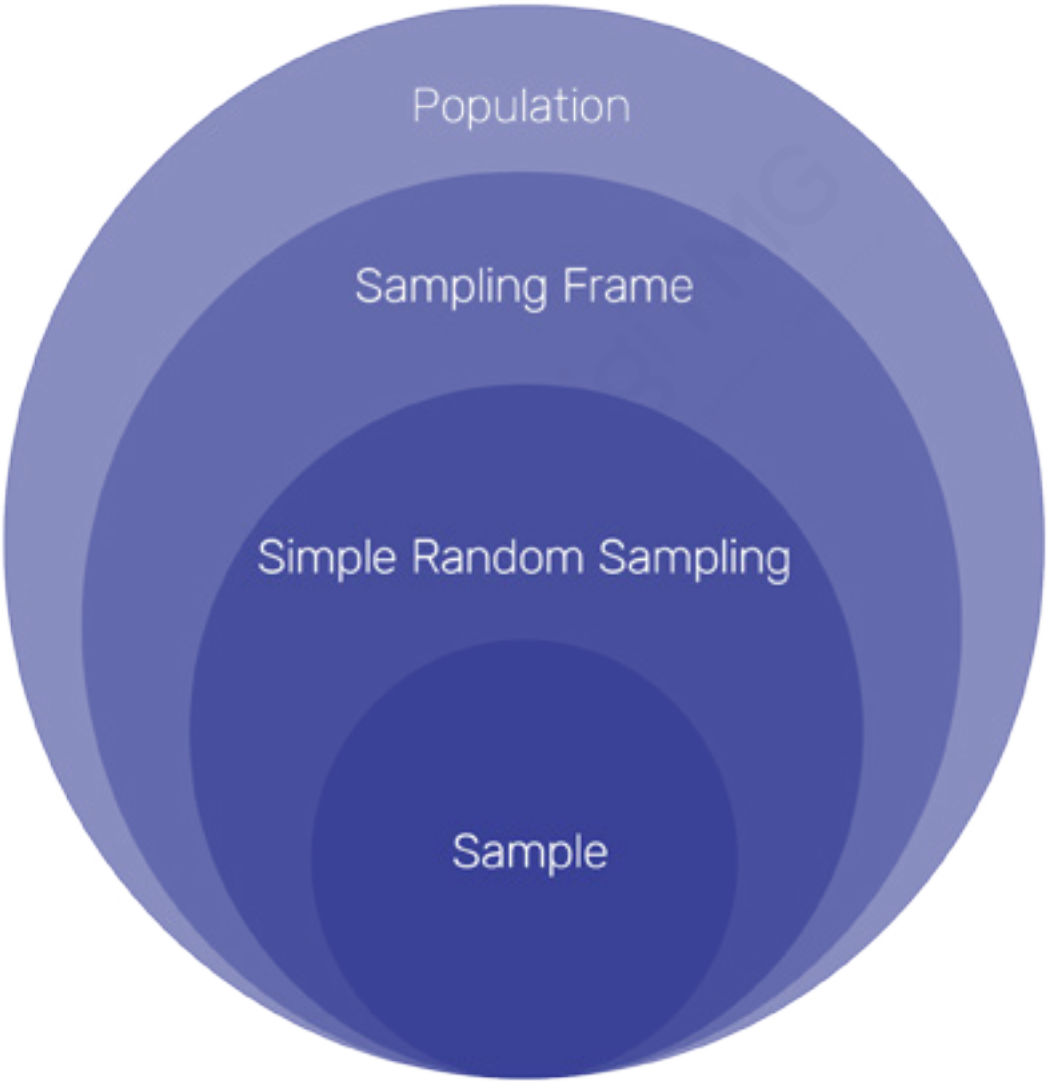


_____ Sampling

_____ Sampling

Inferential Statistics

Inferential statistical is a process of analysing the _____ and deriving statements / properties of a _____.



Sampling Techniques

Sampling Techniques is a technique which is based on convenience, wherein, priority varies for the data that is to be collected to represent the population, these approaches are also known as _____ sampling.

A few examples of Non-Probability Sampling:

1. Convenience Sampling
2. Quota Sampling
3. Judgment Sampling
4. Snowball Sampling

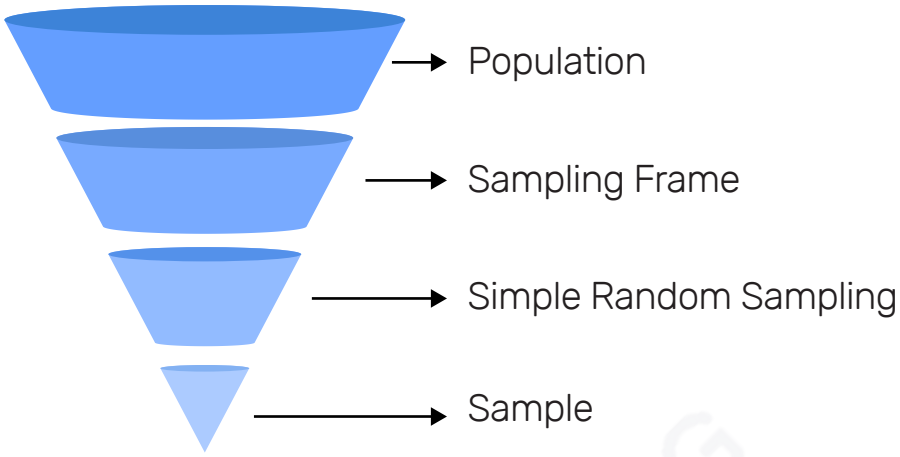
Sampling Techniques

_____ Sampling also known as _____. Sampling is the default approach for inferential statistics. Each data point to be collected will have _____ to get selected.

A few examples of _____ Sampling:

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Clustered Sampling

Sampling Funnel



Population	All Covid-19 cases on the planet
Sampling Frame	• The majority of Covid-19 cases are in the USA, India, and Brazil and hence these 3 countries can be selected as a _____ • _____ does not have any hard and fast rule, It is devised based on business logic
Simple Random Sampling	• Randomly Sample 10% or 20% of the data from the sampling frame using Simple _____ technique • _____ is the gold standard technique used for sampling • _____ is the only sampling technique which has no bias • Other sampling techniques such as Stratified sampling, etc., Also can be used to sample the data but _____ is the best

CRISP - DM Data Cleansing

Data Cleansing / Data Preparation

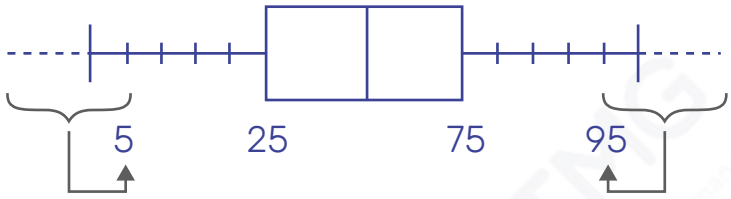
Data Cleansing is also called as _____, _____, _____, _____.

Outlier or_____ - Any value, which is extremely small or extremely large from the remaining data.

Outliers are treated using 3 R technique:

Winsorization Technique

Winsorization is the technique, which modifies the sample distribution of random variables by removing outliers. For example, 90% winsorization means all data below the 5th percentile is set at 5th percentile and all the data above the 95th percentile is set at 95th percentile.



All values below 5th percentile are changed to 5th percentile value

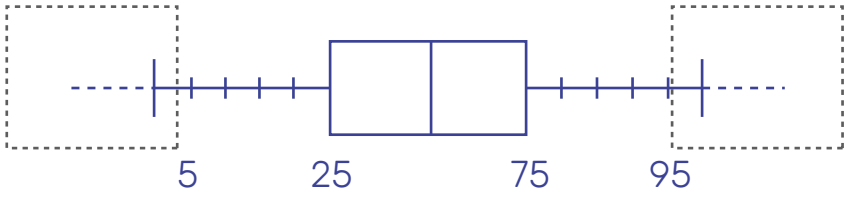
All values above 95th percentile are changed to 95th percentile value

Alpha Trimmed Technique

Alpha Trimmed Technique lets you set an alpha value, for example if alpha = 5%, then all the lower & upper 5% values are trimmed or removed.

Lower 5% values are removed/trimmed

Upper 5% values are removed/trimmed



Missing Values

Missing Values - Fields in the data which might have blank spaces and (or) NA,

_____, _____, _____.

3 Variants of Missing Values

- _____ (MAR)
- Missingness Not At Random (MNAR)
- _____ (MCAR)

Name	Age	Salary
Steve	23	$ 4,000
Raj	33	$ 6,500
Chen	41	
Wilma	37	$ 7,200
Audrey	51	$ 9,300

Imputation

Imputation is a technique used to replace missing values with logical values.

Wide variety of Techniques are available, choosing the one which fits the data is an art:

(Simple Strategies)

- List-Wise Deletion or Complete-Case Analysis
- Available Case Method or Pair-Wise Deletion

Single Imputation Methods

- Mean Imputation
- Median Imputation
- Mode Imputation
- Random Imputation
- Hot deck Imputation
- Regression Imputation
- KNN Imputation

Transformation

_____

Types of transformation

- Logarithmic
- _____
- Square Root
- _____
- Box-Cox
- Johnson

_____ / Binning / Grouping - Converting _____ data to _____

Binarization - Converting continuous data into _____

Rounding - Rounding off the decimals to the nearest integer e.g. 5.6 = 6

Binning - Two types of Binning

- Fixed Width Binning
- Adaptive Binning

Normalization

Normalization/ _____ - Making the data _____ and _____.

$$Z = \frac{X - mean}{stdev}$$

Methods of Normalization / _____ include _____ also called as _____ , _____ also called as _____ or_____.

Standardization has two parts:

- _____ or Mean Subtraction - Mean Normalization will make the mean of the data _____.
- Variance Normalization - _____ will make the variance of the data _____.

$$\frac{X - min(x)}{max(x) - min(x)}$$

Normalization is also called the _____. Normalized data has minimum value = 0 and maximum value = 1 and sometimes when dealing with negative values the range can be in between -1 to +1.

Mix-Max Scaler's disadvantage is that its scaled values are influenced by _____.

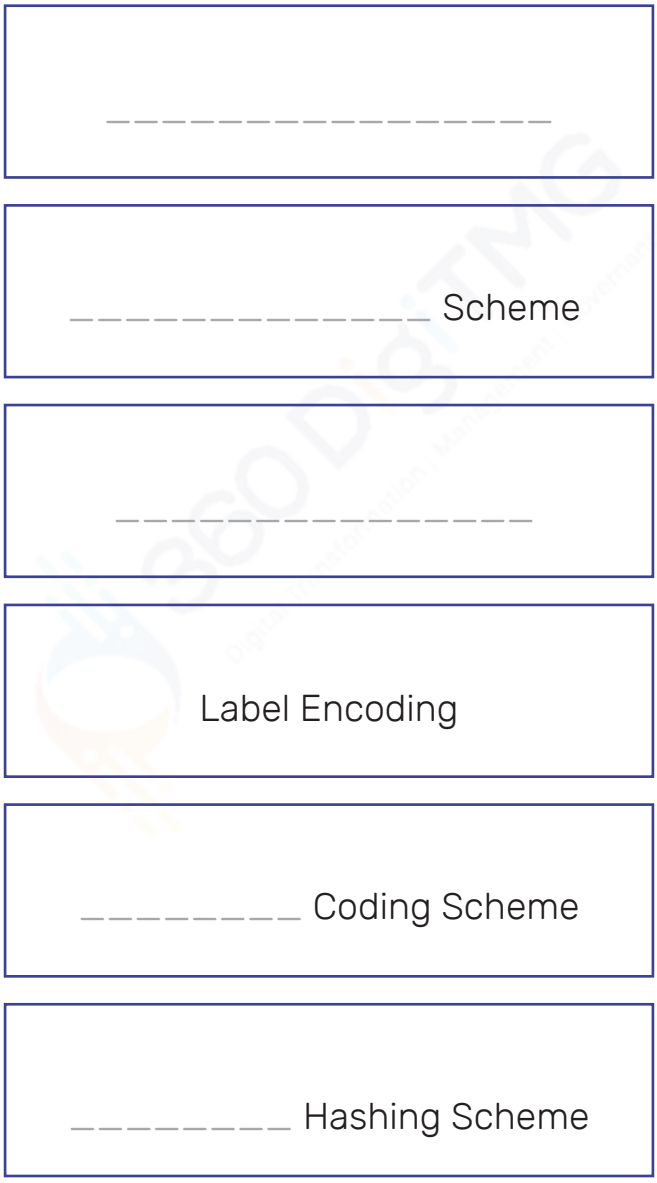
_____ is not influenced by outliers because it considers 'Median' & 'IQR'.

$$\frac{X - median(x)}{IQR(x)}$$

Dummy Variable

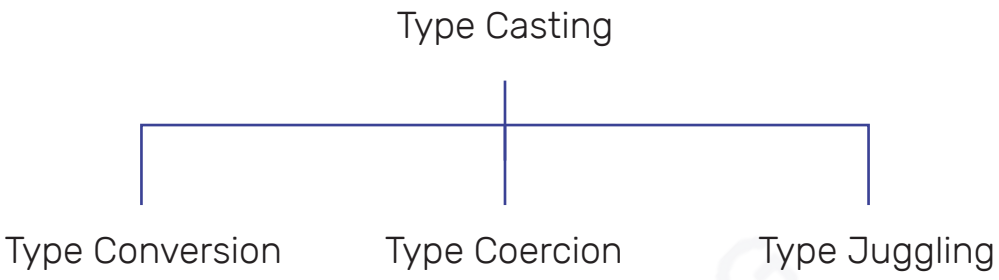
Dummy Variable Creation: Representing/Converting categorical data in numerical data

Techniques for Dummy Variable creation are:



Type Casting

Converting one type to another, for e.g. converting 'Character' type to 'Factor'; 'Integer' type to 'Float'.



Handling Duplicates

Ensures that we get a _____ of _____ from all the various locations.

E.g. A person opens a bank account but his transactions are recorded as John Travolta in a few, John in a few entries and Travolta in a few; however, all 3 are the name of the same person. So we merge all these names into one.

Name	Amount Spent
John Travolta	$ 1,000
Travolta	$ 800
John	$ 1,800

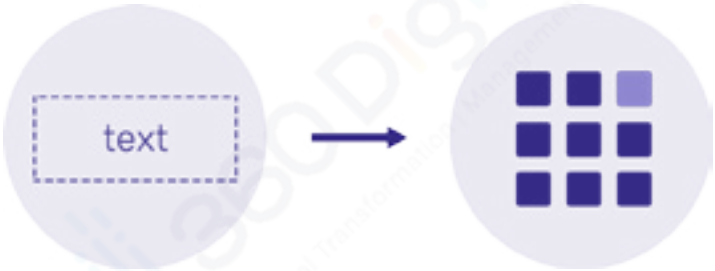
Name	Amount Spent
John Travolta	$ 3,600

Merged because all '3' entries belong to same customer.

String Manipulation

Working with textual data. Various ways of converting unstructured textual data into structured data are:

_____	Stemming	_____
Stopword Removal	_____	



Zero or Near Zero Variance

_____ & Near_____ feature:

Variables which are factors with a single level or majority of the levels are the same. E.g. All the zip code numbers are the same or Gender column has all entries listed as female.

We remove the variables from our analysis which have _____ or _____ variance in features.

CRISP - DM Exploratory Data Analysis (EDA)

Elements in EDA

Measure of Central Tendency is also called as "First Moment Business Decision".

_____ _____

MEDIAN

- Median is the middle value of the dataset
- Median of a dataset does not get influenced by _____

MODE

- Mode is the value, which repeats _____ times
- Mode is applied to _____ data
- If data has _____ mode it is called _____, if the data has _____ called _____ data and more than two modes called _____

Measure of _____, also called as Second Moment Business Decision

Variance

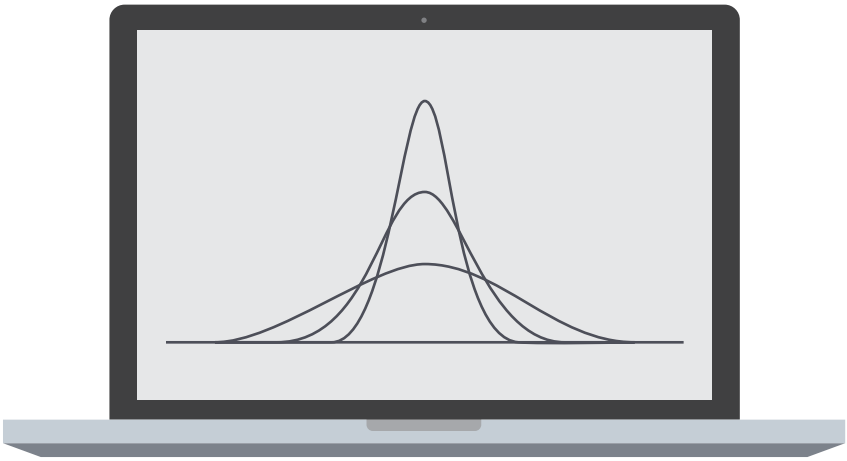
- How far away is each data point from mean/average?
- Units of measurement get squared

Standard Deviation

- Standard deviation is the square root of variance
- Get back the original units, which were squared during variance calculation

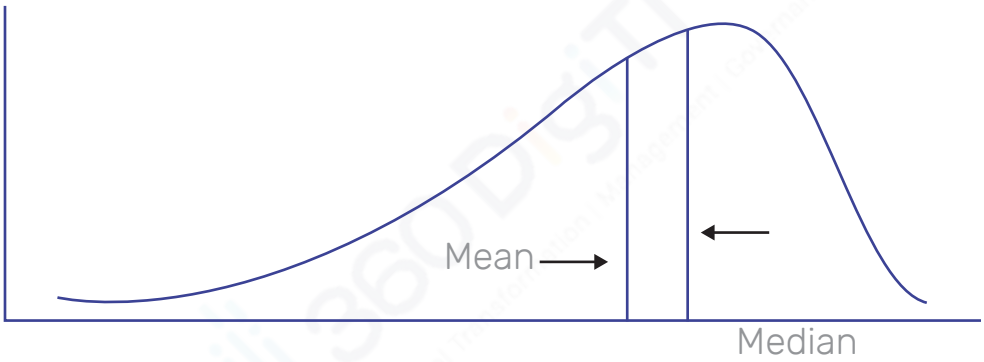
Range

- Represents the boundaries of the data spread
- Maximum - Minimum

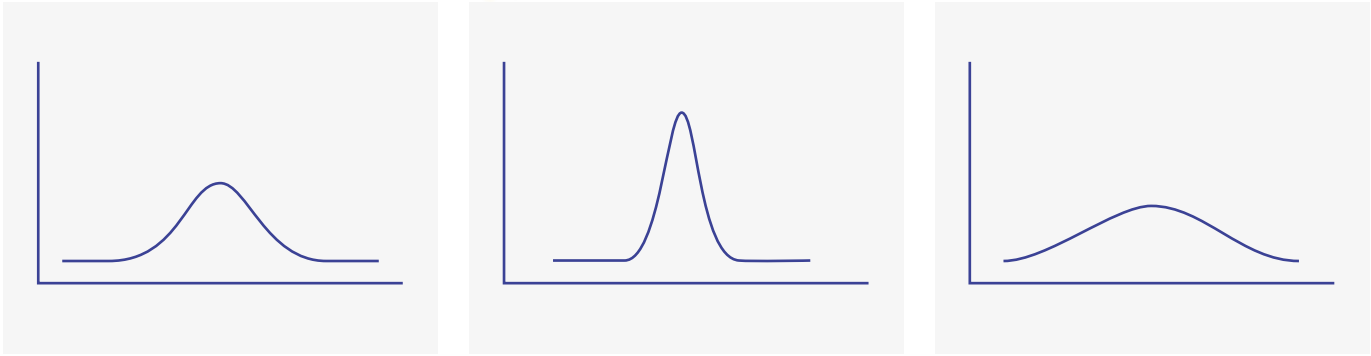


Measure of Skewness

- Data concentrated to the _____ is _____ Skewed also called _____Skewed
- Data concentrated to the _____ is Right Skewed also called _____ Skewed
- Presence of long tails helps in devising interesting business strategies



Measure of Kurtosis



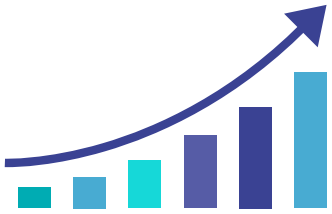
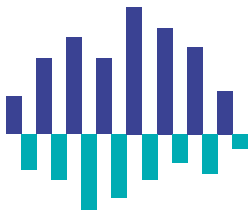
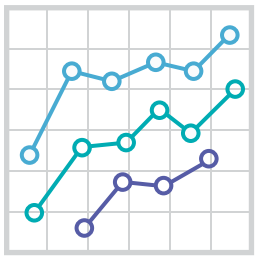
_____ Curve _____ Curve _____ Curve

Graphical Representations

Univariate analysis - Analysis of a single variable is called Univariate Analysis.

Graphs using which we can visualize single variables are:

1. Bar Plot
2. _____
3. _____
4. Strip Plot
5. _____
6. _____
7. Candle Plot
8. _____
9. Time Series Plots
10. _____
11. Density Plot
12. Boxplot or Box & Whisker Plot
13. _____ or _____

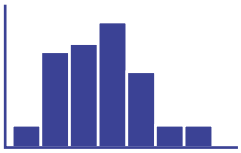
Graphical Representations

Majorly used plots for Univariate Analysis includes Histogram, Box Plot, and Q-Q Plot.

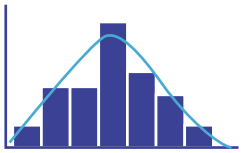
Histogram

Histogram is also called as Frequency Distribution Plot.

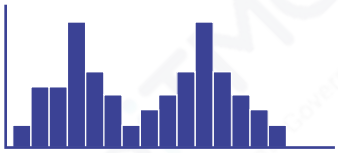
Primary Purpose: Histogram is used to identify the shape of the distribution.



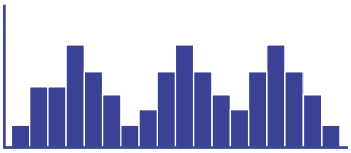
Summarises the data into discrete bins



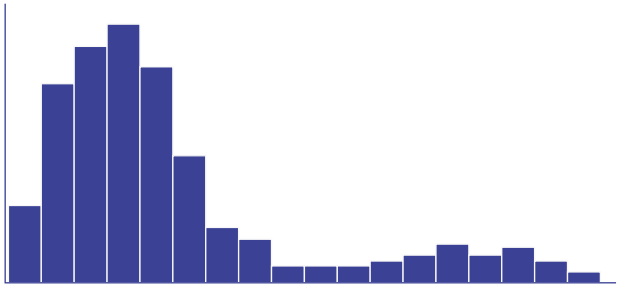
Used to identify the shape of the Distribution

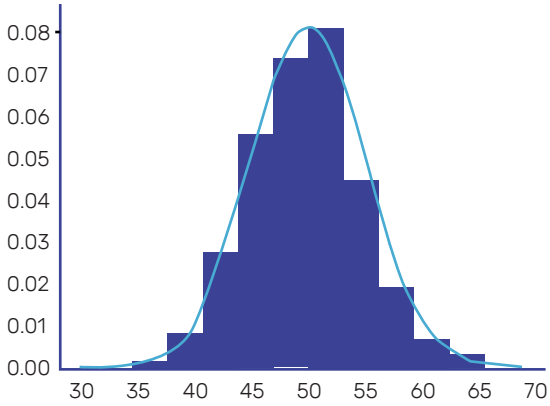


Identify if the data is unimodal, bimodal or multimodal



Secondary Purpose: Histogram is used to identify the presence of Outliers.

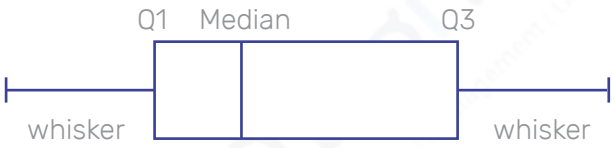




Is used to identify presence of Outliers

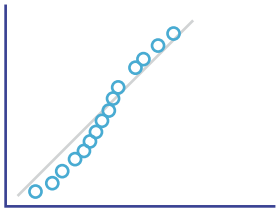
Box Plot is also called as Box and Whisker Plot

- Box Plot gives the 5 point summary, namely, Min, Max, Q1 / First Quartile, Q3 / Third Quartile, Median / Q2 / Second Quartile
- Middle 50% of data is located in the Inter Quartile Range (IQR) = Q3 - Q1
- Formula used to identify outliers is Q1 - 1.5 (IQR) on the lower side and Q3 + 1.5 (IQR) on the upper side
- Primary Purpose of Boxplot is to identify the existence of outliers
- Secondary Purpose of Boxplot is to identify the shape of distribution



Q-Q plot is also called Quantile Quantile Plot

- Q-Q plot is used to check whether the data are normally distributed or not. If data are non-normal then we resort to transformation techniques to make the data normal
- The line in the Q-Q plot connects from Q1 to Q3
- X-axis contains the standardized values of the random variable
- Y-axis contains the random values, which are not standardized
- If the data points fall along the line then data are considered to be Normally Distributed


Q-Q Plot

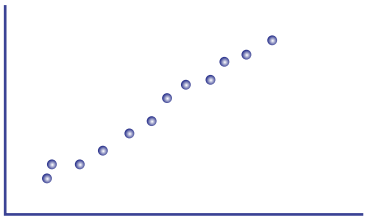
Bivariate Analysis

Bivariate analysis is analyzing two variables.

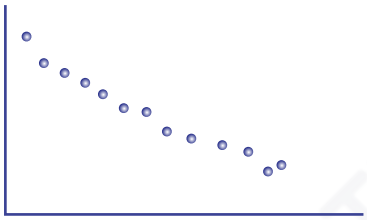
Scatter plot is used to check for the correlation between two variables.

The primary purpose of the Scatter Plot is to determine the following:

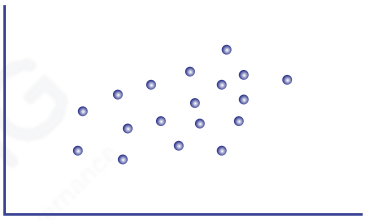
- Direction - Whether the direction is Positive or Negative or No Correlation



Positive Correlation

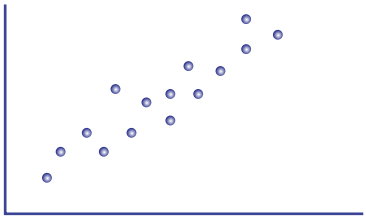


Negative Correlation

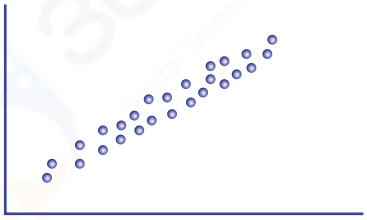


No Correlation

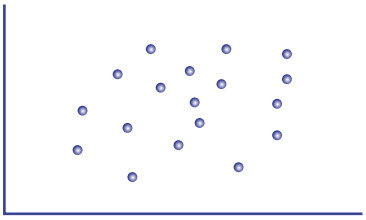
- Strength - Whether the strength is Strong or Moderate or Weak



Moderate Correlation

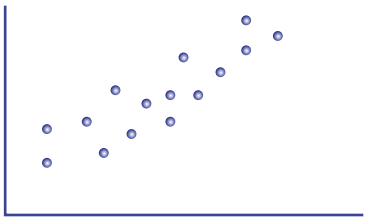


Strong Correlation

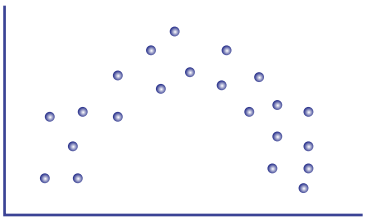


No Correlation

- Check whether the relationship is Linear or Nonlinear



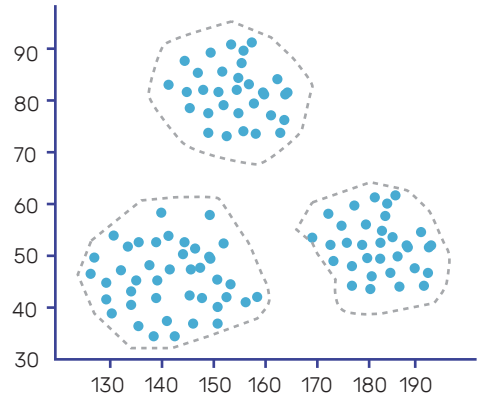
Linear



Nonlinear

The Secondary purpose of the Scatter Plot is to determine

_____.



- Determining strength using a scatter plot is subjective
- Objectively evaluate strength using Correlation Coefficient (r)
- Correlation coefficient value ranges from +1 to -1
- Covariance is also used to track the correlation between 2 variables
- However, Correlation Coefficient normalizes the data in correlation calculations whereas Covariance does not normalize the data in correlation calculation
- $|r| > 0.85$ implies that there is a strong correlation between the variables
- $|r| <= 0.4$ implies that there is a weak correlation
- $|r| > 0.4$ & $|r| <= 0.85$ implies that there is a moderate correlation

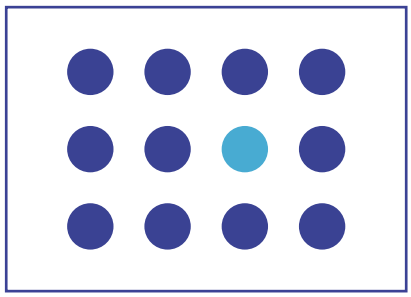
Multivariate Analysis

The two main plots to perform Multivariate analysis are:

- Pair Plot
- Interaction Plot

Data Quality Analysis

Focus of this step is to identify the potential errors, shortcomings, and issues with data.



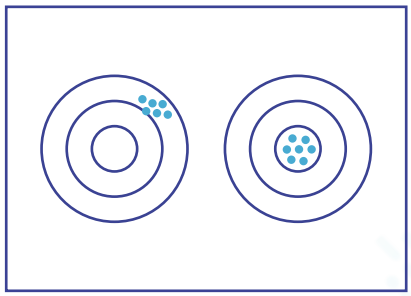
Identify _____

Name	Age
Steve	23
Jeff	37
Clara	28
Peter	41

Identify _____

Date	Date	Date
2001	Jan - 01	1-Jan-01
2001	Jan - 01	17-Jan-01
2002	Feb - 01	8-Feb-02
2003	Jun - 01	12-Jun-03

Identify different levels of granularity



Validation and Reliability

Sales	Region
19,345	North
23,424	West
24,164	East
19,453	South

_____ Data

Name	Age	Salary
Steve	$ 12,000	23
Jeff	$ 4,500	37
Clara	$ 5,200	28

Wrong metadata information

Wrong information due to data errors (manual / automated) - _____ _____

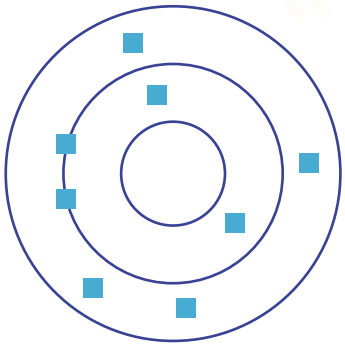
Four errors to be avoided during Data Collection

1. Random Errors - Measurement device (thermometer) faulty or Person measuring does mistakes. Leads to False Positives

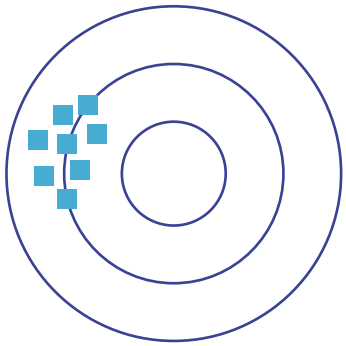
2. Systematic Errors - Social desirability bias of Trump on Twitter. Wearable devices data is of wealthy customers

3. Errors in choosing what to measure - Rather than choosing a person from a top university for a job, maybe we need to look at their social network which guided them through a series of events, which resulted in them joining the top school. High SAT score is not just based on high IQ, it depends on access to good tutors and purchasing good study material. Someone might like a subject and hence got a high GPA, but can we guarantee such a success in other fields

4. Errors of exclusion - Not capturing women data pertaining to cardiovascular diseases. An election in the US, not having data of colored women candidates. Chief Diversity Officer in big firms is a solution!



Random Errors



Systematic Errors

Data Integration

Data Integration is invoked when there are multiple datasets to be integrated or merged

Appending

Multiple datasets with the same attributes/columns.

Name	Age	Salary
Steve	23	$ 4,000
Raj	33	$ 6,500
Chen	41	$ 5,900

Name	Age	Salary
Wilma	37	$ 7,200
Audrey	51	$ 9,300

Name	Age	Salary
Steve	23	$ 4,000
Raj	33	$ 6,500
Chen	41	$ 5,900
Wilma	37	$ 7,200
Audrey	51	$ 9,300

Merging

Multiple datasets having different attributes using a common attribute.

Name	Age	Salary
Steve	23	$ 4,000
Raj	33	$ 6,500
Chen	41	$ 5,900

Name	Designation	Location
Wilma	Manager	Kuala Lumpur
Chen	V.P	NY City

Name	Age	Salary	Designation	Location
Steve	23	$ 4,000	Manager	Kuala Lumpur
Raj	33	$ 6,500	Nan	NaN
Chen	41	$ 5,900	V.P	NY City

Feature Engineering

Attribute Generation is also called Feature Extraction or Feature Engineering. Using the given variables, try to apply domain knowledge to come up with more meaningful derived variables.

Feature Extraction can be performed on:

- for Temporal Data

  - Date Based Features
  - Time-Based Features

- for Numeric Data
- for Categorical Data
- on Text Data
- on Image Data

Feature Extraction

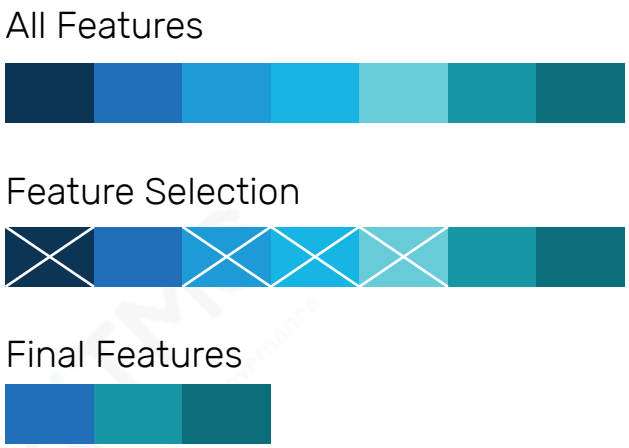
1. Deep Learning is performed using Automatic Extraction
2. Shallow Machine Learning is performed using Manual Extraction
3. Feature Extraction is used to get either Derived Features or Normalized Features

Feature Selection

Feature Selection or Attribute Selection is shortlisting a subset of features or attributes.

It is based on:

- Attribute importance
- Quality
- _____
- Assumptions
- Constraints



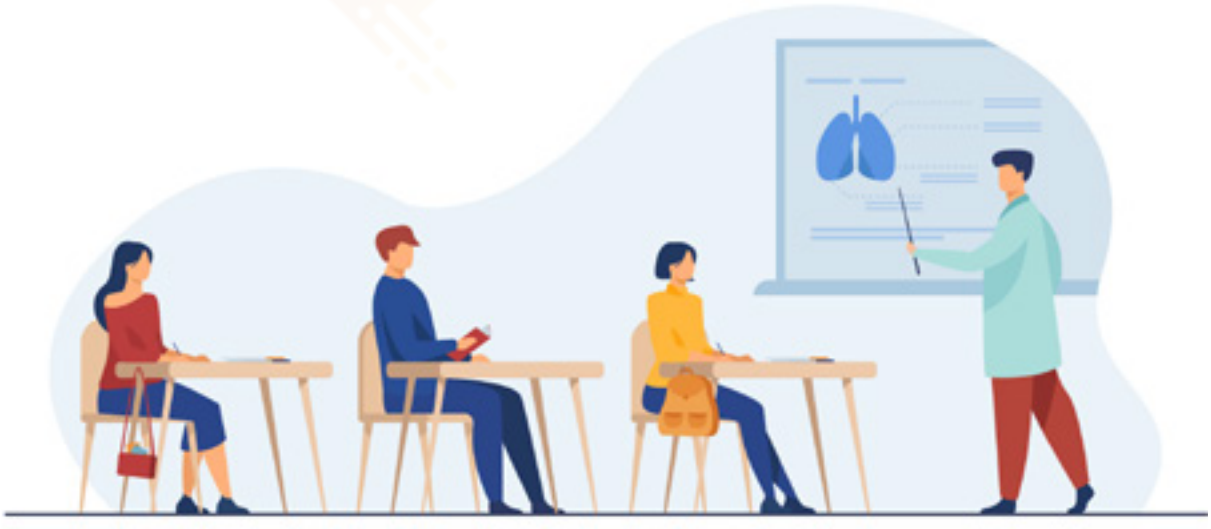
Feature Selection Techniques

- Filter Methods
- Wrapper Methods
- _____
- Threshold-Based Methods
- Statistical Methods
- Hypothesis Testing
- _____
- Model-Based Selection
- _____
- Variable Importance Plot
- Subset Selection Methods includes:
  - _____ (Lasso Regression, Ridge Regression)
  - Forward Stepwise Selection
  - Backward Stepwise Selection

Model Building using Data Mining

Supervised Learning

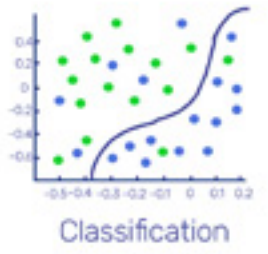
In the historical data if the _____ variable _____ is known, then we apply supervised learning tasks on the historical data. Supervised Learning is also called _____ or Machine Learning.

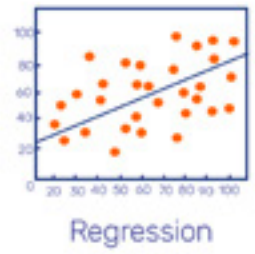


Supervised Learning has four broad problems to solve:

- Predict a _____ class: _____.
  - Example: Does the pathology image show signs of Benign or Malignant tumour
  - Is employee 'X' going to Attrite or Not Attrite
- Predict a _____ value: _____

  (also sometimes called as _____)
  - Example - What will be the stock value tomorrow?
  - How many Samsung mobile phones will we sell next month?
- Predict user _____ from a large pool of options: Recommendation
  - Example - Who will be the best match for getting married on a matrimonial website?
- Predict RELEVANCE of an entity to a "query": Retrieval
  - Example - Return to the most relevant website in Google search?


Classification


Regression


Recommendation


Retrieval

Data Mining-Unsupervised

What is Unsupervised learning?

Algorithms that draw conclusions on _____.

_____ data is data where output variable 'Y' is unknown.

Unsupervised Learning algorithms help in _____ analysis.

A few of the algorithms are:

Clustering	_____
Network Analysis	_____
_____	

Unsupervised Learing - Preliminaries

Distance Calculation:

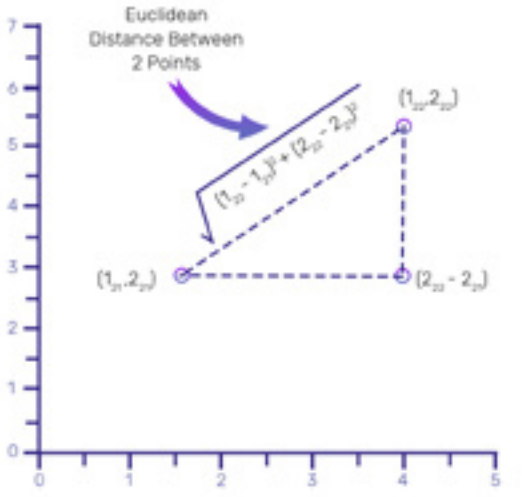
Distance is either calculated between:

Two _____	Between a _____ and a _____	Between two clusters

Distance Properties:

- Should be non-negative (distance > 0)
- Distance between a record to itself is equal to 0
- Satisfies _____ (Distance between records 'i' & 'j' is equal to the distance between records 'j' & 'i')

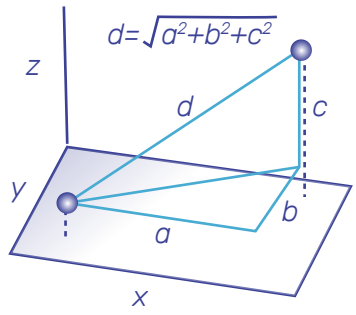
Standardize or _____ the variables before calculating the distance if the variables scale or are of different units.



Distance Calculations

Distance Metrics for Continuous Data

- _____ Distance which is calculated using Correlation Matrix
- _____ Distance, is also called as L1 norm
- Euclidean Distance, is also called as L2 norm



$$d = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2$$

Distance Metrics for Binary Categorical Data

- Binary Euclidean Distance
- Simple Matching Coefficient
- _____ Coefficient

Distance Metrics for Categorical Data (> 2 categories)

- Distance is 0, if both items have same category
- Distance is 1 otherwise

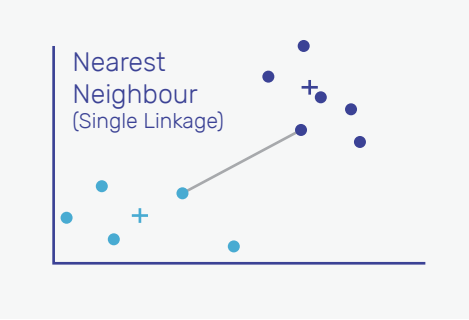
Distance Metrics when both Quantitative Data & Categorical Data exists in a dataset

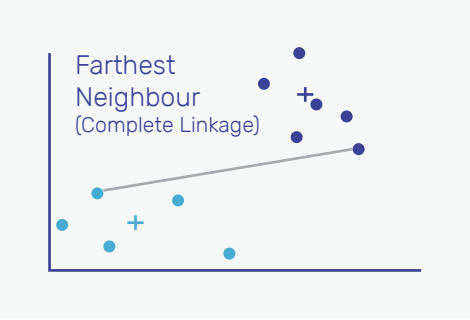
- _____ General Dissimilarity Coefficient

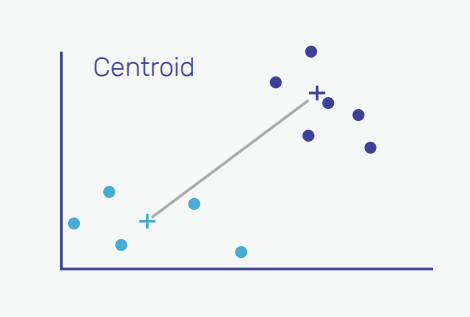
Linkages

Linkages - Distance between a record & a cluster, or between two clusters.

1. Single Linkage - This is the least distance between a record and a cluster, or between two clusters.
    - Single Linkage is also called as _____
    - Emphasis is on close records or regions and not on overall structure of data
    - Capable of clustering non-elliptical shaped regions
    - Gets influenced greatly by outliers or noisy data
2. Complete Linkage - This is the largest distance (diameter) between a record and a cluster, or between two clusters.
    - Complete Linkage is also called as _____
    - Complete Linkage is also sensitive to outliers







3. Average Linkage - This is the average of all distances between a record and a cluster, or between two clusters.
    - Average Linkage is also called Group Average
    - Very expensive because computation takes a lot of time
4. Centroid Linkage - This is the distance between the centroids (centers) of two clusters or between a record and centroid of a cluster.
    - Centroid Linkage is also called Centroid Similarity
5. _____ Criterion - It is the increase in the value of SSE criterion for clustering obtained by merging them into a single cluster.
    - This is also called Ward's Minimum Variance and it minimizes the total within cluster variance
6. G_____ A_____ A_____ C_____ (GAAC)
    - Two clusters are merged based on cardinality of the clusters and centroid of clusters
    - Cardinality is the number of elements in the cluster

Clustering / Segmentation

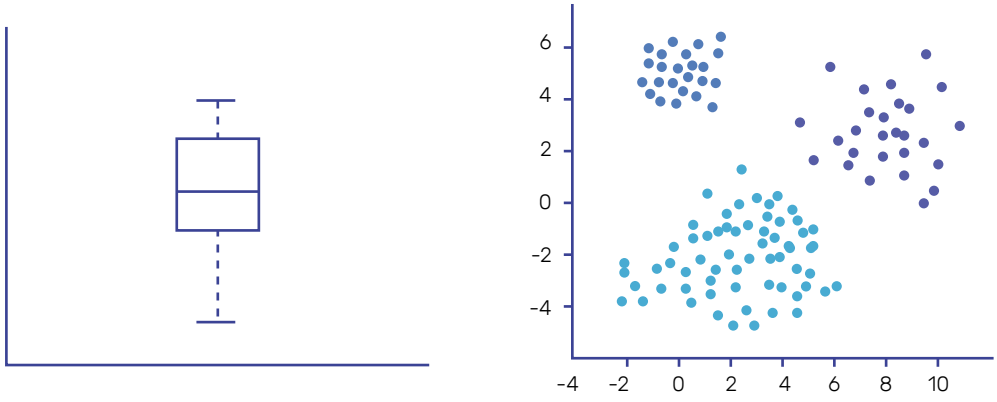
Clustering has two main criteria:

a. Similar records to be grouped together. High _____ similarity

b. Dissimilar records to be assigned to different groups.
   Less _____ similarity

- _____ groups will form _____ groups after clustering exercise
- Clustering is an _____ technique
- Separation of clusters can be of two types:
  _____ (one entry belongs to one cluster)
  vs _____ (one entry belongs to more than one cluster)

When we have a single variable then clustering can be performed by using a simple boxplot. When we have 2 variables then we can perform scatter diagrams.



Clustering / Segmentation

When we have more than 2 variables then there are a lot of other techniques such as:

Partitioning Based Methods:

- K-Means Clustering
- K-Means ++ Clustering
- _____ Clustering
- Genetic K-Means Clustering
- K-Medoids Clustering
- K-Medians Clustering
- K-Modes Clustering
- _____ Clustering
- _____ Clustering

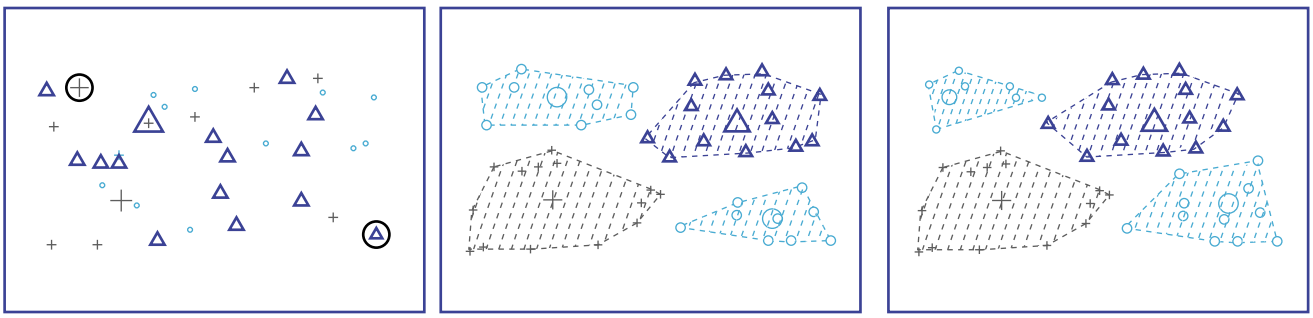
K-Means Clustering

K-Means clustering is called Non-Hierarchical Clustering.

We upfront decide the number of clusters using _____ or Elbow Curve.

Steps for K-Means

1. Decide the number of clusters 'K' based on the elbow curve of scree plot or based on the thumb rule _____. Alternatively, users may intuitively decide upon the number of clusters
2. Dataset is partitioned into K _____ with 'K' centers called centroids. These centroids are randomly chosen as part of _____
3. Each data point of the dataset, which is the closest to one of the centroids will form a cluster with that closest centroid
4. Centroids are again _____ with the data points assigned to each cluster
5. Steps 3 & 4 are repeated iteratively until no _____ of data points to other clusters is possible



Disadvantages of K-Means Clustering

Random initialization of centroids leads to clustering exercise terminating at a _____ (_____ because the objective is to get _____ within the sum of squares).

Solution:

Initialize the algorithm multiple times with different initial partitions.

No defined rule for selecting the K-value, while there are thumb rules, these are not foolproof.

Solution:

Run the algorithm with multiple 'K' values (range) and select the clusters with the least within the sum of squares and highest _____ within Sum of Squares'.

Extremely sensitive to the outliers or extreme values.

Solution:

K-medians, _____ are a few other variants which handle outliers very well.

K-Means clustering works for the data which is continuous in nature.

Solution:

Use _____ for categorical data.

Cannot discover clusters with non-convex shapes.

Solution:

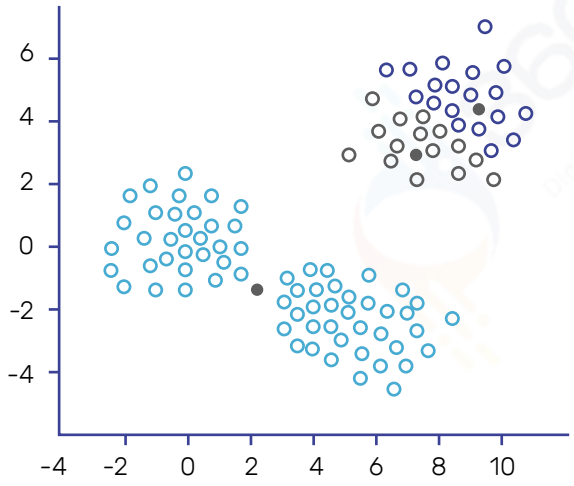
Use _____ clustering and _____ K-Means.

K-Means++ Clustering

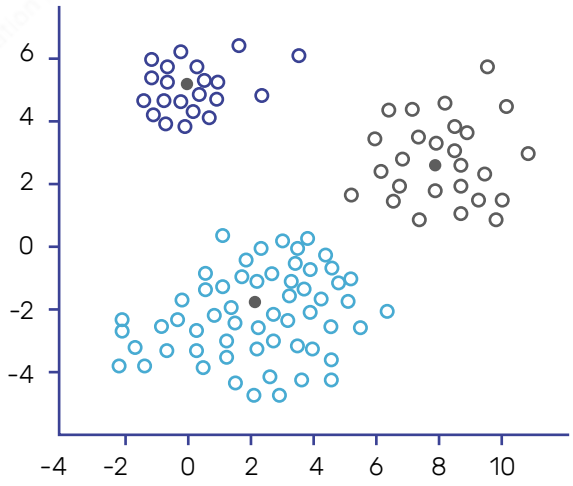
K-Means++ altogether, addresses the problem of different initializations leading to different clusters.

Steps:

1. Decide the number of clusters 'K'
2. First centroid is randomly selected
3. Second centroid is selected such that, it is at the _____
4. Step 2 depends on weighted _____ score criteria
5. This process continues until all 'K' centroids are obtained



K-Means Clustering



K-Means++ Clustering

K-Medians

K-Medians is very good at handling outliers.

L1 Norm is the distance measure used and is also called Manhattan Distance.

Steps are very similar to K-Means except that instead of calculating Mean we calculate Median.

1. K-Means cannot handle categorical data
2. Categorical data can be converted into one-hot encoding but will hamper the quality of the clusters, especially when the dimensions are large
3. K-Modes is the solution and uses modes instead of means and everything else is similar to K-Means
4. Distance is measured using _____
5. If the data has a mixture of categorical and numerical data then the _____ method can be used
6. K-Means can only handle linearly separable patterns and _____. Kernel K-Means clustering works well when the data is in non-convex format (non-linearly separable patterns)
7. _____functions are used to take data to high-dimensional space to make it linear and captures the patterns to form clusters

_____ Functions to be used are:

_____ Kernel	Gaussian Radial Basis Function (RBF) _____	Sigmoid _____

K-Medoids

K-Medoids address the problem of K-Means getting influenced by outliers.

Steps:

1. Choose 'K' data points randomly as medoids
2. Instead of taking the centriod of data points of a cluster, medoids are considered to be the center
3. Find out the distance from each and every data point to the medoid and add them to get a value. This value is called total cost
4. Select any other point randomly as a representative point (any point other than medoid points)
5. Find out the distance from each of the points to the new representative point and add them to get a value. This value is called the total cost of a new representative point
6. If the total cost of step 3 is greater than the total cost of step 5 then the representative point at step 4 will become a new medoid and the process continues
7. If the total cost of step 3 is less than the total cost of step 5 then the algorithm ends

Partitioning Around Medoids (PAM)

Partitioning Around Medoids (PAM) is a classic example of _____ Algorithm.

Steps:

1. Randomly points are chosen to be _____
2. Replace medoids will non-medoids, if the _____
   (Sum of Squared Errors - SSE) of the resulting cluster is improved (reduced)
3. Continue iteratively until the_____ criteria of step 2 is satisfied

PAM is well suited for small datasets but it fails for large datasets.

CLARA - Clustering Large Applications:

1. In the case of large datasets performing clustering by in-memory computation is not feasible. The sampling technique is used to avoid this problem
2. CLARA is a variant of PAM
3. However unlike PAM, the medoids of all the data points aren't calculated, but only for a small sample
4. The PAM algorithm is now applied to create optimal medoids for the sample
5. CLARA then performs the entire process for a specified no of points to reduce bias

CLARANS - Clustering Large Applications based on RANdomised Search:

1. The shortcoming of CLARA is that, it varies based on the sample size
2. CLARANS is akin to double randomization where the algorithm randomly selects the 'K'. And also randomly selects medoids and a non-medoid object (Similar to K-Medoids)
3. CLARANS repeats this randomised process a finite number of times to obtain optimal solution

Hierarchical Clustering

Hierarchical clustering is also called Agglomerative technique (bottom-up hierarchy of clusters) or Divisive technique (top-down hierarchy of clusters).

Agglomerative:

Start by considering each data point as a cluster and keep merging the records or clusters until we exhaust all records and reach a single big cluster.

Steps:

1. Start with 'n' number of clusters where 'n' is the number of data points
2. Merge two records, or a record and a cluster, or two clusters at each step based on the distance criteria and linkage functions

Divisive:

- Start by considering that all data points belong to one single cluster and keep splitting into two groups each time, until we reach a stage where each data point is a single cluster
- Divisive Clustering is more efficient than Agglomerative Clustering
- Split the clusters with the largest SSE value
- Splitting criterion can be Ward's criterion or Gini-index in case of categorical data
- Stopping criterion can be used to determine the termination criterion

Number of clusters are decided after running the algorithm and viewing the Dendrogram. Dendrogram is a set of data points, which appear like a tree of clusters with multi-level nested partitioning.

Disadvantages of Hierarchical Clustering

Work done previously cannot be un-done and cannot work well on _____ datasets.

Types of Hierarchical Clustering

1 BIRCH

B_____ I_____ R_____ and C_____ using H _____

2 CURE

C_____ U_____ RE_____

3 CHAMELEON

Hierarchical Clustering using Dynamic Modeling. This is a _____ approach used in clustering _____ structures

4 P_____ Hierarchical Clustering

5 G_____ Clustering Model

Density Based Clustering: DBSCAN

- Clustering based on a local cluster criterion
- Can discover clusters of random shapes and can handle outliers
- Density parameters should be provided for stopping condition

DBSCAN - D_____B_____ S_____ C_____ of A_____ with N_____

Works on the basis of two parameters:

_____

Maximum Radius of the neighbourhood

_____

Minimum number of points in the Eps-neighbourhood of a point

It works on the principle of

_____

OPTICS

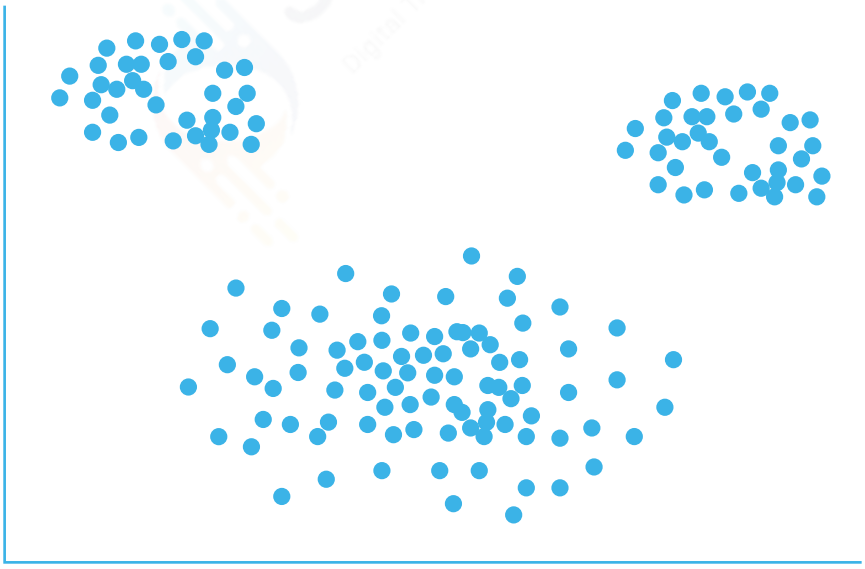
Ordering of Points to Identify Cluster Structure

Works on the principle of varying density of clusters

2 Aspects for Optics

Core Distance	Reachability Distance

"Plot the number of clusters for the image if it was subject to Optics clustering".



_____- Based Clustering Methods

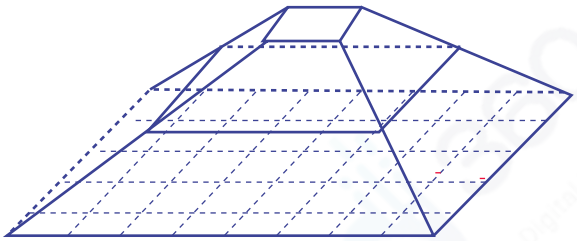
Partition the data space into finite number of cells to form a _____.

Find clusters from the cells in the _____ structure.

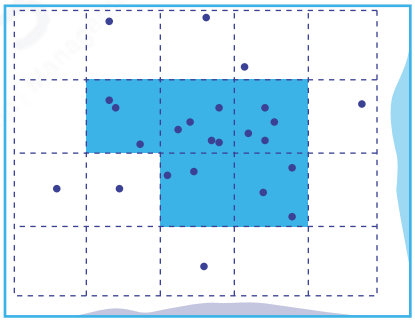
Challenges:

Difficult to handle irregular distribution in the data.

Suffers from the curse of dimensionality, i.e., difficult to cluster high-dimensional data.



STING



CLIQUE

Methods:

STING - ST_____ IN_____ G_____ approach

CLIQUE - CI_____ in QUE_____ - This is both density-based as well as grid-based subspace clustering algorithm.

Three broad categories of measurement in clustering

External	Internal	Relative

_____

Used to compare the clustering output against subject matter expertise (ground truth)

Four criteria for _____ Methods are:

1. Cluster homogeneity - More the purity, better is the cluster formation
2. Cluster completeness - Ground truth of objects and cluster assigned objects belong to same cluster
3. Rag bag better than alien - Assigning heterogeneous object is very different from the remaining points of a cluster to a cluster will be penalized more than assigning it into a rag bag/miscellaneous/other category
4. Small cluster preservation - Splitting a large cluster into smaller clusters is much better than splitting a small cluster into smaller clusters

Most Common _____ Measures

1. _____-based measures

- Purity
- Maximum Matching
- F-measure (Precision & Recall)

2. _____-based measures

- Entropy of Clustering
- Conditional Entropy
- Normalized Mutual Information (NMI)
- Entropy of Partitioning
- Mutual Information

3. Pairwise measures

- True Positive
- False Positive
- _____ Coefficient
- Fowlkes - Mallow Measure
- False Negative
- True Negative
- _____ Statistic

4. Correlation measures

- Discretized _____ Static
- Normalized Discretized _____ Static

Most Common External Measures

_____

Goodness of clustering and an example of same is _____ coefficient

Most common internal measures:

1. Beta-CV measure
2. _____ Cut
3. Modularity
4. Relative measure - _____ Coefficient

_____

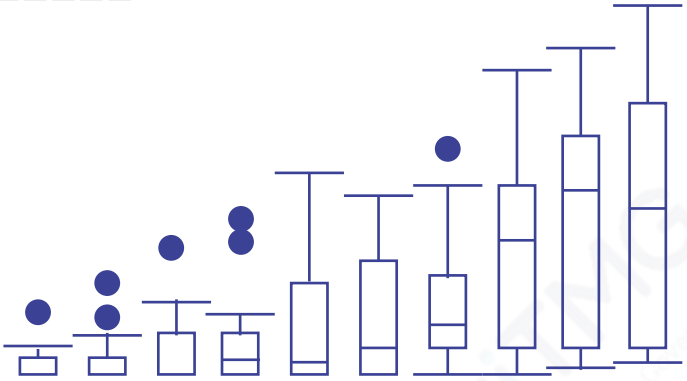
Compare the results of clustering obtained by different parameter settings of the same algorithm.

Clustering Assessment Methods

1. _____ Histogram

2. Distance Distribution

3. _____ Statistic



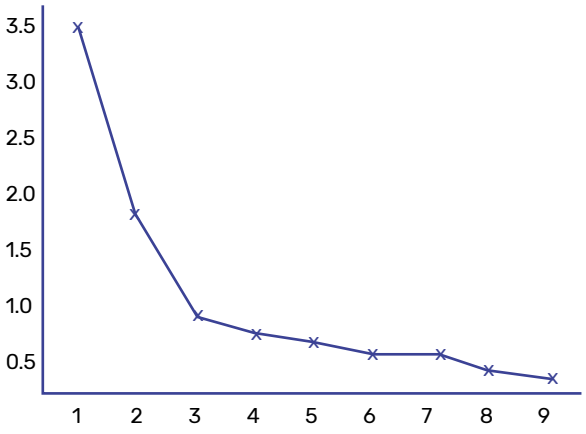
Finding K value in clustering

1. _____ Approach

2. Empirical Method = $\sqrt{\frac{n}{2}}$

3. Elbow Method

4. _____-Validation Method



Mathematical Foundations

Basic Matrix Operations:

1. Matrix _____
2. Matrix Multiplication
3. Matrix _____
4. _____ Matrix
5. _____ and Eigenvalues

Dimension Analysis and _____ _____(PCA, SVD, LDA)

Dimensions are also called as _____, Variables, _____.

Feature extraction of input variables from hundreds of variables is known as _____ Reduction.

Lesser dimensions means easy interpretability, quicker calculations, which also helps in reducing the _____ conditions and also avoiding _____.

Another benefit of dimensionality reduction is _____ the multivariate data on a _____.

Out of the many techniques available, in this book we will discuss the most popular methods:

- PCA - P_____ C_____ A_____
- SVD - S_____ V_____ D_____
- LDA - L_____ D_____ A_____
- _____ Analysis

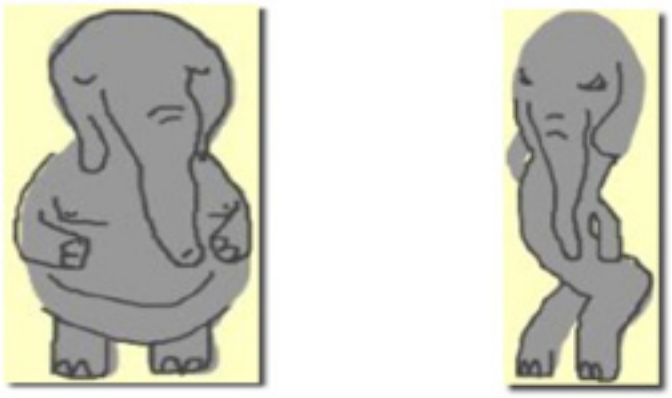
PCA - P_____ C_____ A_____

PCA is applied on Dense data (data, which does not have many zeros) which is quantitative in nature.

PCA is used to convert a large number of features into an equal number of features called P_____ C_____ (PCs).

These PCs capture 100% information, however, the initial set of PCs alone can capture maximum information.

PCA helps us reduce the size of the dataset significantly at the expenses of minimum information loss.



If the original dataset has features, which are all correlated then applying PCA does not help.

Each PC will capture information contained in all the variables of the original dataset.

Benefits of PCA

- Reduction of number of _____ & hence faster processing
- Identify the _____ between multiple columns at one go by interpreting the _____ of PCs
- Visualizing _____ data using a ___ visualization technique
- Inputs being _____ is called as collinearity and this is a problem, which is overcome by PCA because it makes the inputs _____
- Helps in identifying similar columns

The i^{th} principal component is a weighted average of original measurements / columns:

$$PC_i = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 .... + a_{in}x_n$$

Weights (a_{ij}) are chosen such that:

- PCs are ordered by their _____(PC1 > PC2 > PC3, and so on)
- Pairs of PCs have _____ = 0
- For each PC, sum of _____ = 1 (Unit Vector)

Data Normalization / Standardization should be performed before applying PCA.

SVD - S_____ V_____ D_____

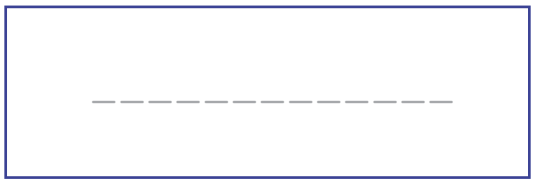
S_____ V_____ D_____ or SVD - is applied to reduce _____ data (data, which has a lot of entries as zeros).

SVD is applied on the images to reduce the size of images and helps immensely in image processing.

SVD is extensively used in _____

It is a _____ decomposition method, represented as:

- diagonal matrix values are known as the singular values of the original matrix X
- U matrix column values are called the _____ of X
- V matrix column values are called the _____ of X



LDA - L_____ D_____ A_____

L_____ D_____ A_____ (LDA) is used to solve dimensionality reduction for data with higher attributes.

Linear Discriminant Analysis is a supervised algorithm as it takes the class label into consideration.

LDA finds a centroid for each class datapoints.

LDA determines a new dimension based on centroids in a way to satisfy two criteria:

1. _____ the distance between the centroid of each class.
2. _____ the variation (which LDA calls scatter and is represented by s2), within each category.

Association Rules

Relationship Mining, _____ Analysis, or _____ Analysis - All mean the same thing, i.e., how are two entities related to each other, is there any dependency between them.

The objective of study of association is to find

Association rules are known as probabilistic '_____' statements. Generating the most ideal statements among all which show true dependencies is done using the following measures.

Support	Confidence	Lift

If part of the statement is called as _____.

Then part of the statement is called _____.

Association Rules

Support:

Percentage / Number of transactions in which IF/_____ & THEN / Consequent appear in the data

$$\text{Support} = \frac{\#\text{ transactions in which A \& C appear together}}{\_\_\_\_\_\ \#\text{ of transactions}}$$

Drawbacks of Support:

1. Generating all possible rules is exponential in the number of distinct items
2. It does not capture the true dependency - How good are these rules beyond the point that they have high support?

Association Rules

_____

Percentage of If/Antecedent transactions that also have the Then/Consequent item set.

P (Consequent | Antecedent) = P(C & A) / ____

$$\text{_____} = \frac{\text{_____}}{\text{\# transactions with A}}$$

Drawbacks of Confidence:

- Carries the same drawback as of Support
- It does not capture the true dependency - How good is the dependency between entities which have high Support?

Lift Ratio is a measure describing the ratio between dependency and independency between entities.

Formula: Confidence / _____

$$\text{Lift} = \frac{\text{Confidence}}{\_\_\_\_\_}$$

Note: _____ assumes independence between antecedent & consequent:

_____

$P(C|A) = P(C \& A) / P(A) = P(C) \times P(A) / P(A) = P(C)$

$$\_\_\_\_\_ = \frac{\text{\# transactions with consequent item sets}}{\text{\# transactions in database}}$$

Threshold - 1:

Lift > 1 indicates a rule that is useful in finding consequent item sets. The rule above is much better than selecting random transactions

Recommender Systems

Recommender Systems is also called _____

Data used for the analysis usually has 'Users' as rows and 'Items' will be columns. The entries within the dataset can be:

(from retail ecommerce context)

Whether a user has purchased or not	Whether user has _____ the product or not
How many products each user has purchased?	What is the rating provided by the user?

Sometimes the values, for example ratings columns, are divided by the _____. _____ refers to the number of customers who have purchased or rated the item. This process is called _____ the ratings.

Generally applied on e-commerce platforms. Customers purchasing patterns are analysed to design Personalized strategies to recommend items which have a high likely chance of getting purchased.

- What is the item most likely to be purchased?
- Can we identify and make suggestions/recommendations upfront?

These are broad two questions that have to be addressed. Recommendations in turn help to gain confidence in the user and make him loyal to the brand.

Types of Recommendation Strategies

1	_____ Recommender system
2	_____ Recommender system
3	Demographic based Recommender system
4	_____ based Recommender system
5	Knowledge based Recommender system
6	_____ Recommender system

Recommender System

Collaborative filtering is the most popular approach and it is based on similarity measures between users.

Similarity Measures:

_____ Based Similarity:

$$Cos(A,B) = A \cdot B / |A| * |B|$$

_____ Based Similarity:

$$Corr(AB) = Covariance(A,B) / Stdev(A) * Stdev(B)$$

Euclidean distance

_____ distance, etc.

What to Recommend:

List out and recommend the items that the person is MOST LIKELY to buy from the list of items that similar customers have already purchased.

Sorting the list of items can be based on:

- How many similar customers purchased it
- Rated by most
- Highest rated, etc.

Recommender System

Disadvantages:

_____ learning and _____ expensive

Compute is very expensive - n^2 similarities calculations

Options to reduce computational burden

1. Discard _____ buyers
2. Discard items that are very popular or very unpopular
3. _____ can reduce the number of rows
4. PCA (dimension reduction) can reduce the number of columns
5. _____ customers

Recommender System

Alternative Approaches

Strategic decision making in terms of

'Better accuracy and _____' vs 'Slightly lower accuracy and

_____recommendations'

Search-Based method is a recommendation based on previous purchases.

A variant of Search-Based method is called Item-to-Item Collaborative Filtering.

Rows will be Items and Columns will be Users.

Disadvantage is that most obvious items are always recommended.

Recommendations vs Association Rules

Association Rules	Recommendation Engine
_____, Common, Generic Strategy	Personalized strategy
_____ is important	_____ is unimportant
Useful for large physical stores	Useful for _____ recommendation

Recommender System

For New Users

- Recommend _____ popular items
- Recommend _____ popular items based on demography
- Recommend based on _____
- Make user login using social network, then look at the user's social media activity and recommend accordingly
- Show a few items and ask user to rate them so that based on the rating, one can be recommended

For New Items

- Recommend _____ to a few users
- Recommend to the tech-geeks (if it is a gadget)
- Identify the most _____ person in the social media graph data and recommend the new item to this influential person

Challenge with Rating Matrix-Based Recommendation:

Rating matrices are huge and sparse (too many empty cells)

_____ is used to handle the sparse rating matrix

Network Analysis

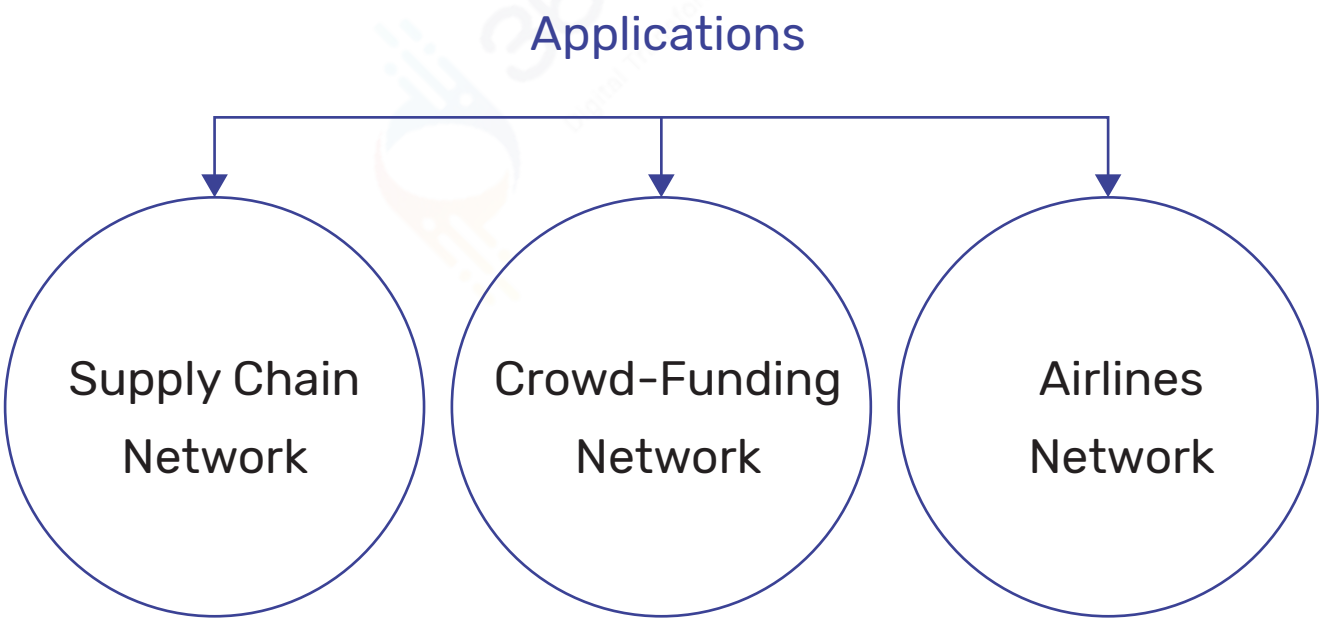
Network Data or _____ Data is a different type of data, which requires different types of analysis.

Key components of a _____ or Network are

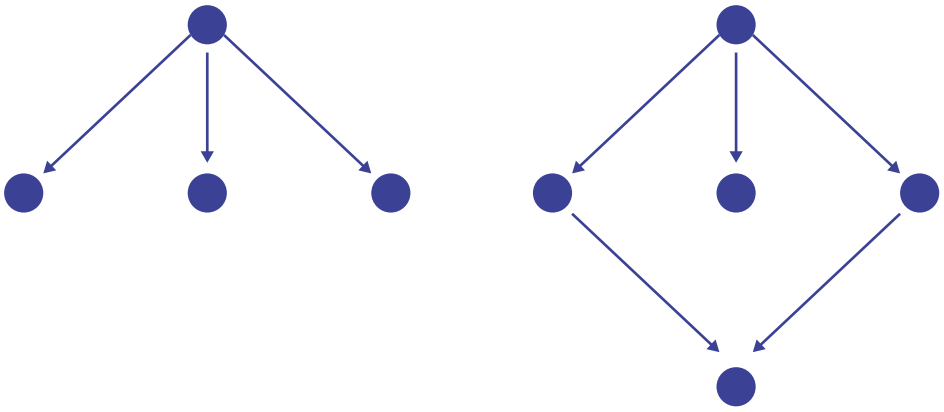
Vertices / _____ and _____ / Links

Network can be represented as Adjacency Matrix. Note: For an undirected graph, the adjacency matrix is symmetric in nature.

Links / _____ between _____ can be either bidirectional or _____



Network Analysis



Node Properties

_____ = Number of direct ties with other nodes

In-Degree = Number of Incoming connections

Out-Degree = Number of Outgoing connections

Degree centrality is a local measure and hence we should look at other measures.

_____ is how close the node is to other nodes in the network

_____

= 1/(sum of distances to all other nodes)

When comparison of two networks arise then normalized _____ should be considered

Normalized _____ = (Total number of nodes - 1) * Closeness

Network Analysis

_____ centrality can be measured for a node or an edge

_____ centrality is how often the node/edge lies on the shortest path between pairs

$$\text{\_\_\_\_\_ centrality} = \sum \frac{\text{\# of shortest paths between a pair it lies on}}{\text{\# of shortest path between a pair}}$$

When two networks are compared then we use normalized _____

$$\text{Normalized \_\_\_\_\_} = \frac{\text{\_\_\_\_\_}}{\text{\# of all possible pairs except the focal node}}$$

Network Analysis

_____ centrality measures who are you connected to and not just how many you are connected to

- Nodes which are connected to high scoring nodes contribute more to the score of that nodes which are connected to low scoring nodes
- _____ is calculated from _____ of adjacency matrix

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} a_{11}\, a_{12} \ldots a_{1n} \\ . \\ . \\ . \\ a_{n1}\, a_{n2} \ldots a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix}$$

$$\lambda x = AX$$

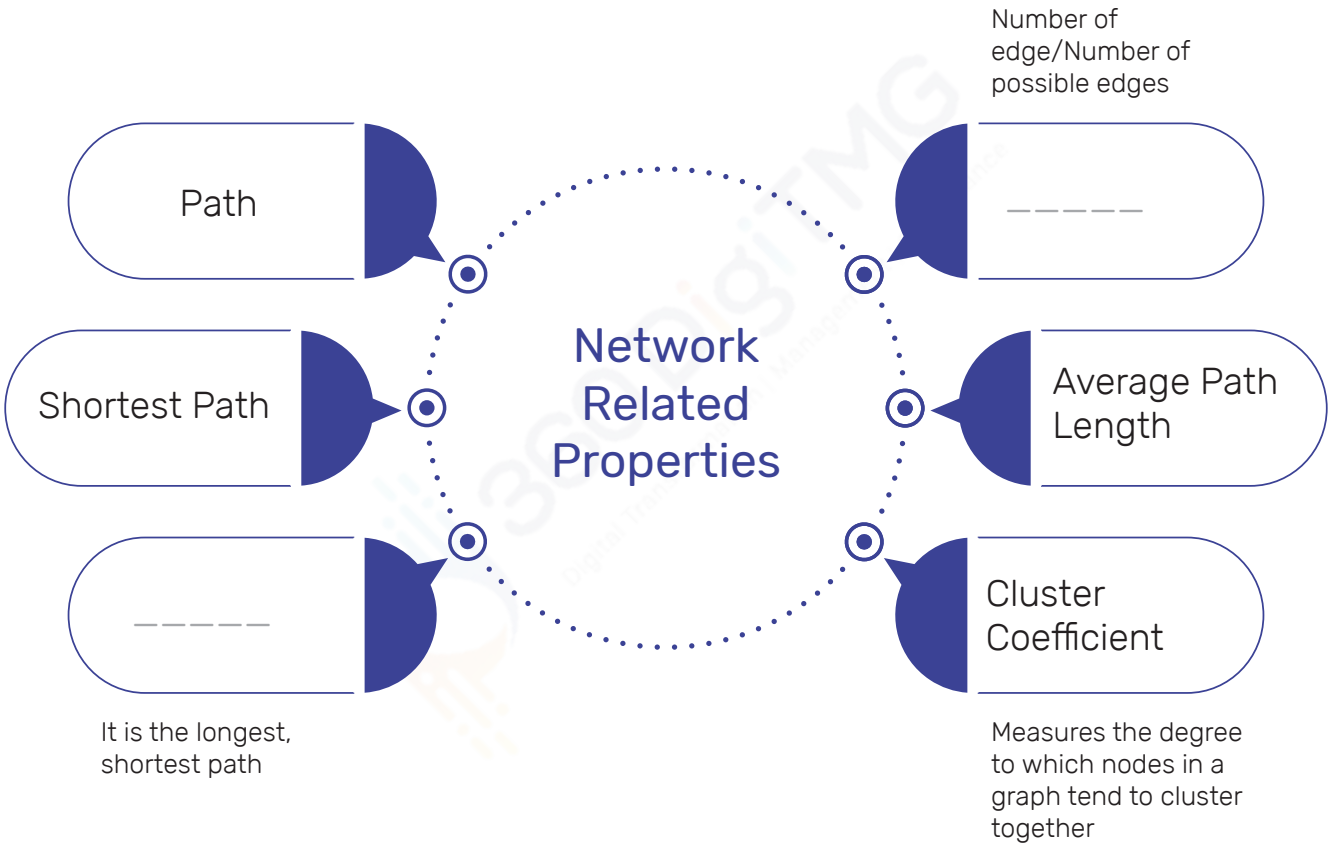
X corresponding to the highest Eigenvalue λ is the vector that consists of the _____ centralities of the nodes.

_____ Centrality is a measure on how likely is a person, who receives the information, going to diffuse the information further.

Network Analysis

Edge / Link Properties

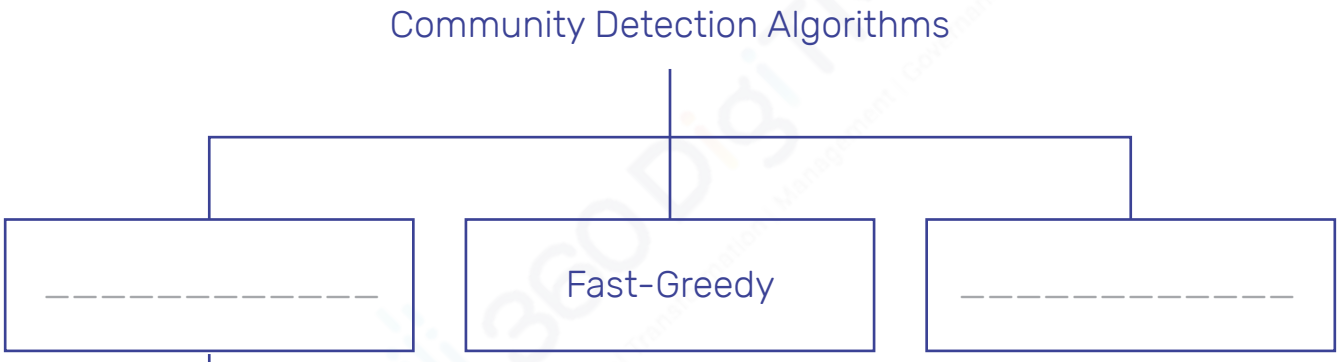
Edge or Link properties are defined based on the domain knowledge and there is no defined rule in defining the same.



Network Analysis

$$\text{Cluster Coefficient of a node} = \frac{\text{\# of links that exist among its neighbors}}{\text{\# of links that could have existed among its neighbors}}$$

Cluster coefficient of a network is average cluster coefficient of nodes in the network.

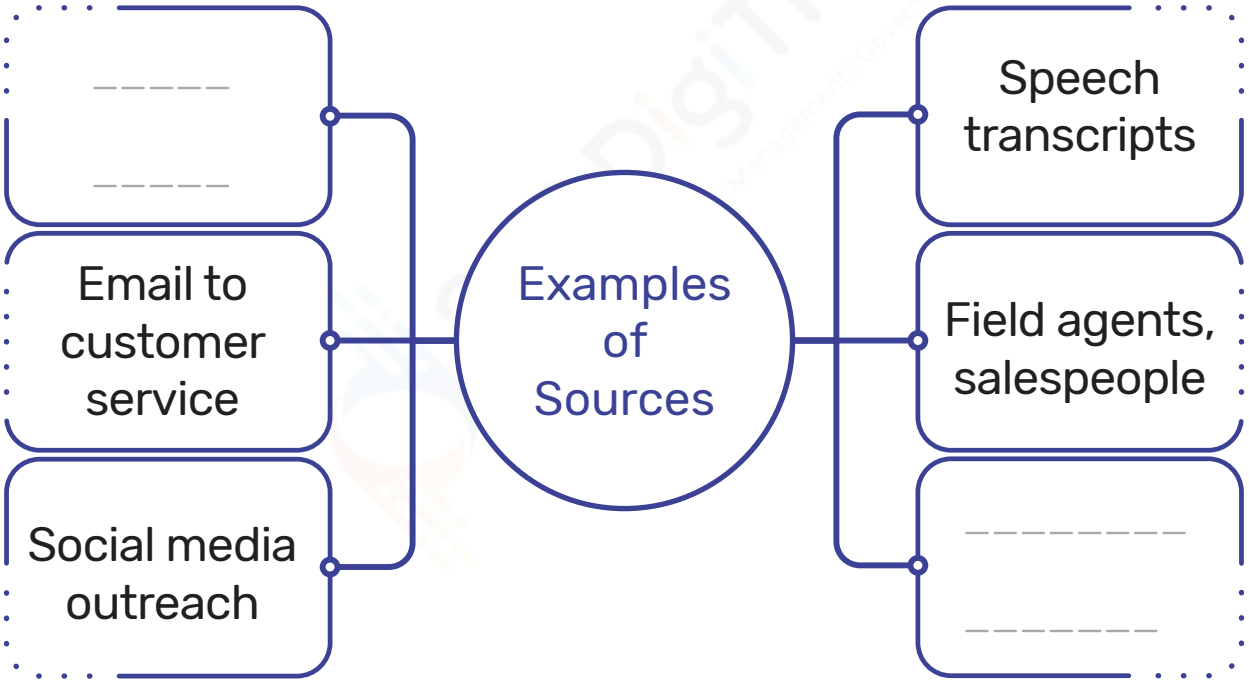


- Also called as Iterative network divisive algorithm
- Steps
  - Calculate edge betweenness of each edge
  - Remove the edge with the highest betweenness
  - Repeat the above mentioned two steps until no edge is remained
- Starts with the _____ and goes on until all nodes are isolated communities

Text Mining

Analyzing _____ Text data by generating _____ data in key-value pair form. Deriving insights from the extracted keywords by arranging the extracted keywords in a plain space with font sizes varying based on their frequency is called _____.

Collect the text data / Extract data from sources.



Text Mining

Pre-process the data

- Typos
- Case - uppercase / lowercase / proper case
- Punctuations & special symbols ('%', '!', '&', etc.)
- Filler words, connectors, pronouns ('all', 'for', 'of', 'my', 'to', etc.)
- Numbers
- Extra White spaces
- Custom words
- Stemming
- Lemmatization
- Tokenization - Tokenization refers to the process of splitting a sentence into its constituent words

Document Term Matrix / Term Document Matrix

Documents arranged in rows and Terms arranged in columns is called as DTM and transpose of DTM is TDM.

Word Cloud

_____ - words present in positive dictionary.

_____- words present in negative dictionary.

_____ - two words repeated together - gives better context of the content.

Natural Language Processing (NLP)

Text Analytics is the method of extracting meaningful insights and answering questions from text data.

N_____ L_____ U_____ (NLU)

A process by which an inanimate object (not alive - machines, systems, robots) with computing power is able to comprehend spoken language.

Example: Humans talk to robot

N_____ L_____ G_____ (NLG)

A process by which an inanimate object (not alive - machines, systems, robots) with computing power is able to manifest its thoughts in a language that humans are able to understand.

Example: Robot responds to human queries

POS tags - Parts of Speech Tagging – Process of tagging words within sentences into their respective PoS and then labelling them.

N_____ E_____ R_____

_____ are usually not present in the dictionaries so we need to treat them separately. People, place, organizations, quantities, percentages, etc.

Topic Modeling Algorithms

LSA/LSI (Latent Semantic Analysis/Latent Semantic)

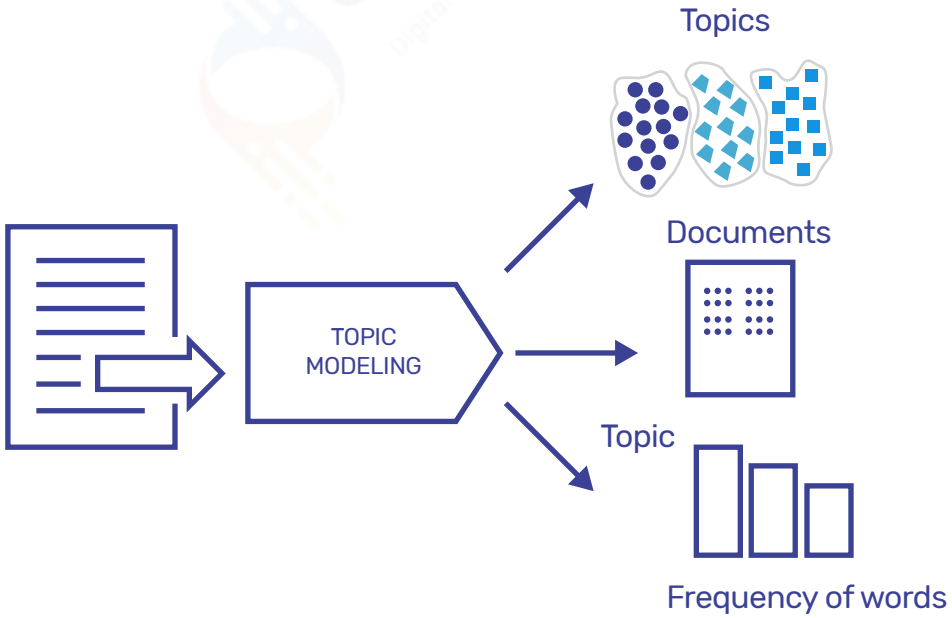
Reducing dimension for classification. LSA assumes that the words will occur in similar pieces of text if they have similar meaning.

LDA (_____)

A topic modelling method that generates topics based on words/expression frequency from documents.

Text Summarization:

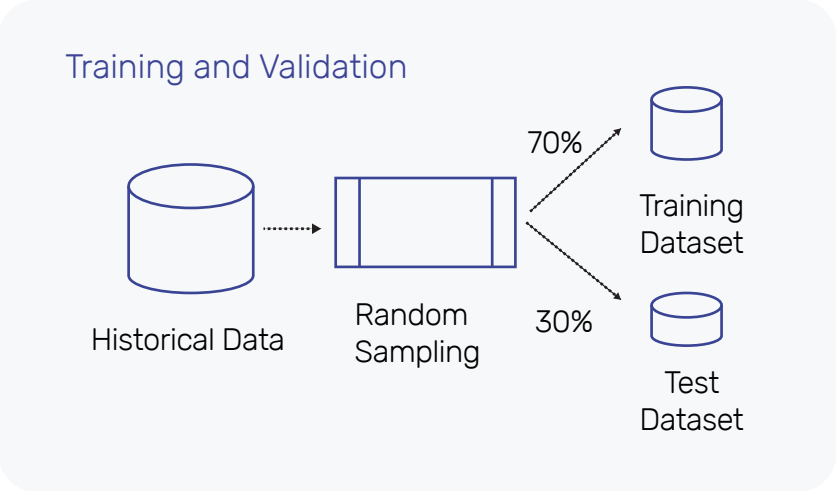
Process of producing concise version of text by retaining all the important information.



Machine Learning Primer

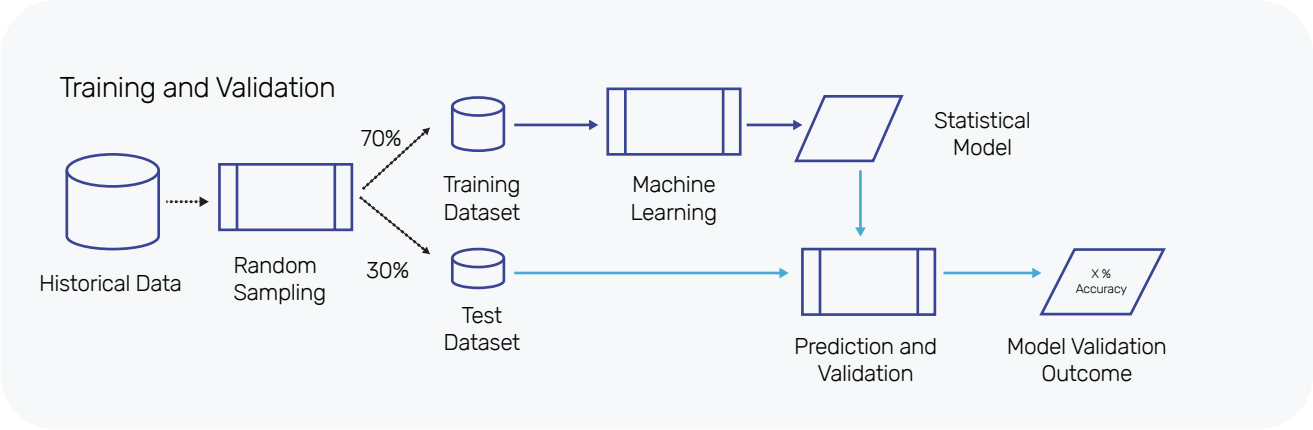
Steps based on Training & Testing datasets

1. Get the _____ / _____ data needed for analysis which is the output of data cleansing
2. Split the data into training data & testing data



   a. Split the data based on random sampling if the data is balanced
   b. Split the data based on other sampling techniques if the data is imbalanced
      (Refer to Step 2 of CRISP-DM to know about imbalance dataset sampling techniques)
   c. _____

Machine Learning Primer



3. _____

4. Test the model on testing data to get the predicted values

5. Compare the _____ and _____ values of testing data to calculate error or accuracy. (Model evaluation techniques are discussed in subsequent sections). This will give us Testing Error or Testing Accuracy

6. Also test the built model on training data

7. Compare the training data predicted values and training data actual values to calculate the error or accuracy. This will give us Training Error or Training Accuracy

Machine Learning Primer

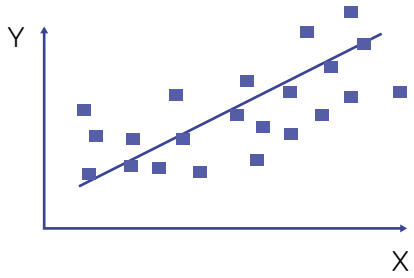
8. Training Error and Testing Error

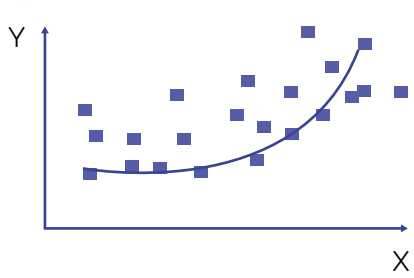
   a. If training error and testing error are small and close to each other then the model is considered to be RIGHT FIT (how low the error values should be is a subjective evaluation. E.g., In healthcare even 1% error might be considered high, whereas in a garment manufacturing process even 8% error might be considered low)

   b. If training error is low and testing error is high then the model is considered to be _____ . _____ is also called _____

   c. If training error is high then testing error also will be high. This scenario is called _____ or _____

   d. If training error is high and testing error is low then something is seriously wrong with the data or model you built. Redo the entire project





_____ _____ _____

Machine Learning Primer

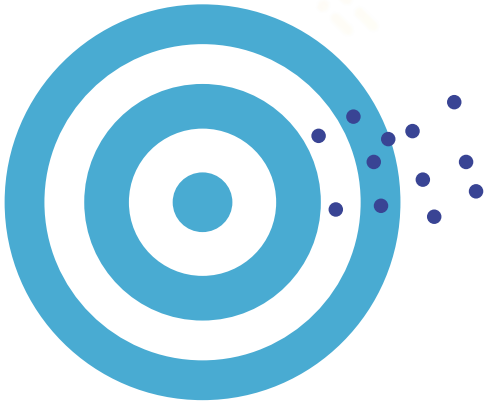
9. _____ is a common problem and also challenging to solve. Different Machine Learning algorithms have different regularization techniques (also called as generalization techniques) to handle

   _____

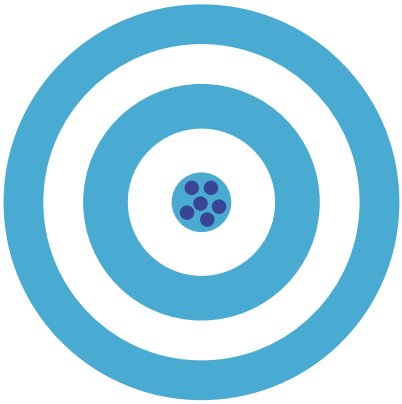
10. _____ problems can be solved easily by increasing the number of datapoints (observations) and/or features (columns). Also proper feature engineering and transformation will address this issue

1. High bias, Low variance

2. Low bias, High variance



3. High bias, High variance

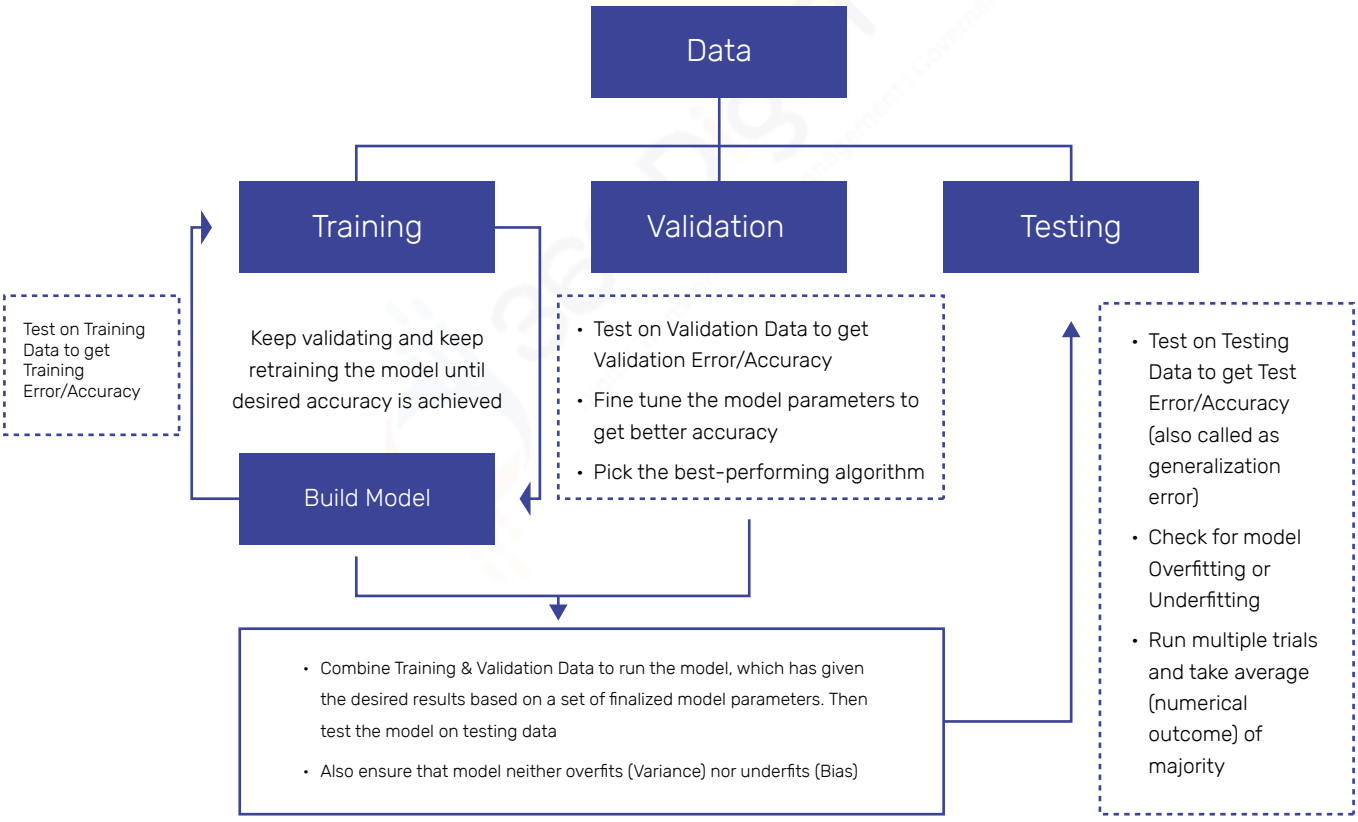


4. Low bias, Low variance

Machine Learning Primer

The challenge of Training & Testing dataset split, which leads to information leak is countered with new school of thought with an idea to split the data into:

- Training Data
- _____ (Development Data)
- _____



Model Evaluation Techniques

If the 'Y' output variable is _____ then we can use the following list of error functions to evaluate the model.

Error = Predicted Value - Actual Value (Actual Value is also called as _____ Value)

- _____(ME)

$$ME = \frac{1}{T}\sum_{t=1}^{n} e_t$$

- _____ (MAE)
  or_____ (MAD)

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

- Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

- Mean Percentage Error (MPE)

$$MPE = \frac{1}{n}\sum_{t=1}^{n} \frac{e_t}{Y_t}$$

- _____
  (MAPE)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{e_t}{Y_t}\right|$$

- Mean Absolute Scaled Error (MASE)

$$MASE = \frac{MAE}{MAE_{\text{in-sample, naive}}}$$

$MAE_{\text{in-sample, naive}}$ is the mean absolute error produced by a naive forecast

Actual Data	Prediction Model 1	Error from Model 1	Prediction Model 2	Error from Model 2
100	101	1	110	10
200	199	-1	190	-10
300	301	1	310	10
400	399	-1	390	-10

- Correlation Coefficient

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

Model Evaluation Techniques

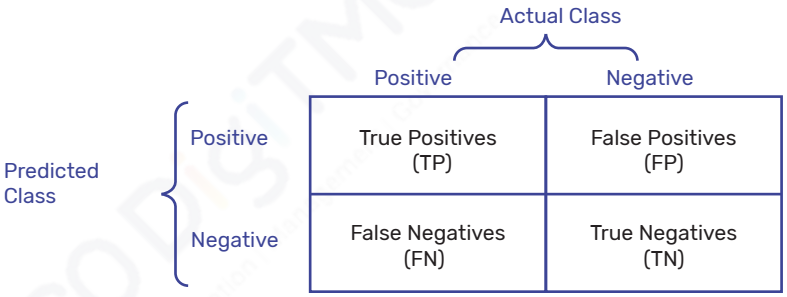
If the 'Y' is Discrete variable (Classification Models) then we can use the following list:

Confusion Matrix:

Can be applied for both _____ classification as well as _____ classification models.

Confusion matrix is used to compare predicted values and actual values.

Binary Classification Confusion Matrix:



True Positive (TP)

- Patient with disease is told that he/she has disease

True Negative (TN)

- Patient with no disease is told that he/she has no disease

False Positive (FP)

- Patient with no disease is told that he/she has disease

False Negative (FN)

- Patient with disease is told that he/she has no disease

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TNP}$$

Error = 1- Accuracy

Accuracy should be greater than % of majority class

Model Evaluation Techniques

Possible Outcomes

Decision Alternatives	Not Your wife's Birthday	Your wife's Birthday
Did not buy Flowers (No Action)	 Status Quo	 Wife Angry
Bought Flowers (Action)	 Wife Suspicious Money Wasted	 Domestic Bliss

Model Evaluation Techniques

- _____ = TP/(TP+FP) = TP/Predicted Positive = Prob. of correctly identifying a random patient with disease as having disease

  _____ is also called as _____ (PPV)

$$Precision = \frac{TP}{TP + FP}$$ (Designers in the formula hide the name precision)

- _____ (Recall or _____ or _____ Rate) = TP/(TP+FN) = TP/Actual Positive = Proportion of people with disease who are correctly identified as having disease

$$Recall = \frac{TP}{TP + FN}$$

- _____ (True negative rate) = TN/(TN+FP) = Proportion of people with no disease being characterized as not having disease
- _____ (Alpha or type I error) = 1 - Specificity
- FN rate (Beta or type II error) = 1 - Sensitivity
- _____ = 2 * ((Precision * Recall) / (Precision + Recall)); F1: 1 to 0 & defines a measure that balances precision & recall

$$F1\ Score = \frac{2\ X\ Precision\ X\ Recall}{Precision + Recall}$$

F1 score is the harmonic mean of precision and recall.

Closer the 'F1' value to 1, better the accuracy.

Confusion Matrix

A Confusion Matrix is also called a Cross Table or _____. Here is an example of a multi-class classification problem.

Activity recognition from video

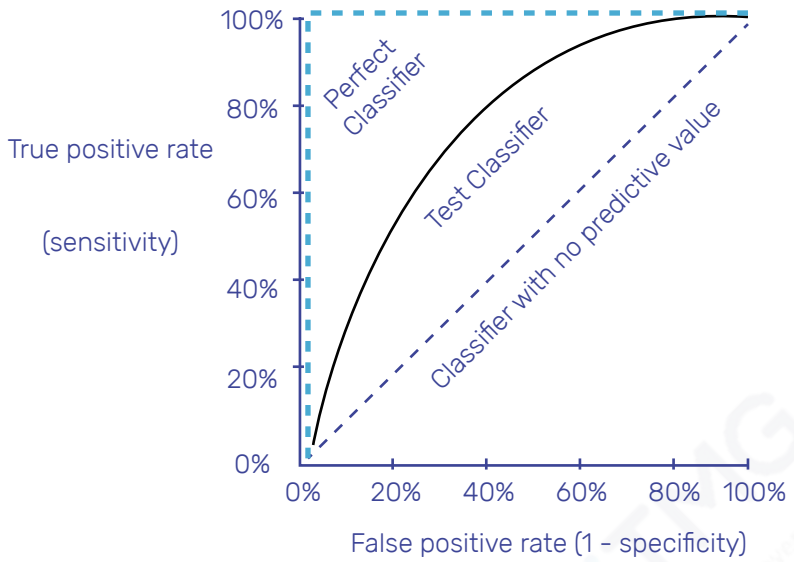
Actual Class

Bend	100	0	0	0	0	0
Jack	0	100	0	0	0	0
Jump	0	0	89	0	0	11
Run	0	0	0	67	0	33
Skip	0	0	0	0	100	0
Walk	0	0	11	33	0	100
	Bend	Jack	Jump	Run	Skip	Walk

Predicted Class

The values along the diagonal are right predictions and the values off the diagonal are incorrect predictions.

ROC Curve



R_____ O_____ C_____ Curve was used right from World War II to distinguish between true signals and false alarms.
The ROC curve has the 'True Positive Rate (TPR)' on the Y-axis and 'False Positive Rate (FPR)' on the X-axis.

ROC curve is used to visually depict accuracy.

ROC curve is also used to find the _____ value
(Example: Risk Neutral: should probability be > 0.5 as cut-off value to categorize a customer under 'will default' category; Risk Taking: should the probability be > 0.8 cut-off to categorize a customer under 'will default' category; or Risk Averse: should the probability be > 0.3 cut-off to categorize a customer under 'will default' category)

ROC Curve

Numerically if one must evaluate the accuracy then AUC (Area Under the Curve) can be calculated.

	Disease Present	Disease Absent
Test Positive	True Positives	False Positives
Test Negative	False Negatives	True Negatives

0.9 - 1.0 = A (outstanding)

0.8 - 0.9 = B (excellent/good)

0.7 - 0.8 = C (acceptable/fair)

0.6 - 0.7 = D (poor)

0.5 - 0.6 = F (no discrimination)

K-Nearest Neighbour

KNN also known as:

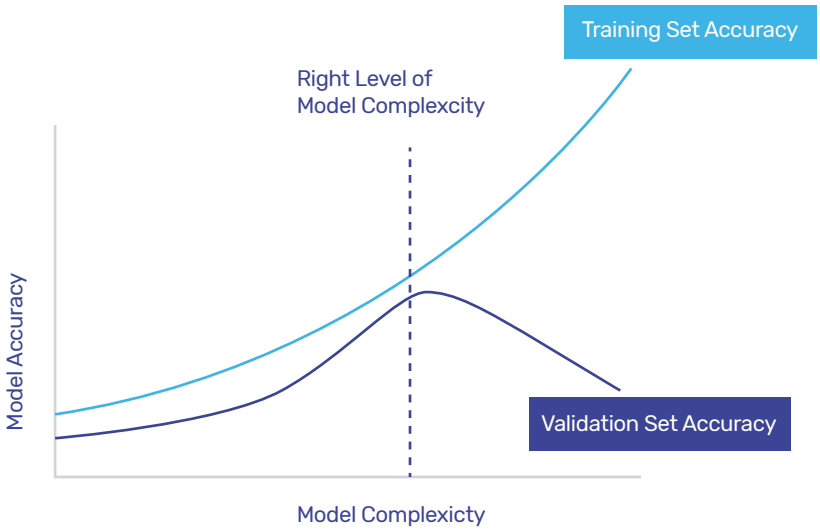
- On-demand or Lazy Learning
- _____-Based Reasoning
- _____-Based Reasoning
- Instance-Based Learning
- Rote Learning
- _____ Reasoning

KNN works for both the scenarios : Y is _____ as well as _____.

KNN is based on calculating distance among the various points. Distance can be any of the distance measures such as Euclidean distance discussed in previous sections.

KNN also has an improved version where _____ _____ are assigned to the neighbors based on their distance from the query point.

In case of continuous output, the final prediction will be the _____ of all output values and in case of categorical output, the final prediction will be the _____ of all the output values.



Choosing 'K' value is critical because it is used to solve the problem of bias-variance tradeoff.

- Low 'K' value is _____
- High 'K' value might introduce data points from _____

Pros (Advantages) and Cons (Disadvantages)

Strengths	Weakness
Does not depend on the underlying data distribution	There is no model produced and hence no interesting relationship among output and inputs is learnt
Testing process will be very fast	Memory requirement is large because distance calculations is saved in memory
	Testing process is slower in comparison to other models
	Categorical Inputs require additional processing
	Suffers from Curse of dimensionality

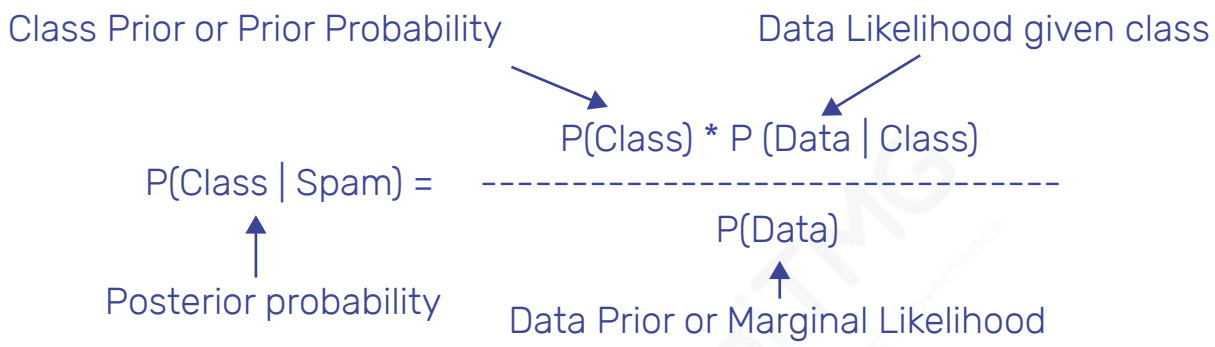
Naive Bayes Algorithm

Naive Bayes is a machine learning algorithm based on the principle of probability.

The relationship between _____ events is described using Bayes Theorem.

Probability of event A given that event B has occurred is called as

_____Probability.



Y (Whether the email is spam or not)	X (Whether the email contains the word lottery or not)
Spam	Lottery
Not Spam	Lottery
Spam	No Lottery
Spam	No Lottery
Not Spam	No Lottery
Not Spam	No Lottery
Not Spam	Lottery
Not Spam	Lottery
Spam	No Lottery
Spam	No Lottery

P(Class) = P(Spam) = No. of times spam appears in the data / Total no. of emails = 5/10

P(Data) = P(Lottery) = No. of times lottery appears in the data / Total no. of emails = 4/10

P(Data | Class) = P(Lottery | Spam) = No. of emails having word lottery given that emails are spam = 1/5. In total there are 5 spam emails and out of which 1 email has the word lottery.

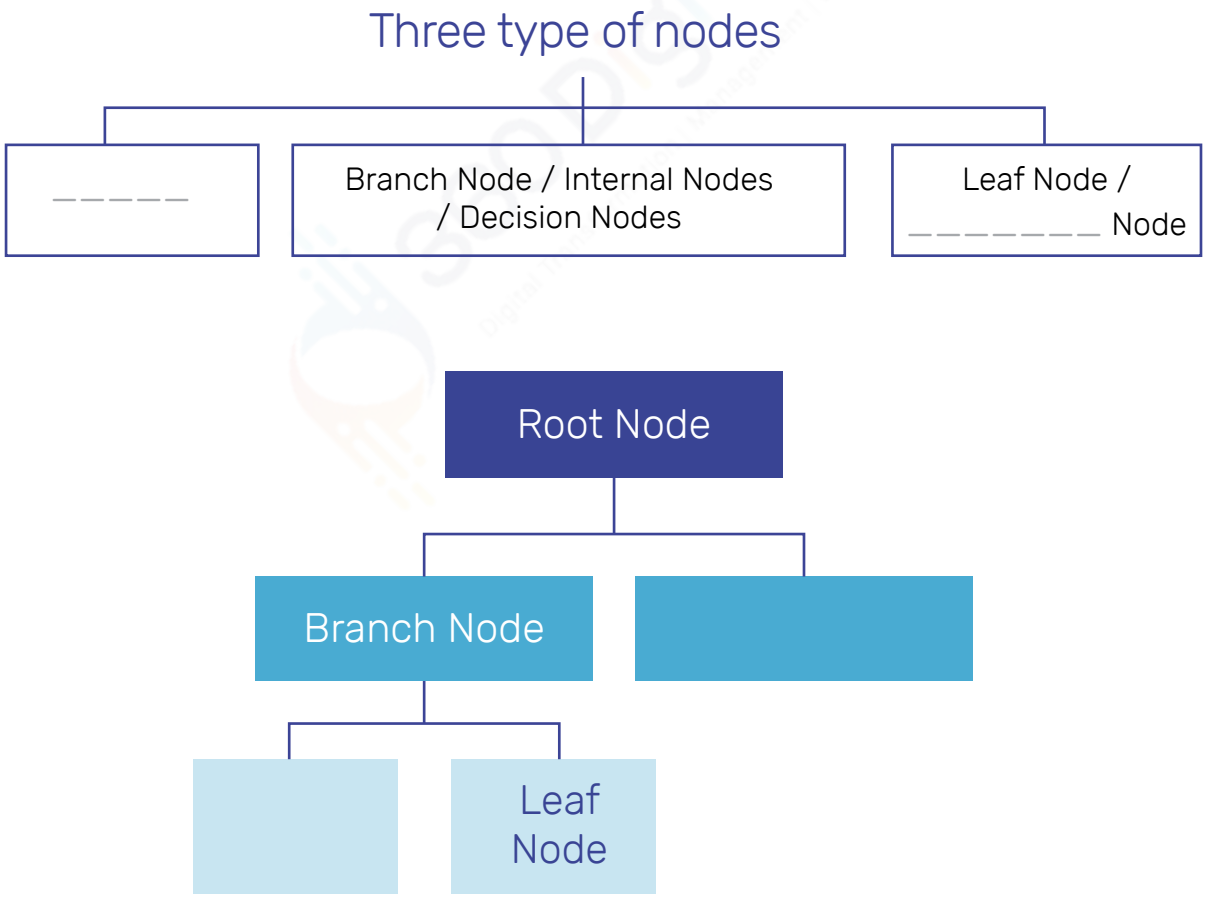
Decision Tree

"_____ Decision Tree"
When output is Categorical

"_____ Decision Tree"
When output is Numerical

Decision trees are

- Nonparametric _____ model, that works on divide & conquer strategy
- Rule-based algorithm that works on the principle of _____.
  A path from root node to leaf node represents a rule
- Tree-like structure in which an internal node represents a test on an attribute, each branch represents outcome of test and each leaf node represents the class label

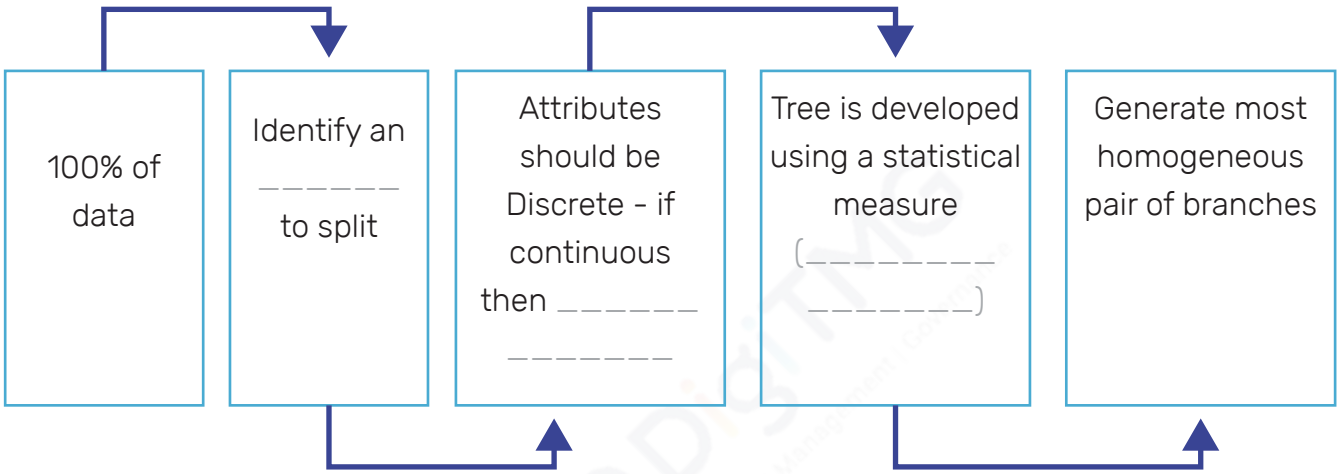


A Greedy Algorithm

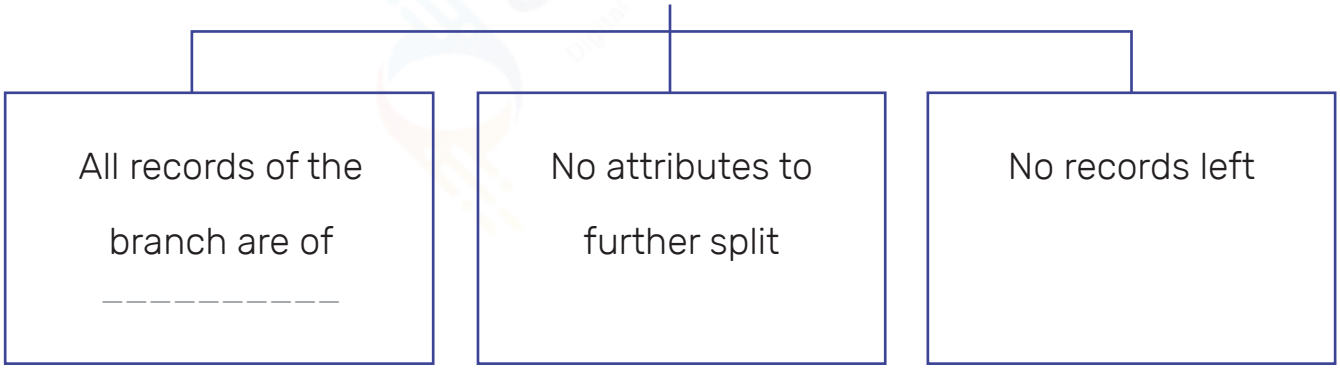
To develop a Decision Tree, consider 2 important questions:

Q1. Which _____

Q2. When to _____



Conditions to Stop



Age	CR	Class
>40	Fair	Yes
>40	Excellent	No
31 ..40	Excellent	Yes
<=30	Fair	Yes

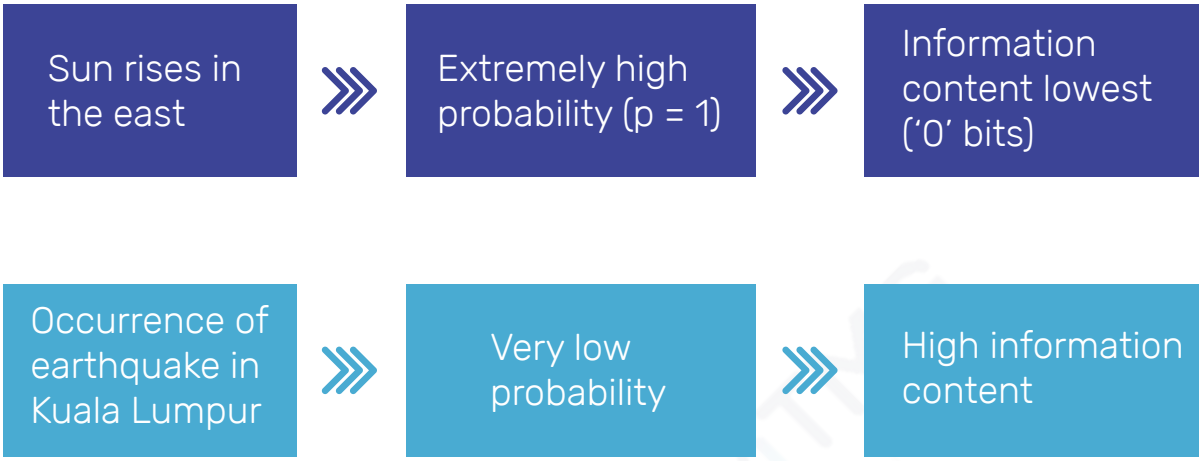
Age	CR	Class
>40	Fair	Yes
<=30	Excellent	Yes

Age	CR	Class
31 ..40	Fair	Yes

Information Theory 101

If the event is very _____, then the _____ content in the event is very low.

Examples



In conclusion "Information Content is proportional to Rarity"

$$I(\text{event}) = \log_2 \left(\frac{1}{\text{Prob(event)}} \right) = -\log_2 \text{Prob (event)}$$

Entropy:

- Entropy is the expected information content of all the events
- Entropy value of 0 means the sample is completely homogeneous
- Entropy value of 1 means the sample is completely heterogeneous

$$H(p = (p_1 \dots p_n)) = \sum_{i=1}^{n} p_i \log \left(\frac{1}{p_i} \right) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Purity = Accuracy = 1 - Entropy

Information Theory 101

In Accuracy we assign the _____ Label to each region.

	5	60	40	40	60	10
Dominant Label	Sky Blue		NA		Royal Blue	
Accuracy - Sky Blue	60/65		40/80		60/70	

Entropy is a measure of disorder or impurity (variation/_____)

Decision trees find attributes which return the most homogeneous branches.

Purity can also be measured using GINI Measure, which is the Expected. Accuracy with _____ Labeling.

	5	60	40	40	60	10
Dominant - Royal Blue	5/65		40/80		10/70	
Accuracy - Sky Blue	60/65		40/80		60/70	

$$\left(\frac{5}{65}\right) x \left(\frac{5}{65}\right) + \left(\frac{60}{65}\right) x \left(\frac{60}{65}\right) \qquad \left(\frac{60}{70}\right) x \left(\frac{60}{70}\right) + \left(\frac{10}{70}\right) x \left(\frac{10}{70}\right)$$

After calculating the measure of _____, one must decide on which feature to split. For this, one must measure the change in _____ resulting from a split on each possible feature. This calculation is known as _____

Information gain of a feature is the difference between entropy in the segment before the split (S_1) and partitions resulting from the split (S_2).

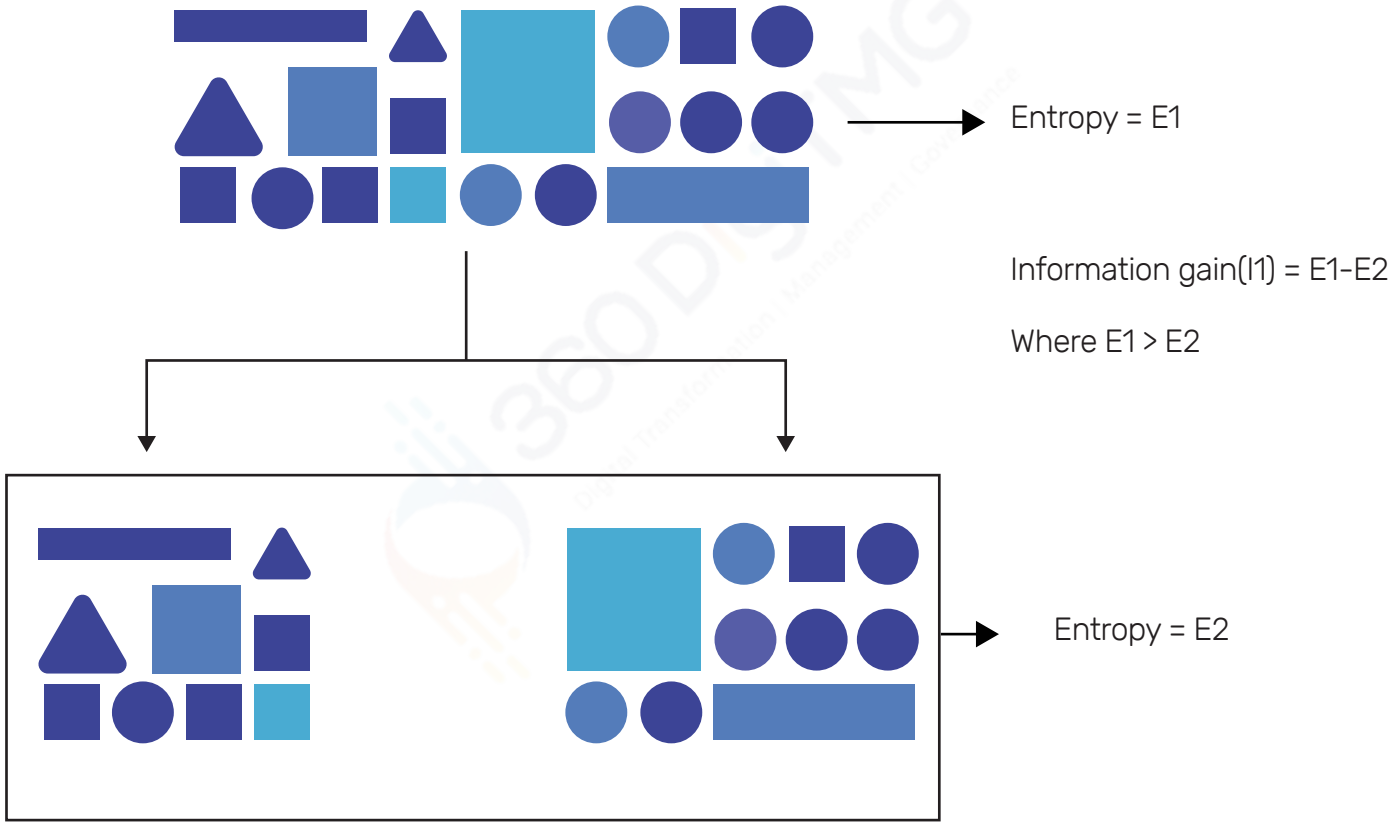
InfoGain (F) = Entropy (S_1) - Entropy (S_2)

Information Theory 101

Less the variation in class labels post the split better the _____

Information gain: Decrease in the _____ (variation) after the dataset is split on an attribute.

Higher homogeneity implies Higher information gain.



Pros and Cons of Decision Tree

Strengths	Weaknesses
Uses the important feature during decision making	Biased towards factors (features), which have a lot of levels
Interpretation is very simple because there is no mathematical background needed	Small changes in the data will result in large changes to decision making

Model overfitting can be addressed using _____ techniques.

_____ is the regularization technique used in the Decision Tree.

Pruning is the process of reducing the size of the tree to generalize the unseen data.

Two _____ techniques are

Pre-_____ or Early Stopping

Stopping the tree from growing once the desired condition is meet.

- Stop the tree from growing once it reaches a certain number of decisions
- Stop the tree from growing if decision nodes contain only a small number of examples

Disadvantage: When to stop the tree from growing. What if an important pattern was prevented from learning?

Post-_____

Grows the tree completely and then apply the conditions to reduce the tree size.

Example, if the error rate is less than 3% then reduce the nodes.

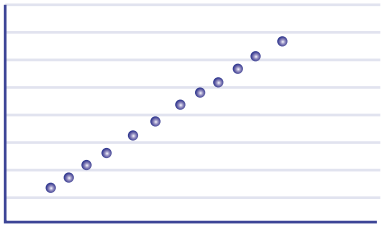
So, the nodes and branches that have less reduction of errors are removed.

This process of grafting branches is known as subtree raising or subtree replacement.

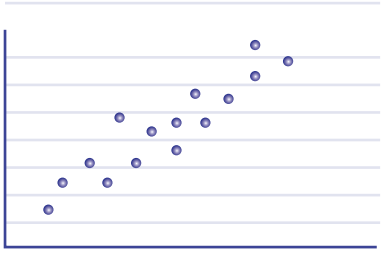
Post _____ is more effective than pre- _____

Continuous Value Prediction

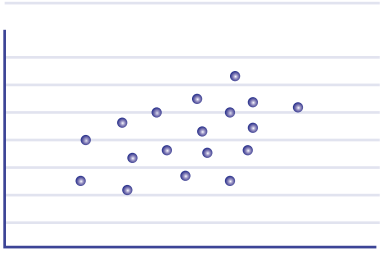
Scatter Diagram - Visual representation of the relationship between two continuous variables



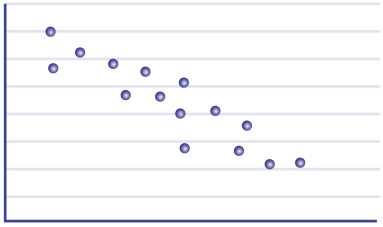
Strong Positive Correlation



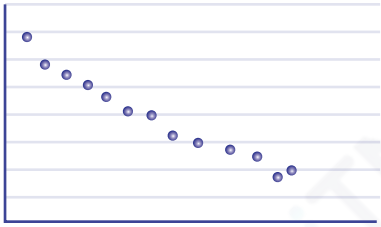
Moderate Positive Correlation



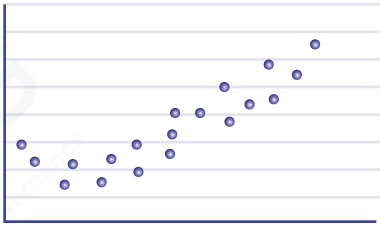
No Correlation



Moderate Negative Correlation

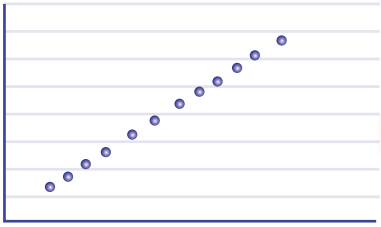


Strong Negative Correlation



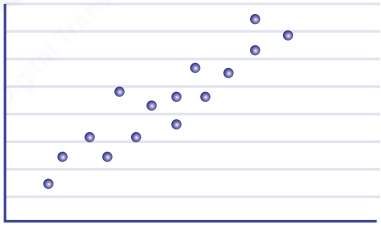
Curvilinear relationship

Correlation Analysis - Measures the correlation between two variables



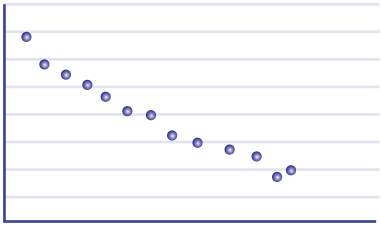
r = +1:

Perfect Positive Correlation



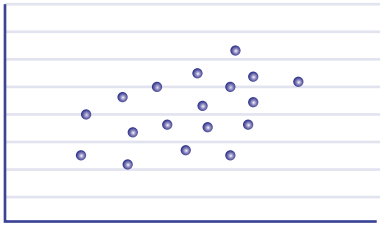
r close to +1:

Strong Positive Association



r close to -1:

Strong Negative Association



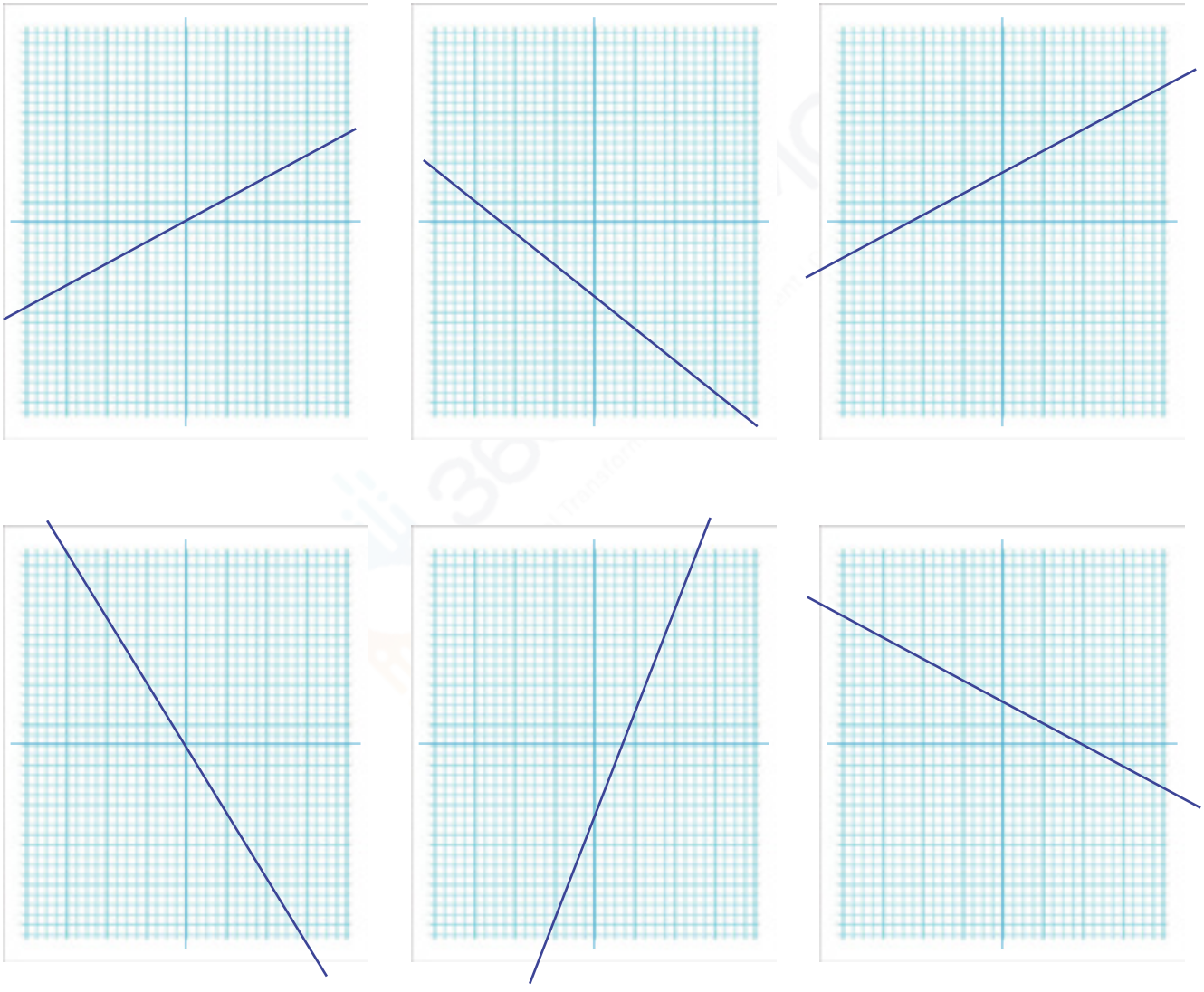
r close to 0:

Weak or No Association

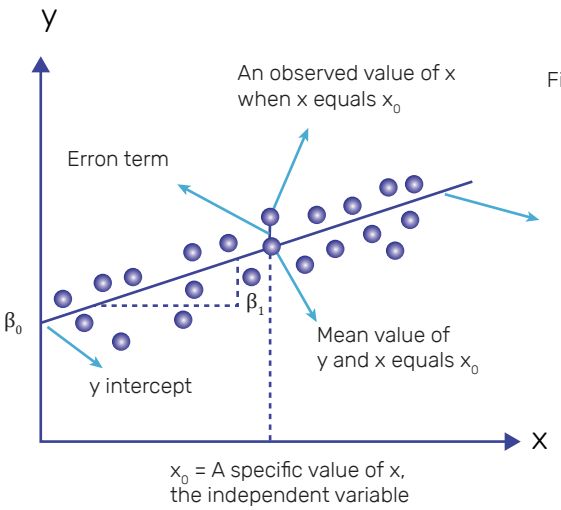
Linear Regression

Equation of straight line that we have learnt in our school days

$$\_\_\_\_\_$$



Ordinary Least Squares



$$y = \beta_0 + \beta_1 x + \varepsilon; \quad \varepsilon = \text{error term}$$

_____ Technique to find the best fit line.

The best fit line is the line which has minimum square deviations from all the data points to the line.

To improve the accuracy, transformations can be applied, this will ensure that the data has a linear pattern with minimum spread.

Coefficient of Determination R^2 - also known as goodness of fit, is the measure of predictability of Y (dependent variable) when X's (independent variable) are given.

It can be interpreted as the % of variability in output (Y) that can be explained with the

_____ (X)

R^2 = SSR/SST = ((SSR)/(SSR + SSE))

$0 \leqq R^2 \leqq 1$

Where,

SSR = $\boldsymbol{\Sigma}$ $(\hat{y} - \bar{y})^2$ (measure of explained variation)

SSE = $\boldsymbol{\Sigma}$ $(y - \hat{y})^2$ (measure of unexplained variation)

SST = SSR + SSE = $\boldsymbol{\Sigma}$ $(y - \bar{y})^2$ (measure of the total variation in y)

Model Assumptions

_____

_____

_____

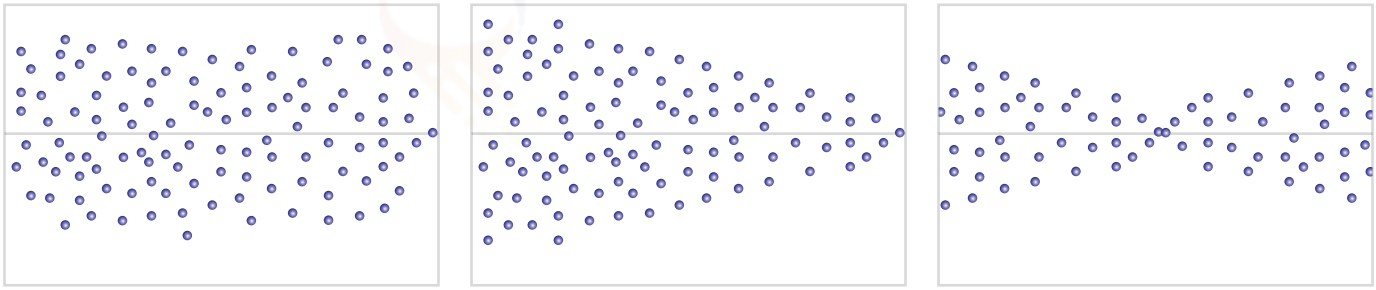
_____

Problems arise while linear regression model training:

_____ : Errors are dependent on each other

_____ : Errors have non-constant variance

_____ : Independent variable pair are linearly dependent on each other

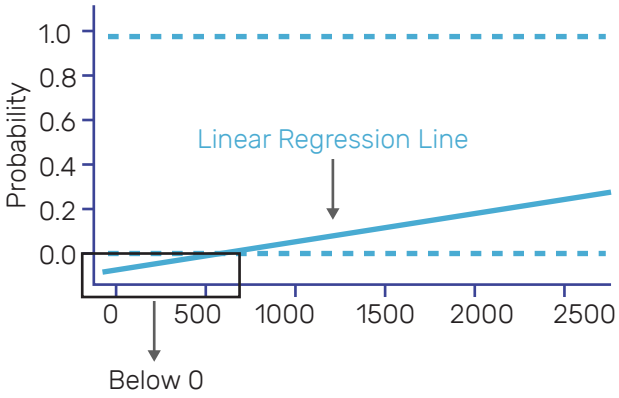


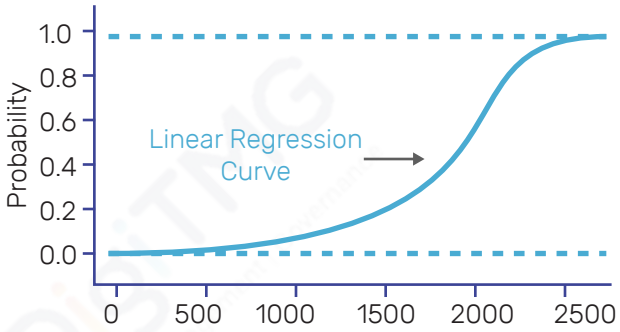
Logistic Regression

Predicts the _____ of the outcome class.

The algorithm finds the linear relationship between independent variables and a link function of these probabilities.

The link function that provides the best goodness-of-fit for the given data is chosen.





The output from logistic regression will lie between 0 to 1.

The logistic regression curve is known as _____ Curve.

Probability values are segregated into binary outcomes using a _____ value. The default cutoff is treated as 0.5 (50%)

- If probability of an event > 0.5; then Event is considered to be True (predicted outcome = 1)
- If probability of an event </= 0.5; then Event is considered to be False (predicted outcome = 0)

Logistic Regression

The logistic regression performed using:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Where,

β_0 = the y is _____

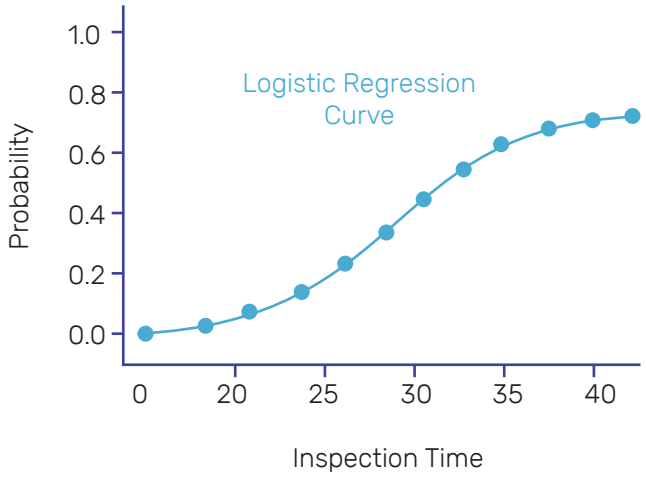
β_i = the model coefficient for the linear effect of variable is on y

e = the random error

The probability function:

$$p = \frac{e^y}{1+e^y}$$; where e = 2.7183

The output of the logistic regression will give a sigmoid curve (also known as S curve)



Interpretation: Probability p indicates that the event has a chance 'p' for a given

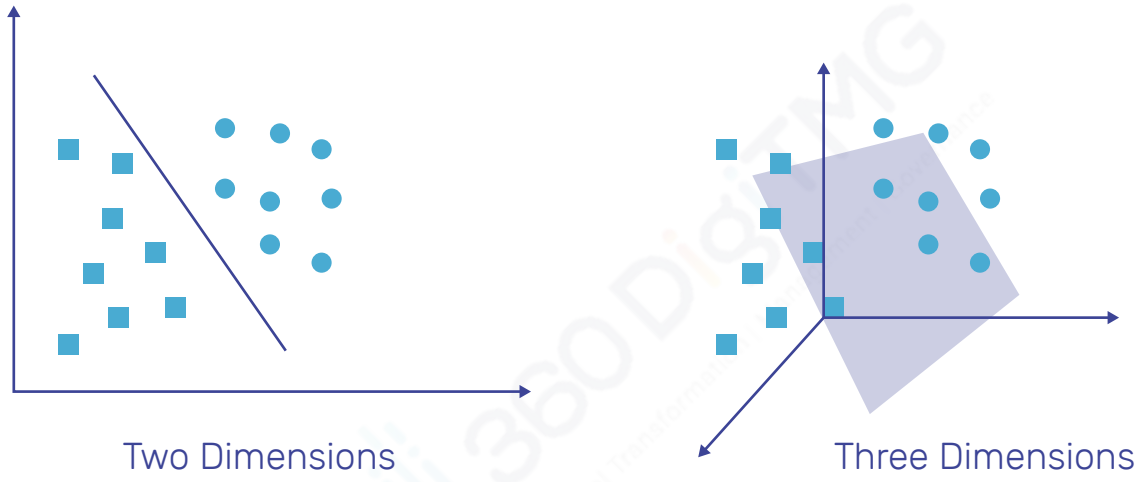
_____

Support Vector Machine

SVMs can be adapted to use with nearly any type of learning task, including both _____ and _____.

SVM is inspired from statistical learning theory.

Other names: Large-margin classifier, Max-margin classifier, _____

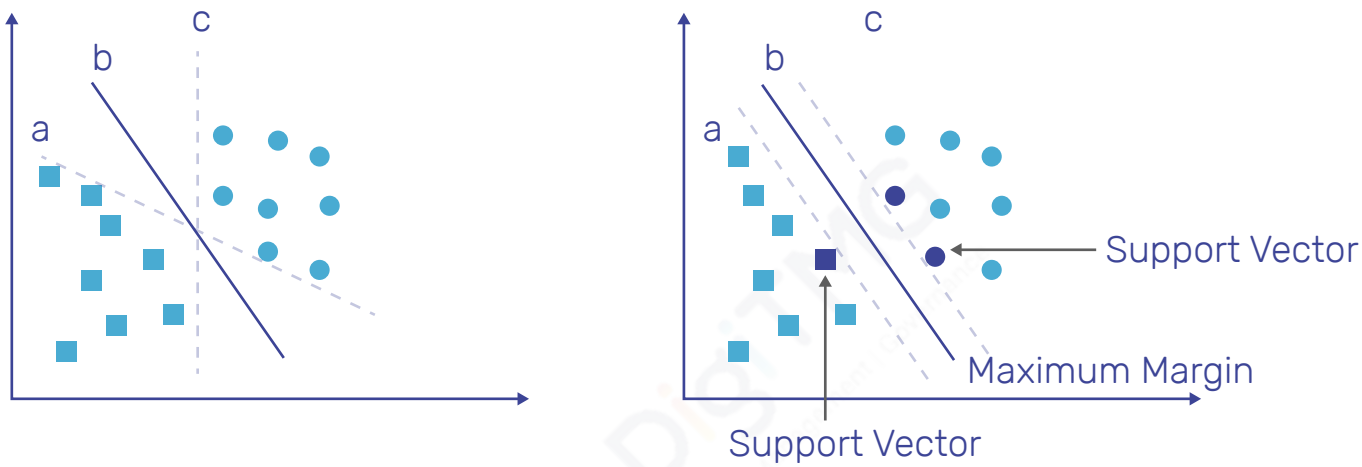


The task of the SVM algorithm is to identify a line that separates the two classes in a binary problem. However, in a multidimensional problem a line cannot separate the classes.

The goal of an SVM is to create a flat boundary called a _____, which divides the space to create _____ partitions.

Support Vector Machine

There is more than one choice of dividing line between the groups of circles and squares.



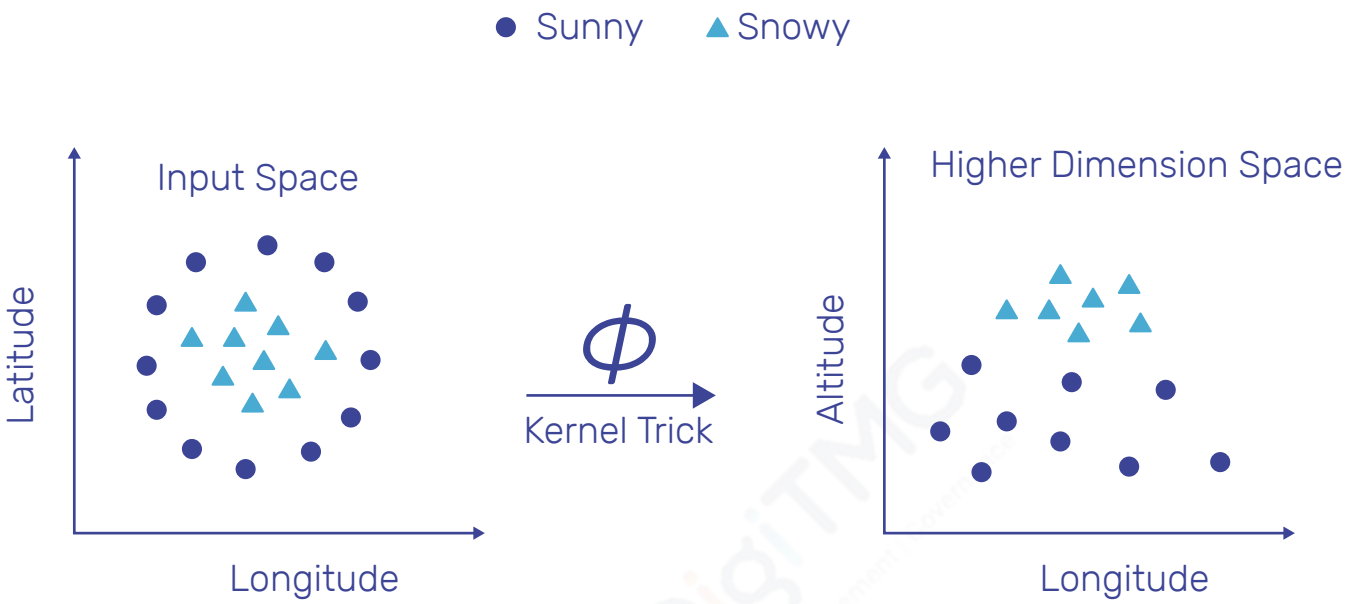
SVM searches for M_____ M_____ H_____ (MMH)

MMH is as far away as possible from outer boundaries (convex hull) of the two groups of data points.

The maximum margin linear classifier is the linear classifier with the maximum margin. This is the simplest kind of SVM (Called an LSVM).

Support Vector Machine

Non-Linear Spaces



Kernel Tricks

A key feature of SVMs is their ability to map the problem into a higher dimension space using a process known as the _____ trick. After the _____ trick has been applied, we look at the data through the lens of a new dimension and a nonlinear relationship may suddenly appear to be quite linear.

Support Vector Machine

Kernel Functions

- The linear kernel does not transform the data at all. Therefore, it can be expressed simply as the dot product of the features:

$$K\left(\vec{x}_i, \vec{x}_j\right) = \left(\vec{x}_i, \vec{x}_j\right)$$

- The_____ kernel results in a SVM model somewhat analogous to a neural network using a sigmoid activation function. The Greek letters kappa and delta are used as kernel parameters

$$K\left(\vec{x}_i, \vec{x}_j\right) = \tanh\left(k\vec{x}_i . \vec{x}_j - \delta\right)$$

- The polynomial kernel of degree d adds a simple non-linear transformation of the data

$$K\left(\vec{x}_i, \vec{x}_j\right) = \left(\vec{x}_i . \vec{x}_j + 1\right)^d$$

- The _____ kernel is similar to a RBF neural network. The RBF kernel performs well on many types of data and is thought to be a reasonable starting point for many learning tasks

$$K\left(\vec{x}_i, \vec{x}_j\right) = e^{\frac{-\left\|\vec{x}_i - \vec{x}_j\right\|^2}{2\sigma^2}}$$

Deep Learning/Neural Network

Artificial Neural Network is used to mimic Biological Neural Network.

Deep Learning is named in this way because it has many _____ _____ to the output.

_____ Models versus _____ Learning Models

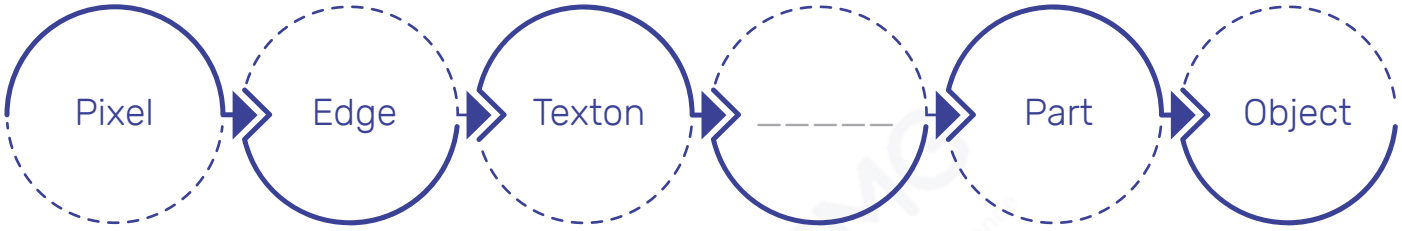
_____ Extraction from the data (Images, Speech, Text, Videos (videos are a subset of images))is automatically performed using Deep Learning models.

Deep Learning/Neural Network

The multiple layers capture compositionality:

Image Recognition

Each layer captures some features, For Example:

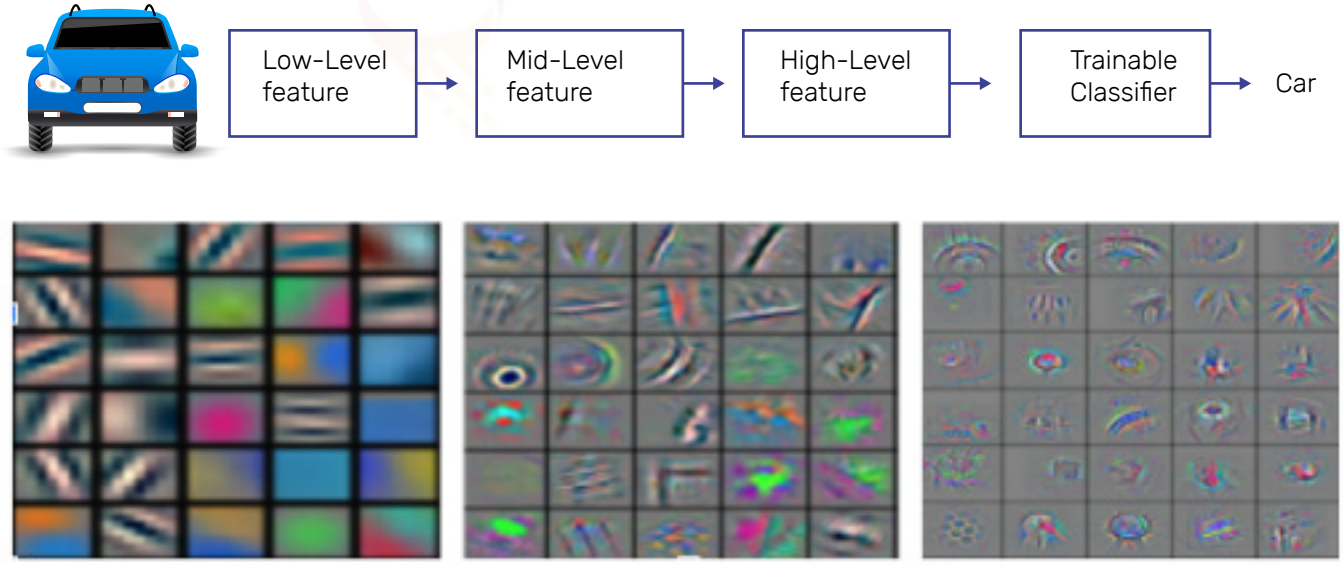


Initial layers capture _____ features

Next layers capture _____ features

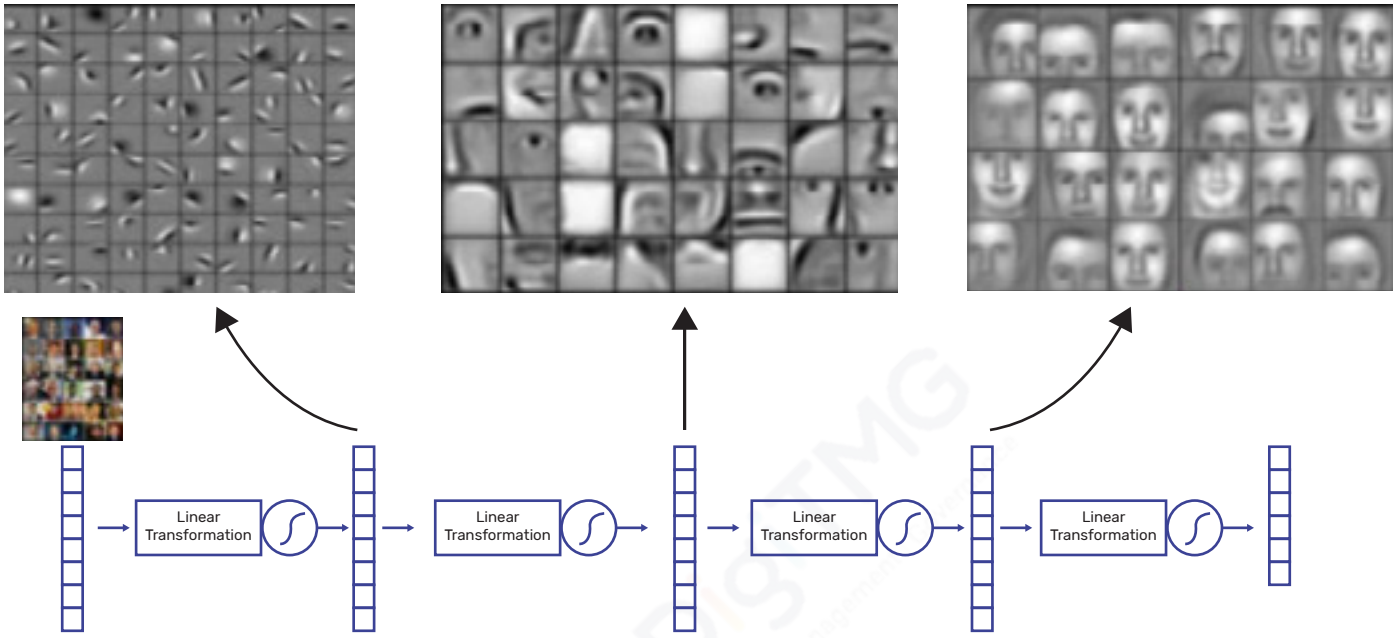
Final layers capture _____ features

At the end, the classifier will predict the output.

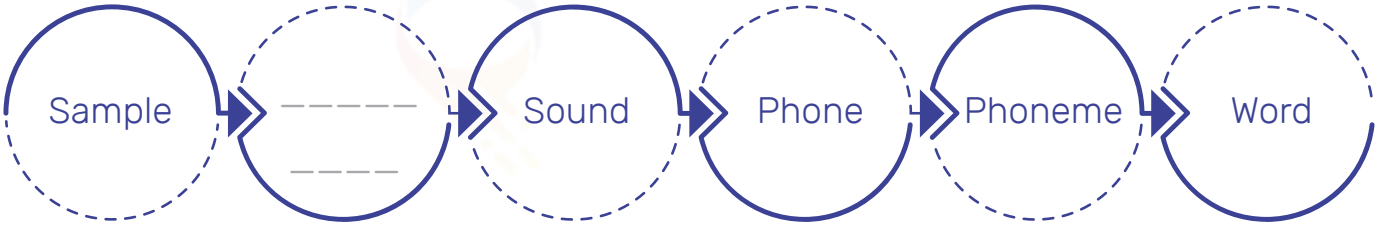


Deep Learning/Neural Network

Deep Learning Learns Layers of Features



Speech data is processed through multiple layers and _____ is captured.



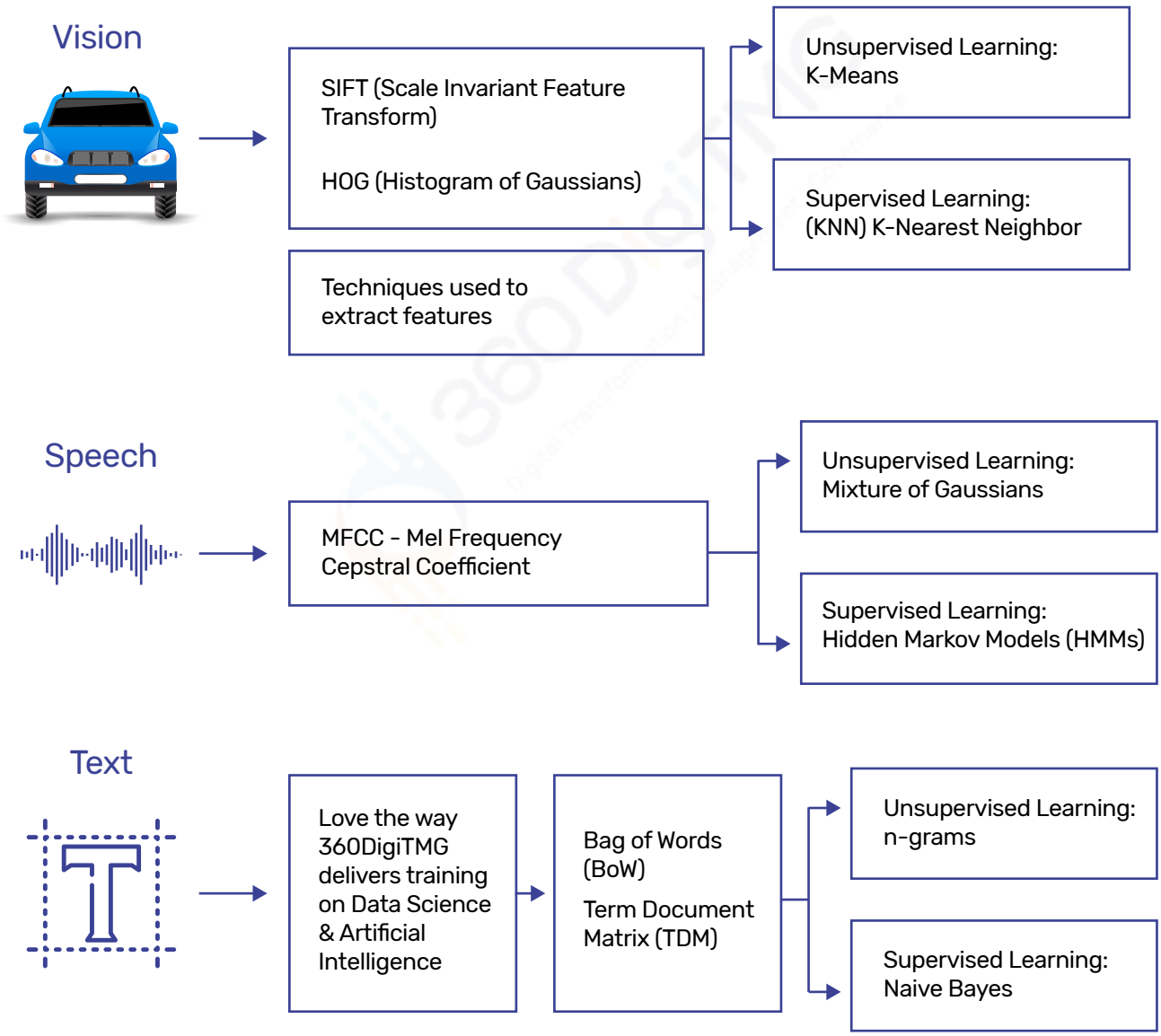
Text data is processed through Deep Learning layers and compositionality is captured.

Deep Learning/Neural Network

Shallow Machine Learning Models:

Feature extraction from the data is performed manually.

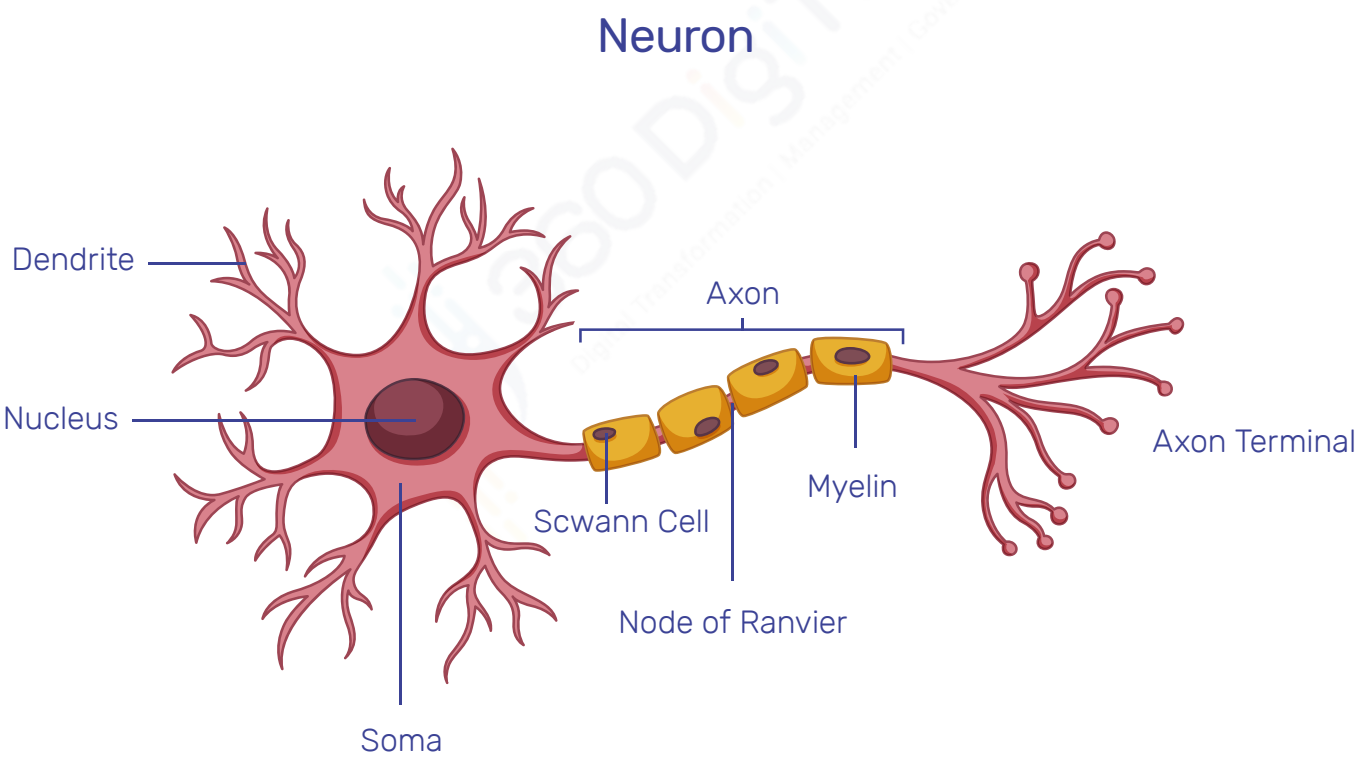
Vision / Images / Videos: Videos are broken into _____
and from each _____, the features are extracted.



Perceptron Algorithm:

Artificial Intelligence is about trying to mimic a human brain.

An Artificial Neural Network (ANN) models the relationship between a set of _____ signals and an _____ signal using a model derived from our understanding of how a biological brain responds to stimuli from sensory inputs. Just as a brain uses a network of interconnected cells called Neurons to create a massive parallel processor, ANN uses a network of artificial neurons or nodes to solve learning problems.



Deep Learning/Neural Network

Simple Neural Network components:

1. Input layer - contains the numbers of _____ equal to the number of input features

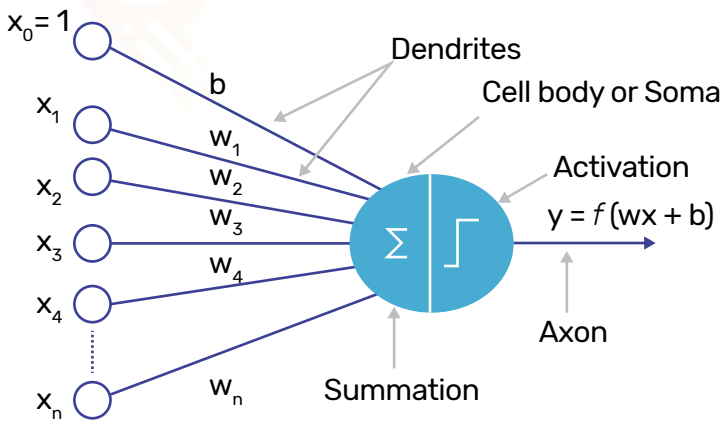
2. Input layer also has one additional neuron called _____, which is equivalent to the 'b' (y-intercept) in the equation of the line y = b + mx

3. 'b', 'w_1', 'w_2', 'w_3',..... are called as weights and are _____ initialized

4. These neurons are also called as nodes and are connected via an edge to the neuron in the next layer

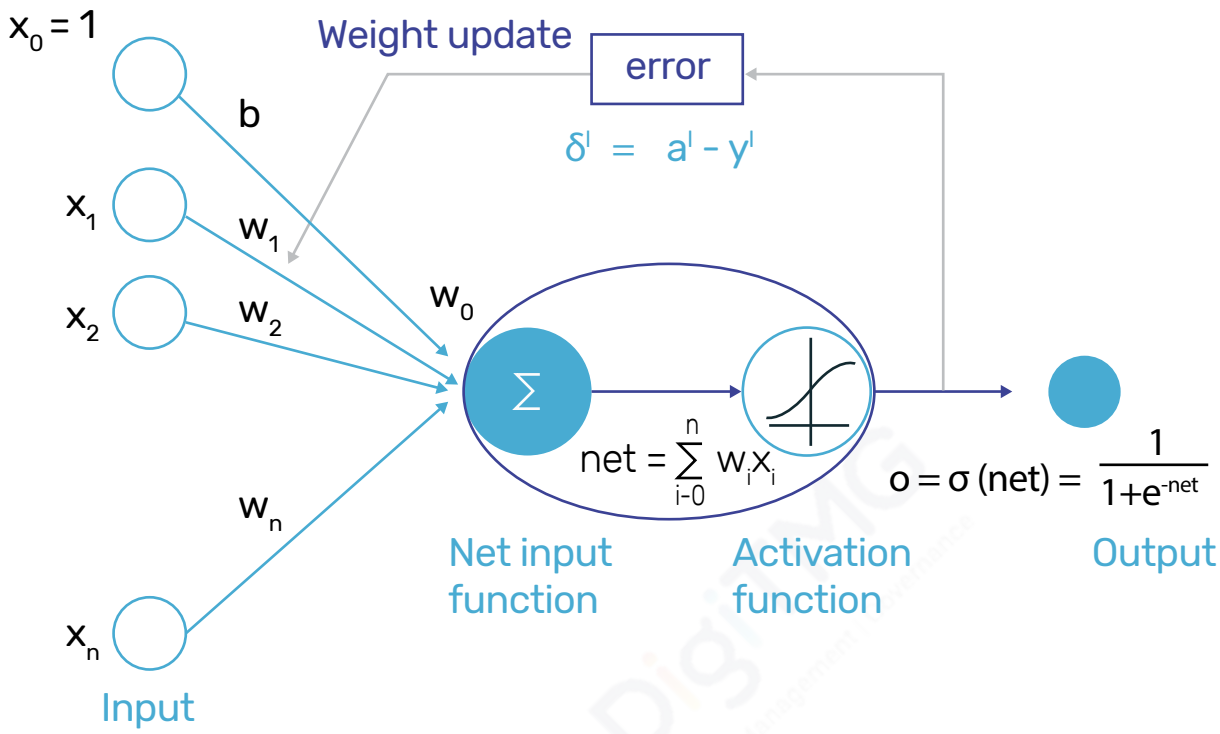
5. _____function (usually summation) is used to _____ all the inputs and corresponding weights
   $f(x) = b + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5$ -> This equation will give a numerical output

6. The output of the integration function is passed on to the _____ component of the neuron



Neural Network

Deep Learning/Neural Network



7. Based on the functioning of _____ function, the final output is predicted

8. Predicted output and actual output are compared to calculate the _____ function / _____ function (error calculated for each record is called as _____ function and combination of all these individual errors is called as cost function)

9. Based on this error, the _____ algorithm is used to go back in the network to update the weights

10. Weights are updated with the objective of minimizing the error and this minimization of error is achieved using _____ Algorithm

Deep Learning/Neural Network

Perceptron Algorithm

The Perceptron algorithm was proposed back in 1958 by Frank Rosenblatt (Cornell Aeronautical Laboratory).

Neural network with no hidden layers and a single output neuron is called a _____ Algorithm.

The _____ algorithm can only handle _____ boundaries. _____ boundaries are handled using the Multi-Layered _____ algorithm.

Weight updation as part of _____ algorithm is done using the following formula:

Randomly initialize weight vector $\vec{w}_0$

Repeat until error is less than a threshold γ or max_iteratios M:

For each training example $(\vec{x}_i, t_i)$:

Predict output y1 using current network weights $\vec{w}_n$

Update weight vector as follows:

$$\vec{w}_{n+1} = \vec{w}_n + \eta * (t_i - y_i) * \vec{x}_i$$

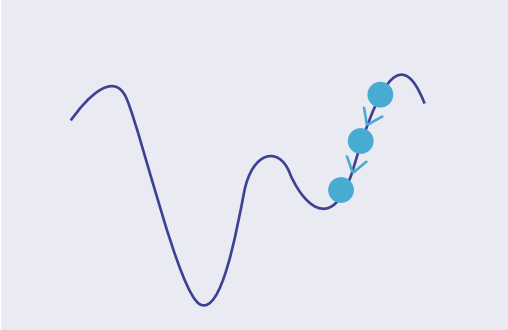
Learning Rate Error

Deep Learning/Neural Network

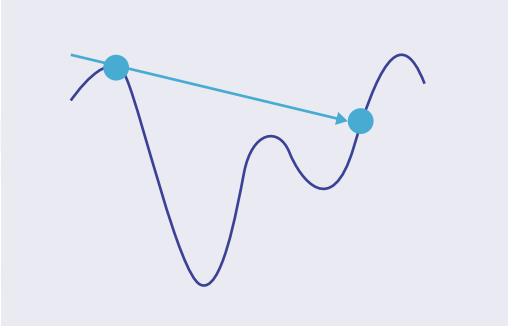
_____ are updated so that the error is minimized.

Learning Rate is also called Eta value and ranges from 0 to 1.

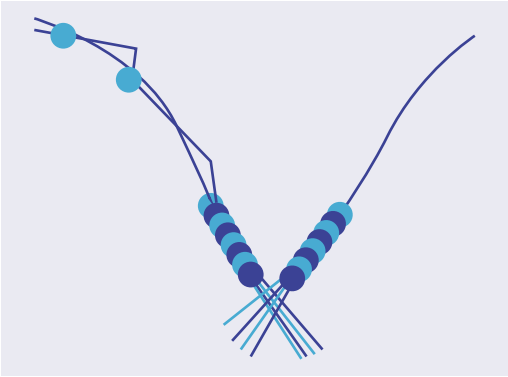
A value close to 0 would mean _____ steps to arrive at the bottom of the error surface.



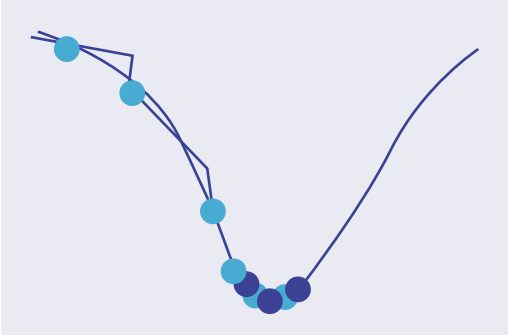
A value close to 1 would mean _____ the bottom of the error surface.



Constant learning rate creates a problem of bouncing around the bowl. The gradient will never reach the bottom of the error surface.

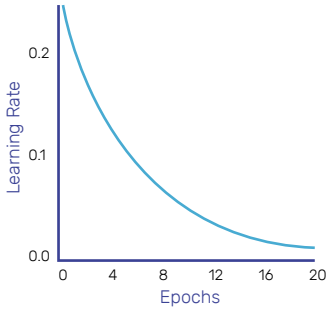


This problem is solved using Changing Learning Rate (_____).

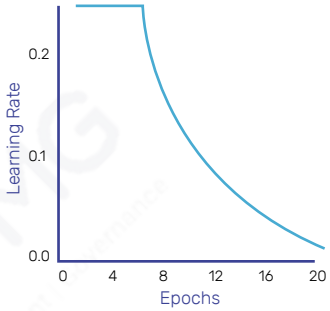


Deep Learning/Neural Network

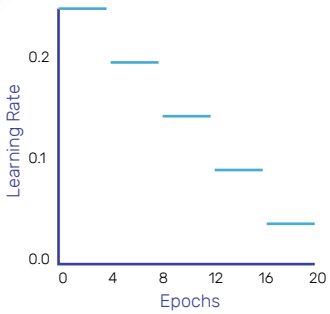
_____- Learning rate is reduced epoch after epoch, until it reaches the end of a defined number of epochs.



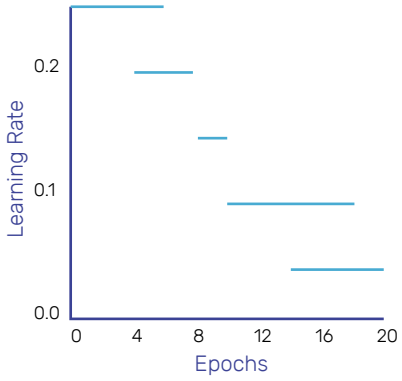
_____ - Learning rate is the same for a fixed number of epochs and then it starts reducing after every epoch until the defined number of epochs reaches to end.

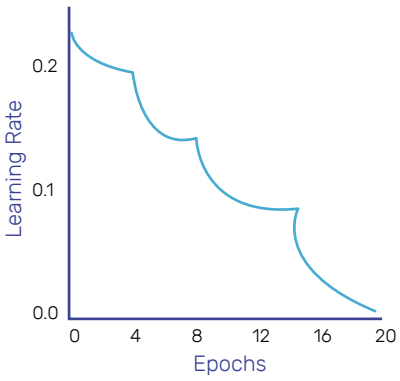


_____ - Learning rate is reduced after a set of fixed number of epochs (for e.g., learning rate will be reduced by 10% after every 5 epochs).



_____- Learning rate is reduced when it is observed that the error stops reducing.





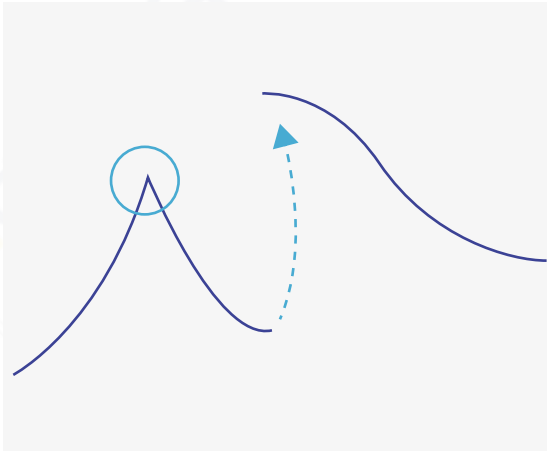
Deep Learning/Neural Network

Gradient Primer:

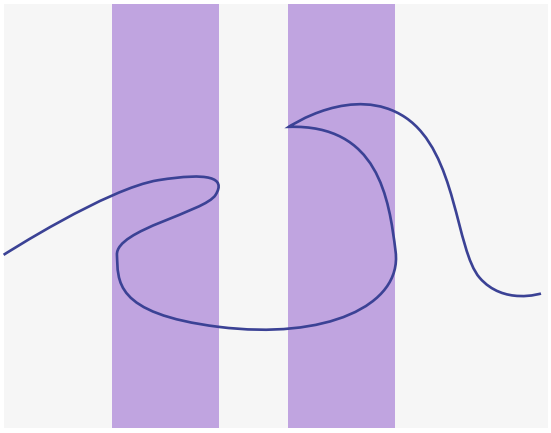
Gradient is also called as

- Rate of change
- _____
- Slope

Curves/Surfaces should be continuous and smooth (_____ /sharp points)



Curves / Surfaces should be _____



Deep Learning/Neural Network

Gradient Descent Algorithms Variants:

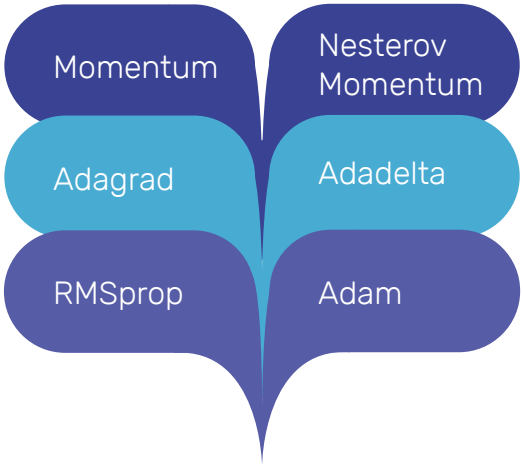
A few definitions:

Iteration: Equivalent to when a weight update is done

Epoch: When entire training set is used once to update the weights

	Batch Gradient Descent	Stochastic Gradient Descent	Mini-batch Stochastic Gradient Descent
Epoch	1	1	1
Example	10000 training records	10000 training records	10000 training records
Iteration	1	10000	100 (if minibatch size is 100). 10000/100 = 100 iterations
Example	Weights are updated once, after all 10000 training records are passed through the network	Weights are updated after each training sample passes through the network. If we have 10000 training samples then weights are updated 10000 times	Weights are updated after every minibatch (100 in this case) of records are passed through the network. Records within minibatch are randomly chosen.

Other advanced variants of Mini-Batch SGD:



Deep Learning/Neural Network

Empirically Determined components are:

1. Number of hidden layers
2. Number of _____ within each hidden layer
3. _____
4. Error/Cost/Loss Functions
5. _____ Methods

Y (output)	No. of neurons in output layer	Activation Function in Output layer	Loss Function
Continuous	1	Linear / Identify	ME, MAE, MSE, etc.
Discrete (2 categories)	1 for binary classification problem	Sigmoid / Tanh	Binary Cross Entropy
Discrete (>2 categories)	10 if we have a 10 class problem	Sigmoid	Categorical Cross Entropy

Note: Hidden layers can have any activation function and majorly _____ activation functions seem to be giving good results.

Multi-Layered Perceptron (MLP) / Artificial Neural Network (ANN)

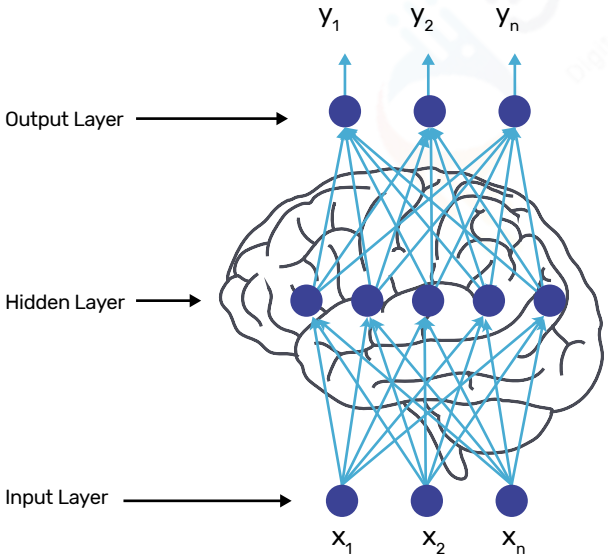
Non-Linear patterns can be handled in two ways: Changing _____ Function:

Quadratic Function

$$f = \sum_{j=1}^{m} w_j x_j^2 - \theta$$

Spherical Function

$$f = \sum_{j=1}^{m} (x_j - w_j)^2 - \theta$$



The presence of hidden layers alone will not capture the _____ pattern. Activation function to be used should be non-linear.

Usage of linear or identify activation functions within the neurons of the hidden layer will only capture linear patterns.

If no activation functions are specified in the layers then by default the network assumes_____.

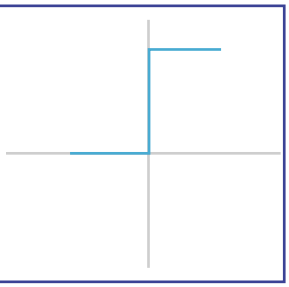
Multi-Layered Perceptron

List of activation functions include

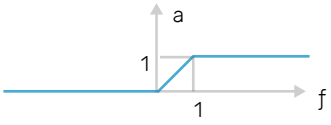
Identity function (Linear function)

$a(f) = a$

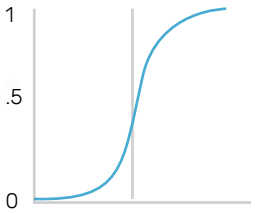
____ function



Ramp function

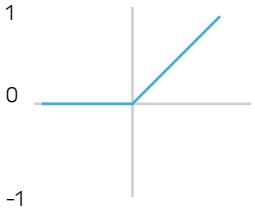


Sigmoid function



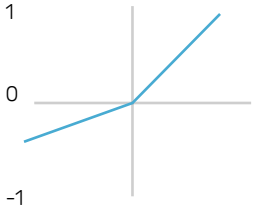
Tanh function

ReLU (Rectified Linear Unit) function

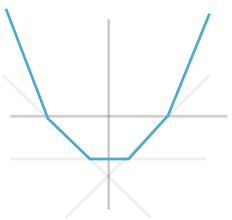


ELU (Exponential Linear Unit)

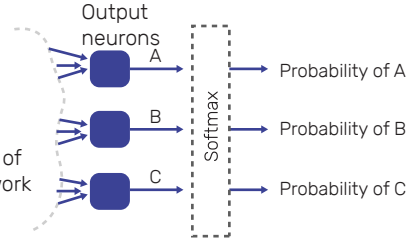
_____ ReLU



Maxout



_____ Rest of network



Multi-Layered Perceptron

Regularization Techniques used for Overfitting

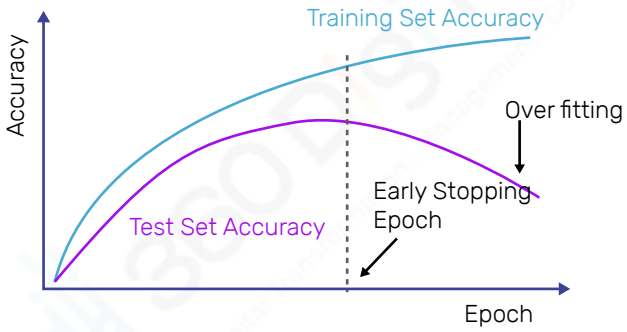
L1 regularization / L1 _____ decay term

L2 regularization / L2 _____ decay term

Weight Decay Term

$$j(\emptyset) = \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\| h_{\emptyset}(x^{(i)}) - y^{(i)}\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n1-1}\sum_{i=1}^{s1}\sum_{j=1}^{s1+1}\left(w_{ji}^{(l)}\right)^2$$

___ stopping:



Error-change criterion

1. Stop when error isn't dropping over a window of, say, 10 epochs
2. Train for a fixed number of _____ after criterion is reached (possibly with lower learning rate)

Weight-change criterion

1. Compare _____ at epochs t-10 & t and test

$$\max_i \left\| w_i^t - w_i^{t-10} \right\| < \rho$$

2. Possibly express as a _____ of the weight

Multi-Layered Perceptron

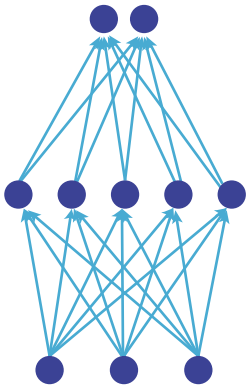
Dropout

It is an interesting way to perform model averaging in Deep Learning

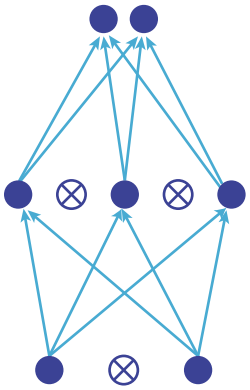
Training Phase: For each hidden layer, for each training sample, for each iteration, ignore (zero out) a _____, p, of nodes (and corresponding activations).

Test Phase: Use all _____, but reduce them by a factor p (to account for the missing activations during training).

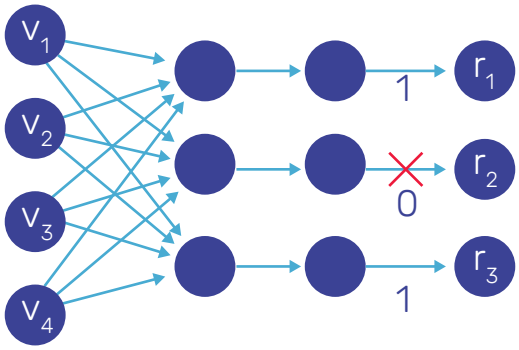
Randomly select a subset of _____ and force their output to _____.



Standard
Neutral Net



After applying
Dropout

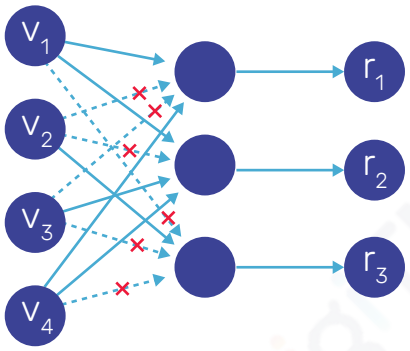


Multi-Layered Perceptron

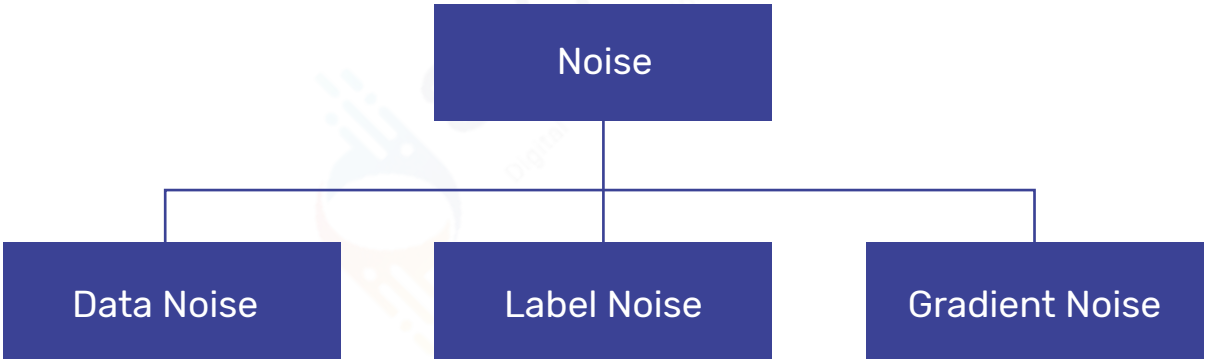
Drop Connect:

Very similar to _____, however, we disable the weights instead of the nodes.

Here the nodes are partially active.



Noise:



- Add noise to data while _____

- Disturb each training sample with the probability
- For each disturbed sample, the label is randomly drawn from a uniform distribution regardless of the true label

- Add _____ to the gradient

Multi-Layered Perceptron

Batch Normalization:

Input: Values of x over a mini - batch: B = { x1...m };

Parameters to be learned: γ, β

Output: $\{y_i = BN_{\gamma,\beta}(x_i)\}$

$$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini - batch } \_\_\_\_\_$$

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B) \qquad \text{// mini - batch } \_\_\_\_\_$$

$$X \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma X_i + \beta = BN_{\gamma,\beta}(x_i) \qquad \text{// } \_\_\_\_\_$$

- Batch Normalization layer is usually inserted before _____ layer (after Fully Connected or Dense Layer)
- Reduces the strong dependence on weight initialization

Shuffling inputs:

- Choose examples with maximum _____ content
- Shuffle the training set so that successive _____ examples never (rarely) belong to the same class
- Present input examples that produce a large error more frequently than examples that produce a small error. Why? It helps to take large steps in the Gradient descent

Multi-Layered Perceptron

Weight Initialization Techniques:

_____ initialization $\quad \text{uniform}\left(-\frac{\sqrt{6}}{\sqrt{fan_{in}+fan_{out}}}, \frac{\sqrt{6}}{\sqrt{fan_{in}+fan_{out}}}\right)$

Caffe implements a simpler version of Xavier's initialization

$$\text{uniform}\left(-\frac{2}{fan_{in}+fan_{out}}, \frac{2}{fan_{in}+fan_{out}}\right)$$

_____ initialization

$$\text{uniform}\left(-\frac{4}{fan_{in}+fan_{out}}, \frac{4}{fan_{in}+fan_{out}}\right)$$

Forecasting

Time Series vs _____ Data

Time Series Data:

Data that is collected over equal spaced time intervals and the time interval is also an essential part of the data.

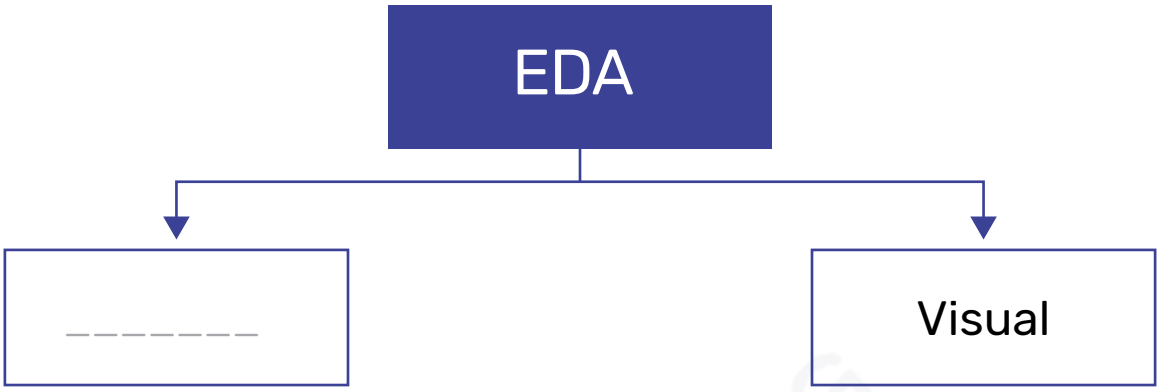
_____ Data:

Data that can be collected at a single point of time.

Forecasting is the use of various modeling techniques to predict a future outcome on the basis of historical time series data.

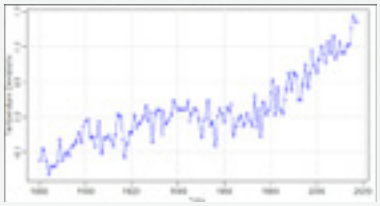
Forecasting

EDA - Components of Time Series

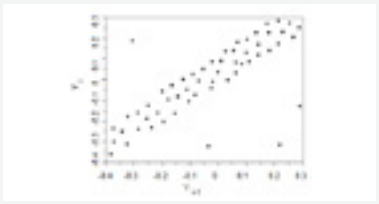


EDA in time series is mostly visual.

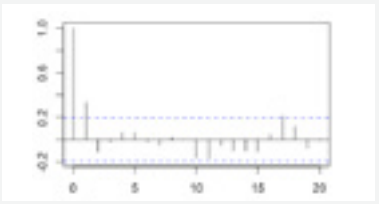
Elements of visualization in time series:



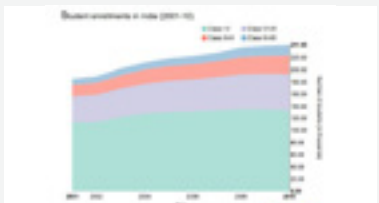
Time plot



Lag Scatter Plot

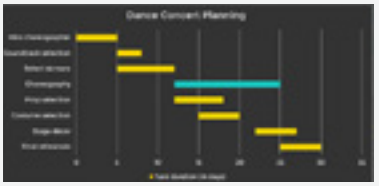


_ _ _ _ _ Plot

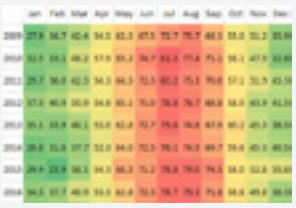


Stacked Area Chart

for multi variable time series data



_ _ _ _ _ _ Chart



Heat Map

Forecasting

Data Partition

Time series should be split in _____ order

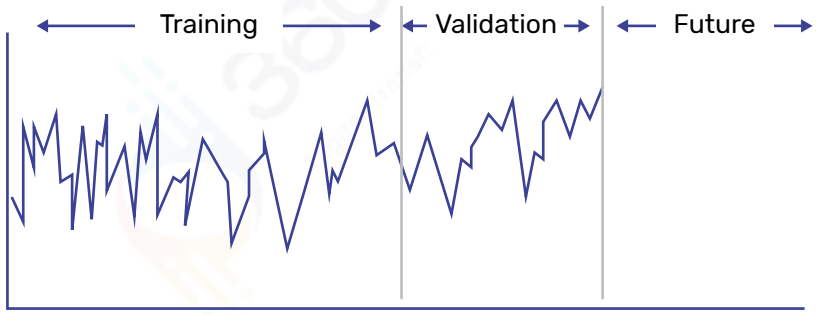
Most Recent period data will be chosen as Validation data.

Training Data

Fit the model only to Training period

Validation Data

Assess performance on Validation period



Conditions to choose the validation period:

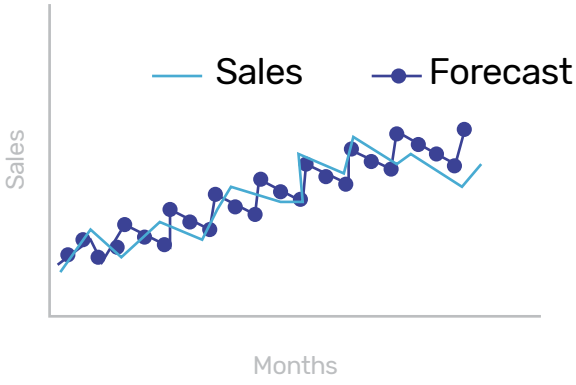
- Forecast Horizon
- _____
- Length of Time series

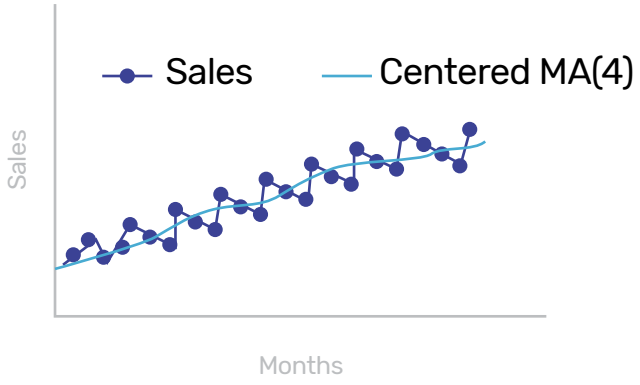
Forecasting

Forecasting Model

There are predominantly 2 approaches to forecasting

Model-Driven	Data-Driven
_____ similar to future	_____ similar to future
1. Linear Regression	1. Naïve forecasts
2. Autoregressive models	2. _____
3. ARIMA	3. Neural nets
4. _____	
5. _____	





Forecasting

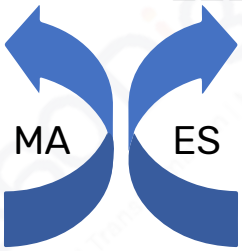
Smoothing Techniques

Moving Average

- _____ Moving Average
- _____ Moving Average

Exponential Smoothing

- Simple Exponential Smoothing
- Holt's Method/
  Double Exponential Smoothing
- _____ Method



Moving Average	Exponential Smoothing
Assigns equal weights to all past observations	Assigns _____ to recent observations than past observations
Better to forecast when data & environment is not _____	Better to forecast when data & environment is _____
Window width is key to success	Smoothing constant (α, β, γ) value is key to success ($0 < \alpha, \beta, \gamma \leq 1$)

Forecasting

De-Trending and De-Seasoning

Regression

- To remove _____ and/or _____, fit a regression model with trend and/or seasonality
- Series of forecast errors should be de-trended & deseasonalized

Differencing

- Simple & popular for removing trend and / or seasonality from a time series
- Lag-1 difference: Yt – Yt-1 (For removing _____); Lag-M difference: Yt – Yt-M (For removing _____)
- Double differencing: difference the differenced series

Moving Average

- Uses moving average to remove _____
- Generates seasonal indexes as a byproduct

Accreditation to international certification bodies

Project Management Institute. Registered Education Provider

TÜV SÜD

Confederation of Indian Industry

FICCI

mongoDB

AiSPRY

irtics Industry Revolution 4.0. Data Analytics

City & Guilds ASSURED

UTM Universiti Teknologi Malaysia — School of Professional and Continuing Education (SPACE)

IBM

Panasonic ED-TECH SOLUTIONS

CAREEREX

360DigiTMG

Digital Transformation | Management | Governance

For further details, call us at

1800-212-654321

enquiry@360digitmg.com

360digitmg.com

2-56/2/19, 3rd Floor, Vijaya Towers, Ayyappa Society Road, Madhapur, Hyderabad, Telangana 500008

USA | INDIA | MALAYSIA | ROMANIA | SOUTH AFRICA | DUBAI | BAHRAIN