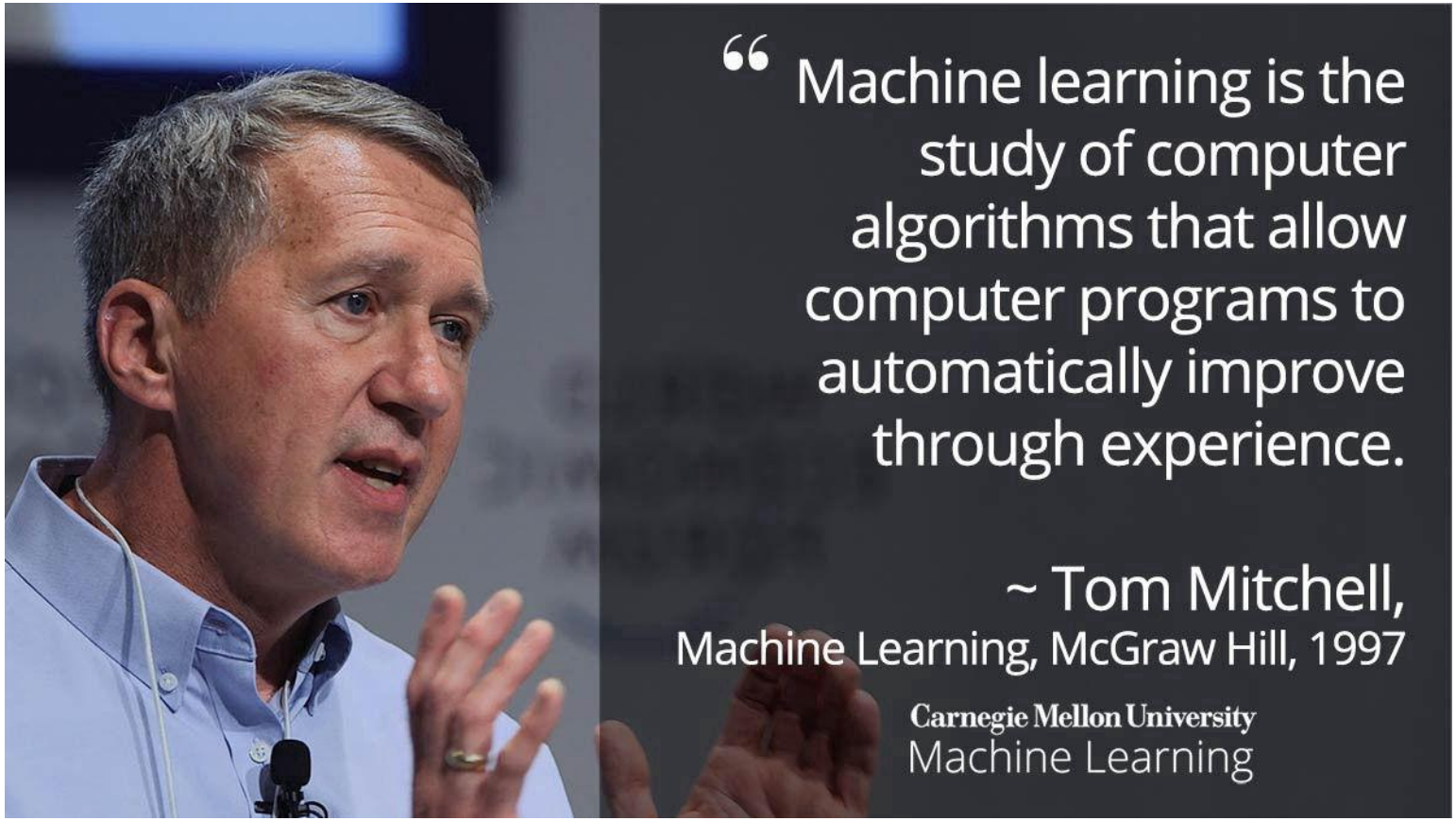
Syllabus

Unit-I

Why Machine learning, Examples of Machine Learning Problems, Structure of Learning, Learning versus Designing, Training versus Testing, Characteristics of Machine learning tasks, Predictive and descriptive tasks, Machine learning Models: Geometric Models, Logical Models, Probabilistic Models. Features: Feature types, Feature Construction and Transformation, Feature Selection.

What is Machine Learning?



In, 1959 Arthur Samuel defined machine learning as a

“**Field of study that gives computers the ability to learn without being explicitly programmed**”

The Traditional Programming Paradigm

Consider Activity Detection

```
if(speed<4){
    status=WALKING;
}
```

```
if(speed<4){
    status=WALKING;
} else {
    status=RUNNING;
}
```

```
if(speed<4){
    status=WALKING;
} else if(speed<12){
    status=RUNNING;
} else {
    status=BIKING;
}
```

```
// ???
```


Answers
Data
Machine Learning
Rules

The Machine Learning Paradigm

```
01010010101001010101
10010101010010111101
01001010100101011001
01010010101001010101
```

Label = WALKING

```
10101001010010101010
01010100100100010001
00100111110101011111
10101001001111010111
```

Label = RUNNING

```
10010100111110101010
11010101110101011110
10101011110101011011
11111100011110101010
```

Label = BIKING

```
11111111110100111010
00111111010111110101
01011101010101011110
10101010101001111110
```

Label = GOLFING

The Machine Learning Paradigm

The Machine Learning Paradigm

Matching X to Y

X = { -1, 0, 1, 2, 3, 4 }

Y = { -3, -1, 1, 3, 5, 7 }

Make a guess!

Y = 3X - 1

X = { -1, 0, 1, 2, 3, 4 }

Y = { -4, -1, 2, 5, 8, 11 }

How good is the guess?

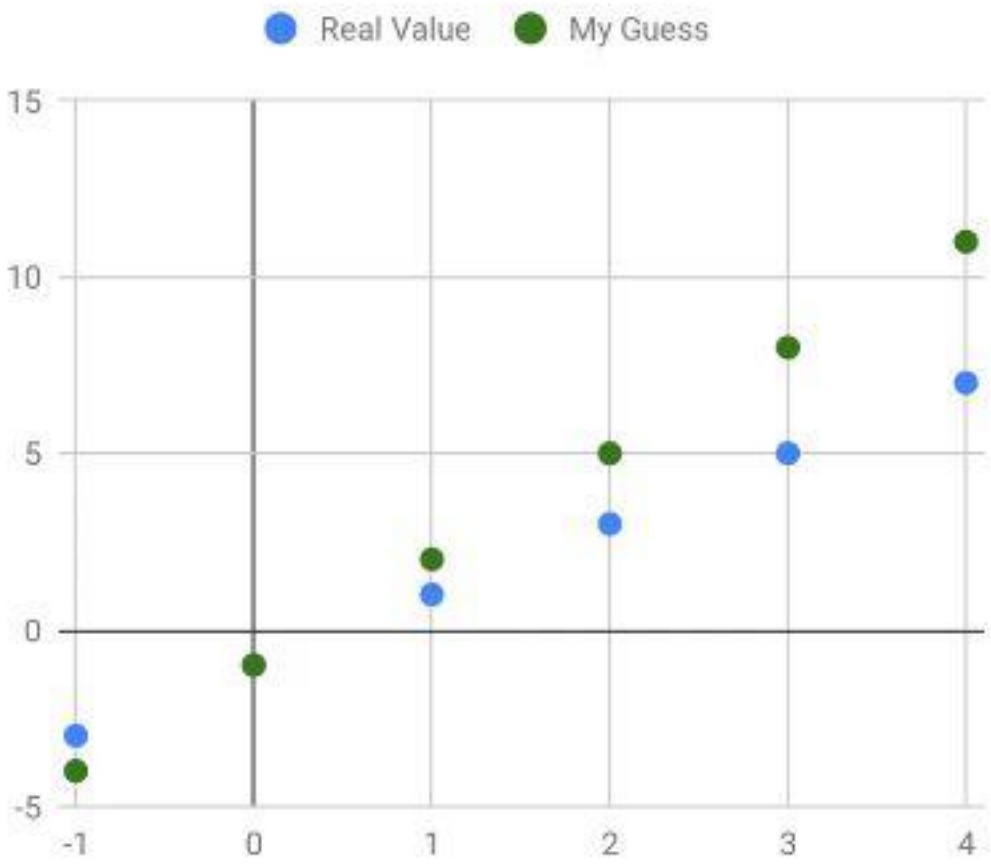
Y = 3X - 1

X = { -1, 0, 1, 2, 3, 4 }

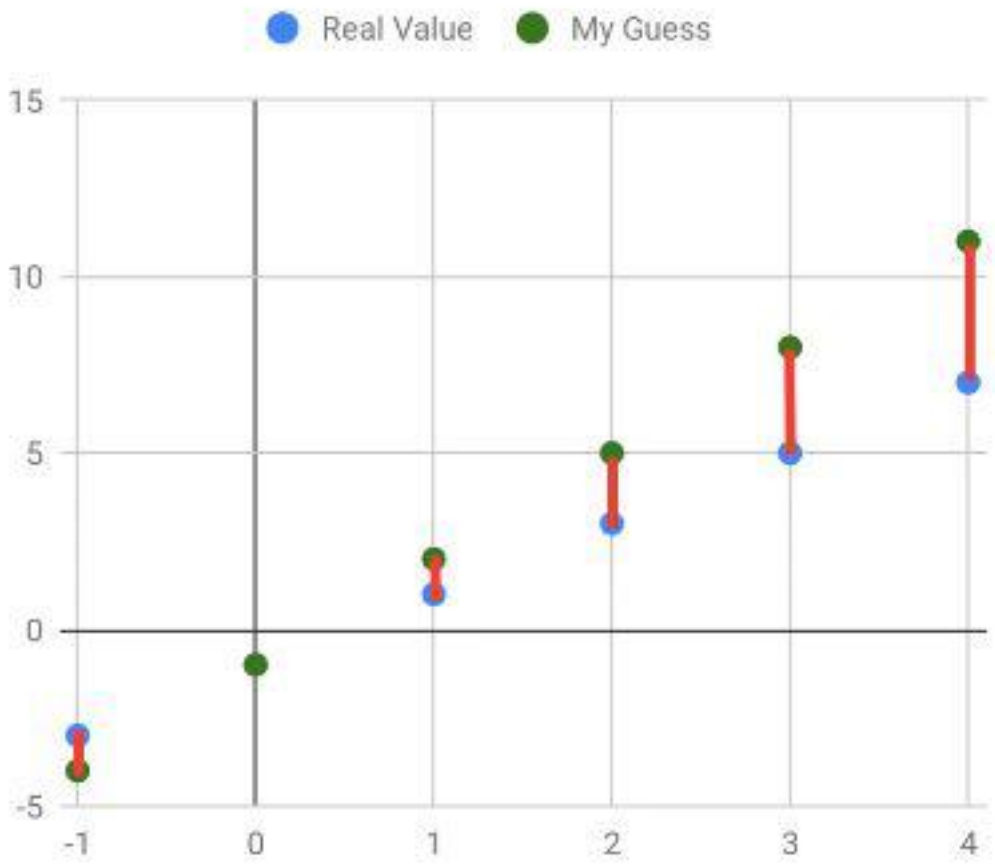
My Y = { -4, -1, 2, 5, 8, 11 }

Real Y = { -3, -1, 1, 3, 5, 7 }

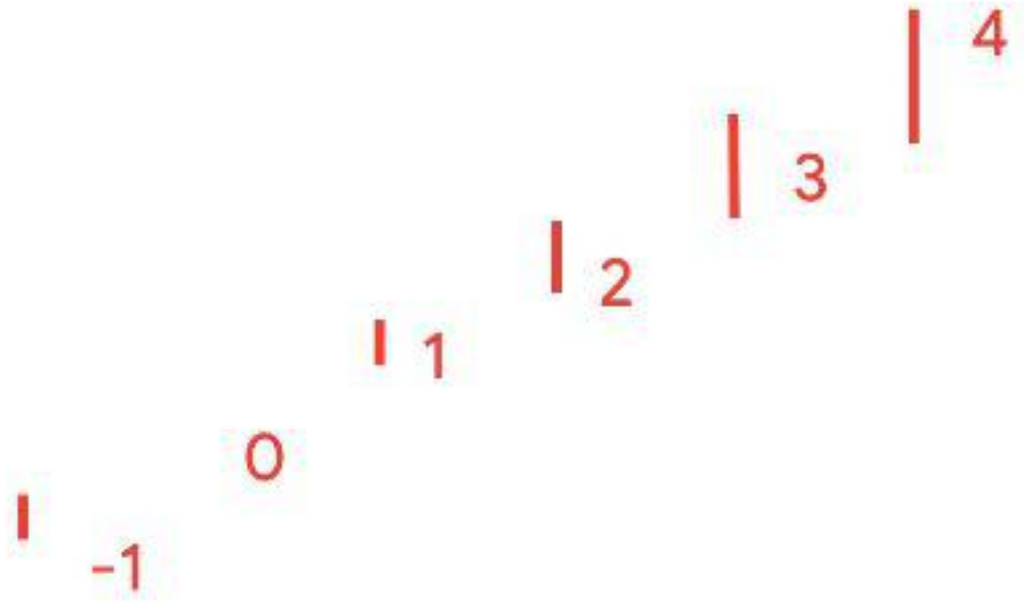
Let's measure it!



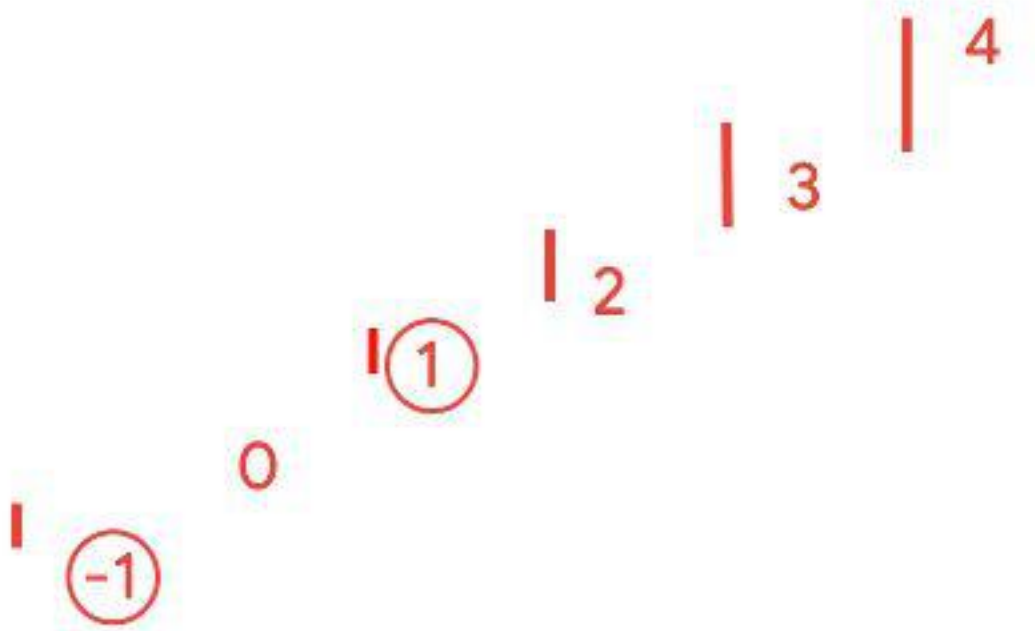
Let's measure it!



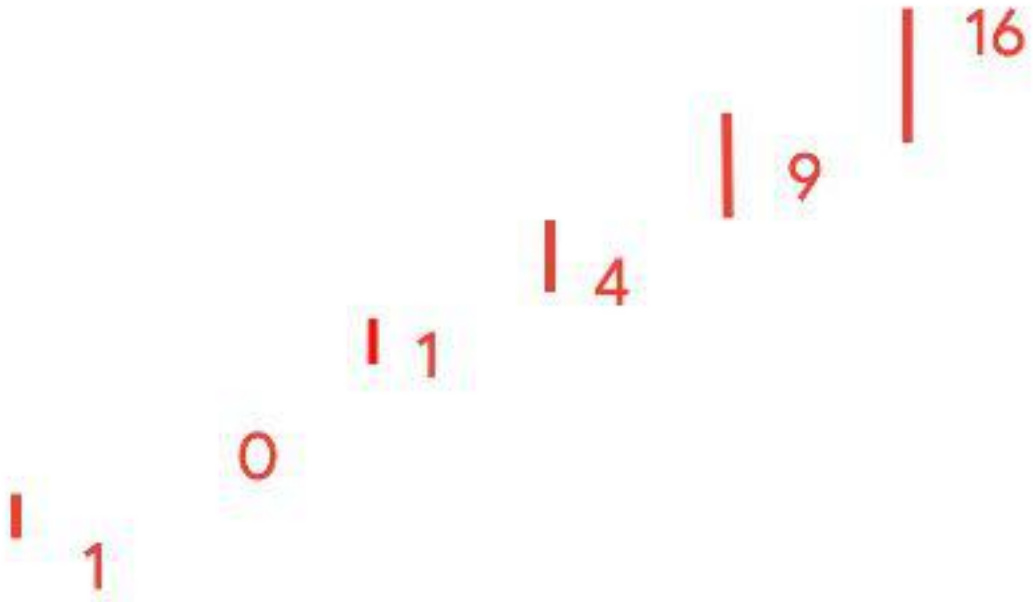
Let's measure it!



Houston, we have a problem!



What if we **square**^{2} them?



Total that (Σ) and take the square root √

sqrt(1 + 1 + 4 + 9 + 16)

= sqrt(31)
= 5.57

Make another guess!

Y = 2X - 2

X = { -1, 0, 1, 2, 3, 4 }

My Y = { -4, -2, 0, 2, 4, 6 }

Real Y = {-3, -1, 1, 3, 5, 7}

$Diff^2$ = {1, 1, 1, 1, 1}

Get the same difference, repeat the same process.

sqrt(1 + 1 + 1 + 1 + 1)

= sqrt(5)
= 2.23

Make another guess!

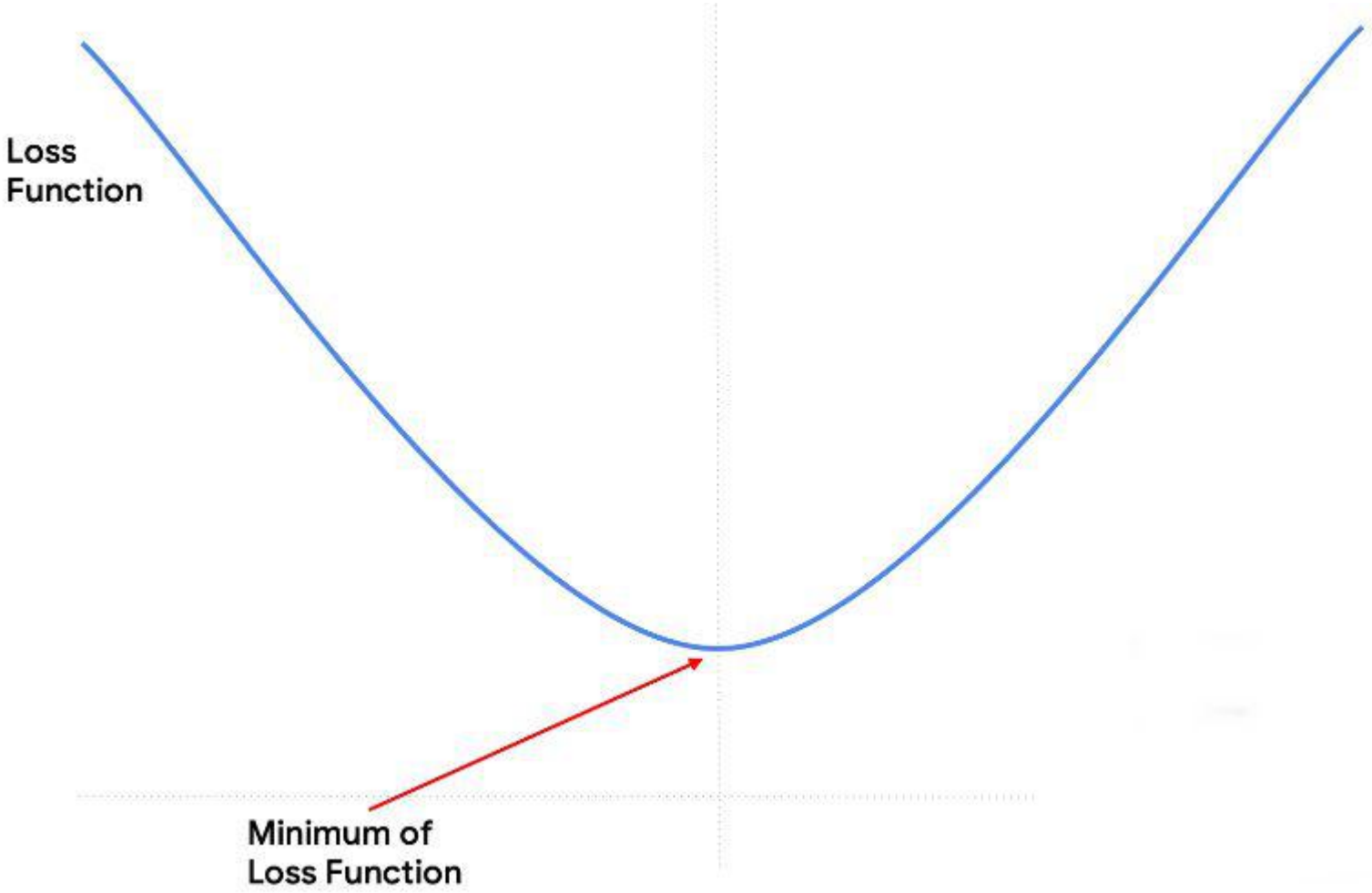
Y = 2X - 1

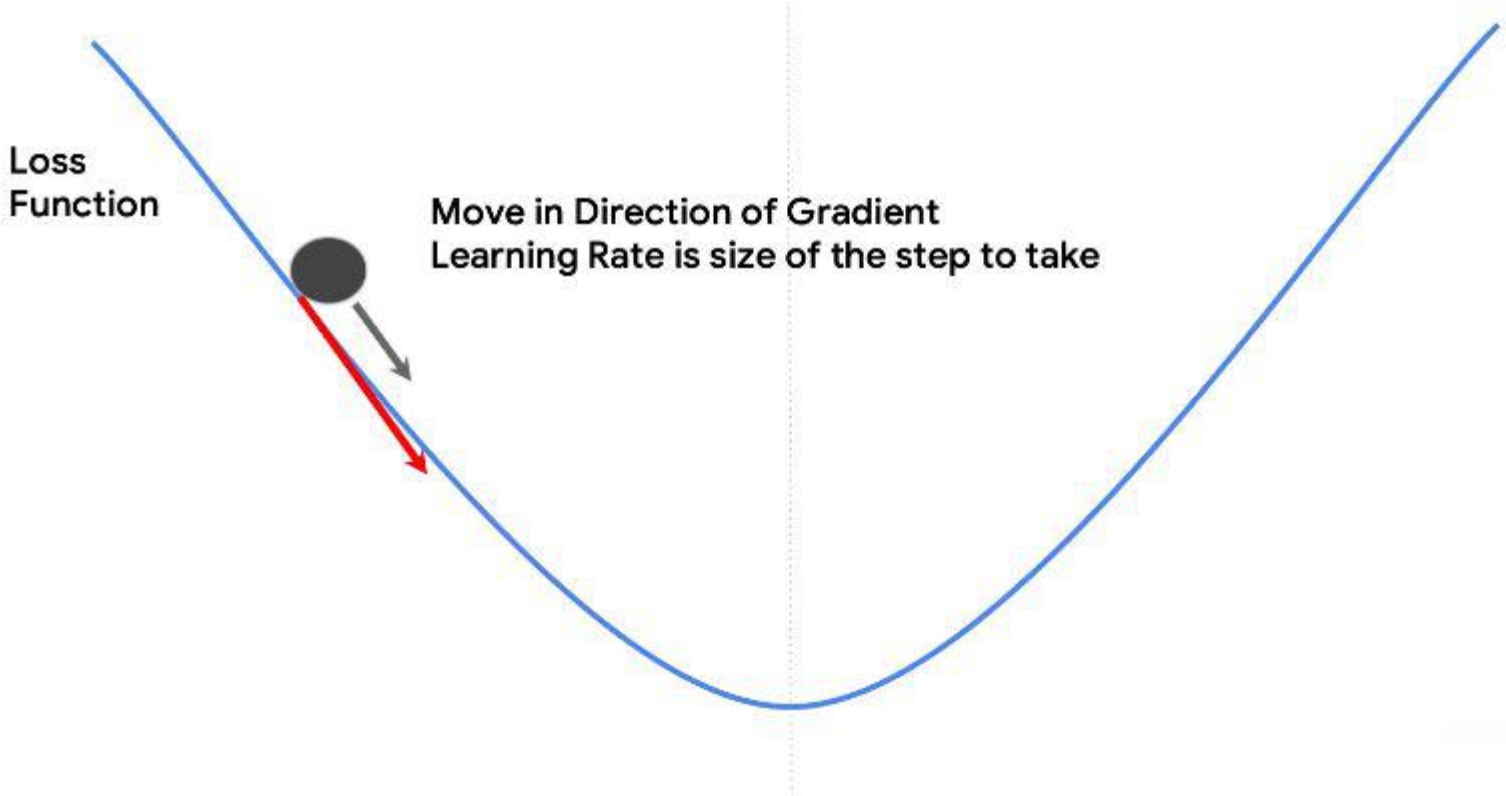
X = { -1, 0, 1, 2, 3, 4 }

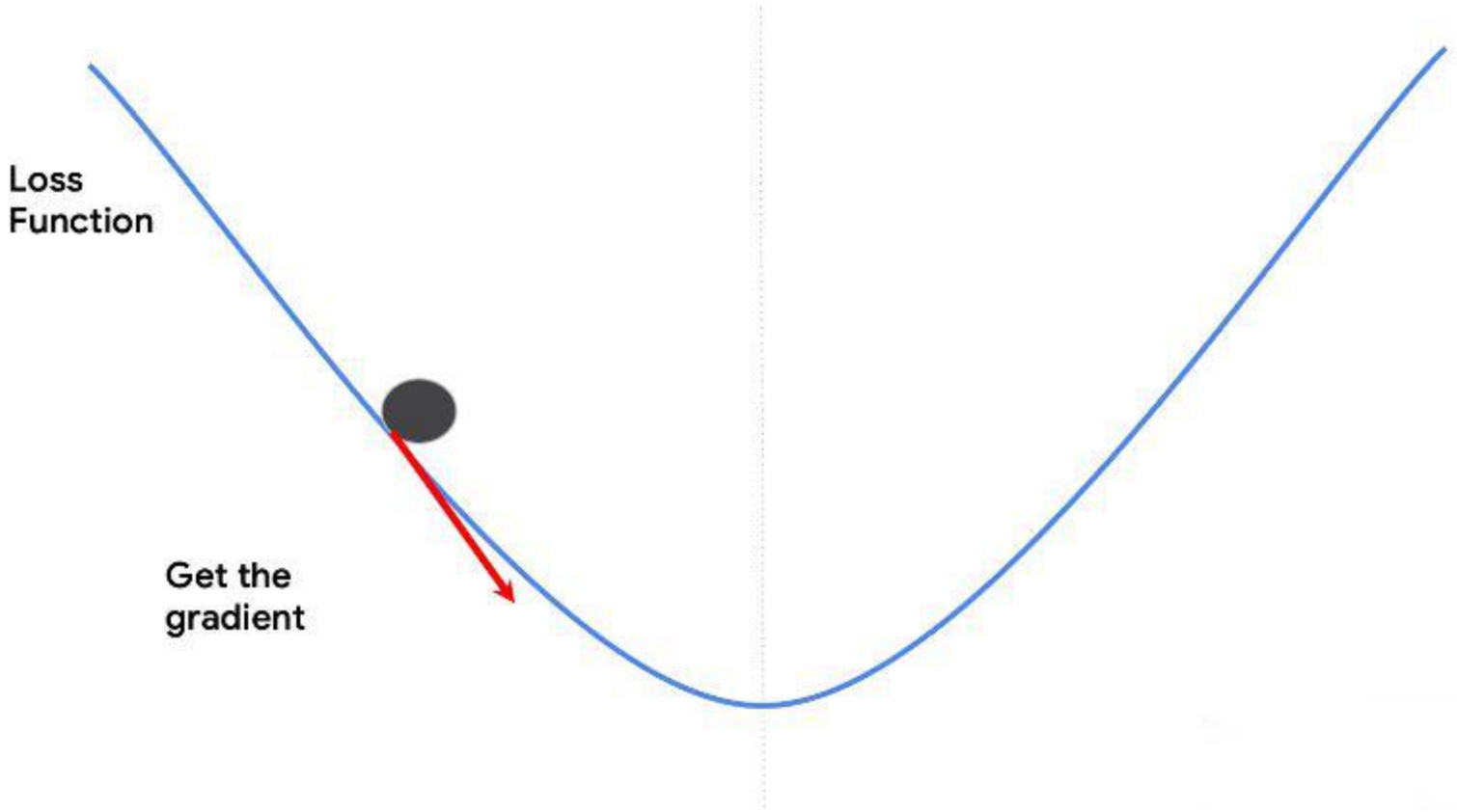
My Y = { -3, -1, 1, 3, 5, 7 }

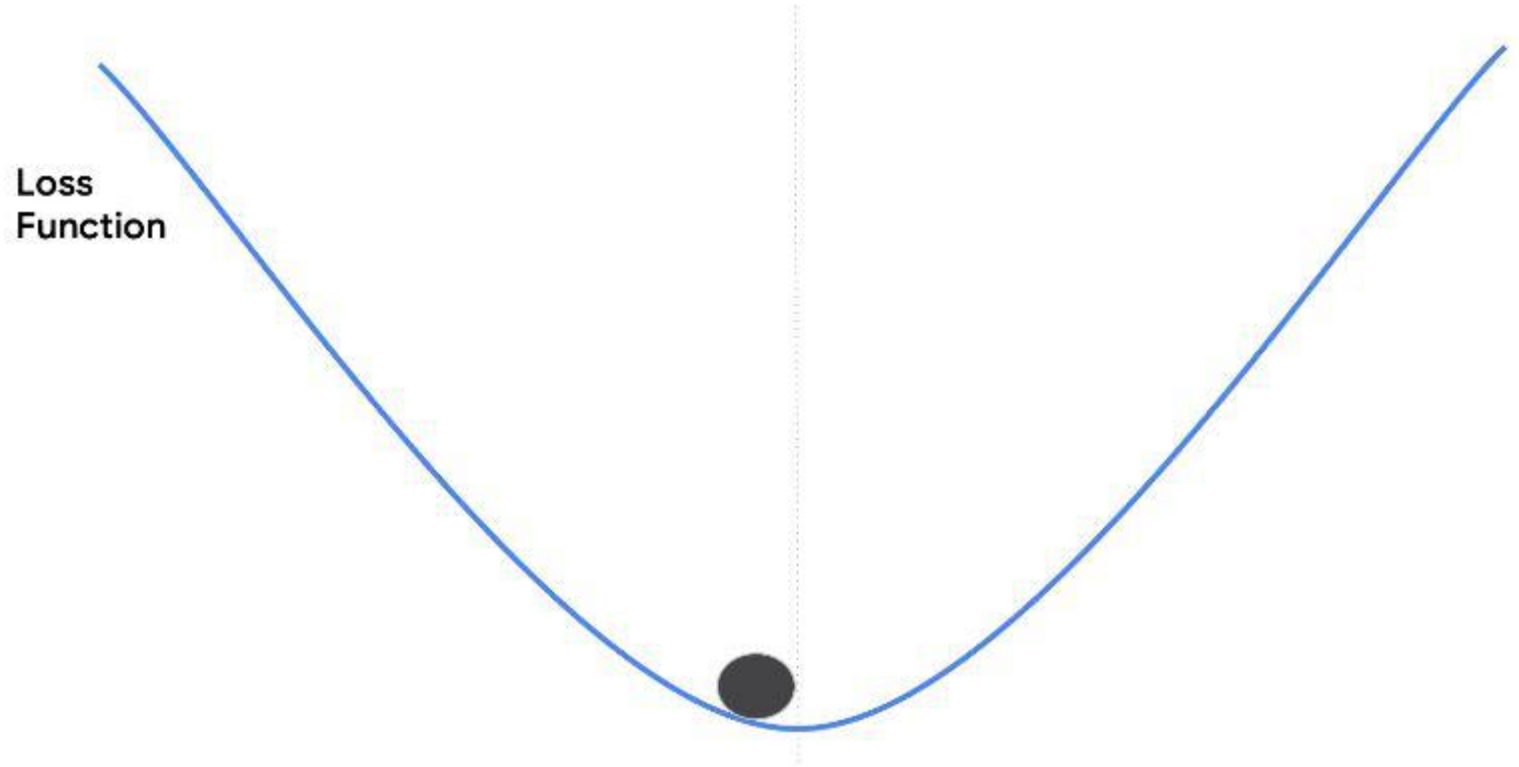
Real Y = {-3, -1, 1, 3, 5, 7}

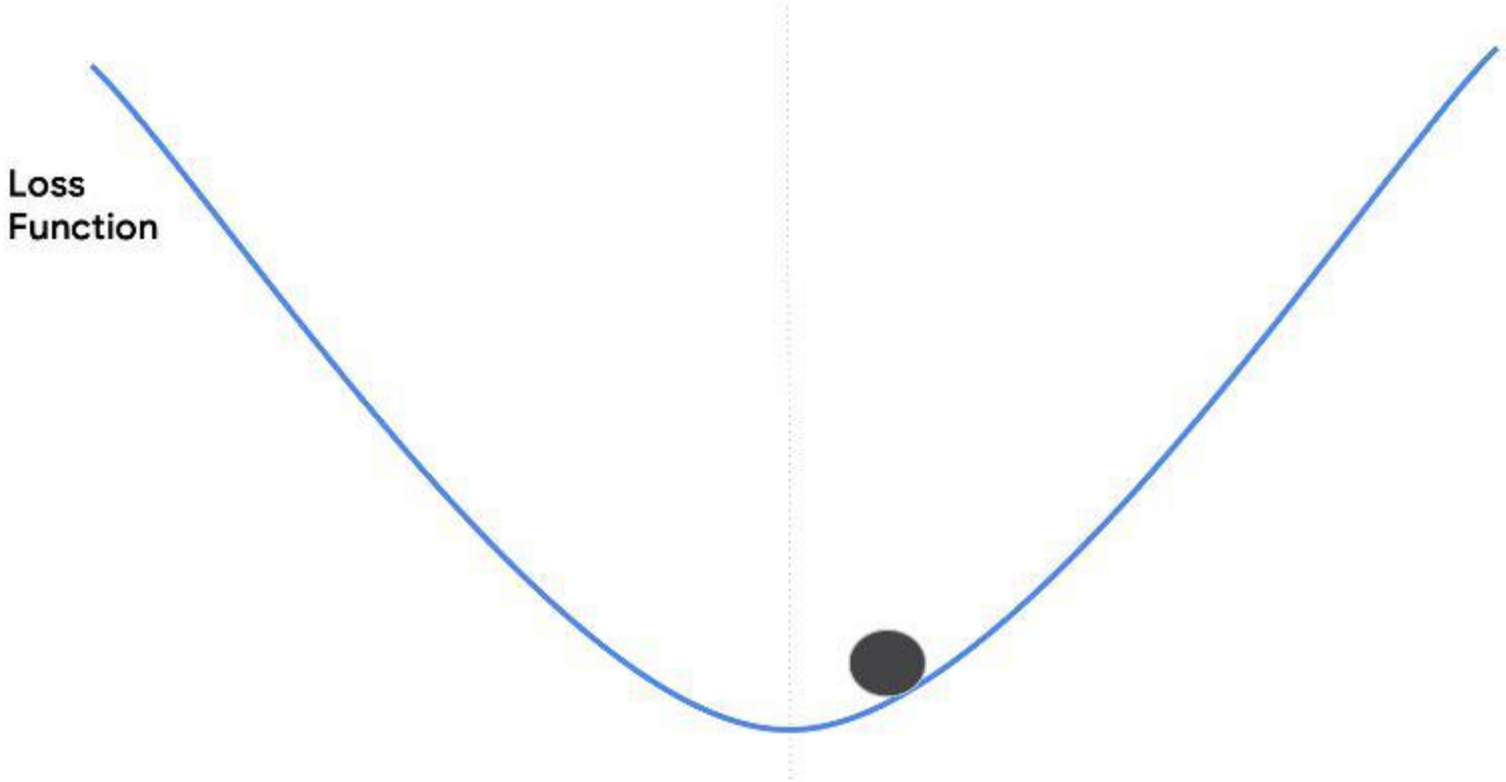
$Diff^2$ = {0, 0, 0, 0, 0}

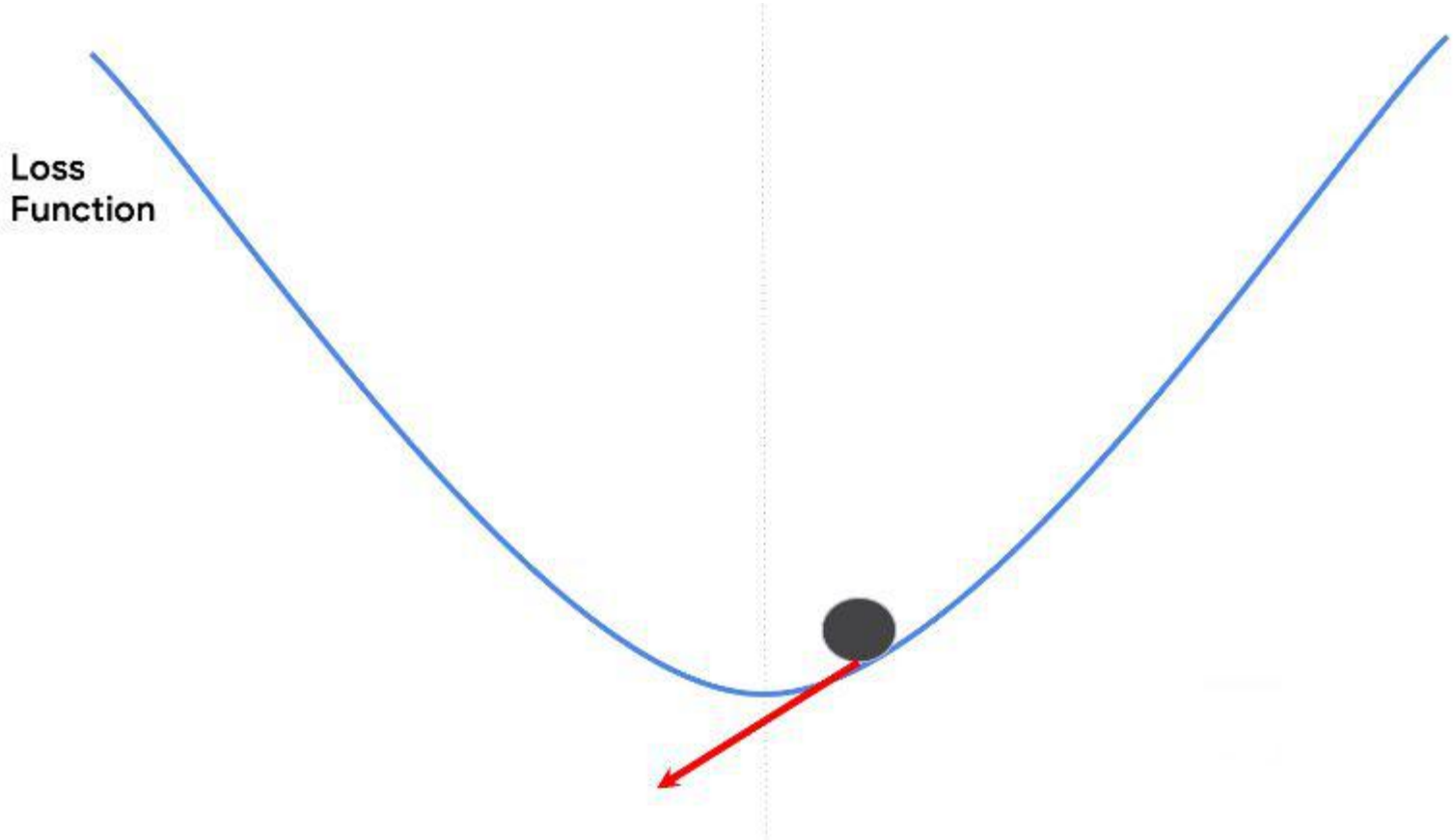

Loss
Function
Minimum of
Loss Function

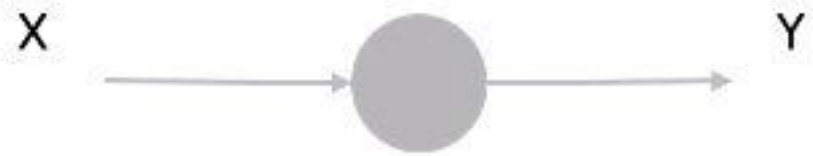

Loss
Function
Move in Direction of Gradient
Learning Rate is size of the step to take


Loss
Function
Get the
gradient

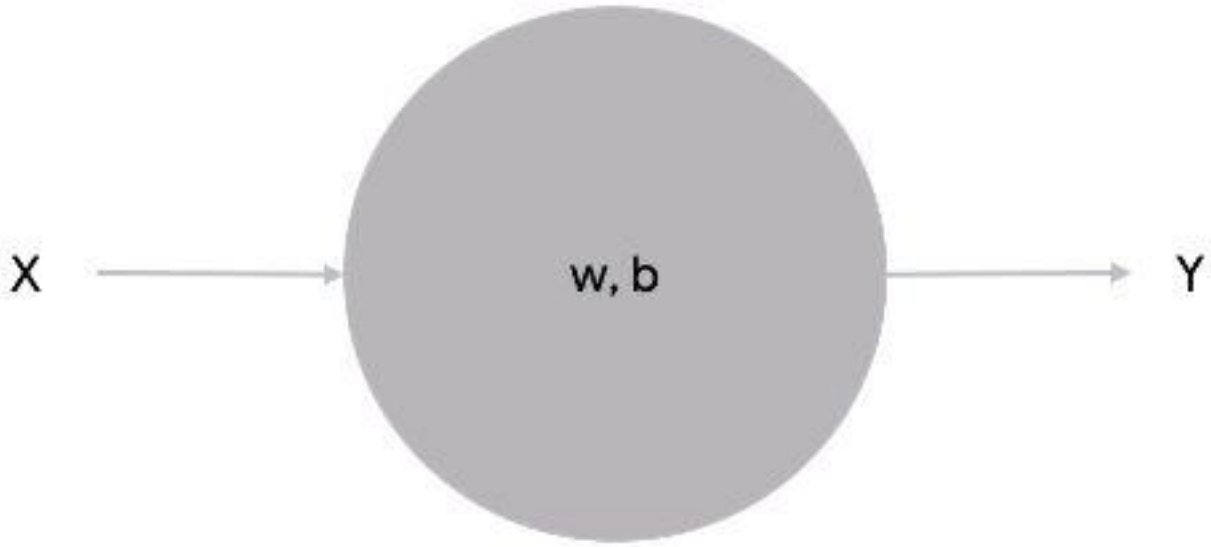

Loss
Function


Loss
Function

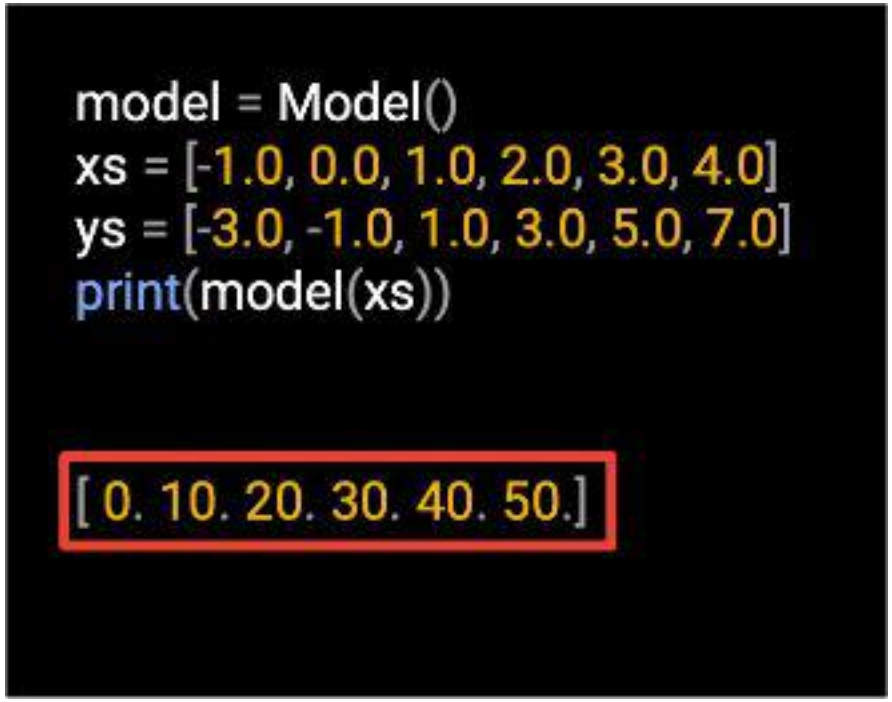

Loss
Function



$$y = f(x) = wx+b$$




```
class Model(object):
  def __init__(self):
    self.w = tf.Variable(10.0)
    self.b = tf.Variable(10.0)

  def __call__(self, x):
    return self.w * x + self.b
```



y = wx+b for w, b =10, 10
y1 = 10 *-1.0 +10 = 0
y2 = 10*0+1- = 10 and so on…

This is very different from ys

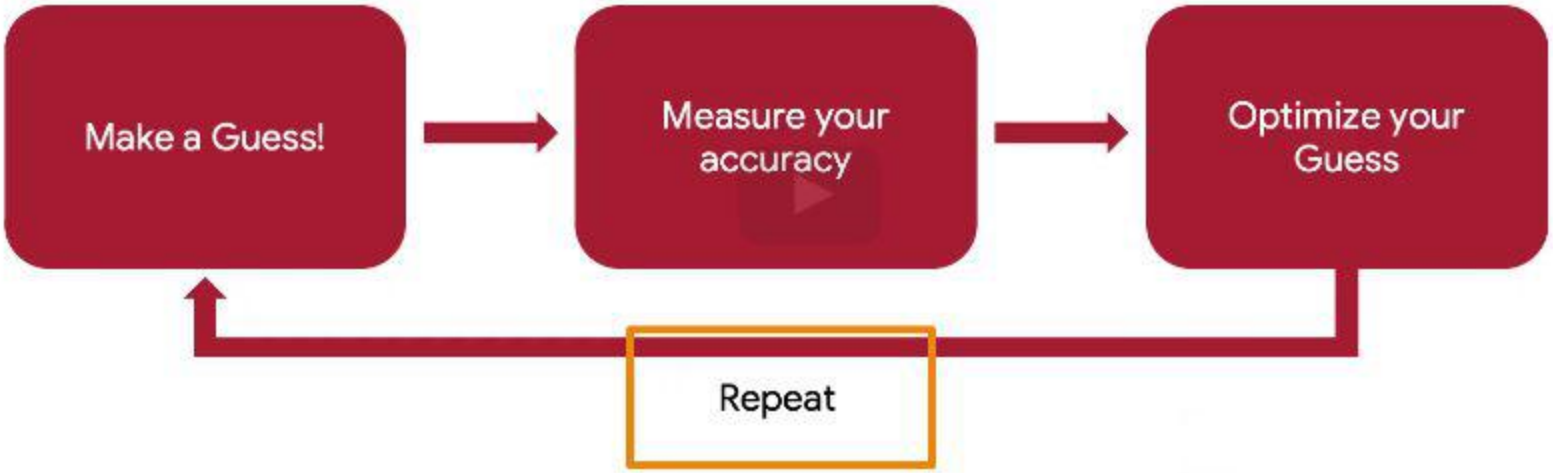
The Machine Learning Paradigm


```
def loss(predicted_y, target_y):
  return tf.reduce_mean(tf.square(predicted_y - target_y))
```

Underlying Training Code

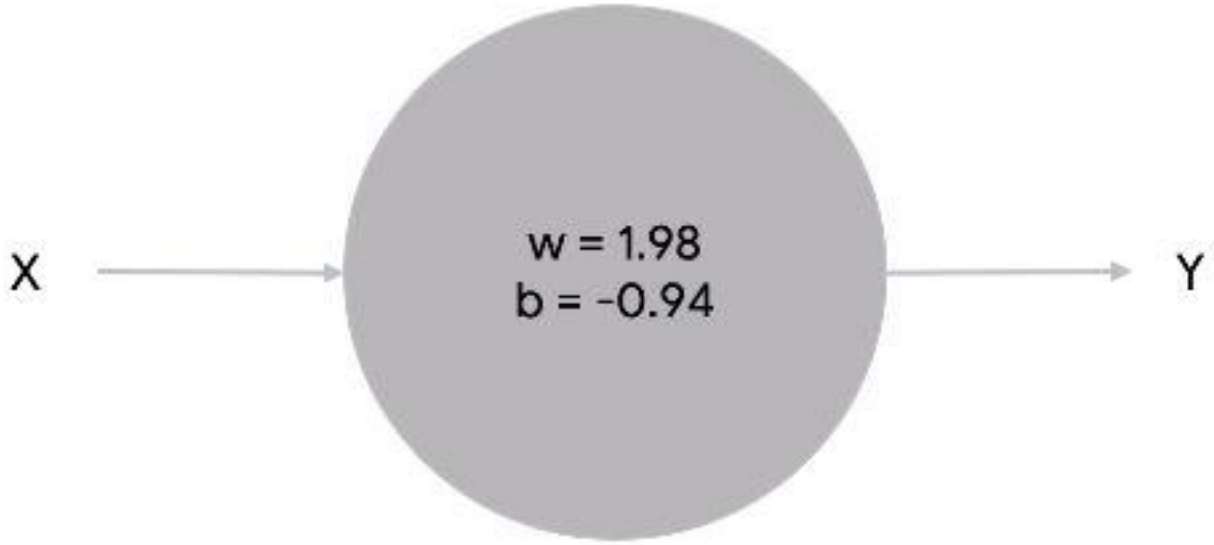
```
def train(model, xs, ys, learning_rate):
  with tf.GradientTape() as t:
    current_loss = loss(model(xs), ys)

  dw, db = t.gradient(current_loss, [model.w, model.b])
  model.w.assign_sub(learning_rate * dw)
  model.b.assign_sub(learning_rate * db)
  return current_loss
```

The Machine Learning Paradigm




```
for epoch in range(50):
  current_loss = train(model, xs, ys, learning_rate=0.1)
```

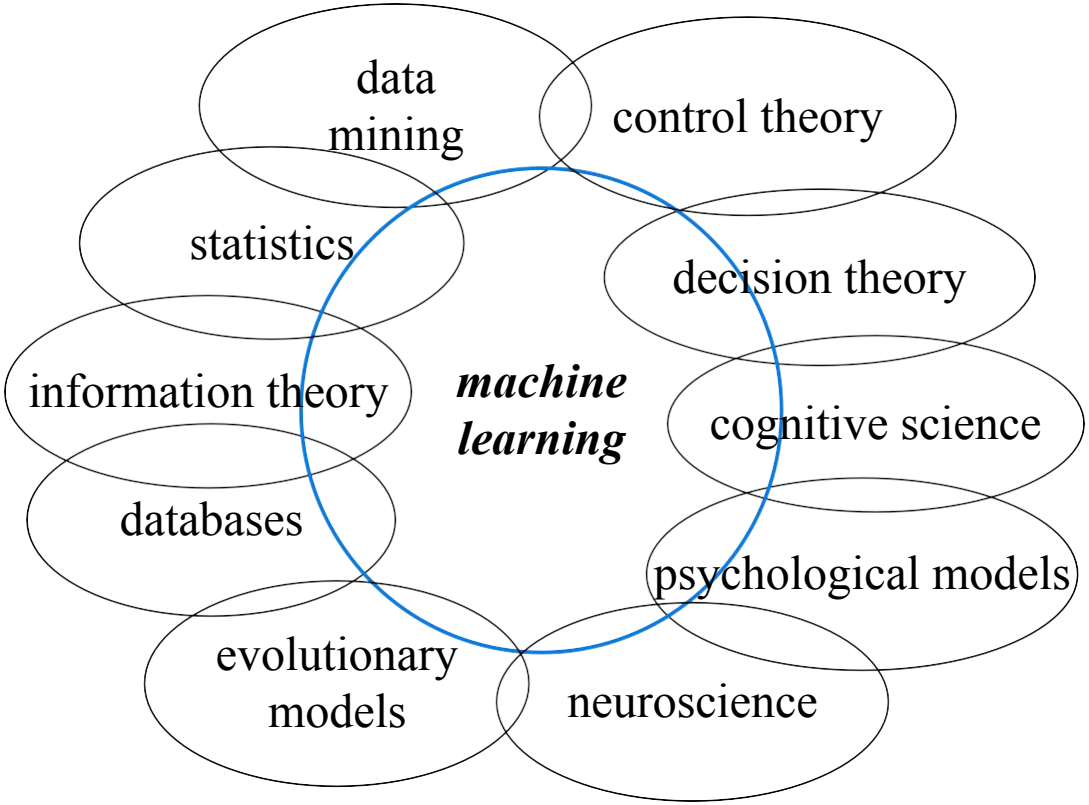
Finally we get:



Why Machine Learning?

- No human experts
  - industrial/manufacturing control
  - mass spectrometer analysis, drug design, astronomic discovery
- Black-box human expertise
  - face/handwriting/speech recognition
  - driving a car, flying a plane
- Rapidly changing phenomena
  - credit scoring, financial modeling
  - diagnosis, fraud detection
- Need for customization/personalization
  - personalized news reader
  - movie/book recommendation

Related Fields

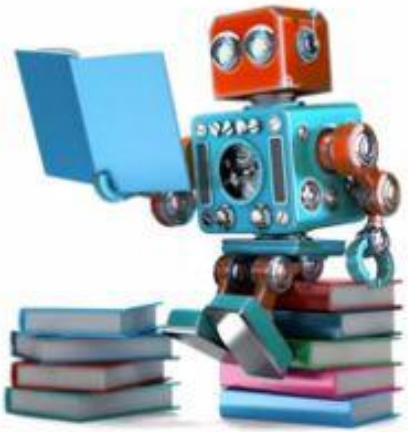


- Machine learning is primarily concerned with the accuracy and effectiveness of the er system

What is Machine Learning?

- Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed
- It focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data

Machine learning uses the data to detect patterns in a dataset, create a model and adjust program actions accordingly



Machine Learning defines Data Science

- While in data analysis you just can report the finding and perhaps visualize the findings
- However Data Science take it further with machine learning. You build a predictive model so that when we have new data with unknown outcome, we can predict the likely outcome and automate the downstream process

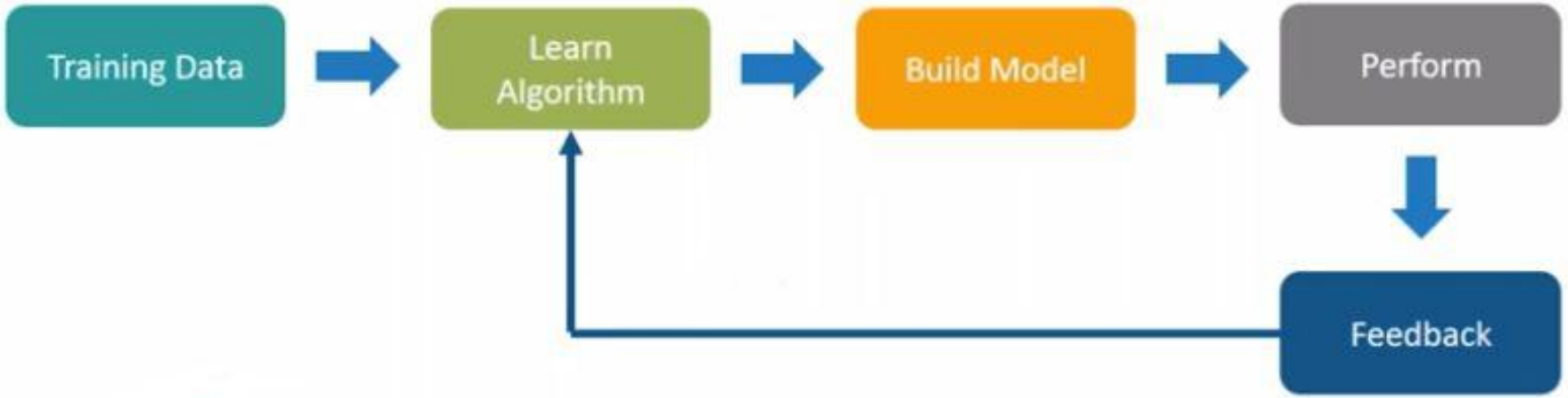
Data analysis :

You analyze credit card fraud in the historical data and report your findings to senior management

Data science :

You can also build a predictive model from the data you analyzed with machine learning, which flags likely fraudulent transactions as they come in and route them to call center for follow-up

Machine Learning Process Flow



Using the data, the system learns an algorithm, and then uses it to build a predictive model. The system then performs the recommended task and uses feedback data to tune the model to be more accurate

Learning Versus Design

- Machine learning is a powerful tool that drives everything from curated content recommendations to optimized user interfaces
- Machine learning answers questions about user behavior
- Machine learning customizes interfaces to users needs
- Digital product designers need to get familiar with machine learning
- Many warn that designers who don't start learning about ML will be left behind. But I haven't seen one that has explored what design and machine learning have to offer each other
- Design and machine learning function like a flywheel: when connected, each provides value to the other. Together, they open up new product experiences and business value
- Design helps machine learning gather better data

Learning Versus Design

- Machine learning is a hungry beast. To deliver the best results, learning algorithms need vast amounts of detailed data, clean of any confounding factors or built-in biases
- Designers can help create user experiences that eliminate noise in data, leading to more accurate and efficient ML-powered applications
- Design helps set expectations and establish trust with users

Error and Noise

Error

Error measures are a tool in ML that quantify the question "how wrong was our estimation". It is a function that compares the output of a learned hypothesis with the output of the real target function. What this means in practice is that we compare the prediction of our model with the real value in data. An error measure is expressed as E(h, f) (a hypothesis $h \in H$, and f is the target function). E is almost always pointwise. It is defined by the difference at two points, therefore, we use the pointwise definition of the error measure e() to compute this error in the different points: e(h(x), f(x)).

Examples:

Squared error: $e(h(x), f(x)) = (h(x)- f(x))^2$

Binary error: $e(h(x), f(x)) = [\![h(x) \neq f(x)]\!]$ (the number of wrong classifications)

Error and Noise

Noise

It refers to the irrelevant information or randomness in a dataset. We can express noisy target as follows:

Noisy target= deterministic target + noise = $\mathbb{E}[y|x] + \varepsilon$ where $\varepsilon = (y - f(x))$ is the difference between the outcome and the predicted value.

$\mathbb{E}[y|x]$ is the expected value of y knowing x, y is our prediction using the target function h(x) and f(x) is the real value of the data point.

We introduced P(y|x) into our learning scheme to account for the fact that there will always be noise in the relationship between x and y while P(x) represents the random variable x and is necessary for us to use Hoeffding‘s inequality.

Training versus Testing

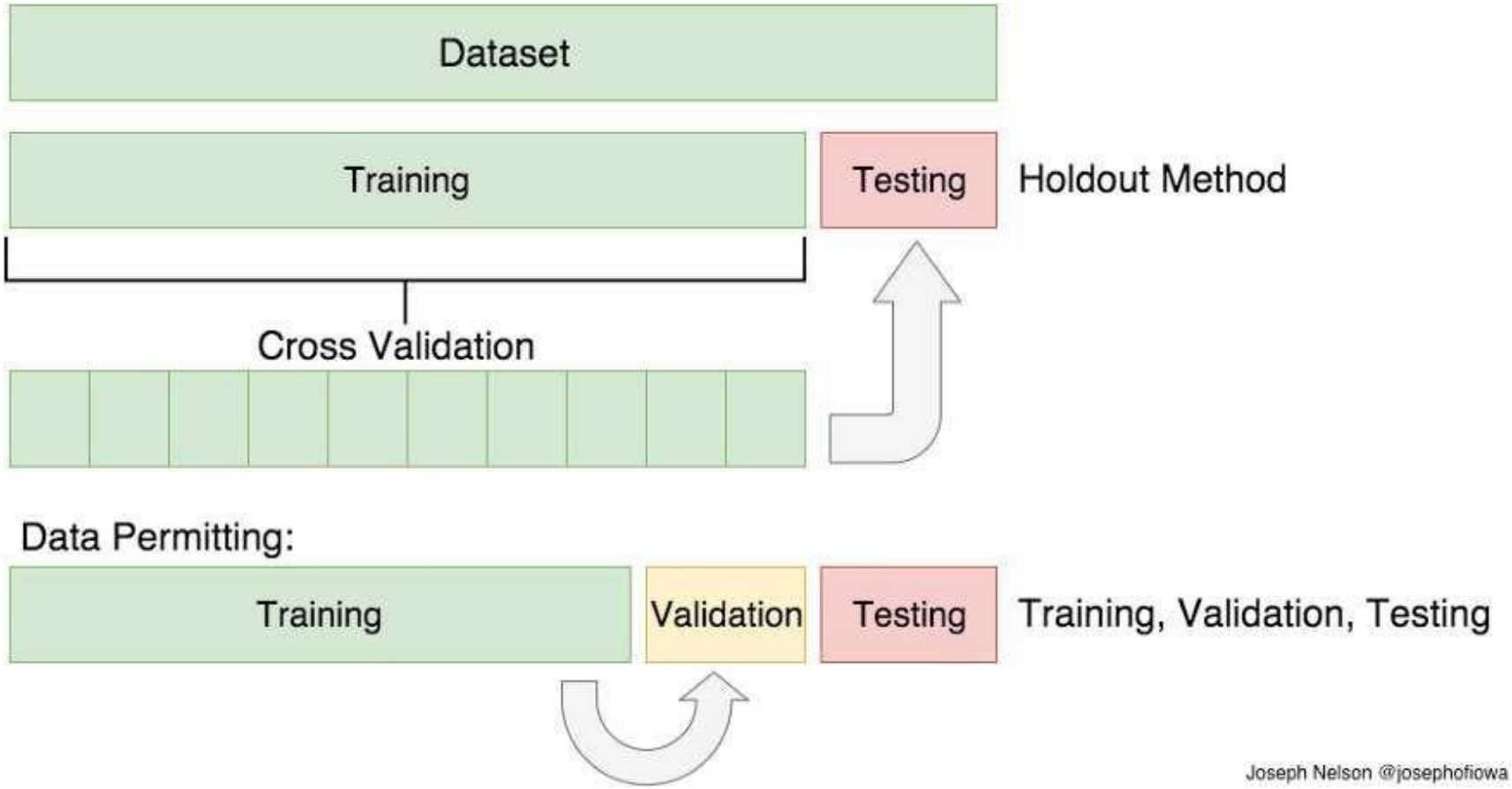
- In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set
- In Machine Learning, we basically try to create a model to predict the test data
- Usually, a dataset is divided into a training set, a validation set (some people use ‘test set’ instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration

Training versus Testing

Sets:

- **Training Set:** Here, you have the complete training dataset. You can extract features and train to fit a model and so on
- **Validation Set:** This is crucial to choose the right parameters for your estimator. We can divide the training set into a train set and validation set. Based on the validation test results, the model can be trained(for instance, changing parameters, classifiers)
- **Testing Set:** Here, once the model is obtained, you can predict using the model obtained on the training set

Training versus Testing



ML Use Case – Siri

- Siri is an intelligent personal assistant
- Apple claims that the software adapts to the user's individual preferences over time and personalizes results
  - It figures out which apps to use for which requests
  - It plays the songs you want to hear
  - It gives you directions
  - It wakes you up

ML Use Case - Financial services

- Financial industry use machine learning technology for two key purposes: to identify important insights in data, and prevent fraud
  - The insights can identify investment opportunities, or help investors know when to trade
  - Data mining can also identify clients with high-risk profiles, or use cybersurveillance to pinpoint warning signs of fraud

ML Use Case - Health care

- Machine learning is a fast-growing trend in the health care industry, thanks to the advent of wearable devices and sensors that can use data to assess a patient's health in real time

  - The technology can also help medical experts analyze large amount of data to identify trends or red flags that may lead to improved diagnoses and treatment

ML Use Case - Retail

- Websites recommending items you might like based on previous purchases are using machine learning to analyze your buying history – and promote other items you'd be interested in
  - This ability to capture data, analyze it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail

Secure | https://data.gov.in

YouTube | (1) Facebook | CRCV | Center for Res | 8 simple chemistry ex | KryssTal : Acids, Base | Drinking Water Clean | Robotics for Children | Untitled form - Goog | LEGO WeDO 2.0 Mor

GOVERNMENT OF INDIA | A Digital India Initiative

Skip to navigation | Skip to main content | DataGov States/ULB | A- A A+ A

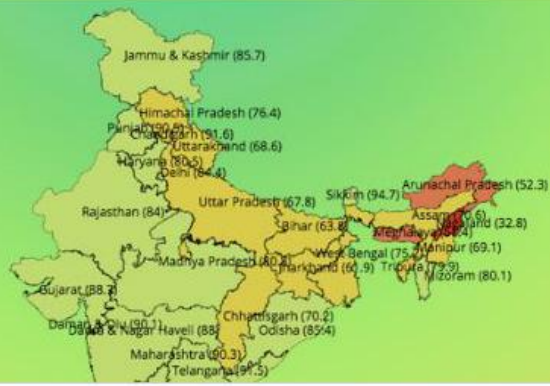
data.gov.in
Open Government Data (OGD) Platform India

Type search keyword

LOG IN | REGISTER

VISUALIZATION OF THE DAY
27 NOV 2017

STATE/UT-WISE INSTITUTIONAL BIRTHS
DURING 2015-16 (NFHS-4)



ANALYTICS

- 140,553 RESOURCES
- 4,238 CATALOGS
- 107 DEPARTMENTS
- 14.06 M TIMES VIEWED
- 5.28 M TIMES DOWNLOADED

CATALOG

HIGH VALUE DATASETS

TRANSPORT TIMETABLES	GOVERNMENT BUDGET	COMPANY REGISTER
NATIONAL STATISTICS	LEGISLATION	ENVIRONMENT

www.who.int/gho/database/en/

YouTube (1) Facebook CRCV | Center for Re 8 simple chemistry e KryssTal : Acids, Base Drinking Water Clean Robotics for Children Untitled form - Goog LEGO WeD

Sign up for WHO updates

عربي 中文 English Français Русский Español

World Health Organization

Health topics | Data | Media centre | Publications | Countries | Programmes | Governance | About WHO | Search

Global Health Observatory (GHO) data

- Global Health Observatory data
- Data repository
- Reports
- Country statistics
- Map gallery
- Standards

The data repository

Browse the GHO data repository

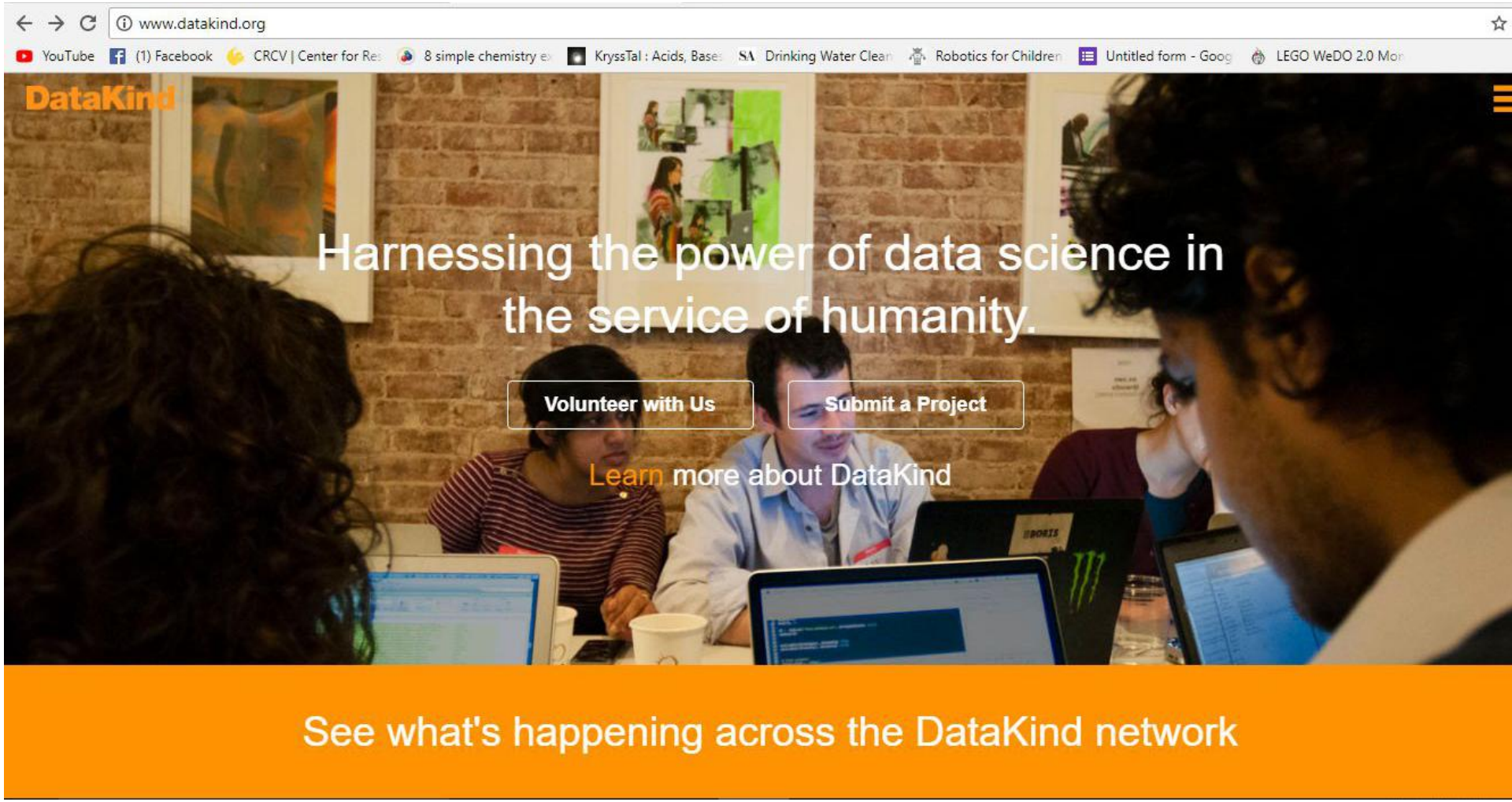
The GHO data repository contains an extensive list of indicators, which can be selected by theme or through a multi-dimension query functionality. It is the World Health Organization's main health statistics repository.

Contact us

Please send us your comment or question by e-mail.

Global Health Observatory (GHO) data > The data repository

lving host...


www.datakind.org
YouTube
(1) Facebook
CRCV | Center for Re
8 simple chemistry e
KryssTal : Acids, Base
Drinking Water Clean
Robotics for Children
Untitled form - Goog
LEGO WeDO 2.0 Mor
DataKind
Harnessing the power of data science in the service of humanity.
Volunteer with Us
Submit a Project
Learn more about DataKind
See what's happening across the DataKind network

archive.ics.uci.edu/ml/index.php

YouTube | (1) Facebook | CRCV | Center for Re | 8 simple chemistry e | KryssTal : Acids, Base | Drinking Water Clean | Robotics for Children | Untitled form - Goog | LEGO WeDO 2.0 Mor

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

Repository Web

Google

View ALL Data Sets

Welcome to the UC Irvine Machine Learning Repository!

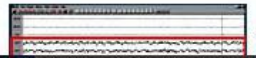
We currently maintain 398 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians. We have also set up a mirror site for the Repository.

Supported By: In Collaboration With: Rexa.info

Latest News:

04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010: Note from donor regarding Netflix data
10-16-2009: Two new data sets have been added.
09-14-2009: Several data sets have been added.
07-23-2008: Repository mirror has been set up.
03-24-2008: New data sets have been added!
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: EEG Database



Data Type: Multivariate, Time-Series
 # Attributes: 4
 # Instances: 122

Newest Data Sets:

11-21-2017: Daily Demand Forecasting Orders
11-17-2017: Z-Alizadeh Sani
11-17-2017: extention of Z-Alizadeh sani dataset
10-23-2017: Paper Reviews
08-28-2017: Burst Header Packet (BHP) flooding attack on Optical Burst Switching (OBS) Network

Most Popular Data Sets (hits since 2007):

1573331: Iris
1025004: Adult
780222: Wine
670436: Car Evaluation
601062: Breast Cancer Wisconsin (Diagnostic)

Its not a new thing

- Analyzing data has a long history!
- There have been many terms that have been used to describe such endeavors:
  - Statistics
  - Artificial Intelligence
  - Business Intelligence
  - Data analytics

Concept behind data science is nothing new, just the terminology is new

Deliver to rashi
kanpur 208002

Shop by
Category

All

Hello, rashi
Your Orders

Try
Prime

Your
Lists

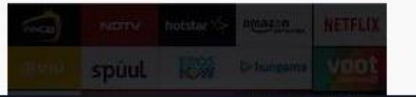
Cart

Disclaimer*

EC/09/2017/10364

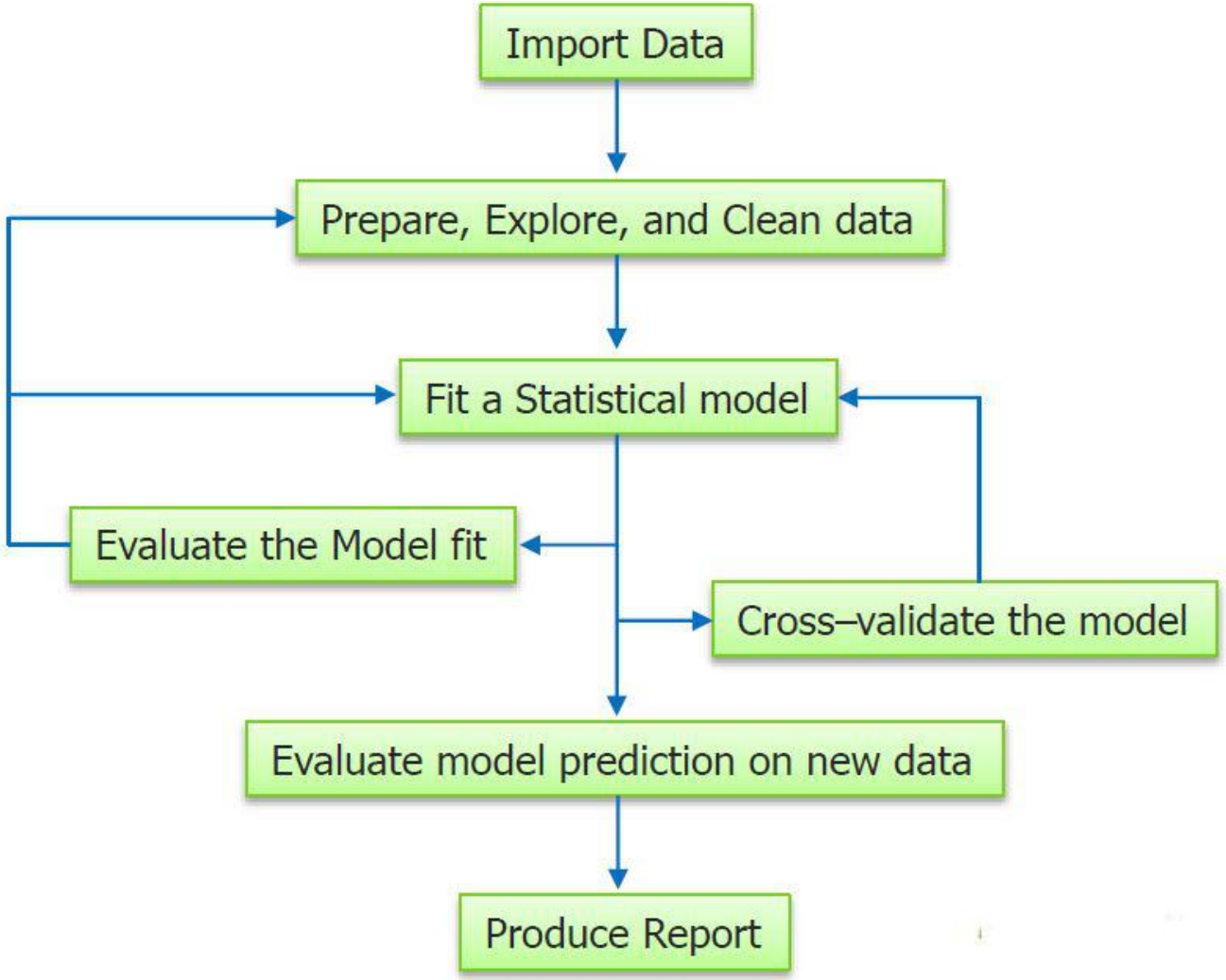
Ad feedback

Inspired by your shopping trends



https://www.amazon.in/dp/ref=PC-bb-Lenovo/B0774LT2GD?pf_rd_p=8b89a328-b8e3-4b34-9fc9-115e34b95fae&pf_rd_r=SG3J0THYS3JYZ1AWF6DG

Typical steps in Machine Learning



Data Acquisition/Data Collection – Use Case

- **Data Acquisition**
- Data Preparation
- Hypothesis & Modelling
- Evaluation & Interpretation
- Deployment
- Operation & Optimization

➢ Now to help John let's see what data we can collect from different locations and how it affects the pricing of an apartment.

Price	Apartment name/no.	No of bedrooms	Floor number	Criminal rate per year	Pollution level	Distance to nearby Educational institution
30L	xv	3	2	3	15	900 m
20L	cs	2	4	2		2 km
28L	df	2	G	5	13	1.5 km
25L	re	1	3	1	12	1.7 m
30L	sd	2	0	3	13	700 m

Data Acquisition/Data Collection

- **Data Acquisition**
- Data Preparation
- Hypothesis & Modelling
- Evaluation & Interpretation
- Deployment
- Operation & Optimization

- Data acquisition involves acquiring data from all the identified internal and external sources that can help answer the business question.
- This data could be,
  - logs from webservers
  - social media data
  - census datasets
  - data streamed from online sources via APIs

Data Preparation/Data Wrangling – Use Case

- Data Acquisition
- **Data Preparation**
- Hypothesis & Modelling
- Evaluation & Interpretation
- Deployment
- Operation & Optimization

- The data we have collected is not clean, there are some errors which need to cleansed.
- Also we may need to change the values of columns as per requirements.

Price	Apartment name/no.	No of bedrooms	Floor number	Criminal rate per year	Pollution level	Educational institution within 1km radius
30L	xv	3	2	3	15	Yes
20L	cs	2	4	2		No
28L	df	2	G	5	13	No
25L	re	1	3	1	12	No
30L	sd	2	0	3	13	Yes

Data Preparation/Data Wrangling

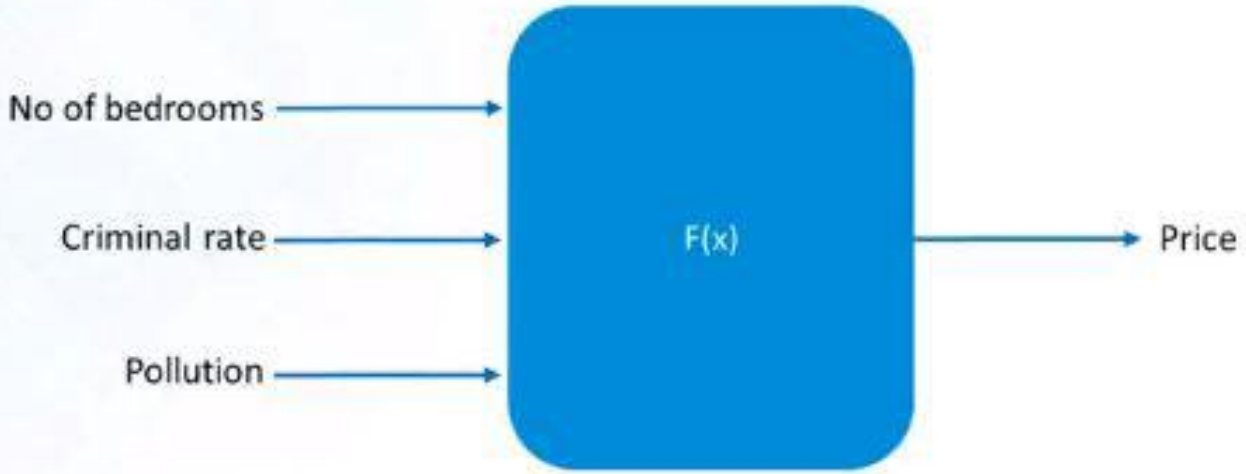
- Data Acquisition
- **Data Preparation**
- Hypothesis & Modelling
- Evaluation & Interpretation
- Deployment
- Operation & Optimization

- Data Wrangling is the process of cleaning and unifying messy and complex data sets.
- Data after reformatting can be converted to JSON, CSV or any other format that makes it easy to load into one of the data science tools.

Hypothesis and Modelling – Use Case

- Data Acquisition
- Data Preparation
- **Hypothesis & Modelling**
- Evaluation & Interpretation
- Deployment
- Operation & Optimization

➢ Based on the requirements, a model is created using the dataset.



Hypothesis and Modelling

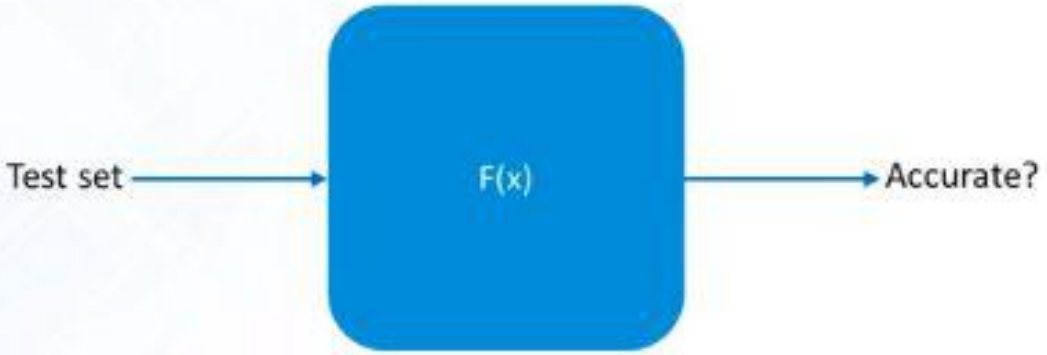
- Data Acquisition
- Data Preparation
- **Hypothesis & Modelling**
- Evaluation & Interpretation
- Deployment
- Operation & Optimization

- Involves forming and testing hypotheses about the data and the processes that generate it.
- Requires writing, running and refining the programs to analyze and derive meaningful business insights from data.
- Mostly written in languages like Python, R, Spark.

Evaluation and Interpretation – Use Case

- Data Acquisition
- Data Preparation
- Hypothesis & Modelling
- **Evaluation & Interpretation**
- Deployment
- Operation & Optimization

- This model is evaluated using test data set.

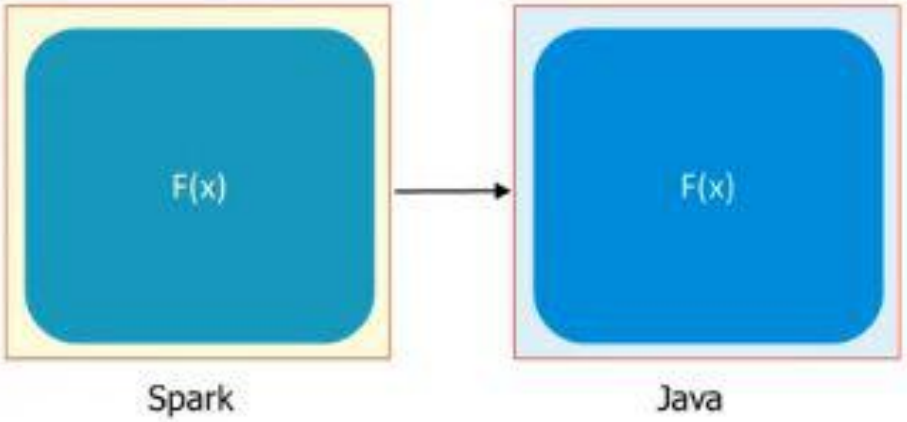


If accuracy is low, the above steps are repeated until a good model is found.

Deployment – Use Case

- Data Acquisition
- Data Preparation
- Hypothesis & Modelling
- Evaluation & Interpretation
- **Deployment**
- Operation & Optimization

- Data scientist might have done this in python or spark, but if the production environment supports only Java then he needs to recode it.



Deployment

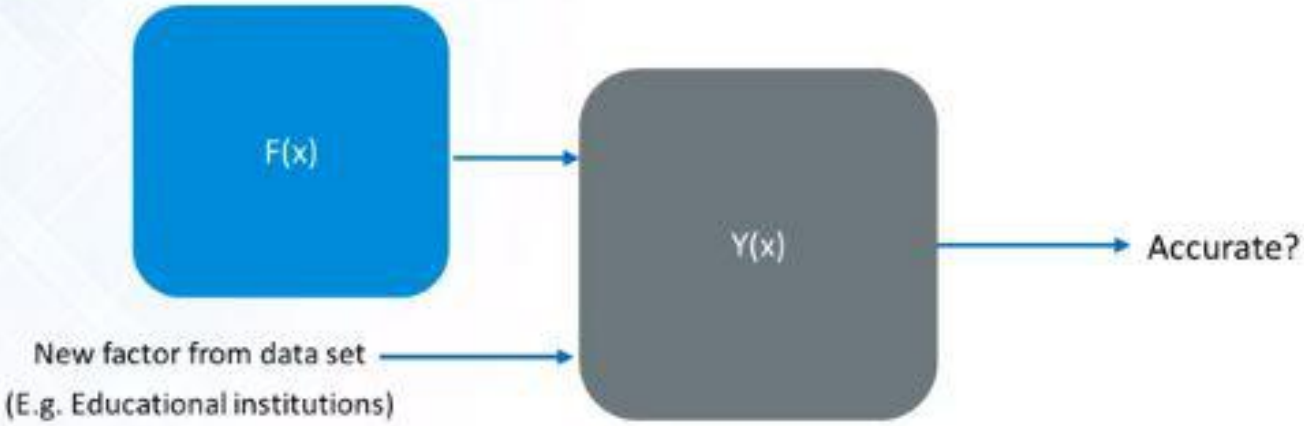
- Data Acquisition
- Data Preparation
- Hypothesis & Modelling
- Evaluation & Interpretation
- **Deployment**
- Operation & Optimization

- In this step the model we created is deployed in to the market.
- Models generally have to be recoded before deployment (e.g., data scientists may favor Python, but production environments may require Java)

Operation and Optimization – Use Case

- Data Acquisition
- Data Preparation
- Hypothesis & Modelling
- Evaluation & Interpretation
- Deployment
- **Operation & Optimization**

➢ Retraining the model using new factor from data set.



After the model is retrained we evaluate the model and deploy it.

Components of Learning

- Collecting and preparing data
- Choosing and training a model
- Evaluating a model
- Hyperparameter tuning and Prediction

Types of Learning

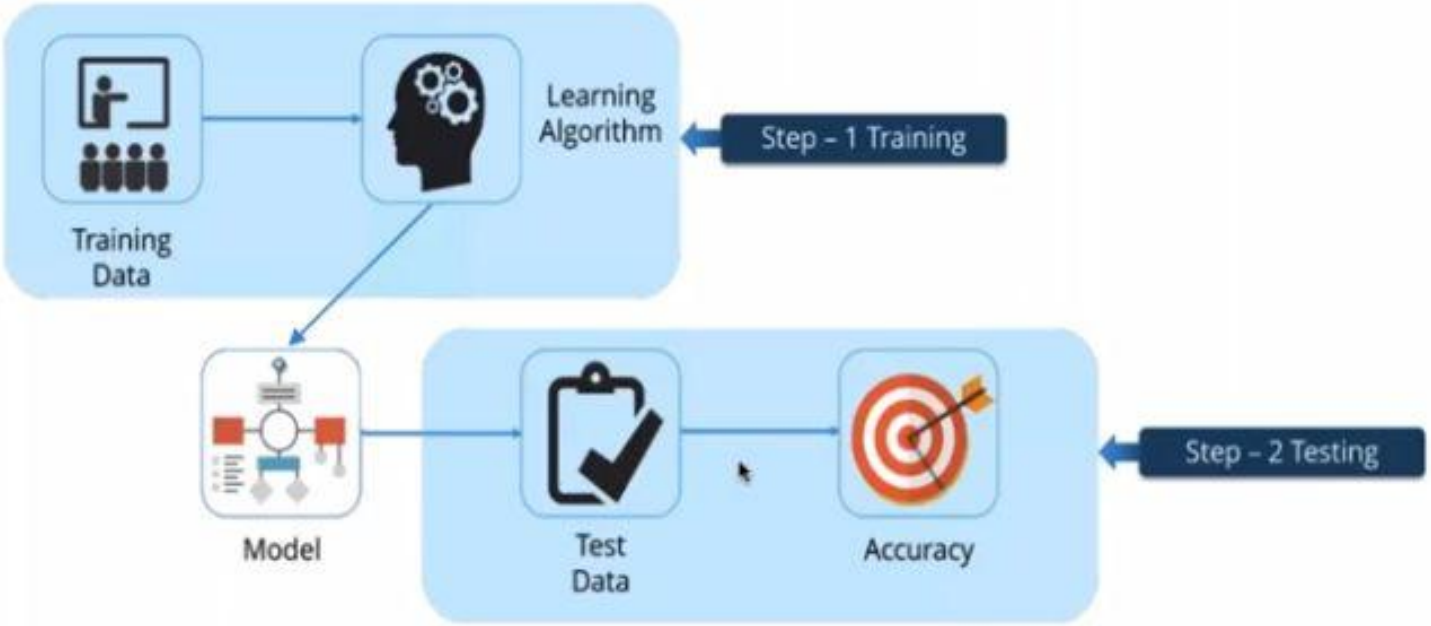
- Supervised learning
- Unsupervised learning
- Reinforcement learning

SUPERVISED LEARNING

Supervised Learning

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

- Supervised Learning is where you have input variables (X) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.



Supervised Learning Types

Types of Learning

Supervised learning

In this type of learning, the data set on which the machine is trained consists of labelled data or simply said, consists both the input parameters as well as the required output.

Supervised Machine Learning Algorithms can be broadly divided into two types of algorithms; Classification and Regression.

- **Classification:** Supervised learning problem that involves predicting a class label
- **Regression:** Supervised learning problem that involves predicting a numerical label

Examples: Linear Regression, Logistic Regression, KNN classification, Support Vector Machine (SVM), Decision Trees, Random Forest, Naive Bayes' theorem.

Machine learning and our focus

- Like human learning from past experiences.
- A computer does not have "experiences".
- A computer system learns from data, which represent some "past experiences" of an application domain.
- Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk.
- The task is commonly called: Supervised learning, classification, or inductive learning.

REGRESSION

- Regression analysis is the major method for numeric prediction
- **Regression analysis** model the relationship between
  - one or more **independent** or **predictor variables** and
  - a **dependent** or **response** variable
- **Regression analysis** is a good choice when all of the **predictor variables** are **continuous** valued as well.

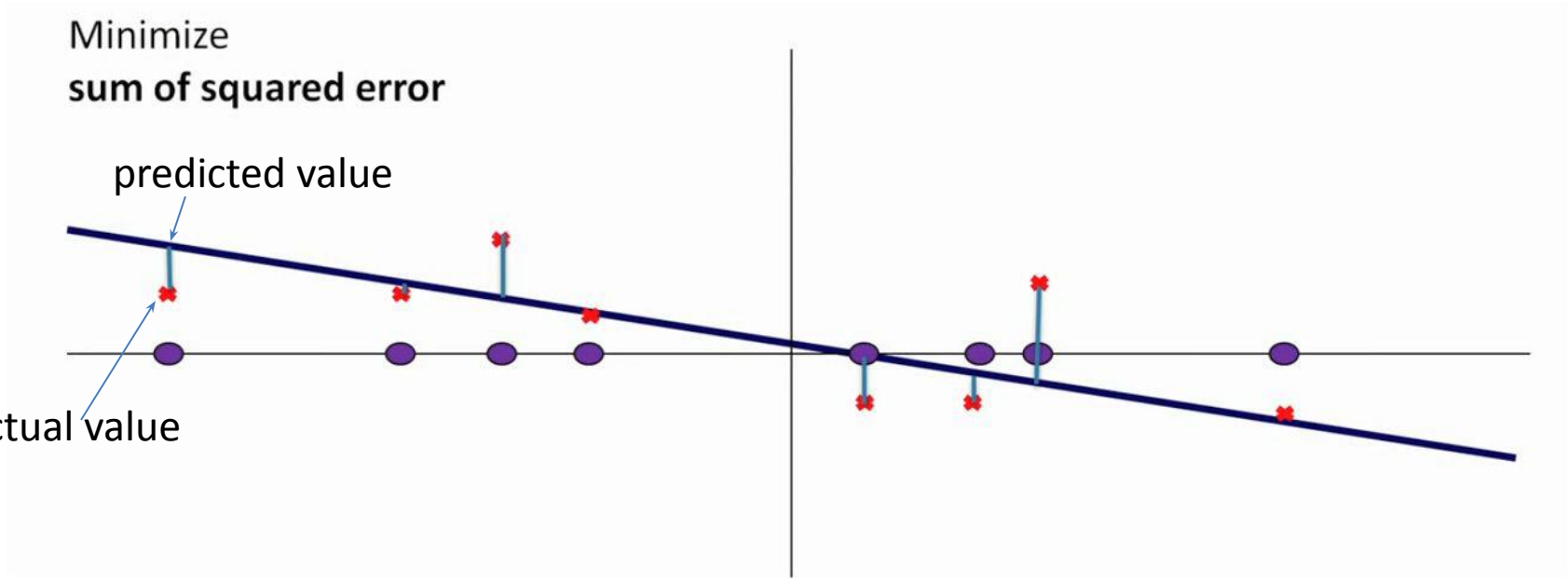
- In the context of data mining
  - The **predictor variables** are the **attributes** of interest describing the instance that are known.
  - The response variable is what we want to predict
- Some classification techniques can be adapted for prediction, e.g.
  - Backpropagation
  - k-nearest-neighbor classifiers
  - Support vector machines

- **Regression analysis methods**:
  - Linear regression
    - Straight-line linear regression
    - Multiple linear regression
  - Non-linear regression
  - Generalized linear model
    - Poisson regression
    - Logistic regression
  - Log-linear models
  - Regression trees and Model trees

- **Numerical prediction** is similar to **classification**
  - construct a model
  - use model to predict continuous or ordered value for a given input
- **Numeric prediction** vs. **classification**
  - Classification refers to predict categorical class label
  - Numeric prediction models continuous-valued functions

LINEAR REGRESSION

Linear Regression



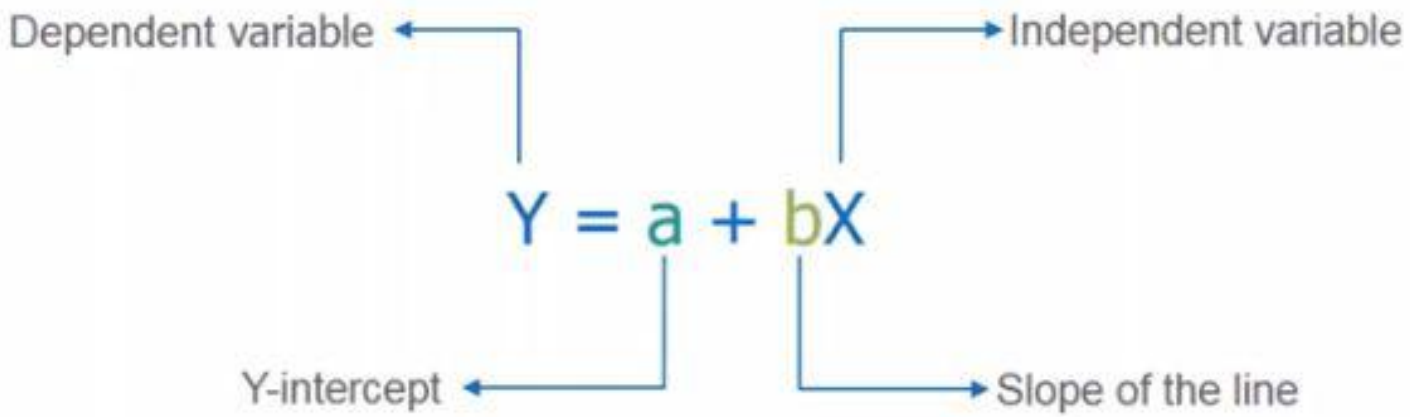
Linear Regression

- **Straight-line linear regression**:
  - involves a response variable y and a single predictor variable x

$$y = w_0 + w_1 x$$

  - w_0 : y-intercept
  - w_1 : slope
  - w_0 & w_1 are **regression coefficients**

Equation for a Regression Line



- Y-intercept (a) is that value of the Dependent Variable(y) when the value of the Independent Variable(x) is zero. It is the point at which the line cuts the y-axis

- Slope (b) is the change in the Dependent Variable for a unit increase in the Independent Variable. It is the tangent of the angle made by the line with the x-axis

Linear Regression

- Consider the following example:
  - A company is facing high Churn out this year, and they are in a process of finding out the reason behind it. Salary hike being one of the major reason, let us consider a company's data where we try to find out the relationship between these two variables

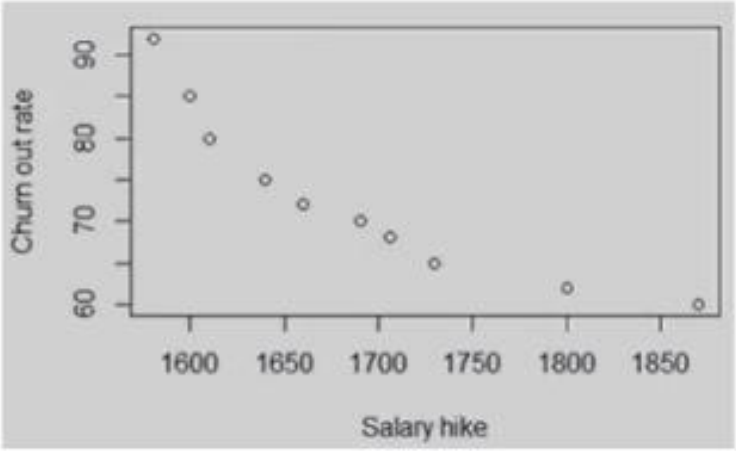
This analysis can be performed using Linear Regression

Salary_hike	Churn_out_rate
1580	92
1600	85
1610	80
1640	75
1660	72
1690	70
1706	68
1730	65
1800	62
1870	60

Relation between DV and IDV

- In order to know how these variables are related to each other, plot a graph

The x-axis = Salary_hike
y-axis = Churn_out_rate



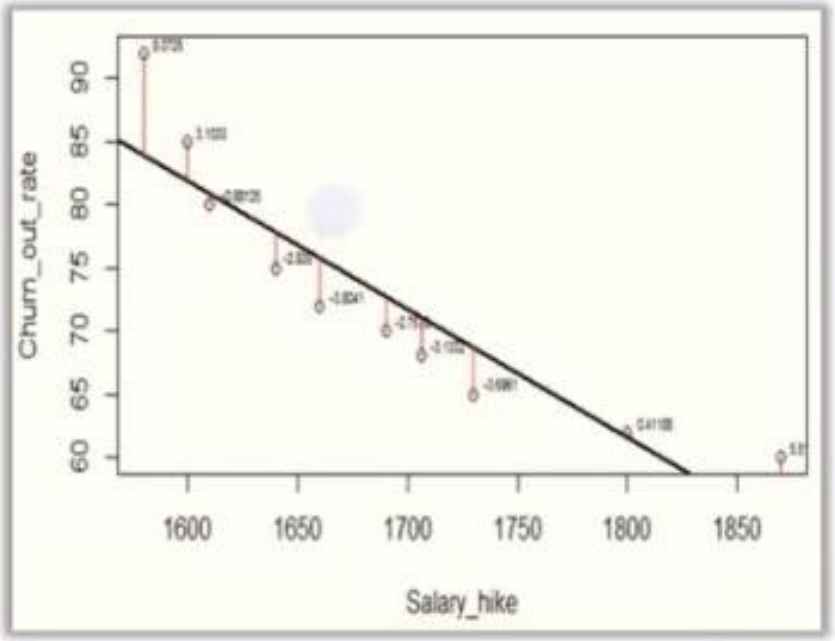
Conclusion:

From the graph, we can see that as the Salary hike decreases, the Churn out rate increases.

Linear Regression Model

- This is the final graph obtained after applying the Linear Regression Technique
- From the graph, the red line shows the deviation from the regression line plotted

When Salary_hike = 1600, Churn_out_rate =83. The deviation at this point from the Regression line = 3.1033



What is Classification?

- Classification is the problem of identifying to which set of categories a new observation belongs
- The goal of classification is to find boundaries that best separate different categories of data. These 'decision boundaries' then allow you to differentiate between classes of data, and classify any new value

Applications of Classification

- Medical Diagnostics
  - predict whether a patient is sick or not
- Animal Recognition
  - Classifying a set of animal images
- Machine vision
  - Classify faces based on patterns(face detection)
- Market segmentation
  - Predict if customer will respond to promotion or not
- Bioinformatics
  - Classify proteins according to their function

Types of Classifiers

- Some of the popular classifiers used in Machine Learning are:

Types of Learning

Unsupervised learning

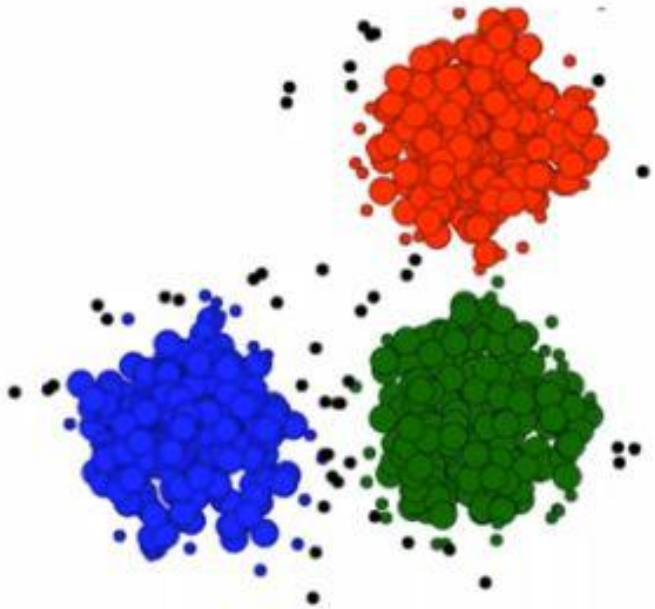
Unlike supervised learning algorithms, where we deal with labelled data for training, the training data will be unlabelled for Unsupervised Machine Learning Algorithms. The clustering of data into a specific group will be done on the basis of the similarities between the variables.

- **Clustering:** Unsupervised learning problem that involves finding groups in data
- **Density estimation:** Unsupervised learning problem that involves summarizing the distribution of data
- **Visualization:** Unsupervised learning problem that involves creating plots of data
- **Projection:** Unsupervised learning problem that involves creating lower-dimensional representations of data

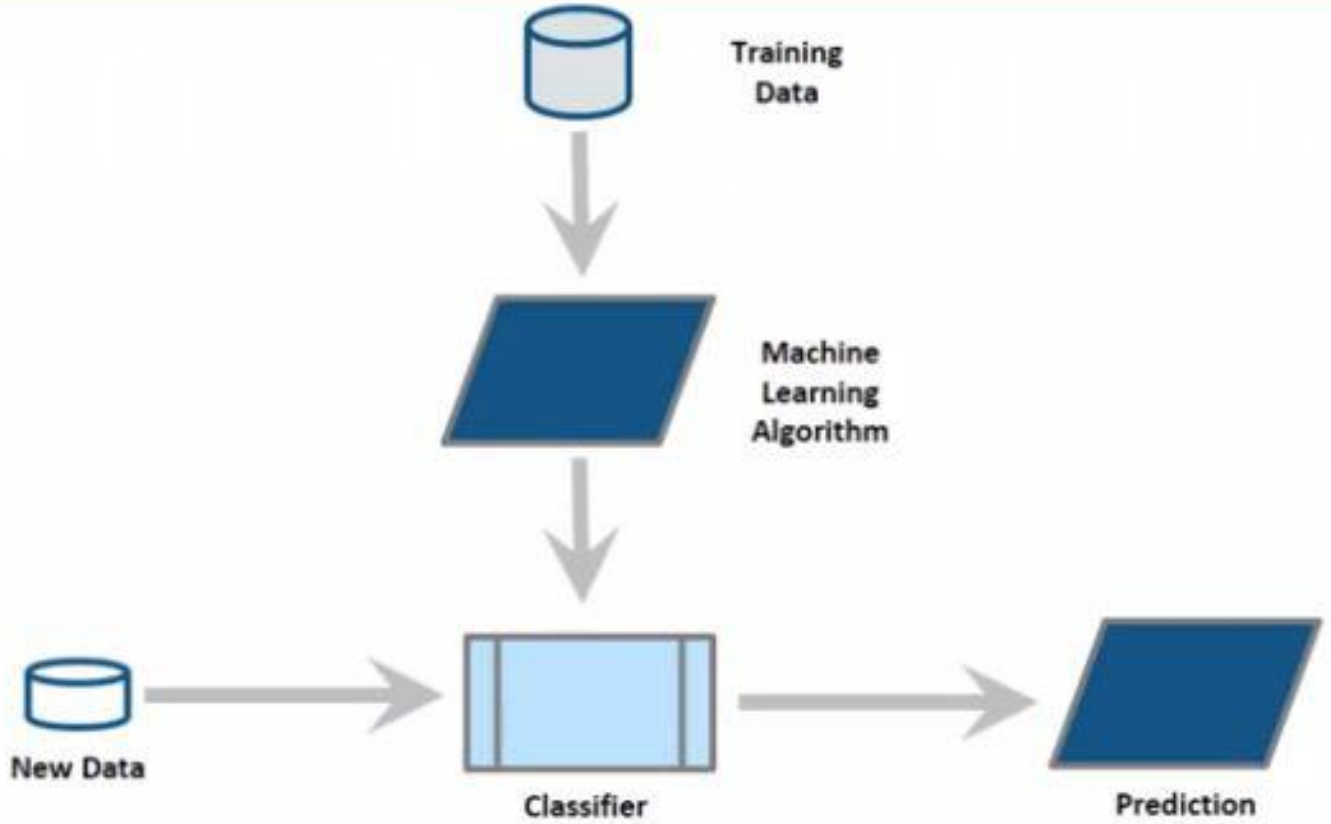
Examples: K-means clustering, neural networks

Unsupervised Learning

- Unsupervised learning is the training of a model using information that is neither classified nor labelled
- This model can be used to cluster the input data in classes on the basis of their statistical properties



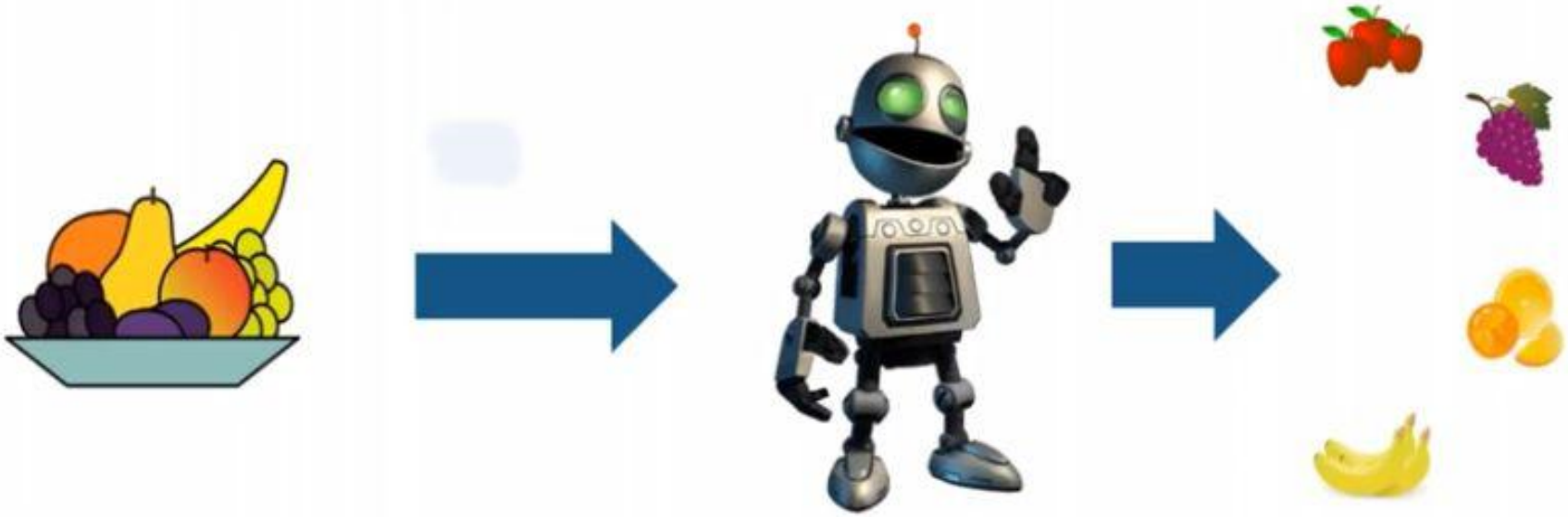
Unsupervised Learning - Process Flow



The machine learns from training data and classifies new data based on it

Unsupervised Learning - Process Flow

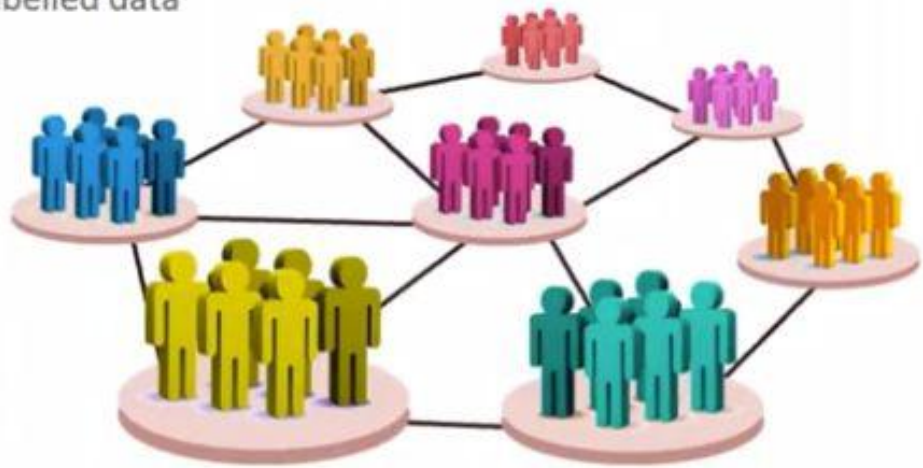
- A set of fruit images is first fed into the system
- The system identifies different fruits using features like color, size, surface type etc, and it categorizes them
- When a new fruit is shown, it analyses its features and puts it into the category having similar featured items



CLUSTER ANALYSIS

What is Clustering?

- Clustering means grouping of objects based on the information found in the data, describing the objects or their relationship
- The goal is that objects in one group will be similar to one other and different from objects in another group
- It deals with finding a structure in a collection of unlabelled data
- Some Examples of clustering methods are :
  - K-means Clustering
  - Fuzzy/ C-means Clustering
  - Hierarchical Clustering



Why Clustering?

- The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data
- Organizing data into clusters shows internal structure of the data
- Sometimes partitioning is the goal
- Some Examples are
  - finding groups of customers with similar behavior
  - classification of animals given their features, example classification of animals into reptiles, mammals, birds, fish etc.

The purpose of clustering algorithm is to make sense of and extract value from large sets of structured and unstructured data

Clustering Use Cases

Marketing
Discovering distinct groups in customer databases, such as customers who make lot of long-distance calls.

Insurance
Identifying groups of crop insurance policy holders with a high average claim rate. Farmers crash crops, when it is "profitable".

Search Engine
Better the clustering algorithm used, better are the chances of getting the required result on the front page.

Seismic studies
Identifying probable areas for oil/gas exploration based on seismic data

Types of Data used for Clustering

- data matrix
  - the "classic" data input

attributes/dimensions

tuples/objects

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- distance or dissimilarity matrix
  - The desired input to some clustering algorithms like Hierarchical clustering

objects

objects

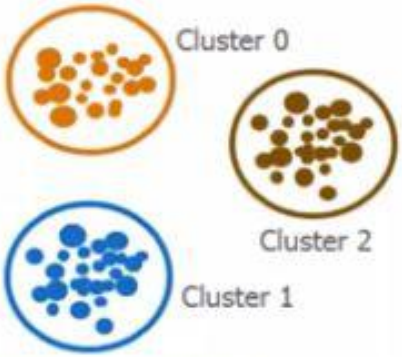
$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix}$$

Types of Clustering

Clustering

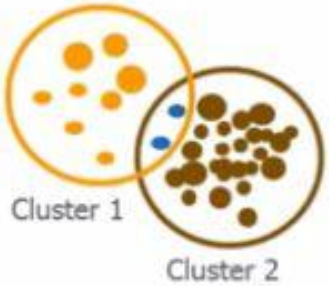
Exclusive Clustering

Here, an item belongs exclusively to one cluster, not several. K-means does this sort of exclusive clustering.



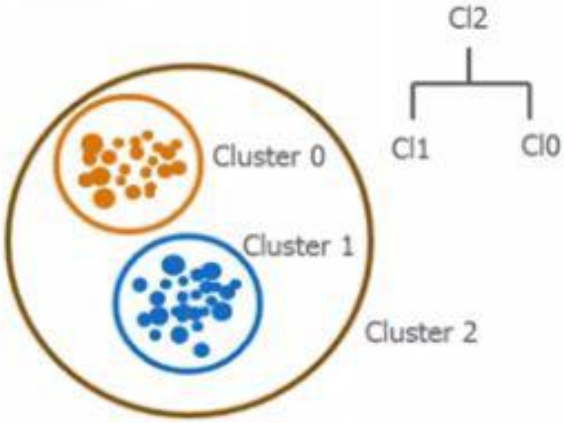
Overlapping Clustering

Here, an item can belong to multiple clusters and its degree of association with each cluster is shown. fuzzy/c-means is of this type.



Hierarchical Clustering

When two cluster have a parent-child relationship or a tree-like structure then it is Hierarchical clustering.



Types of Learning

Reinforcement learning

Reinforcement Learning is a type of Machine Learning in which the machine is required to determine the ideal behaviour within a specific context, in order to maximize its rewards. It works on the rewards and punishment principle which means that for any decision which a machine takes, it will be either be rewarded or punished. Thus, it will understand whether or not the decision was correct. This is how the machine will learn to take the correct decisions to maximize the reward in the long run.

Examples: Q-learning, temporal-difference learning, and deep reinforcement learning.

REINFORCEMENT LEARNING

Learning types

- Learning types
  - *Supervised learning*:

    a situation in which sample (input, output) pairs of the function to be learned can be perceived or are given
    - You can think it as if there is a kind teacher
  - *Reinforcement learning*:

    in the case of the agent acts on its environment, it receives some evaluation of its action (reinforcement), but is not told of which action is the correct one to achieve its goal

Reinforcement learning

- Task

  Learn how to behave successfully to achieve a goal while interacting with an external environment

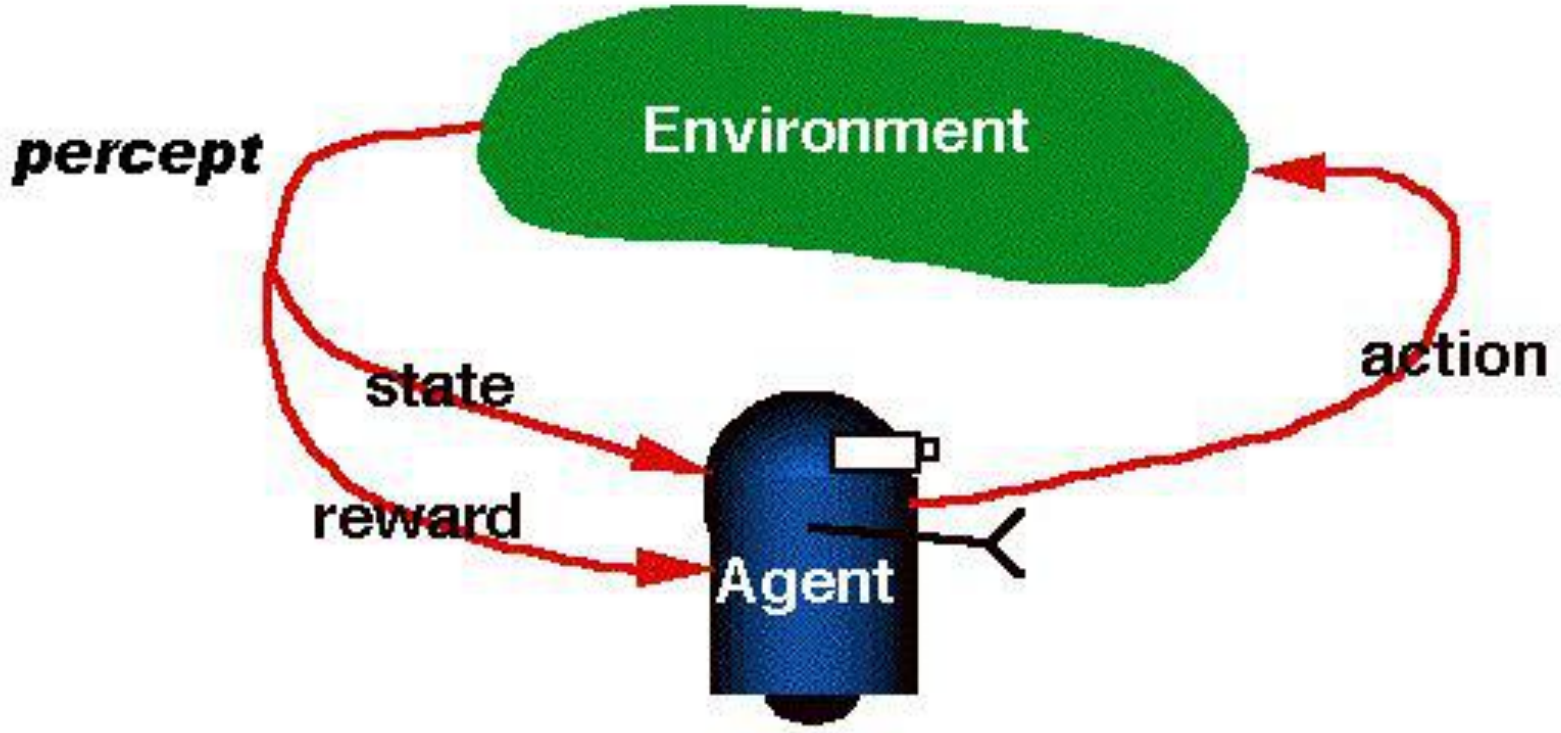
  - *Learn via experiences!*
- Examples
  - Game playing: player knows whether it win or lose, but not know how to move at each step
  - Control: a traffic system can measure the delay of cars, but not know how to decrease it.

Reinforcement Learning

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

- Reinforcement Learning (RL) is learning by interacting with a space or an environment.
- It selects its actions on basis of its past experiences (exploitation) and also by new choices (exploration).

RL is learning from interaction

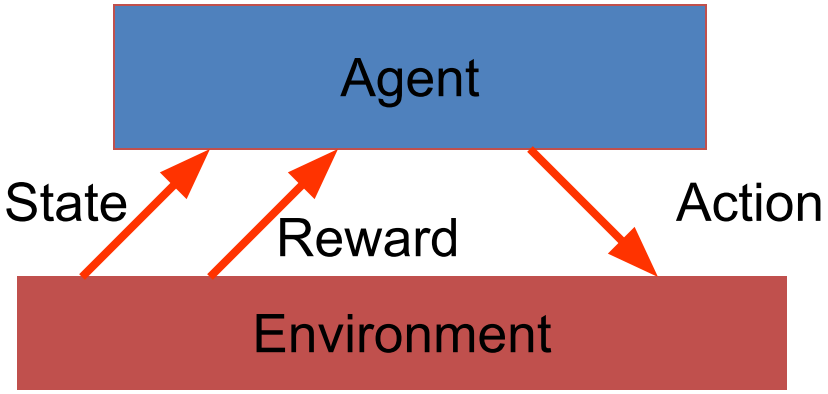


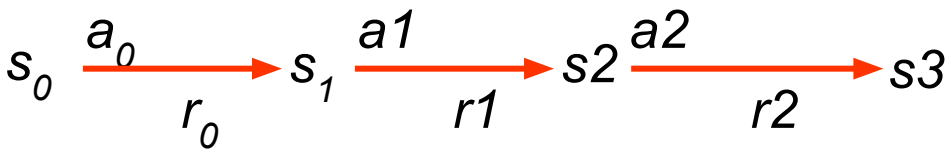
RL model

- Each percept(*e*) is enough to determine the State(the state is accessible)
- The agent can decompose the Reward component from a percept.
- The agent's task: to find a optimal policy, mapping states to actions, that maximize long-run measure of the reinforcement
- Think of reinforcement as reward
- Can be modeled as MDP (Markov Decision Problem) model!

Review of MDP model

- MDP model <S,T,A,R>





- S– set of states
- A– set of actions
- T(s,a,s') = P(s'|s,a)– the probability of transition from *s* to *s'* given action *a*
- R(s,a)– the expected reward for taking action *a* in state *s*

$$R(s,a) = \sum_{s'} P(s' \mid s,a) r(s,a,s')$$

$$R(s,a) = \sum_{s'} T(s,a,s') r(s,a,s')$$

INGREDIENTS OF ML

- TASKS
- MODEL
- FEATURES

Select best features

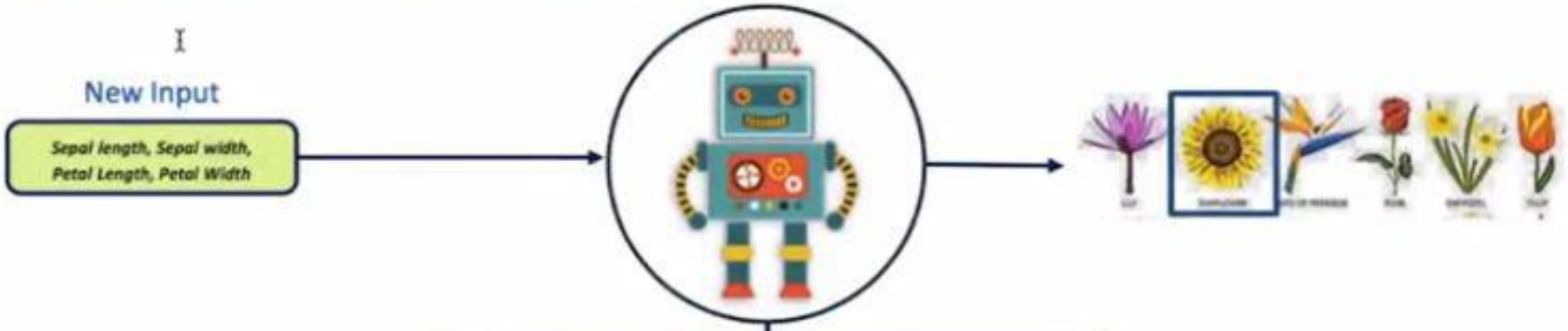
To design a model

To perform right set of tasks.

Machine Learning

- Machine Learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.

Problem Statement: Determine the species of the flower



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Use Case - 1

Patient: I want to get diagnosed for Diabetes
Doctor: Ok let me check the reports

No.of_tim	glucose_conc	blood_pressure	skin_fold_thickne	2-Hour_serum_insuli	BMI	Diabetes_Age	
6	148	72	35	0	33.6	0.627	50

Not Diabetic

Diabetic

After analysing the patient's report, the Doctor with the help of his experience can diagnose whether the patient has Diabetes or not

Use Case - 1

We now want to train a Machine to do the doctor's task.
For this purpose we have to train the machine with the same experience/knowledge using the historical data.

No.of_times	glucose_conc	blood_pressure	skin_fold_thickness	2-Hour_serum_insulin	BMI	Diabetes_Age	Is_Diabetic
6	148	72	35	0	33.6	0.627	50 YES
1	85	66	29	0	26.6	0.351	31 NO
8	183	64	0	0	23.3	0.672	32 YES
1	89	66	23	94	28.1	0.167	21 NO
0	137	40	35	168	43.1	2.288	33 YES
5	116	74	0	0	25.6	0.201	30 NO
3	78	50	32	88	31	0.248	26 YES
10	115	0	0	0	35.3	0.134	29 NO
2	197	70	45	543	30.5	0.158	53 YES
8	125	96	0	0	0	0.232	54 YES
4	110	92	0	0	37.6	0.191	30 NO
10	168	74	0	0	38	0.537	34 YES
10	139	80	0	0	27.1	1.441	57 NO
1	189	60	23	846	30.1	0.398	59 YES
5	166	72	19	175	25.8	0.587	51 YES

Machine Learning from the historical data of 50000 patients

Use Case – 1 Characteristics

- The problem has following characteristics:

1. Labelled learning data and output available.
2. Machine can learn a mapping function. Using the data machine can find a relationship between input and output. After the model is trained upon input of new data it can predict the output.
3. Output classes are predefined.

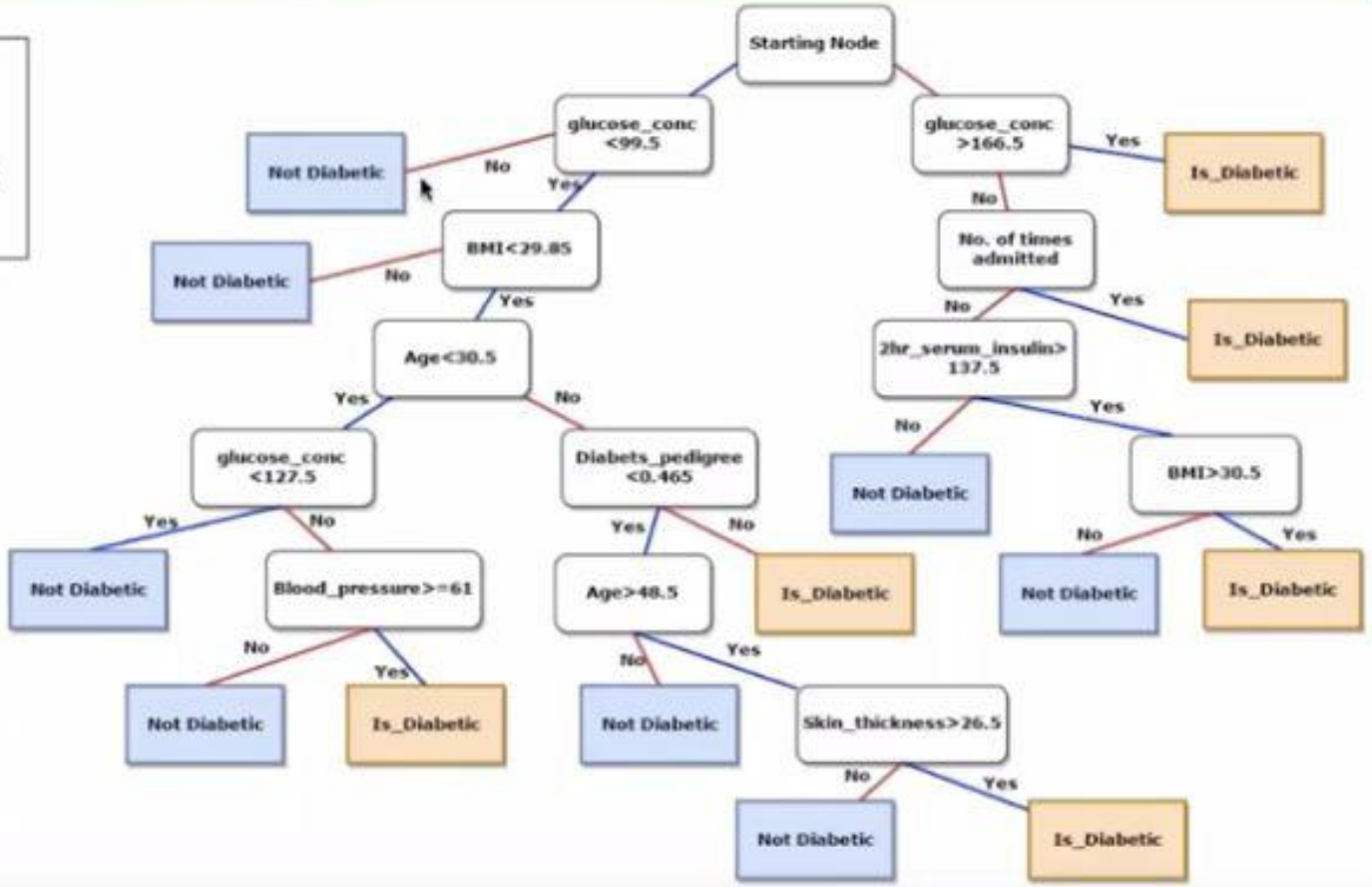
   i.e. Diabetic or not Diabetic

Output

No.of_tim	glucose_conc	blood_pressure	skin_fold_thickne	2-Hour_serum_insuli	BMI	Diabetes_	Age	Is_Diabetic
6	148	72	35	0	33.6	0.627	50	YES
1	85	66	29	0	26.6	0.351	31	NO
8	183	64	0	0	23.3	0.672	32	YES
1	89	66	23	94	28.1	0.167	21	NO
0	137	40	35	168	43.1	2.288	33	YES
5	116	74	0	0	25.6	0.201	30	NO
3	78	50	32	88	31	0.248	26	YES
10	115	0	0	0	35.3	0.134	29	NO
2	197	70	45	543	30.5	0.158	53	YES
8	125	96	0	0	0	0.232	54	YES
4	110	92	0	0	37.6	0.191	30	NO
10	168	74	0	0	38	0.537	34	YES
10	139	80	0	0	27.1	1.441	57	NO
1	189	60	23	846	30.1	0.398	59	YES
5	166	72	19	175	25.8	0.587	51	YES

Flow Graph – Illustrating Rules Created

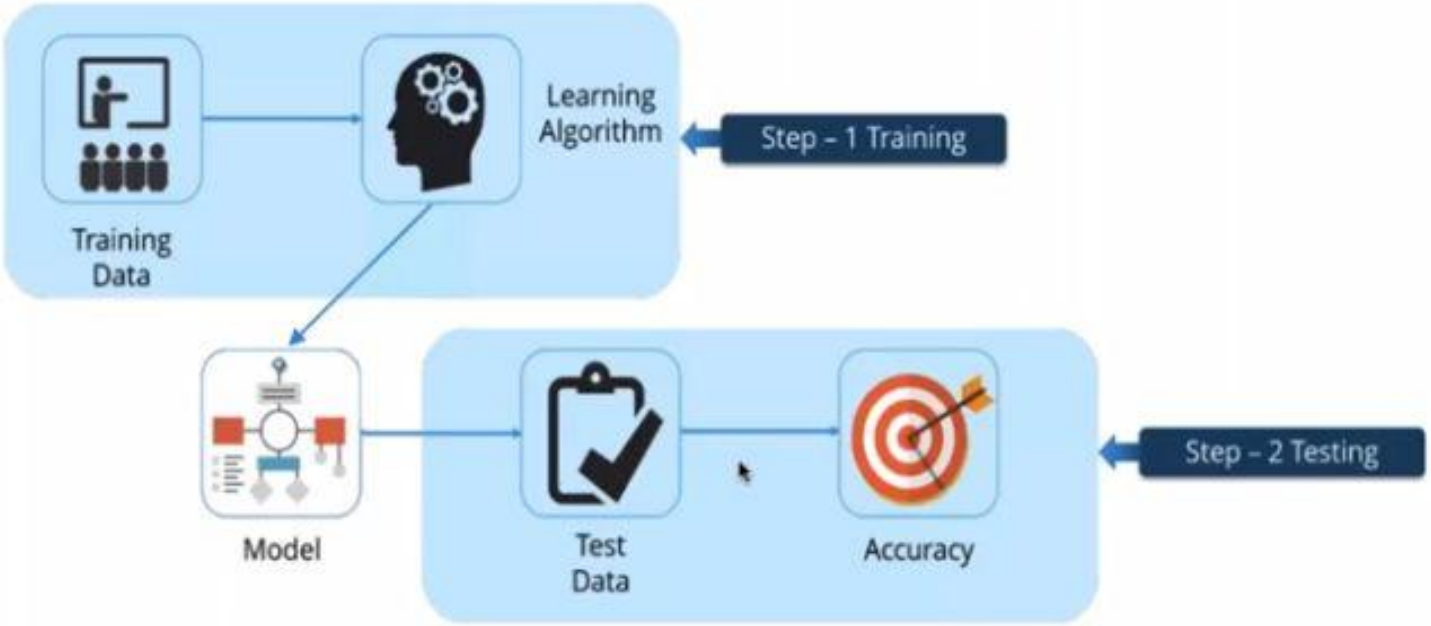
After learning from the training data, machine will create certain rules/model, as depicted by this flow graph.



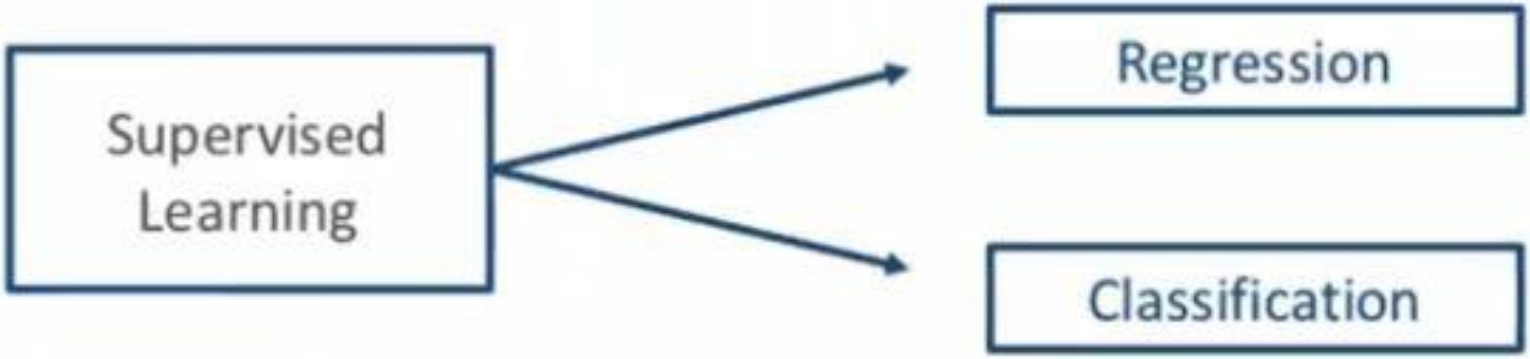
Supervised Learning

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

- Supervised Learning is where you have input variables (X) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.



Supervised Learning Types



Use Case – 1.2

Let's take an example:

- A Real Estate Consultation firm has the data of the price of apartments in Boston. This data contains values such as crime rate, age, accessibility, population etc.
- Based on this data company wants to decide on the price of new apartments.
- This problem can be solved by a linear regression model.

The Boston Data looks like this:

X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9

Data Set Description

- The dataset columns such as:
- Based on these columns we have to predict the house pricing that is medv.
- This data set is already present in python,
- To load it we can use the following code:

```
from sklearn.datasets import load_boston
boston = load_boston()
```

Housing Values in Suburbs of Boston

Description

The `Boston` data frame has 506 rows and 14 columns.

Usage

```
Boston
```

Format

This data frame contains the following columns:

`crim`
per capita crime rate by town.

`zn`
proportion of residential land zoned for lots over 25,000 sq.ft.

`indus`
proportion of non-retail business acres per town.

`chas`
Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

`lstat`
lower status of the population (percent).

`medv`

Linear Regression

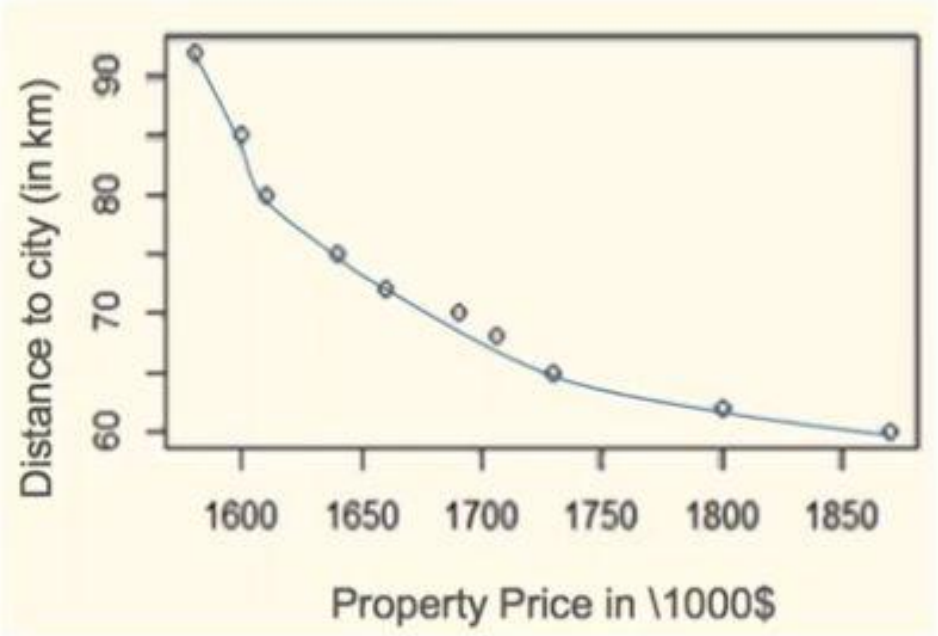
- A Dependent Variable(DV) is the variable to be predicted or explained.
- An Independent Variable(IDV) is the variable related to the dependent variable in a equation.

Crime rate is an independent variable in this case.

Property price (medv) is a dependent variable, it depends on number of other variables.

X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9

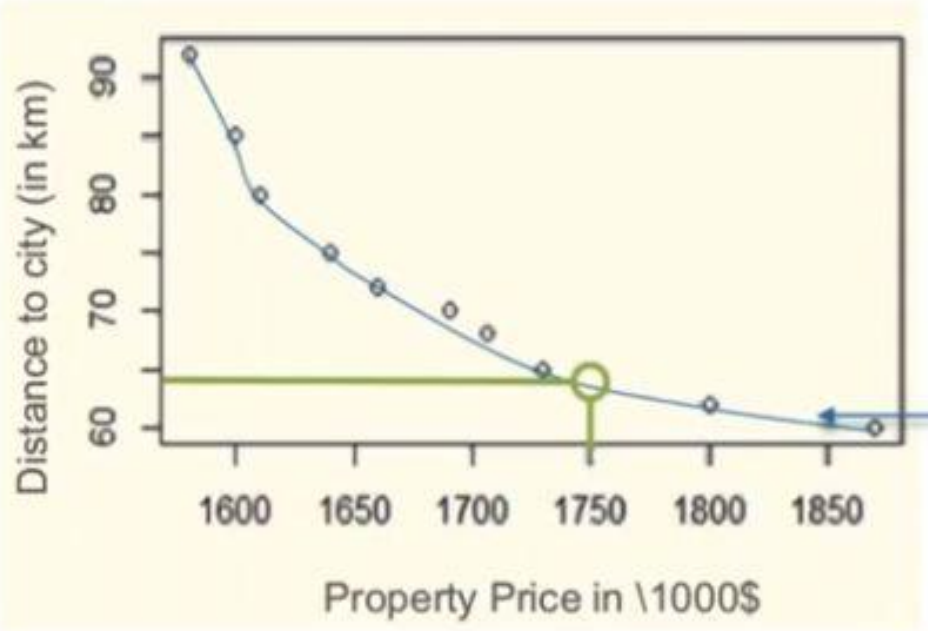
Predicting The Outcome



Average of all the values is calculated first, then a line/curve is plotted which goes through all the average values.

Predicting The Outcome

- The management has planned to open a society about 64 kms from the main city, let's predict possible Property price through this curve

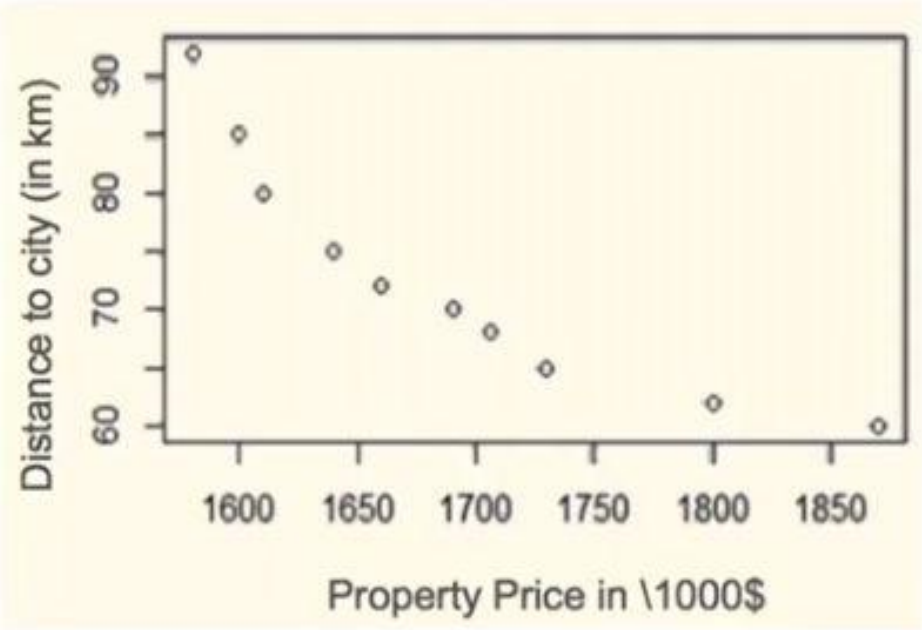


Conclusion:
From the curve, we can predict a possible Property price to be 1750

This line is called Linear regression line.
Let's understand what is Linear Regression.

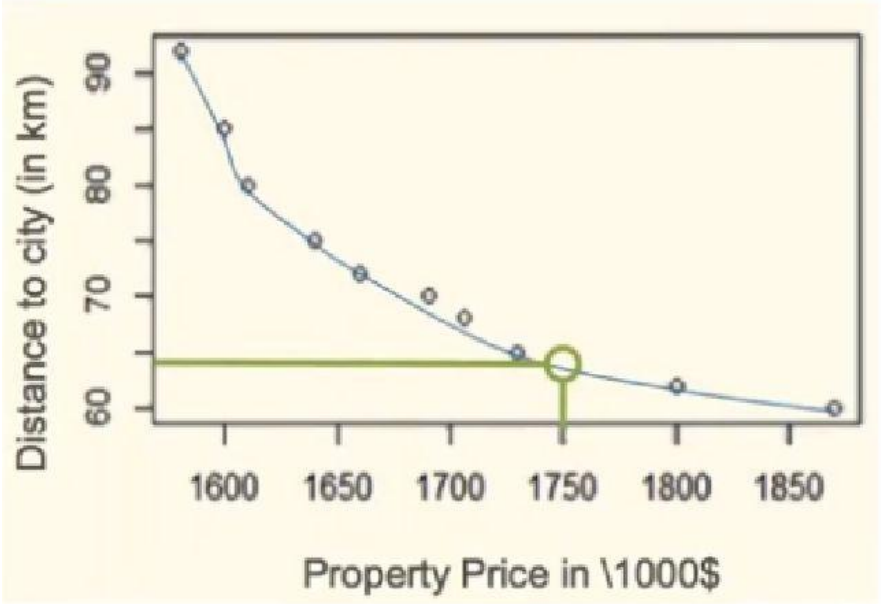
Relation Between Variables

- In order to know how these variables (i.e. Distance to city (in km) and Property Price in \1000$) are related to each other, let's plot a graph.



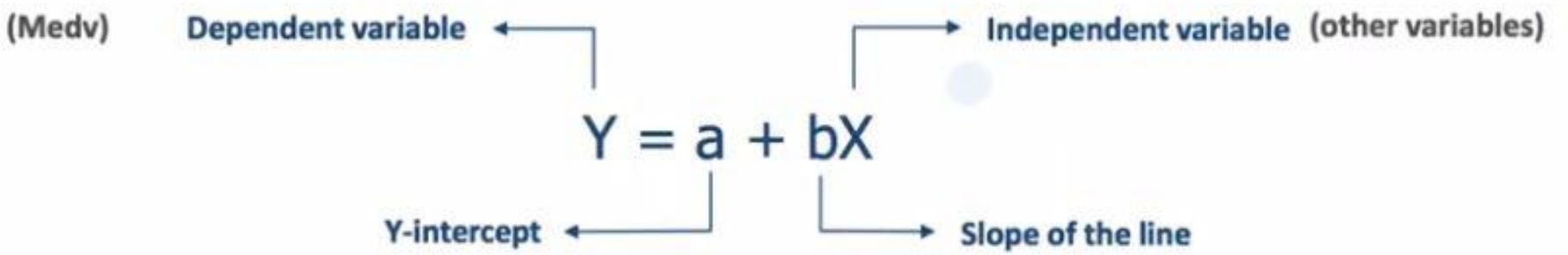
Conclusion:
From the graph, we can see that as the Distance to city (in km) decreases, the Property Price in \1000$ increases.

Predicting The Outcome



Similarly there exist relationship of other variables with price variable that is medv, for complex tasks involving so many variables, we need machine learning.

Equation For Model



- Y-intercept (a) is that value of the Dependent Variable(y) when the value of the Independent Variable(x) is zero. It is the point at which the line cuts the y-axis.
- Slope (b) is the change in the Dependent Variable for a unit increase in the Independent Variable. It is the tangent of the angle made by the line with the x-axis.

Linear Model With All Variables

- In actual use we will be using all variables for the model , i.e.
- **medv** vs (crim + zn + indus + chas + nox + rm + age + dis + rad +tax + ptratio + black + lstat)
- Equation will be like as shown below:

$$Y = a + bX_1 + b_1X_2 + b_3X_3...$$

- Let's learn what Linear Regression is in more detail.

Use Case - 3

We want to train a robotic dog which can learn from it's exploration and experiences just as a real dog does, the sample space and the data are not predefined in this case.

Use Case - 3 Characteristics

- The problem has following characteristics:

1. Output is based on past experience as well as exploration of new choices.
2. Machine learns from it's experiences and choices.
3. Machine's main aim is to get reward and reduce penalty.
4. No predefined choices are available, machine interacts with available environment based upon it's learning.

Learning Models

What is being learned from the data in order to solve the Task?

1. Geometric Models
2. Probabilistic Models
3. Logical Models

All possible instances = Instance Space

Geometric Learning Models

In Geometric models, features could be described as points in two dimensions (x- and y-axis) or a three-dimensional space (x, y, and z). Even when features are not intrinsically geometric, they could be modelled in a geometric manner (for example, temperature as a function of time can be modelled in two axes). In geometric models, there are two ways we could impose similarity.

- We could use geometric concepts like lines/ planes/ distance to segment (classify) the **instance space**. These are called **Linear models**
- Alternatively, we can use the geometric notion of distance to represent similarity. In this case, if two points are close together, they have similar values for features and thus can be classed as similar. We call such models as **Distance-based models**

Geometric Models
(Using lines)

- Easy to visualise
- Linear
- If there exists a linear decision boundary between 2 classes, then those 2 classes are linearly separable

$$w.x = t$$

w= vector perpendicular to the decision boundary
x = arbitrary point on the decision boundary
t = decision threshold

Linear model

w is a vector drawn perpendicular from the center of mass of one class to the center of mass of another class

w.x = t

If p is the center of mass of blue and n is the center of mass of red points

w = p-n

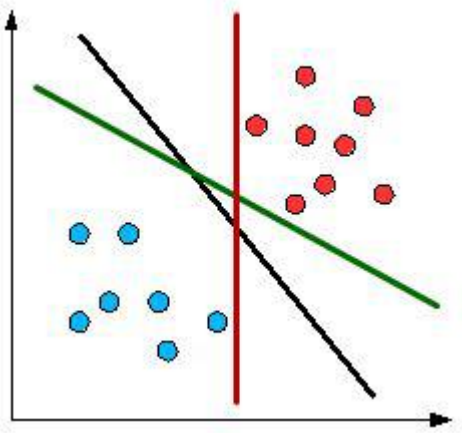
x= (p+n)/2 (average of the two)

Eqn. becomes

(p-n) (p+n)/2 = t

$\|\mathbf{p}\|^2 - \|\mathbf{q}\|^2 = t$

II**p**II is the number of blue points and II**q**II is the number of red points



Failure of Linear Models

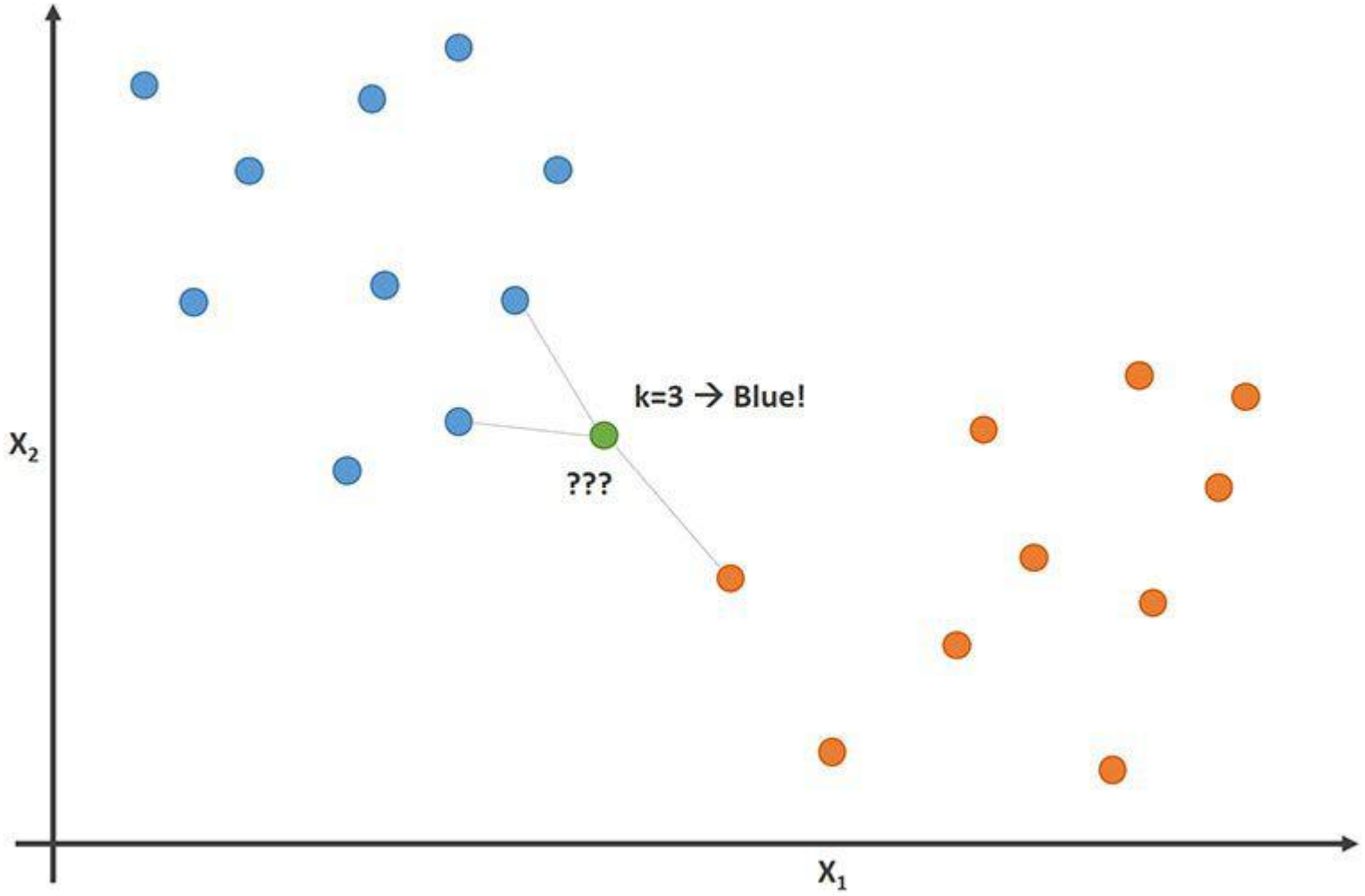
Linear classification fails when the instance space is mostly empty.

E.g. We are trying to find whether a sentence is present in a vocabulary of 10,000 words.
Here most of the sample space is empty, hence linear classification will fail.
Here we will use **large marginal classifiers (e.g. SVM)**

The decision boundary is based on particular instance.

Distance based LM

Nearest neighbor classifier is based on distance measure.



Distance Measures

Euclidean	$d(x,y) = \sqrt{\sum (x_i - y_i)^2}$
Squared Euclidean	$d(x,y) = \sum (x_i - y_i)^2$
Manhattan	$d(x,y) = \sum \lvert (x_i - y_i) \rvert$
Canberra	$d(x,y) = \sum \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$
Chebychev	$d(x,y) = \max(\lvert x_i - y_i \rvert)$
Bray Curtis	$d(x,y) = \frac{\sum \lvert x_i - y_i \rvert}{\sum x_i + y_i}$
Cosine Correlation	$d(x,y) = \frac{\sum (x_i y_i)}{\sqrt{\sum (x_i)^2 \sum (y_i)^2}}$
Pearson Correlation	$d(x,y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (y_i - \overline{y})^2} \sqrt{\sum (y_i - \overline{y})^2}}$
Uncentered Peason Correlation	$d(x,y) = \frac{\sum x_i y_i}{\sqrt{\sum (y_i - \overline{y})^2} \sqrt{\sum (y_i - \overline{y})^2}}$
Euclidean Nullweighted	Same as Euclidean, but only the indexes where both x and y have a value (not NULL) are used, and the result is weighted by the number of values calculated. Nulls must be replaced by the missing value calculator (in dataloader).

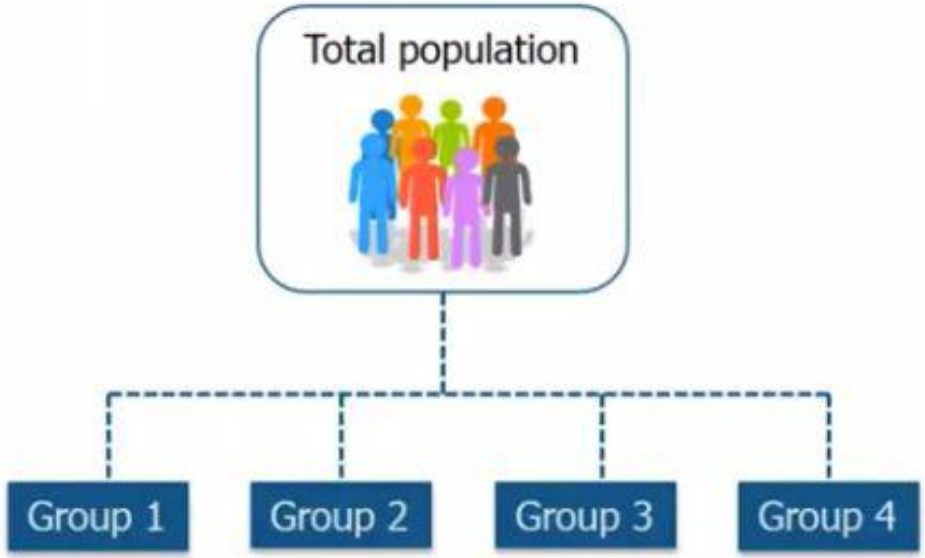
GEOMETRIC MODELS

UNSUPERVISED LEARNING

K-Means Clustering

- The process by which objects are classified into a predefined number of groups so that they are as much dissimilar as possible from one group to another group, but as much similar as possible within each group

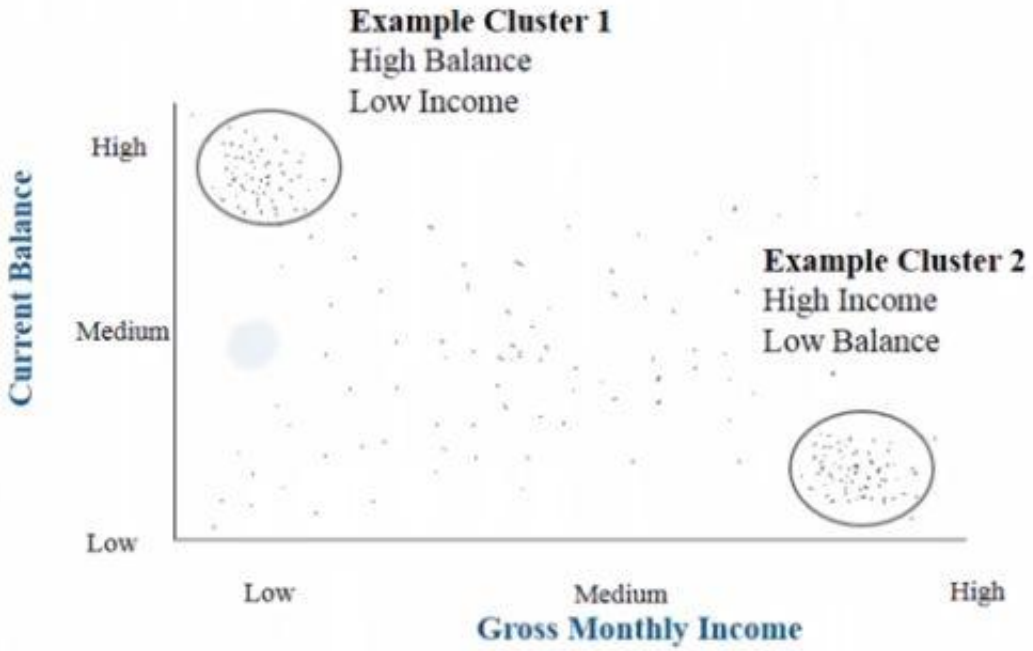
- The objects in group 1 should be as similar as possible
- But there should be much difference between an object in group 1 and group 2
- The attributes of the objects are allowed to determine which objects should be grouped together



K-Means Clustering

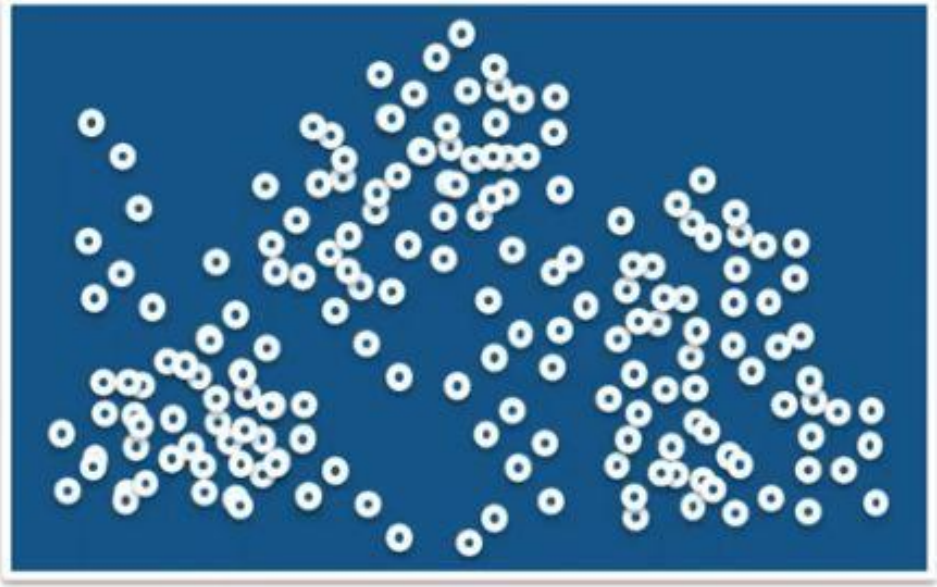
- Consider the following example

- The objects in Cluster 1 have similar characteristics (High Income and Low balance)
- Also the objects in Cluster 2 have the same characteristic (High Balance and Low Income)
- But there are much differences between an object in Cluster 1 and an object in Cluster 2



Example

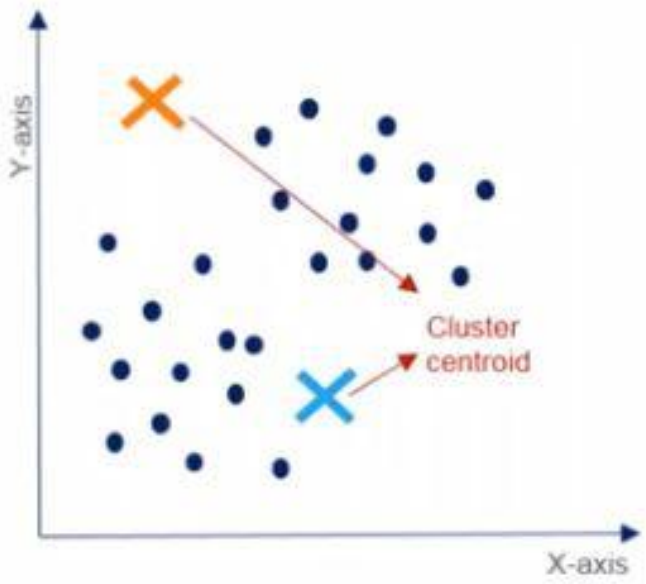
- The plot of students in an area is as given below

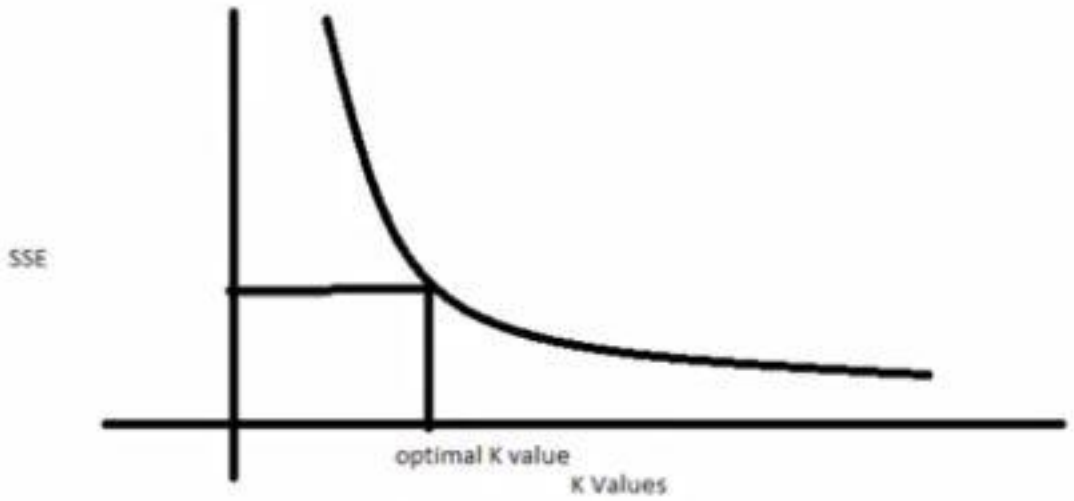


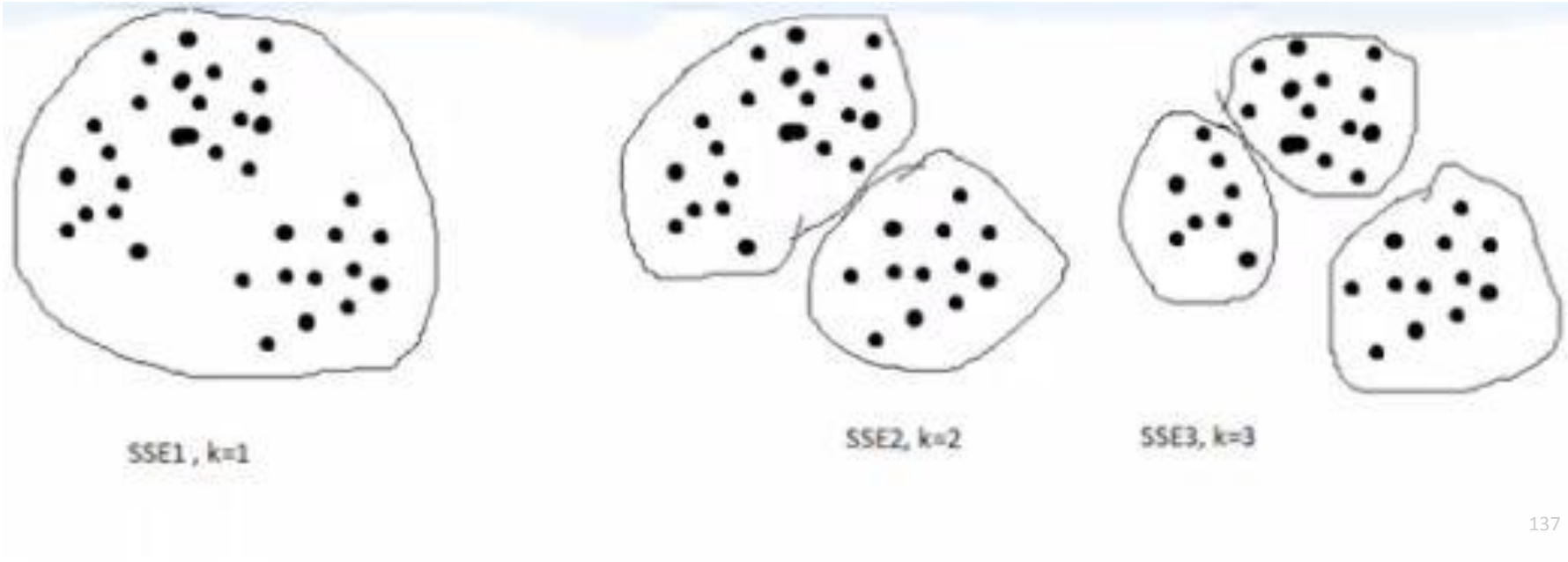
Initialization

1. Initialization
2. Cluster assignment
3. Move centroid
4. Optimization
5. Convergence

- Randomly initialize k points called the cluster centroids
  Here, k = 2
- Value of k(number of clusters) can be determined by the elbow curve




SSE
optimal K value
K Values


SSE1 , k=1
SSE2, k=2
SSE3, k=3

Cluster assignment

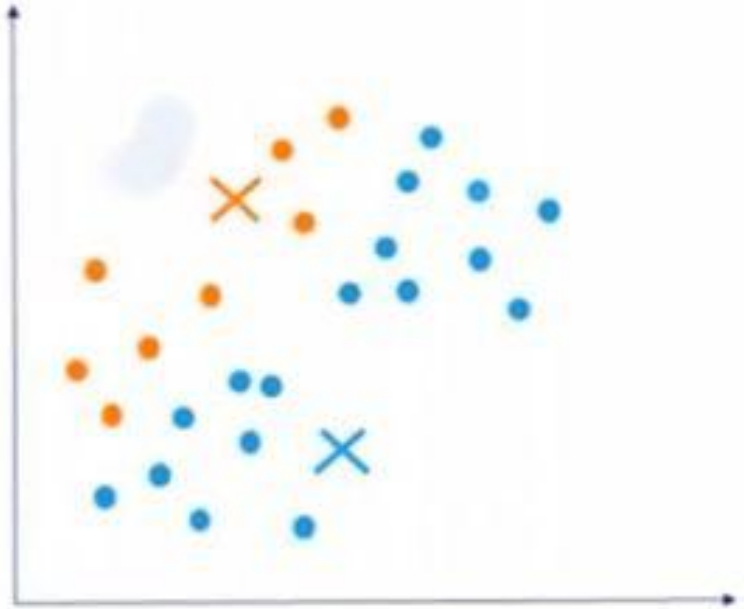
1. Initialization
2. Cluster assignment
3. Move centroid
4. Optimization
5. Convergence

- Compute the distance between the data points and the cluster centroid initialized
- Depending upon the minimum distance, data points are divided into two groups

Move centroid

1. Initialization
2. Cluster assignment
3. Move centroid
4. Optimization
5. Convergence

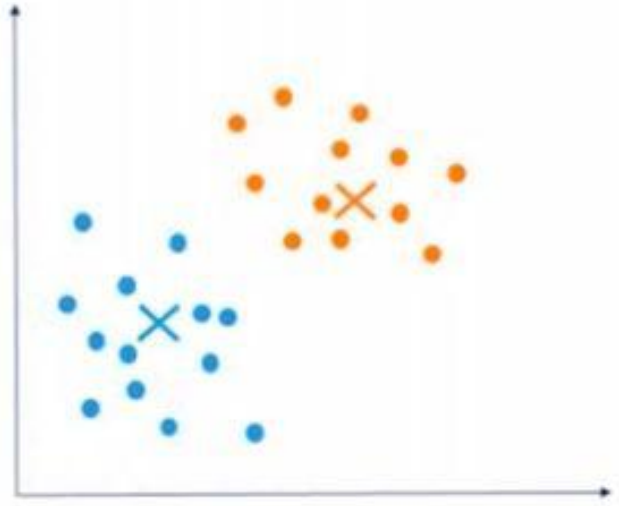
- Compute the mean of blue dots
- Reposition blue cluster centroid to this mean
- Compute the mean of orange dots
- Reposition orange cluster centroid to this mean



Optimization

1. Initialization
2. Cluster assignment
3. Move centroid
4. Optimization
5. Convergence

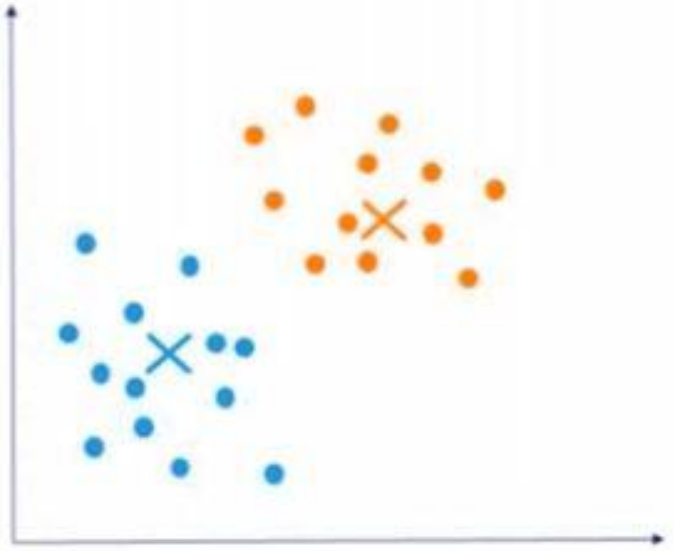
- Repeat previous two steps iteratively till the cluster centroids stop changing their positions



Convergence

1. Initialization
2. Cluster assignment
3. Move centroid
4. Optimization
5. Convergence

- Finally, k-means clustering algorithm converges
- Divides the data points into two clusters clearly visible in orange and blue



K-Means: Hands on

- Here we will use k-means clustering to group a set of flowers using their features
- We will be using iris dataset for this purpose

sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa

PROBABILISTIC MODELS

Probabilistic Learning Models

Probabilistic models see features and target variables as random variables. The process of modelling represents and manipulates the level of uncertainty with respect to these variables. There are two types of probabilistic models: Predictive and Generative.

- **Predictive probability models** use the idea of a conditional probability distribution P (Y |X) from which Y can be predicted from X. We can predict the value of Y only if we know the value of X.
- **Generative models** estimate the joint distribution P (Y, X). Once we know the joint distribution for the generative models, we can derive any conditional or marginal distribution involving the same variables

Probabilistic models use the idea of probability to classify new

entities Naïve Bayes is an example of a probabilistic classifier.

Predictive probability models

Suppose we have 2 words, X = (lottery, prize)
Now I have to predict whether the email is spam or ham

Y= output is a class spam or ham

p(Y/X) = posterior probability of Y given features X has values assigned

p(Y/ lottery = 0, prize =1) gives high probability for spam class

Models based on this:
Bayes Rule Classification

Probability

- Probability is the measure of how likely something will occur.
- It is the ratio of desired outcomes to total outcomes.

  (# desired) / (# total)

- Probabilities of all outcomes sums to 1.

Example:

- ✓ If I roll a dice, there are six total possibilities. (1,2,3,4,5,6)
- ✓ Each possibility only has one outcome, so each has a PROBABILITY of 1/6.
- ✓ For instance, the probability of getting a numeric 2 is 1/6, since there is only a single 2 on the dice.

Bayes Theorem

- Bayes' theorem (also known as Bayes' rule) is a useful tool for calculating conditional probabilities.

$$Bayes'\,theorem:$$

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- For independent events P(A|B) = P(A), so by rearranging the formula we can see,

$$P(A \text{ and } B) = P(A) \times P(B)$$

Probability Distribution

- A probability distribution assigns a probability to each measurable subset of the possible outcomes of a random experiment

One Toss	Head	Tail
Probability	0.5	0.5

Two tosses	Head-Head	Tail-Tail	Head-Tail	Tail-Head
Probability	0.25	0.25	0.25	0.25

- Rules:

1. The outcomes listed must be disjoint
2. Each probability must be between 0 and 1
3. The probabilities must sum to 1

Consider a school with a total population of 100 persons. These 100 persons can be seen either as ‘Students’ and ‘Teachers’ or as a population of ‘Males’ and ‘Females’.

With below tabulation of the 100 people, what is the conditio[illegible]e school is a [illegible] ‘Man’?

	Female	Male	Total
Teacher	8	12	20
Student	32	48	80
Total	40	60	100

$$P(Teacher \mid Male) = \frac{P(Teacher \cap Male)}{P(Male)} = 12/60 = 0.2$$

This can be represented as the intersection of Teacher (A) and Male (B) divided by Male (B). Likewise, the conditional probability of B given A can be computed. The Bayes Rule that we use for Naive Bayes, can be derived from these two notations.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad (1)$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad (2)$$

Bayes Rule is a way to go from P (X | Y) to find P (Y | X)

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$ 1

Known

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$$ 2

UnKnown

P (Evidence | Outcome)
(Known from training data)

P (Outcome | Evidence)
(To be predicted for test data)

Bayes Rule

$$P(Y \mid X) = \frac{P(X \mid Y) * P(Y)}{P(X)}$$

When there are multiple X variables, we simplify it by *assuming the X's are independent*, so the **Bayes** rule

$$P(Y=k \mid X) = \frac{P(X \mid Y=k) * P(Y=k)}{P(X)}$$

where, k is a class of Y

becomes, Naive **Bayes**

$$P(Y=k \mid X1..Xn) = \frac{P(X1 \mid Y=k) * P(X2 \mid Y=k) \ldots * P(Xn \mid Y=k) * P(Y=k)}{P(X1) * P(X2) \ldots * P(Xn)}$$

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

$$P(Y=k \mid X1..Xn) = \frac{P(X1 \mid Y=k) * P(X2 \mid Y=k) \ldots * P(Xn \mid Y=k) * P(Y=k)}{P(X1) * P(X2) \ldots * P(Xn)}$$

can be understood as ..

$$\text{Probability of Outcome I Evidence (Posterior Probability)} = \frac{\text{Probability of Likelihood of evidence} * \text{Prior}}{\text{Probability of Evidence}}$$

Probability of Evidence is same for all classes of Y

Logical Learning Models

Logical models use a logical expression to divide the instance space into segments and hence construct grouping models. A logical expression is an expression that returns a Boolean value, i.e., a True or False outcome. Once the data is grouped using a logical expression, the data is divided into homogeneous groupings for the problem we are trying to solve.

There are two types of logical models: Tree models and Rule models.

- **Rule models** consist of a collection of implications or IF-THEN rules. For tree-based models, the 'if-part' defines a segment and the 'then-part' defines the behaviour of the model for this segment. Rule models follow the same reasoning
- **Tree models** can be seen as a particular type of rule model where the if-parts of the rules are organised in a tree structure.

Both Tree models and Rule models use the same approach to supervised learning

Tree Model

Features are used in construction of trees. They are represented by ellipses.

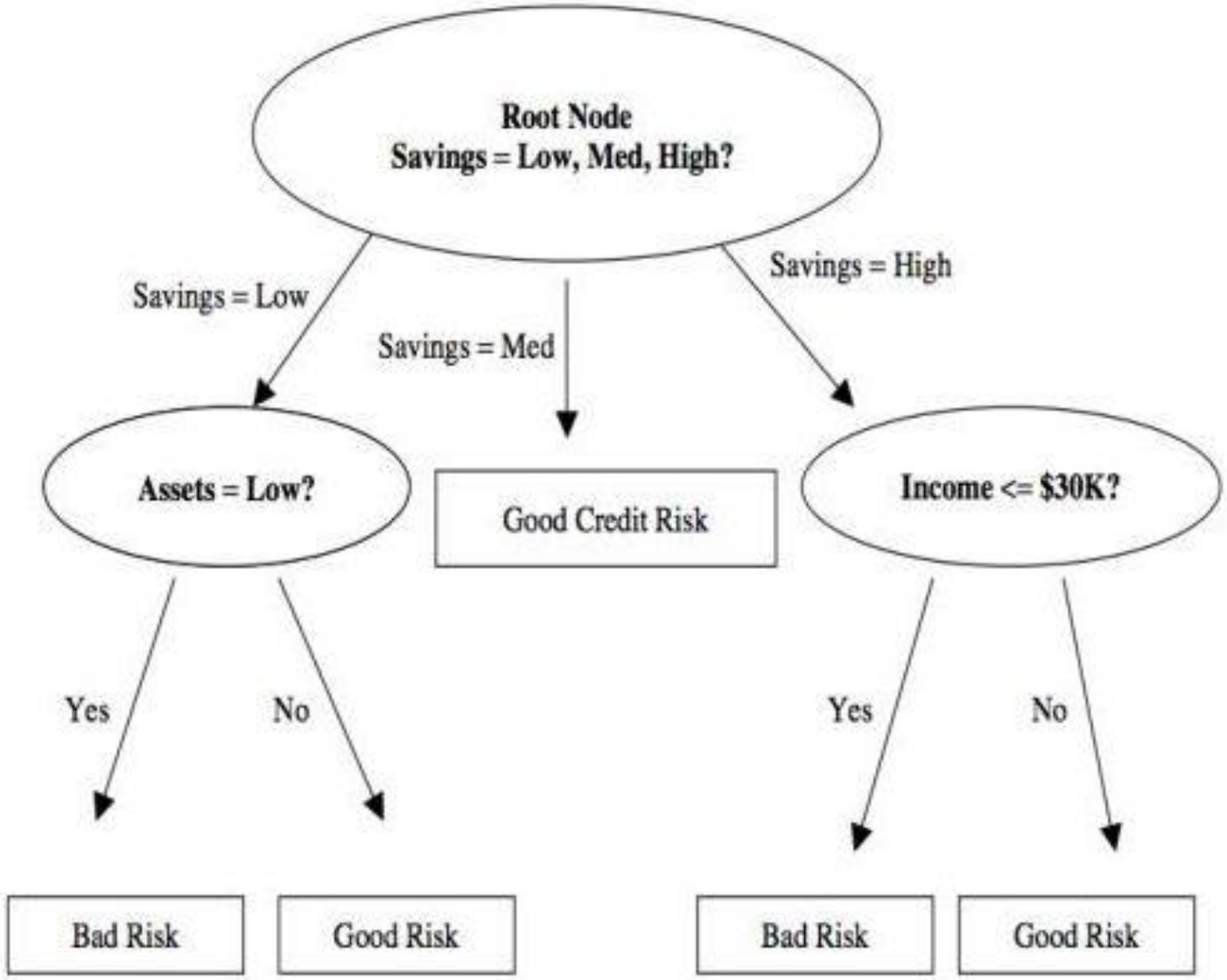
Leaves are rectangular and they contain class/ real value/ probabilities.

By using features present in the instance space, a feature tree is constructed.

- If value is a class then the feature tree is Decision Tree

We use Divide and Conquer approach for making a tree. We choose an attribute/ feature which gives us the best possible split (with least entropy)

Decision Tree



Feature List

We convert a Feature tree into a Feature list by using **if-else** statements

```
if(saving = medium): then credit risk = good
else
      if (saving =high):
            Then
                  if(income <300$): then credit risk = good
                  Else credit risk = bad
      Else
            if(asset=low):
                  Then credit risk =bad
            Else credit risk = good
```

Feature List

if(saving = medium): y = good

else if(saving = low) Λ (asset = low): y = bad

…

And so on…

This is known as **Rule Learning**

This is also learned in top-down fashion

It works on separate and conquer fashion

What is feature selection?

- Consider our training data as a matrix where each row is a vector and each column is a dimension.
- For example consider the matrix for the data x_1=(1, 10, 2), x_2=(2, 8, 0), and x_3=(1, 9, 1)
- We call each dimension a feature or a column in our matrix.

Feature selection

- Useful for high dimensional data such as genomic DNA and text documents.
- Methods
  - Univariate (looks at each feature independently of others)
    - Pearson correlation coefficient
    - F-score
    - Chi-square
    - Signal to noise ratio
    - And more such as mutual information, relief
  - Multivariate (considers all features simultaneously)
    - Dimensionality reduction algorithms
    - Linear classifiers such as support vector machine
    - Recursive feature elimination

Feature selection

- Methods are used to rank features by importance
- Ranking cut-off is determined by user
- Univariate methods measure some type of correlation between two random variables. We apply them to machine learning by setting one variable to be the label (y_i) and the other to be a fixed feature (x_{ij} for fixed j)

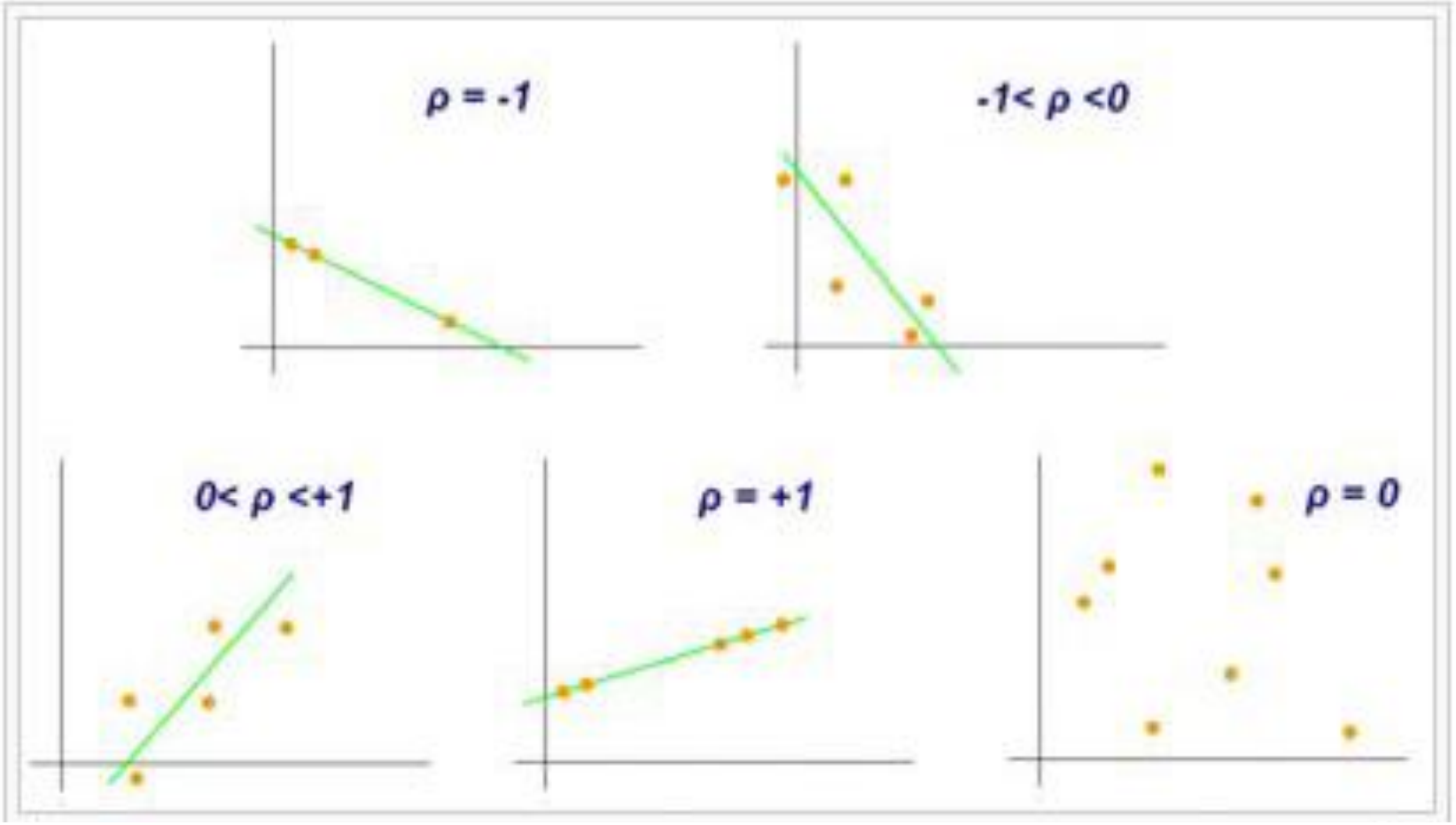
Pearson correlation coefficient

- Measures the correlation between two variables
- Formulas:
  - Covariance(X,Y) = $E((X-\mu_X)(Y-\mu_Y))$
  - Correlation(X,Y)= Covariance(X,Y)/$\sigma_X\sigma_Y$
  - Pearson correlation =

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- The correlation r is between -1 and 1. A value of 1 means perfect positive correlation and -1 in the other direction

Pearson correlation



From Wikipedia

F-score

F-score is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors $\boldsymbol{x}_k, k = 1, \ldots, m$, if the number of positive and negative instances are n_+ and n_-, respectively, then the F-score of the ith feature is defined as:

$$F(i) \equiv \frac{\left(\bar{\boldsymbol{x}}_i^{(+)} - \bar{\boldsymbol{x}}_i\right)^2 + \left(\bar{\boldsymbol{x}}_i^{(-)} - \bar{\boldsymbol{x}}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{\boldsymbol{x}}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{\boldsymbol{x}}_i^{(-)}\right)^2}, \qquad (4)$$

where $\bar{\boldsymbol{x}}_i$, $\bar{\boldsymbol{x}}_i^{(+)}$, $\bar{\boldsymbol{x}}_i^{(-)}$ are the average of the ith feature of the whole, positive, and negative data sets, respectively; $x_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,i}^{(-)}$ is the ith feature of the kth negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative. Therefore, we use this score as a feature selection criterion.

Fr, Kanpur

Dr. Rashi Agarwal, HBTU, Kanpur

Yu-Wei, Chang and Chen, Feature extraction, 2006

Chi-square test

- We have two random variables:
  - Label (L): 0 or 1
  - Feature (F): Categorical
- Null hypothesis: the two variables are independent of each other (unrelated)
- Under independence
  - P(L,F)= P(D)P(G)
  - P(L=0) = (c1+c2)/n
  - P(F=A) = (c1+c3)/n
- Expected values
  - E(X1) = P(L=0)P(F=A)n
- We can calculate the chi-square statistic for a given feature and the probability that it is independent of the label (using the p-value).
- Features with very small probabilities deviate significantly from the independence assumption and therefore considered important.

Contingency table

	Feature=A	Feature=B
Label=0	Observed=c1 Expected=X1	Observed=c2 Expected=X2
Label=1	Observed=c3 Expected=X3	Observed=c4 Expected=X4

Signal to noise ratio

- Difference in means divided by difference in standard deviation between the two classes
- $S2N(X,Y) = (\mu_X - \mu_Y)/(\sigma_X - \sigma_Y)$
- Large values indicate a strong correlation

Multivariate feature selection

- Consider the vector w for any linear classifier.
- Classification of a point x is given by wTx+w0.
- Small entries of w will have little effect on the dot product and therefore those features are less relevant.
- For example if w = (10, .01, -9) then features 0 and 2 are contributing more to the dot product than feature 1. A ranking of features given by this w is 0, 2, 1.

Multivariate feature selection

- The w can be obtained by any of linear classifiers we have see in class so far
- A variant of this approach is called recursive feature elimination:
  - Compute w on all features
  - Remove feature with smallest w_i
  - Recompute w on reduced data
  - If stopping criterion not met then go to step 2

Feature selection in practice

- NIPS 2003 feature selection contest
  - Contest results
  - Reproduced results with feature selection plus SVM
- Effect of feature selection on SVM
- Comprehensive gene selection study comparing feature selection methods
- Ranking genomic causal variants with SVM and chi-square

Limitations

- Unclear how to tell in advance if feature selection will work
  - Only known way is to check but for very high dimensional data (at least half a million features) it helps most of the time
- How many features to select?
  - Perform cross-validation