Example numerical on KNN classification

Name	Acid durability	Strength	Class
Type-1	7	7	Bad
Type-2	7	4	Bad
Type-3	3	4	Good
Type-4	1	4	Good

Test data → Acid durability = 3
Strength = 7
Class = ?

Name	Acid durability	Strength	class	Distance
Type-1	7	7	Bad	$\sqrt{(7-3)^2+(7-7)^2} = 4$
Type-2	7	4	Bad	$\sqrt{(7-3)^2+(7-4)^2} = 5$
Type-3	3	4	Good	$\sqrt{(3-3)^2+(7-4)^2} = 3$
Type-4	1	4	Good	$\sqrt{(3-1)^2+(7-4)^2} = 3.6$

Write the Rank for the records based on the distance between them.

Type	Acid durability	Strength	class	distance	Rank
Type-1	7	7	Bad	4	3
-2	7	4	Bad	5	4
-3	3	4	Good	3	1
-4	1	4	Good	3.6	2

For.

K = 1

Test data belongs to class "Good".

K = 2

Test data belongs to class "Good".

CSE319

Machine Learning Module 1

CONTENTS

- Introduction to Machine learning
- Types of Machine Learning
- Models selection and generalization
- Machine Learning concept work flow
- Applications

Introduction to Machine learning

What is Learning?

- Herbert Simon: “Learning is any process by which a system improves performance from experience.”
- What is the task?
  - Classification
  - Categorization/clustering
  - Problem solving / planning / control
  - Prediction

Why “Learn” ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

What We Talk About When We Talk About"Learning"

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

  People who bought "Blink" also bought "Outliers" (www.amazon.com)

- Build a model that is *a good and useful approximation* to the data.

Data Mining

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Control, robotics, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Spam filters, intrusion detection
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- …

What is Machine Learning?

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

It provides us statistical tools to explore and analyses the data

Applications

- Association
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

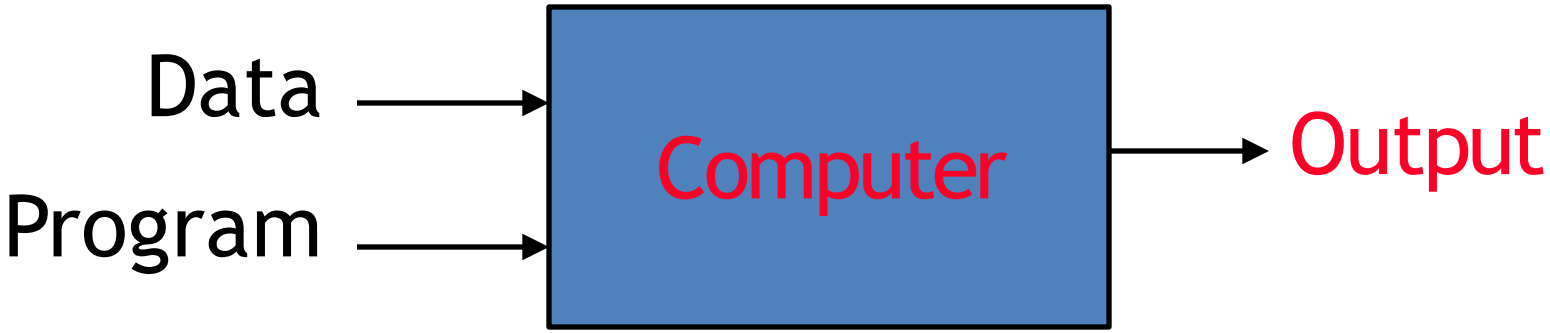
Why Study Machine Learning?
Developing Better Computing Systems

- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (*knowledge engineering bottleneck*).
- Develop systems that can automatically adapt and customize themselves to individual users.
  - Personalized news or mail filter
  - Personalized tutoring
- Discover new knowledge from large databases (*data mining*).
  - Market basket analysis (e.g. diapers and beer)
  - Medical text mining (e.g. migraines to calcium channel blockers to magnesium)

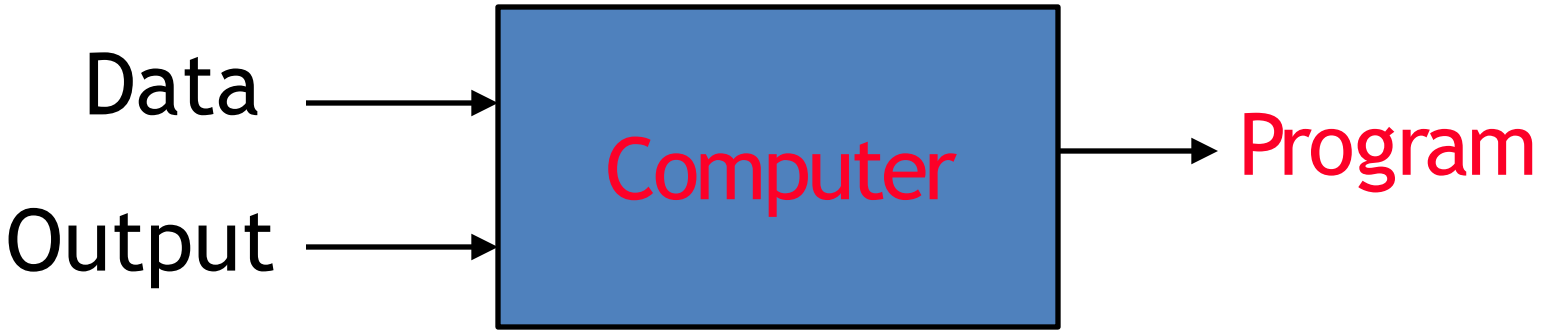
Related Disciplines

- Artificial Intelligence
- Data Mining
- Probability and Statistics
- Information theory
- Numerical optimization
- Computational complexity theory
- Control theory (adaptive)
- Psychology (developmental, cognitive)
- Neurobiology and many more

Traditional Programming



Machine Learning



Human Learning

- It is a process of gaining information through observation, to solve with real world scenarios.
- Human learning happens in one of the three ways:
  - An expert in the subject directly teaches us
  - We build our own notion indirectly based on what we have learnt in the past
  - We learn ourselves, may be after multiple attempts.

Types of Human Learning

- Learning under expert guidance
- Learning guided by knowledge gained from experts
- Learning by self

MACHINE LEARNING

A computer program is said to learn from experience ‘E’ with respect to some class of tasks ‘T’ and performance measure ‘P’, if its performance at tasks in ‘T’, as measured by ‘P’, improves with experience ‘E’.

- *Tom M. Mitchell*

History of Machine Learning

- 1950s
  - Samuel's checker player
- 1960s:
  - Neural networks: Perceptron
  - Pattern recognition
  - Learning in the limit theory
  - Minsky and Papert prove limitations of Perceptron
- 1970s:
  - Symbolic concept induction
  - Winston's arch learner
  - Expert systems and the knowledge acquisition bottleneck
  - Quinlan's ID3
  - Michalski's AQ and soybean diagnosis

History of Machine Learning (cont.)

- 1980s:
  - Advanced decision tree and rule learning
  - Explanation-based Learning (EBL)
  - Learning and planning and problem solving
  - Utility problem, Analogy
  - Cognitive architectures
  - Resurgence of neural networks (connectionism, backpropagation)
- 1990s
  - Data mining
  - Adaptive software agents and web applications
  - Text learning
  - Reinforcement learning (RL)
  - Inductive Logic Programming (ILP)
  - Ensembles: Bagging, Boosting, and Stacking
  - Bayes Net learning

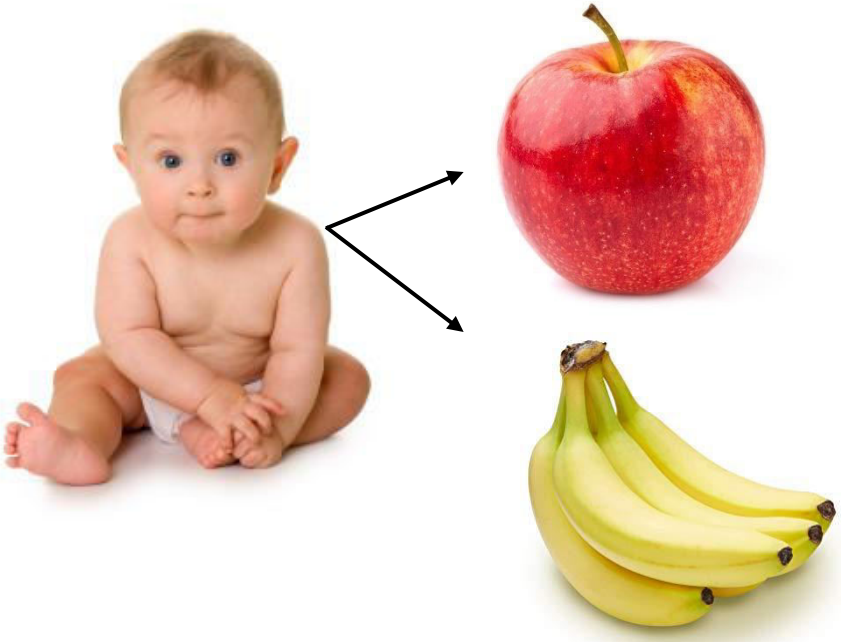
TYPES of MACHINE LEARNING

Types of Machine Learning

- **Supervised learning:** A machine predicts the class of unknown objects based on prior class-related information of similar objects. Also called predictive learning.
- **Unsupervised/clustering learning:** A machine finds patterns in unknown objects by grouping similar objects together. Also called descriptive learning.
- **Reinforcement learning:** A machine learns to act on its own to achieve the given goals.

SUPERVISED LEARNING

- The major motivation of supervised learning is to learn from past information.
- But how do machine learns?
  - By TRAINING DATA using labels.



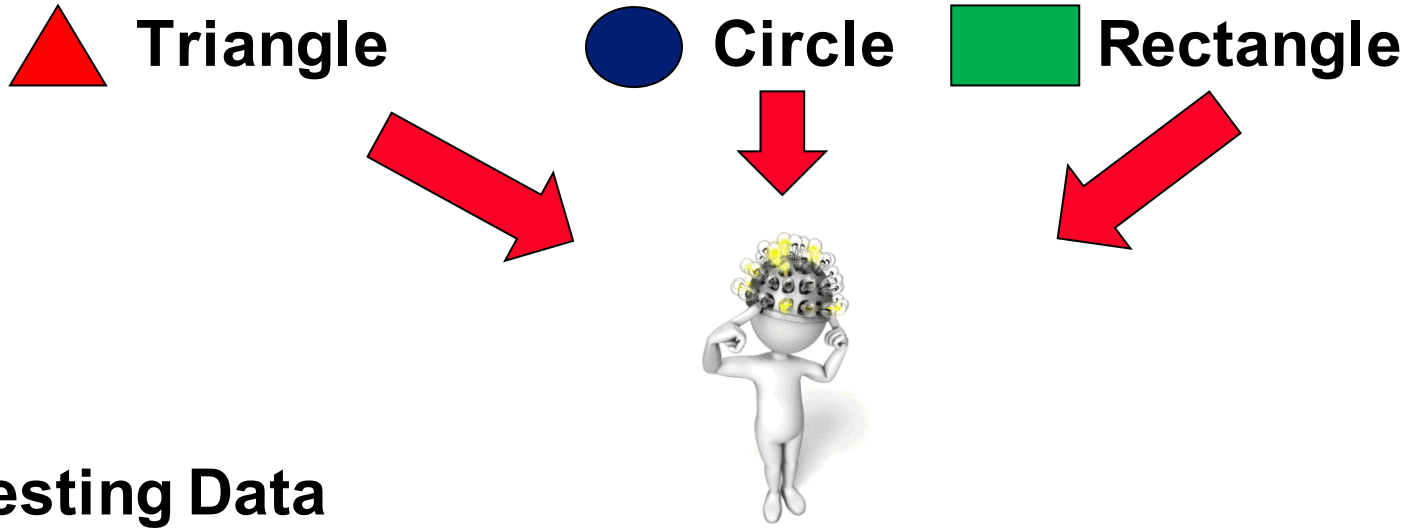
Apple

Banana

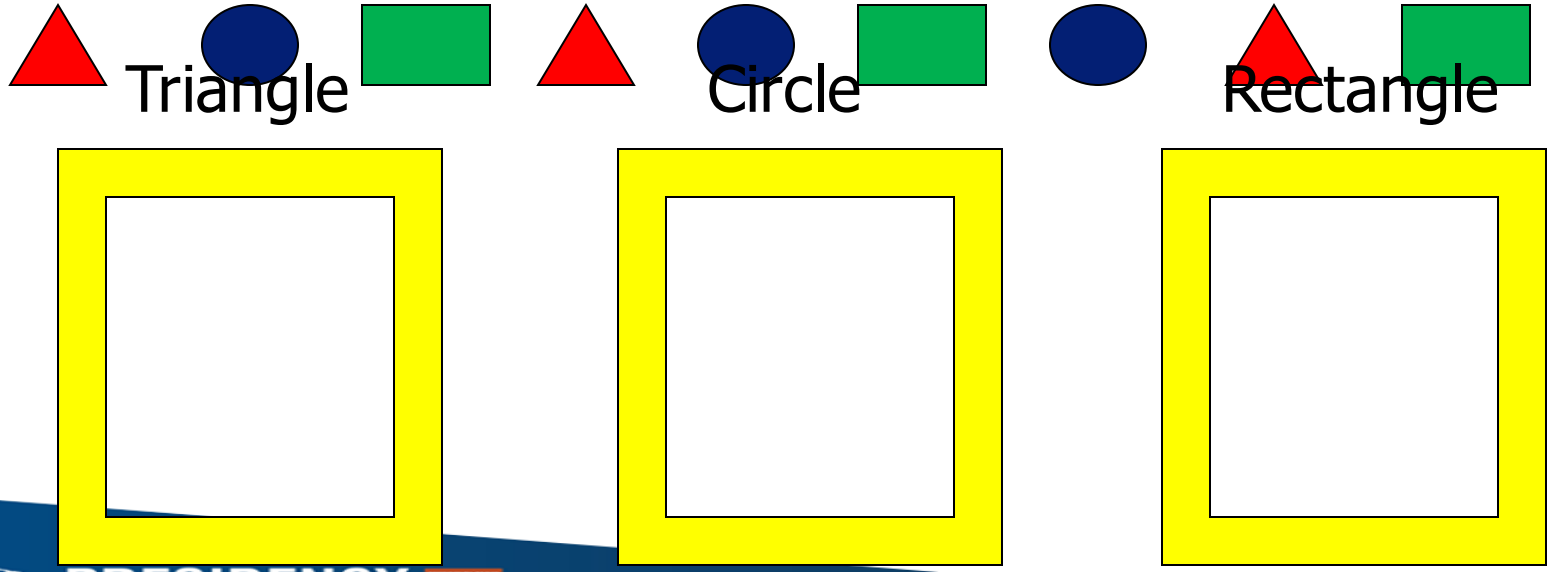
What's this?

It's an Apple

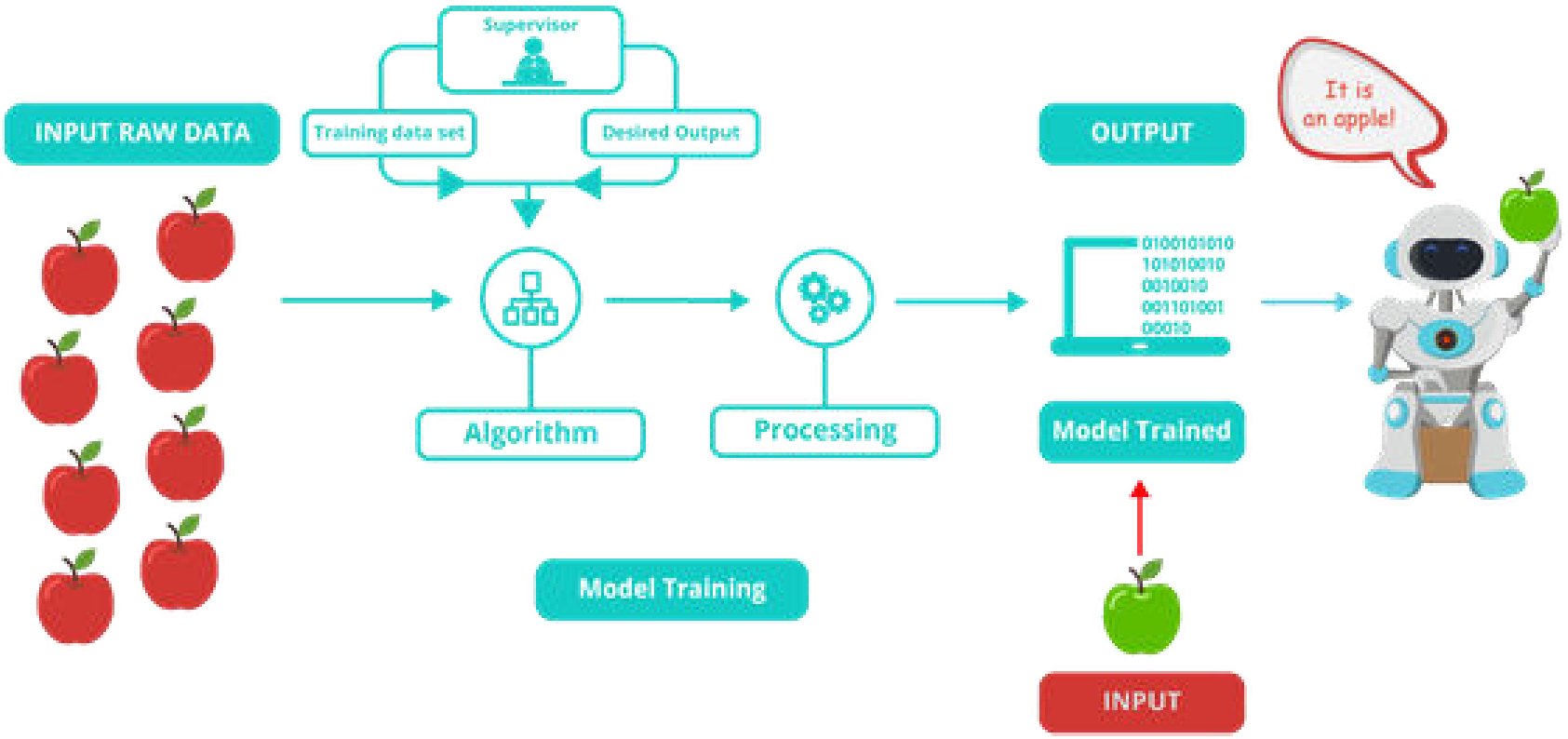
Training Data



Testing Data



Supervised Learning



Examples of Supervised Learning

- Predicting the results of a game.
- Predicting a tumour is malignant or benign.
- Predicting the piece of domains like real estate, stocks, etc.
- Classify texts such as classifying a set of emails as spam or non-spam.

Classification & Regression

- When we are trying to predict a categorical or nominal variable, the problem is known as a classification problem.
  - Ex: Identify Cat or Dog
- When we are trying to predict a real-valued variable, the problem falls under the category of regression.
  - Ex: Predict the value of a property.

Unsupervised Learning

- Unsupervised learning(USL) is a type of self-organized learning that helps find previously unknown patterns in data set without pre-existing labels.
- The objective is to take a dataset as input and try to find natural grouping or patterns within the data elements or records.
- Hence, USL is termed as Descriptive Model and the process of USL is called Pattern or Knowledge Discovery.
- In unsupervised learning, the system is presented with unlabeled, uncategorized data and the system's algorithms act on the data without prior training. The output is dependent upon the coded algorithms.

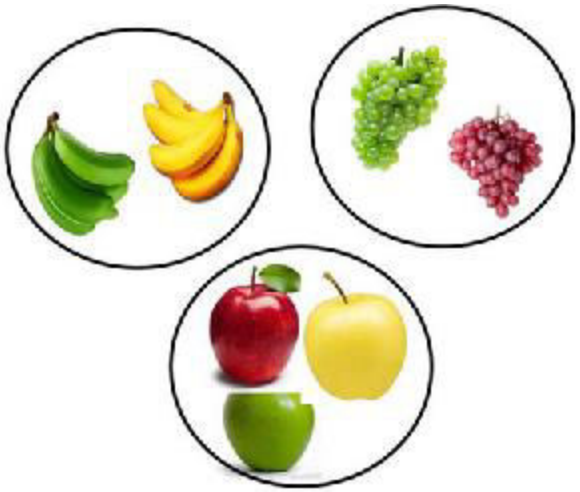
Unsupervised Learning

- Clustering is the main type of Unsupervised Learning.
- Clustering groups similar objects together.

Input Data

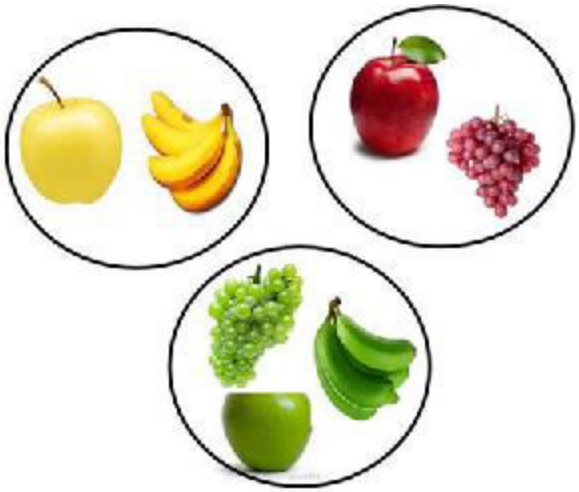
(a)

Cluster by type



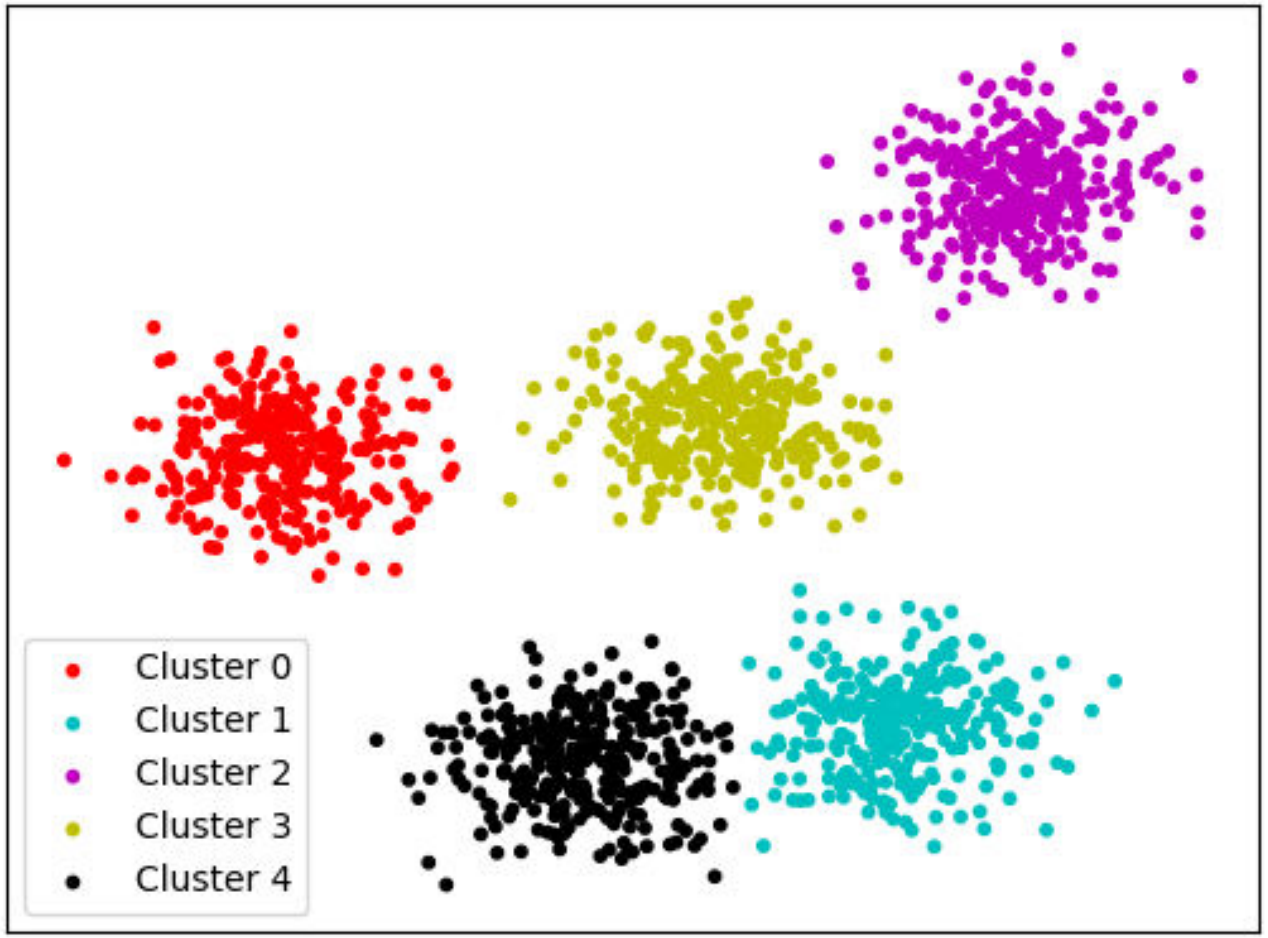
(b)

Cluster by color

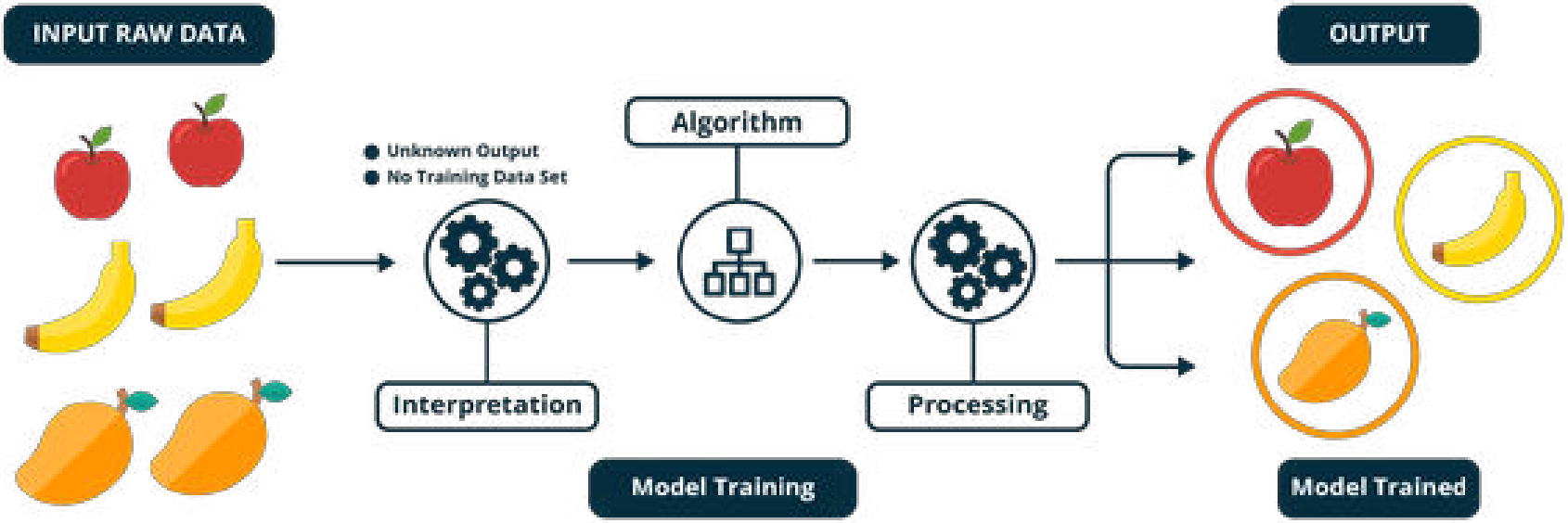


(c)

Sample Clustering



Unsupervised Learning



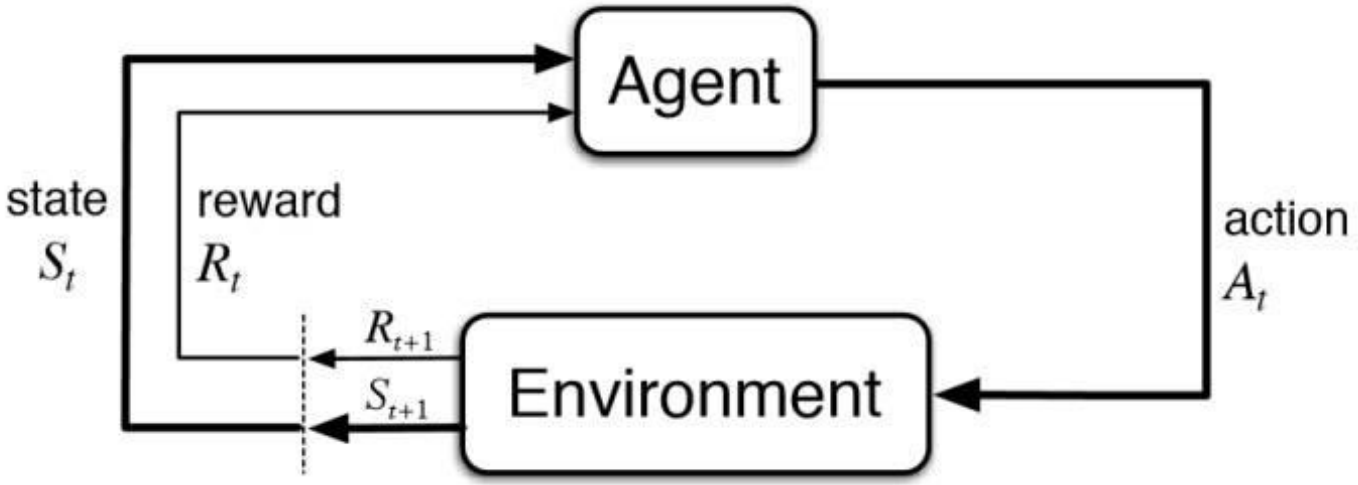
Unsupervised Learning Example: Categorize Cats and Dogs

Reinforcement Learning

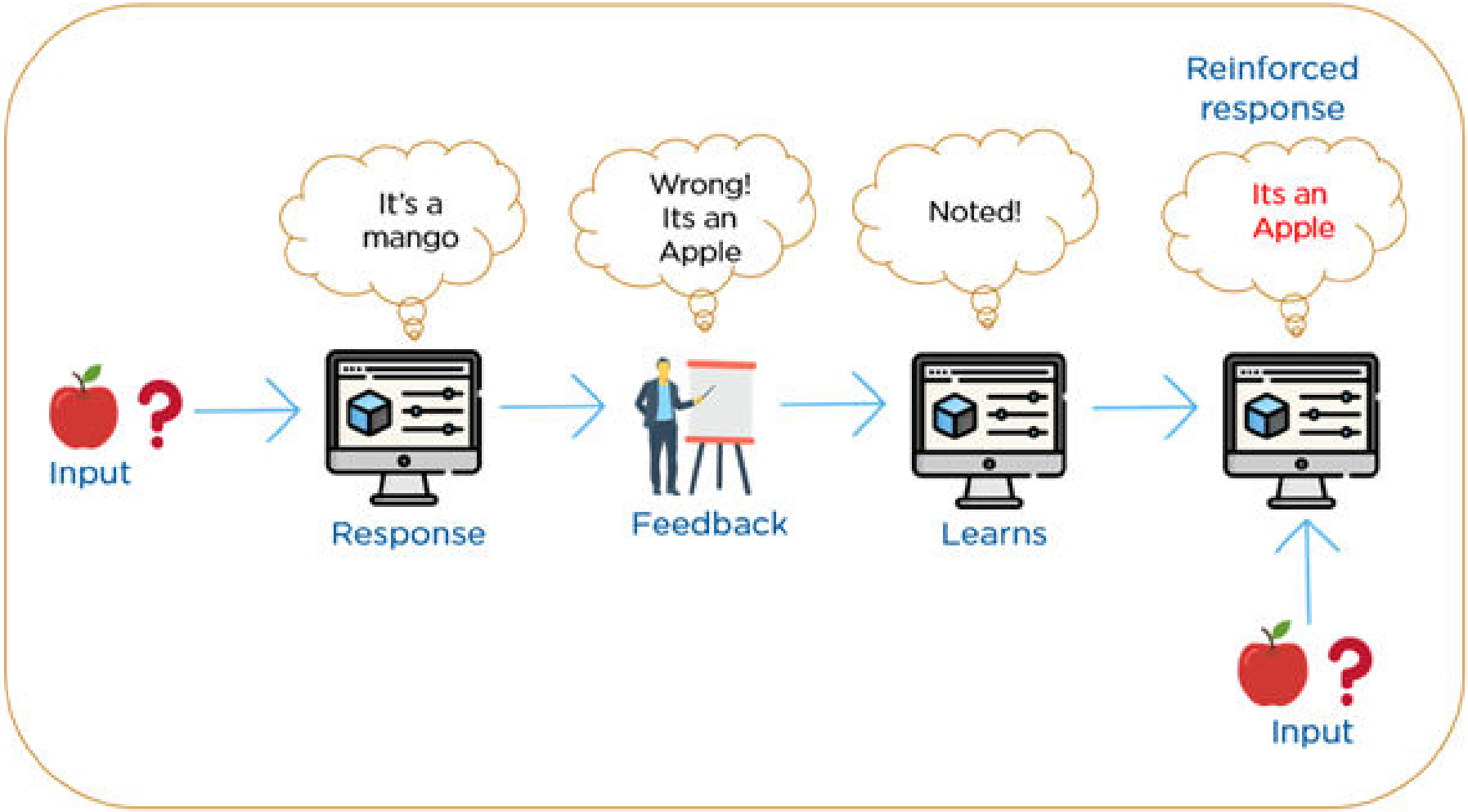
- It is about taking suitable action to maximize reward in a particular situation.
- It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation.
- Close to human learning.

Cont'd

- Algorithm learns a policy of how to act in a given environment.
- Every action has some impact in the environment, and the environment provides rewards that guides the learning algorithm.



Reinforcement Learning



Different Varieties of Machine Learning

- Concept Learning
- Clustering Algorithms
- Connectionist Algorithms
- Genetic Algorithms
- Explanation-based and Transformation-based Learning
- Reinforcement and Case-based Learning
- Macro Learning
- Evaluation Functions
- Cognitive Learning Architectures
- Constructive Induction
- Discovery Systems

Languages or Tools for Machine Learning

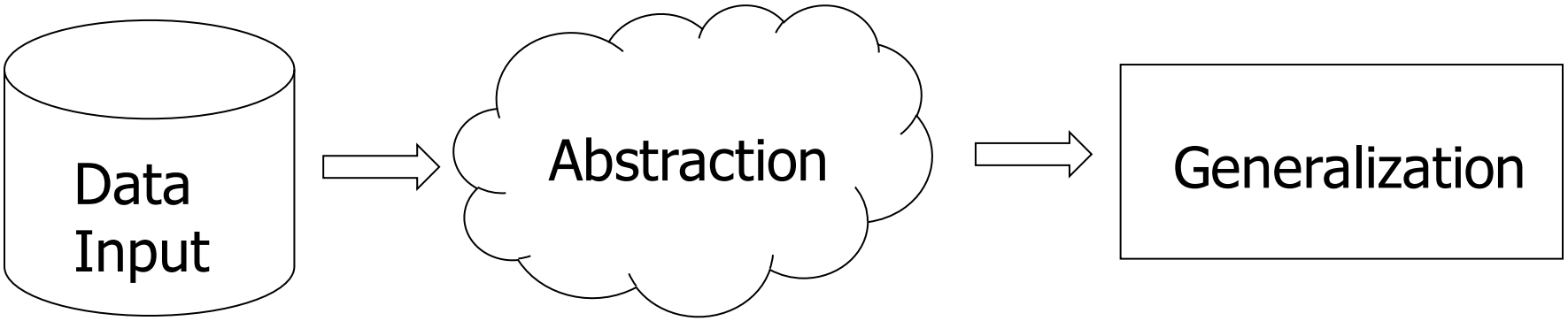
- **Python –** Open source programming language adopted for machine learning.
- **R –** Open source software. Used for statistical computing and data analysis
- **Matlab -** Developed by MathWorks. Licensed version. Used for variety of applications.
- **SAS –** Statistical Analysis System, was developed and licensed by SAS Institute provides strong support for ML.
- **Others-**
  - **SPSS(Statistical Package for the Social Sciences) – IBM**
  - **Julia – MIT(Massachusetts Institute of Technology)**

History of Machine Learning (cont.)

- 2000s
  - Support vector machines
  - Kernel & Graphical models
  - Statistical relational and Transfer learning
  - Sequence labeling
  - Collective classification and structured outputs
  - Computer Systems Applications
    - Compilers
    - Debugging
    - Graphics
    - Security (intrusion, virus, and worm detection)
  - E-mail management
  - Personalized assistants that learn
  - Learning in robotics and vision

MODEL SELECTION and GENERALIZATION

How do Machines Learn?



1. **Data Input: Past data or information is utilized as a basis for future decision-making**
2. **Abstraction: The input data is represented in a broader way through the underlying algorithm.**
3. **Generalization: The abstracted representation is generalized to form a framework for making decisions.**

Abstraction

- The data, given as input, cannot be used in the original shape and form.
- Abstraction helps in deriving a conceptual map based on the input data.
- The model may be in any one of forms:
  - Computational blocks like if/else rules
  - Mathematical equations
  - Specific data structures like trees or graphs
  - Logical groupings of similar observations

Abstraction Cont'd

- The choice of the model is human specific.
- Selection of model is based on:
  - The type of problem to be solved.
  - Nature of the input data.
  - Domain of the problem.

Generalization

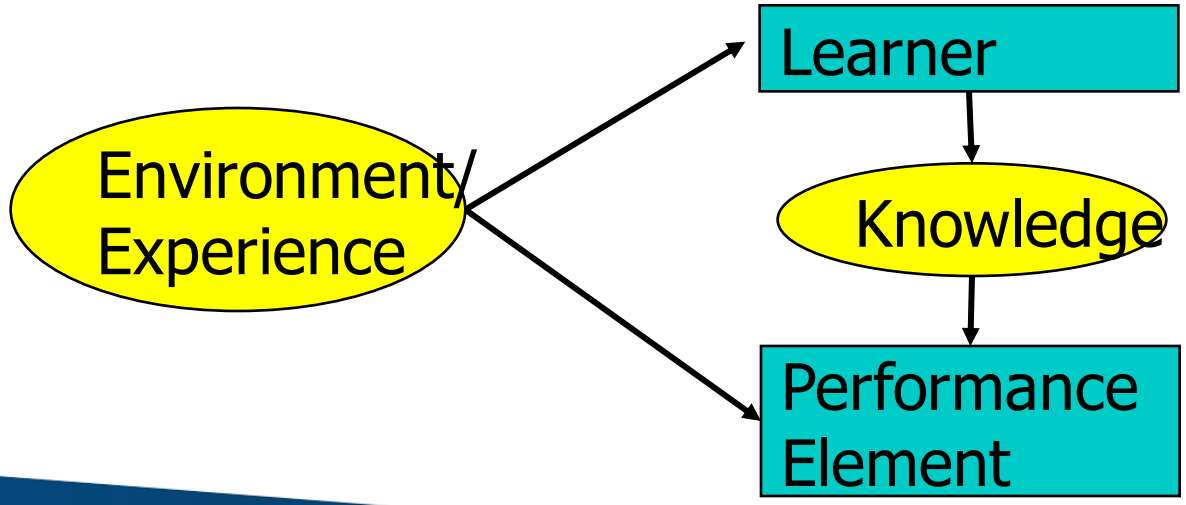
- Generalization is used for taking the decisions after training the model.
- The model is trained for a limited set of data. If we want to apply the model to take decision on a set of unknown data, we may encounter following problems:
  - The trained model is aligned with training data too much, hence may not represent the actual trend.
  - The test data possess certain characteristics apparently unknown to the training data.

Well-posed learning problem

- A framework can be designed for deciding whether a problem can be solved using ML. The framework should answer:
  - What is the problem?
  - Why does the problem need to be solved?
  - How to solve the problem?

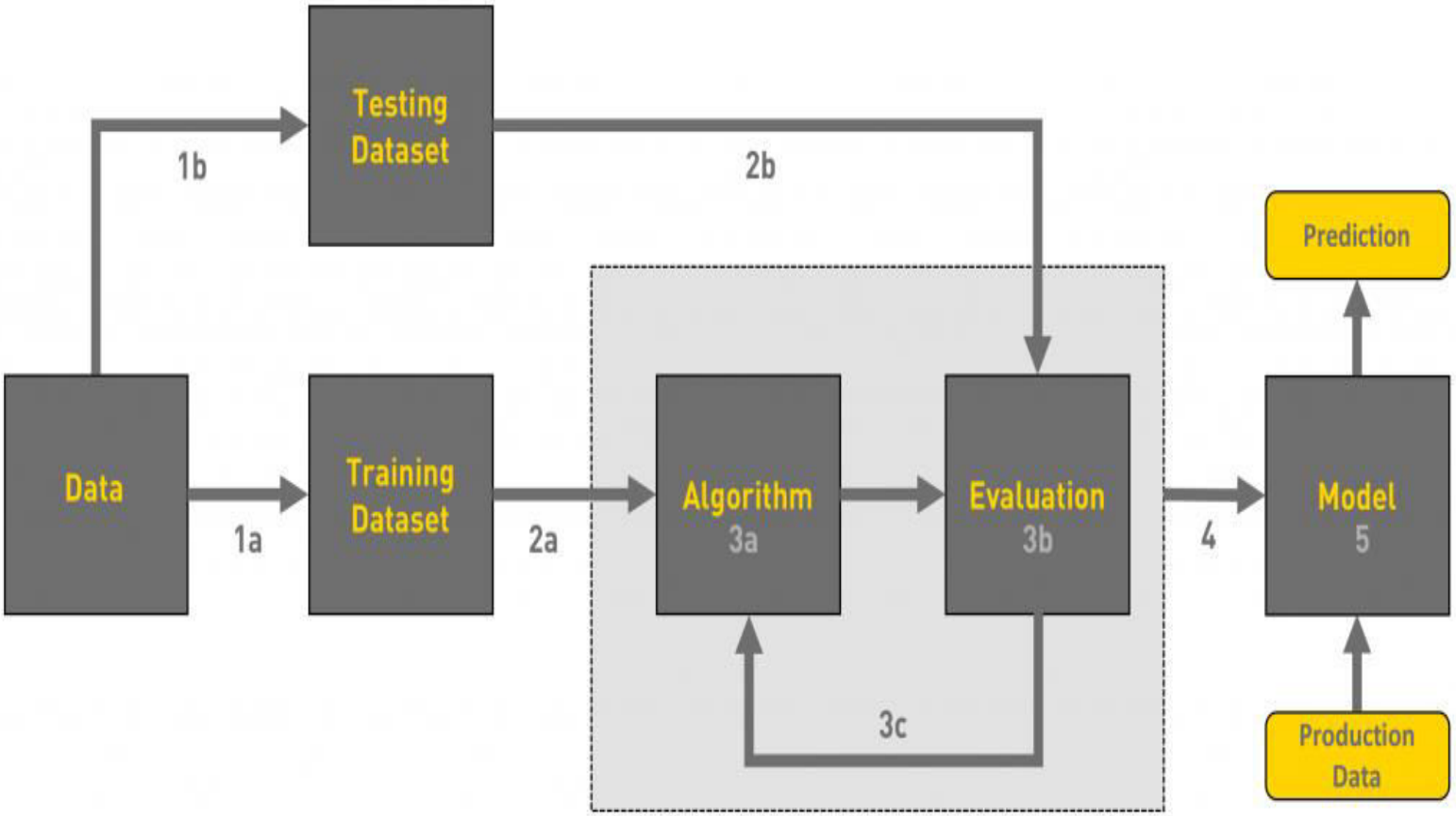
Designing a Learning System

- Choose the training experience
- Choose exactly what is too be learned, i.e. the *target function*.
- Choose how to represent the target function.
- Choose a learning algorithm to infer the target function from the experience.



The Machine Learning Workflow

ML WorkFlow


Testing Dataset
1b
2b
Prediction
Data
1a
Training Dataset
2a
Algorithm
3a
Evaluation
3b
4
Model
5
3c
Production Data

ML WorkFlow

We can define the machine learning workflow in 3 stages.

1. Gathering data
2. Data pre-processing
3. Researching the model that will be best for the type of data
4. Training and testing the model
5. Evaluation

Gathering Data

- The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

- We can also use some free data sets which are present on the internet. **Kaggle** and **UCI Machine learning Repository** are the repositories that are used the most for making Machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms, they also host competitions in which people can participate and get to test their knowledge of machine learning.

Data pre-processing

- Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.
- **What is data pre-processing?**
- Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.
- **Why do we need it?**
- As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

Data pre-processing

- Most of the real-world data is messy, some of these types of data are:
- 1. **Missing data:** Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).
- 2. **Noisy data:** This type of data is also called outliners, this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.
- 3. **Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of Data

- 1. Numeric e.g. income, age
- 2. Categorical e.g. gender, nationality
- 3. Ordinal e.g. low/medium/high

Data pre-processing

- **How can data pre-processing be performed?**

These are some of the basic pre — processing techniques that can be used to convert raw data.

1. **Conversion of data:** As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

2. **Ignoring the missing values:** Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in the dataset.

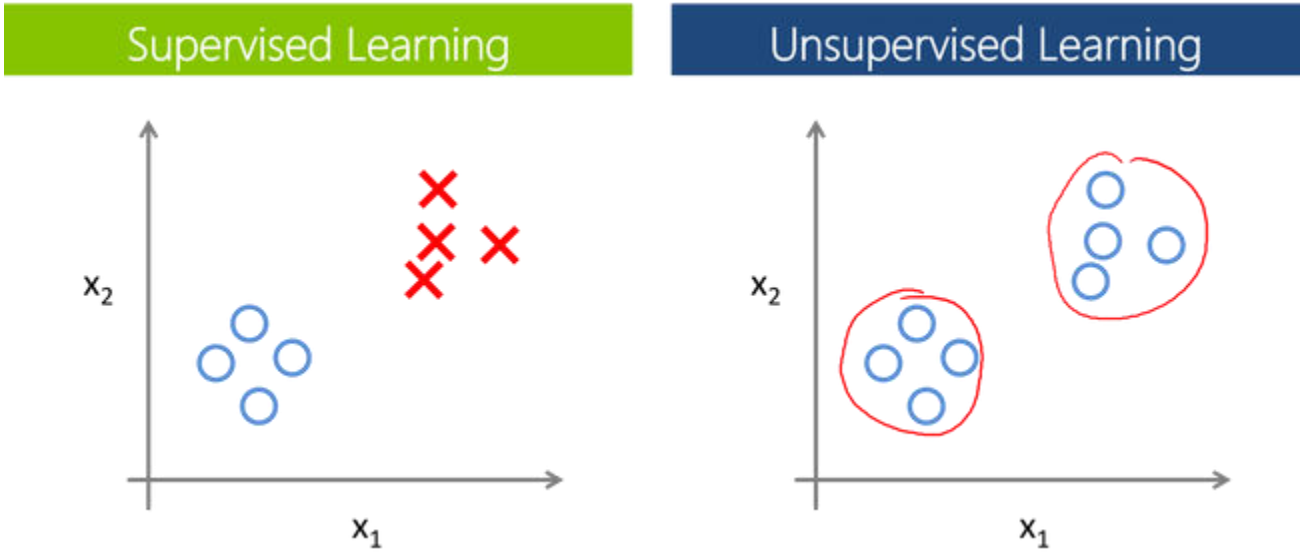
3. **Filling the missing values:** Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

4. **Machine learning:** If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.

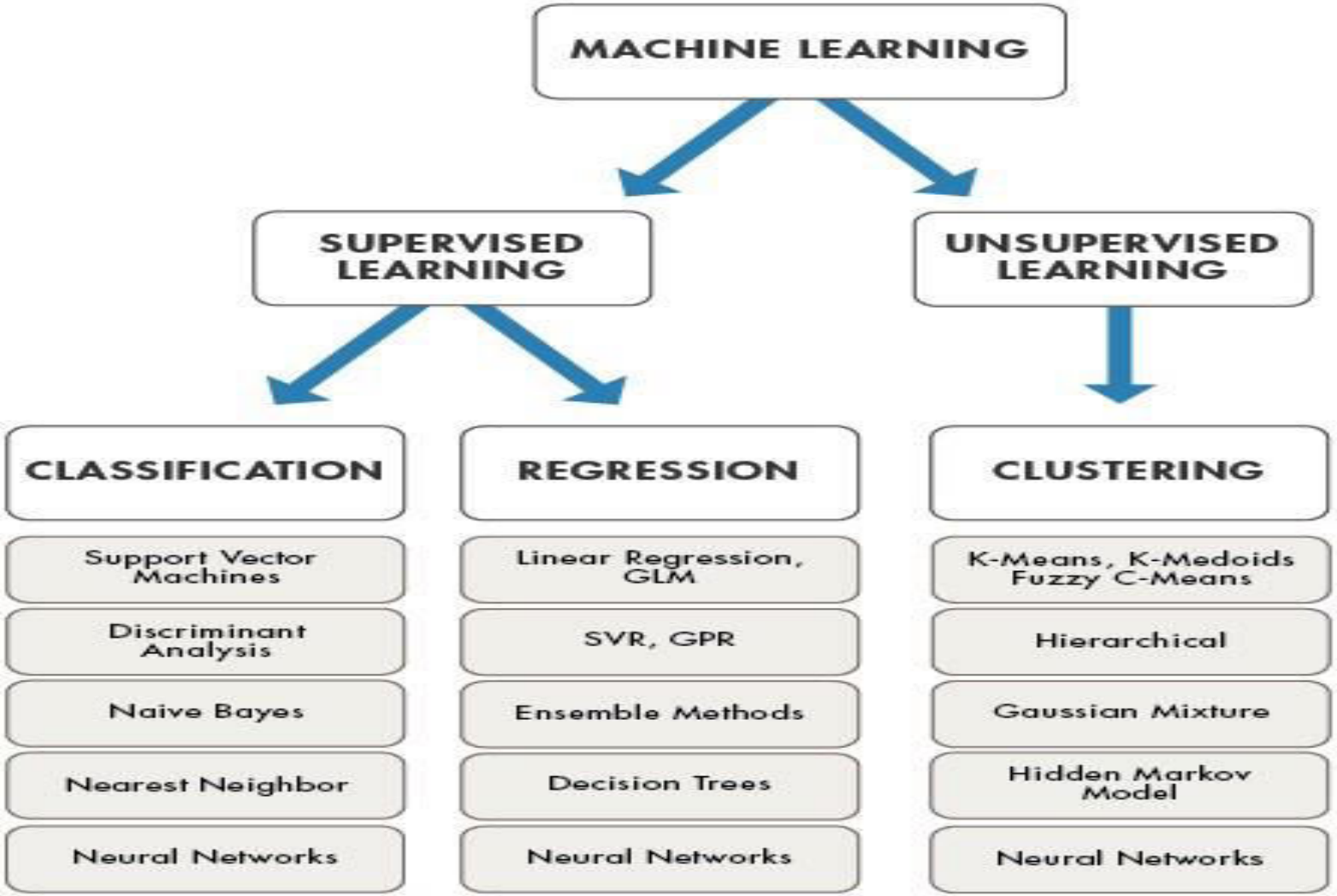
5. **Outliers detection:** There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

Researching the model that will be best for the type of data

- Our main goal is to train the best performing model possible, using the pre-processed data.

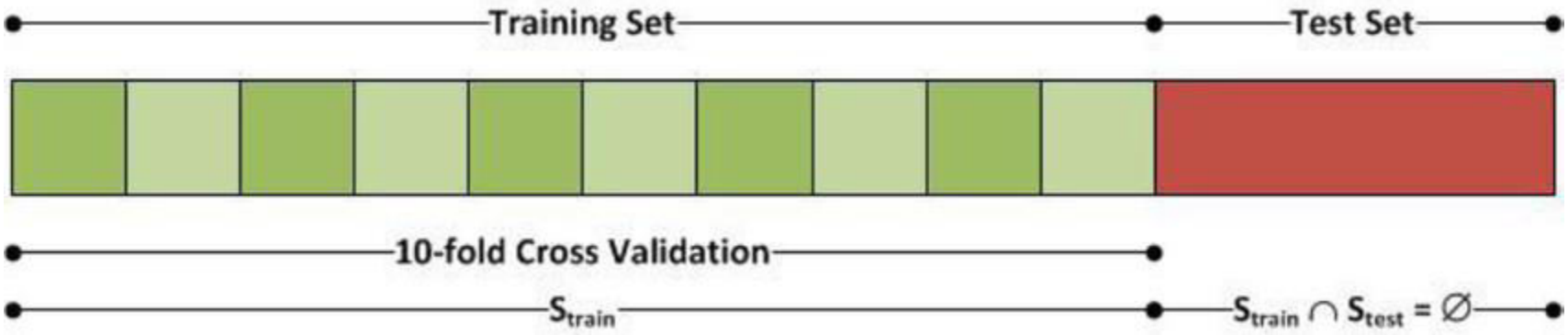


Researching the model that will be best for the type of data



Training and testing the model

- For training a model we initially split the model into 3 three sections which are '**Training data**' ,'**Validation data**' and '**Testing data**'.
- You train the classifier using '**training data set**', tune the parameters using '**validation set**' and then test the performance of your classifier on unseen '**test data set**'. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier



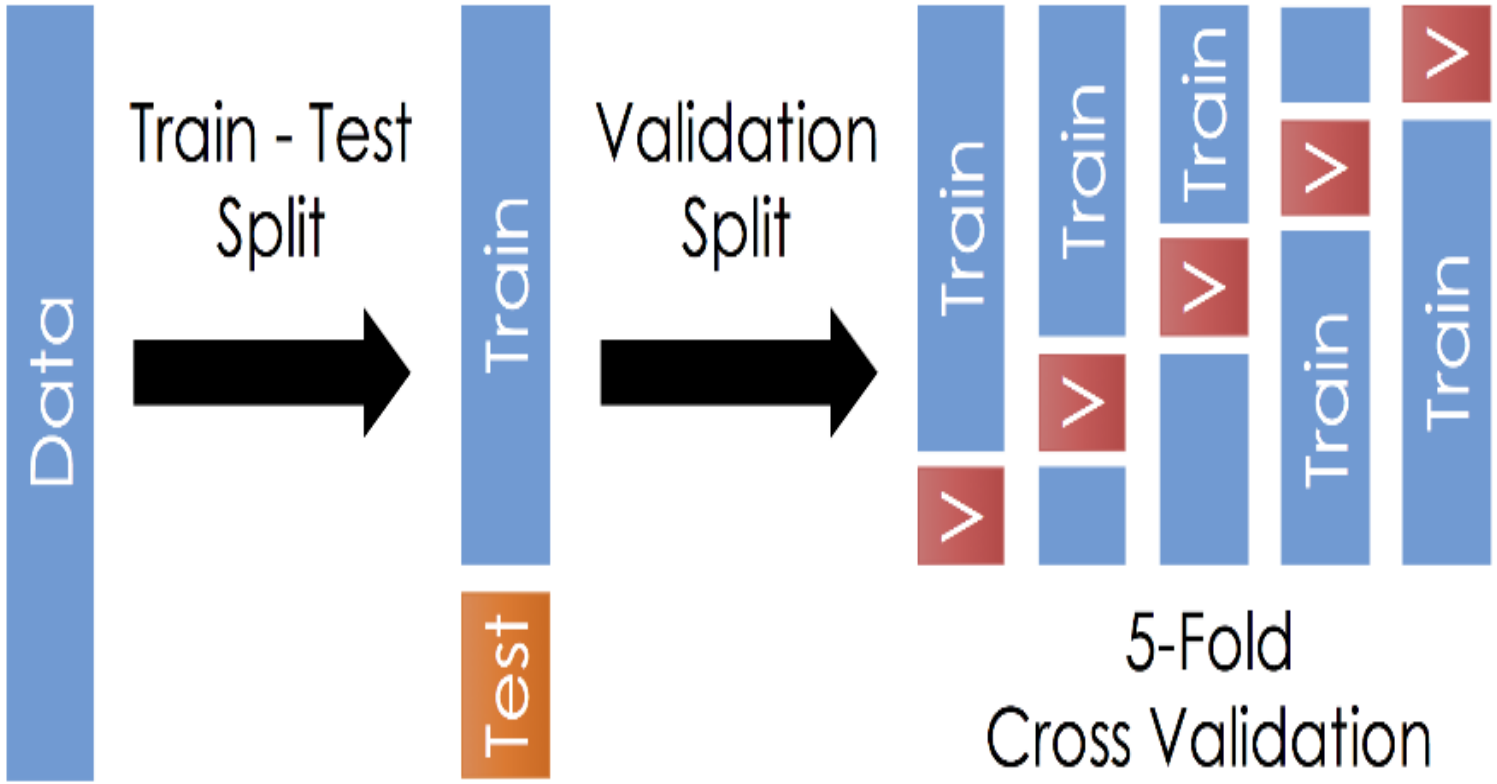
Training and testing the model

- **Training set:** The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning, that is to fit the parameters of the classifier.
- **Validation set:** Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.
- **Test set:** A set of unseen data used only to assess the performance of a fully-specified classifier

Once the data is divided into the 3 given segments we can start the training process.

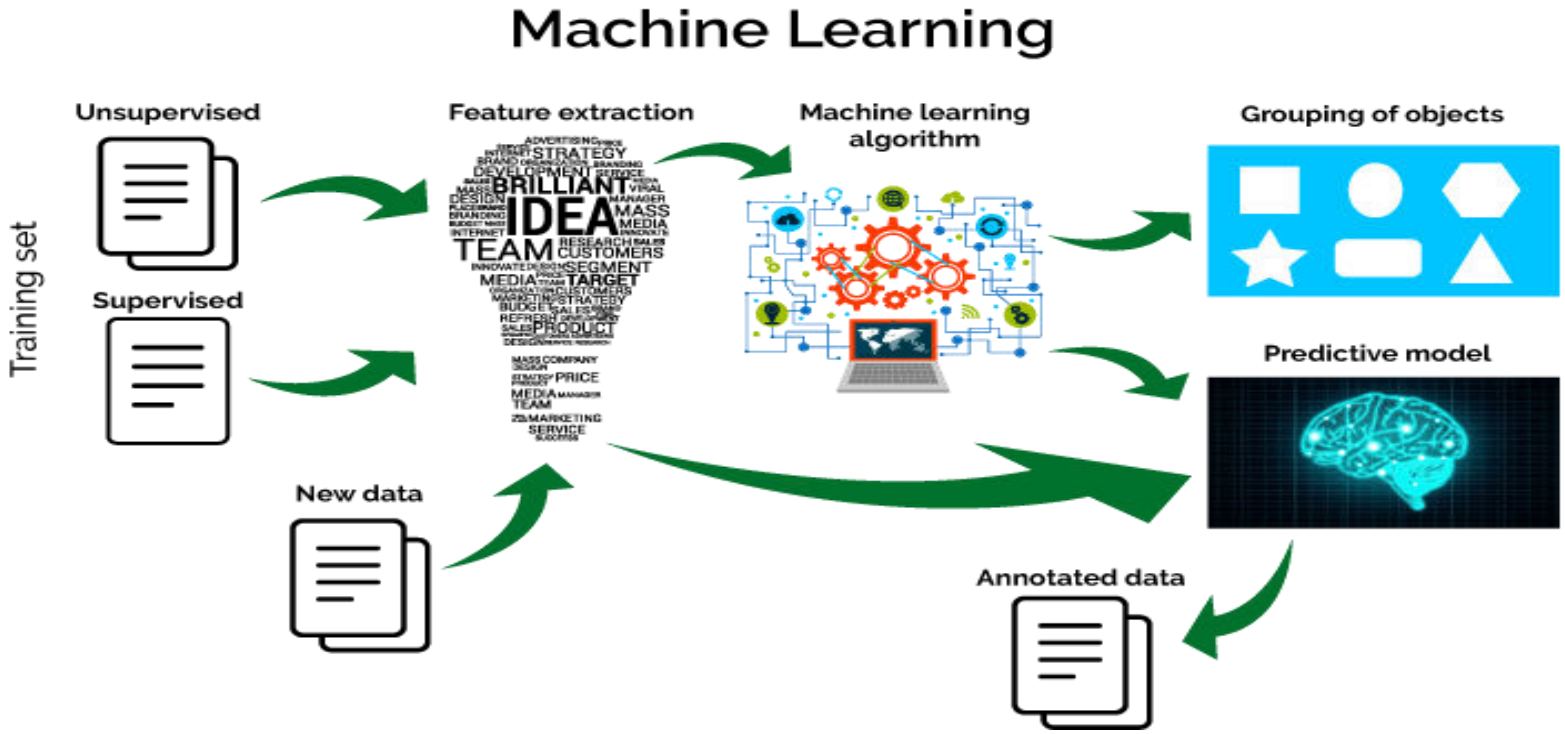
In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use ‘test set’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

Training and testing the model



EVALUATION

- Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives.



Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Debugging software
- [Your favorite area]

Supervised Learning Classification

- **Example: Cancer diagnosis**

Patient ID	# of Tumors	Avg Area	Avg Density	Diagnosis
1	5	20	118	Malignant
2	3	15	130	Benign
3	7	10	52	Benign
4	2	30	100	Malignant

Training Set

- Use this training set to learn how to classify patients where diagnosis is not known:

Patient ID	# of Tumors	Avg Area	Avg Density	Diagnosis
101	4	16	95	?
102	9	22	125	?
103	1	14	80	?

Test Set

Input Data Classification

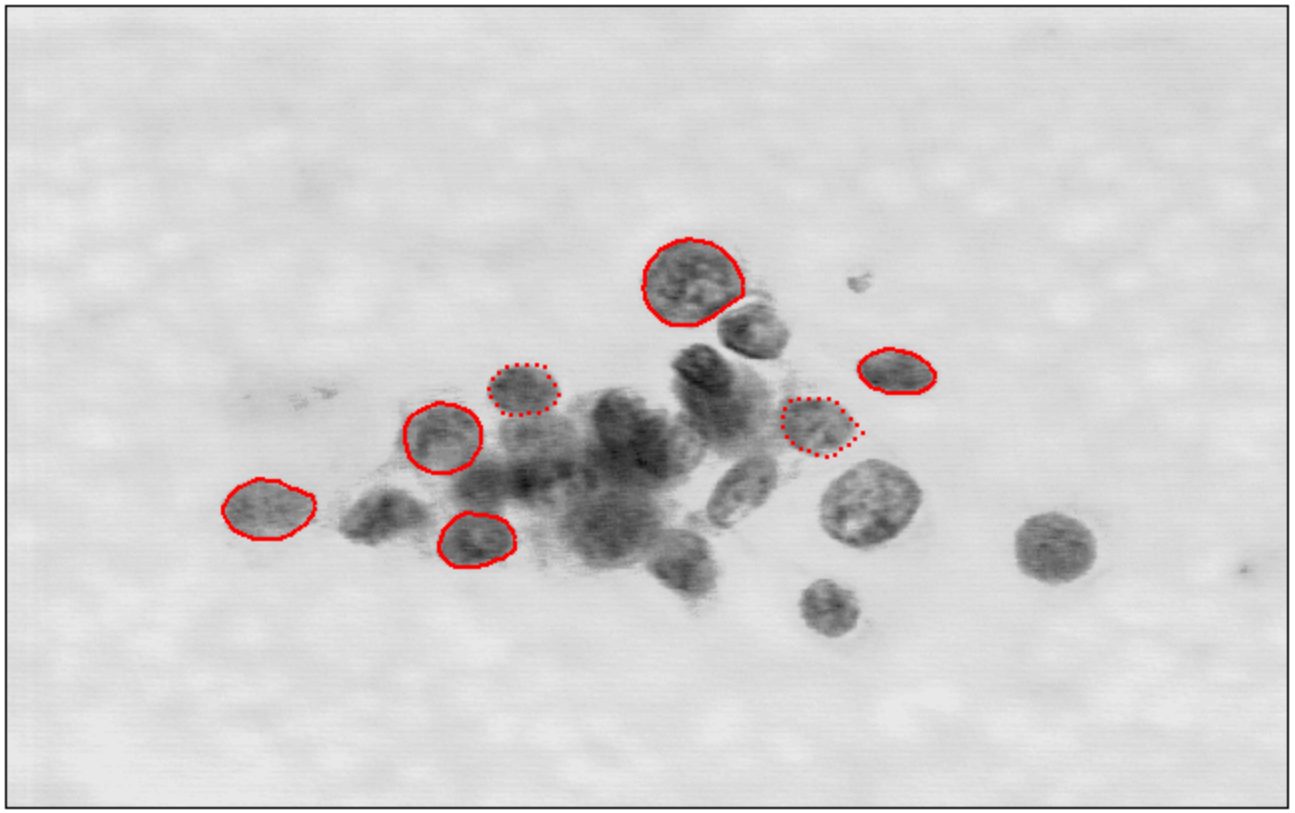
- The input data is often easily obtained, whereas the classification is not.

Classification Problem

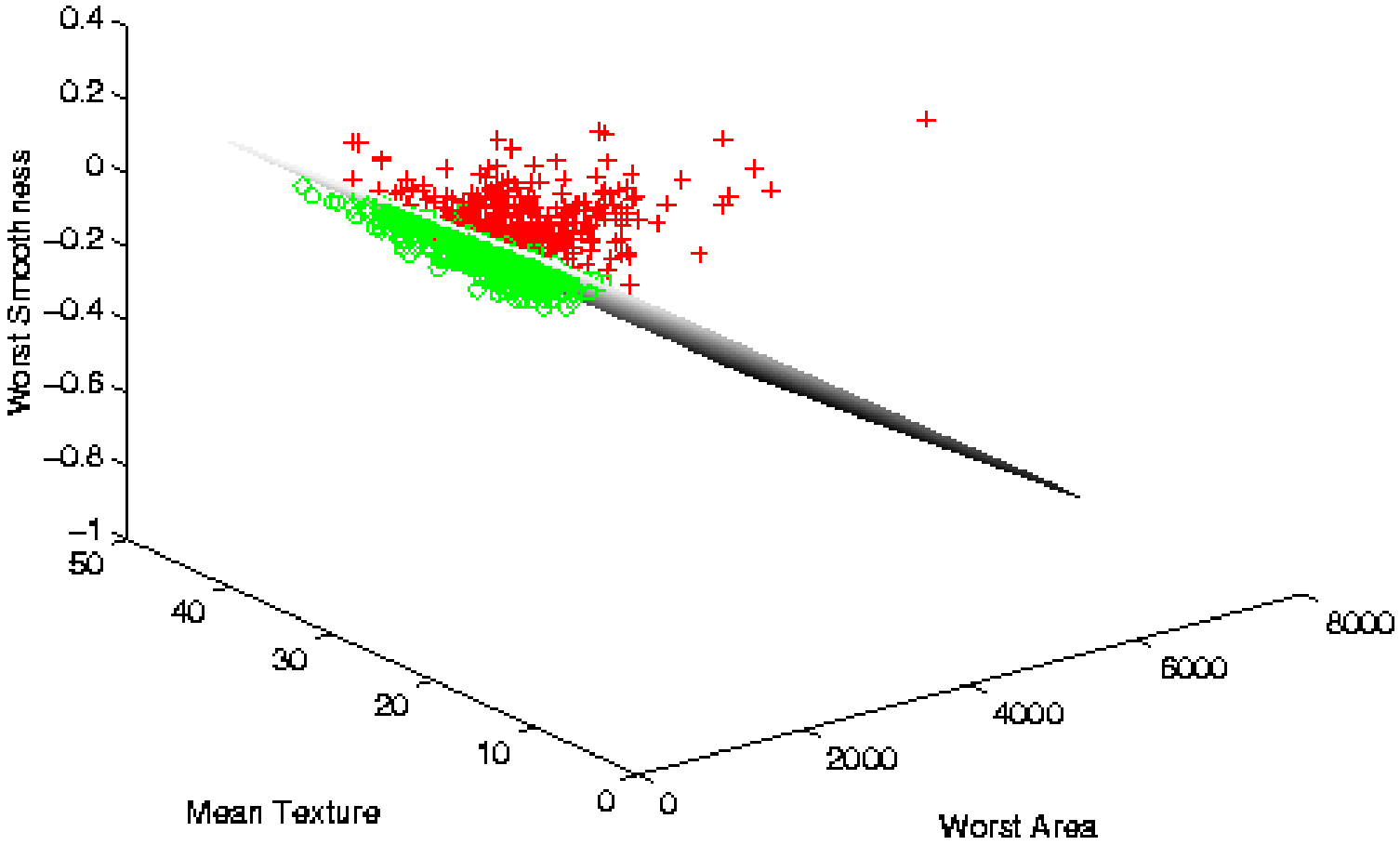
- **Goal: Use training set + some learning method to produce a predictive model.**
- **Use this predictive model to classify new data.**
- **Sample applications:**

Application	Input Data	Classification
Medical Diagnosis	Noninvasive tests	Results from invasive measurements
Optical Character Recognition	Scanned bitmaps	Letter A-Z
Protein Folding	Amino acid construction	Protein shape (helices, loops, sheets)
Research Paper Acceptance	Words in paper title	Paper accepted or rejected

Application: Cancer Diagnosis



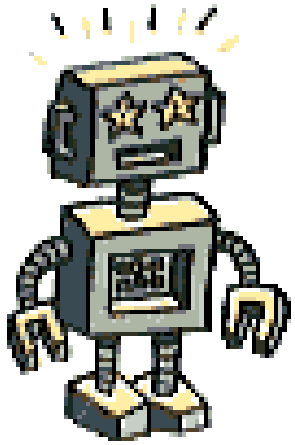
Cancer Diagnosis Separation



Robotics and ML

Areas that robots are used:

- Industrial robots
- Military, government and space robots
- Service robots for home, healthcare, laboratory



Why are robots used?

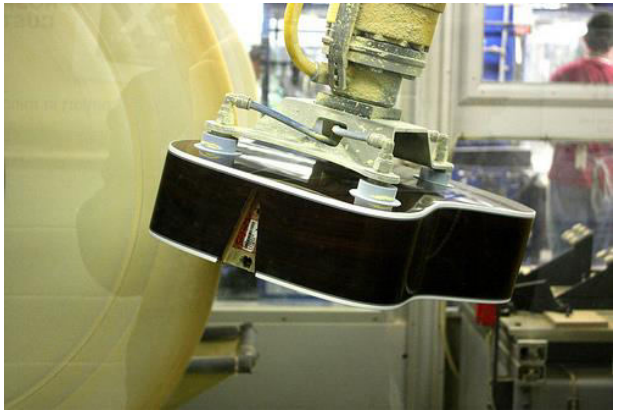
- Dangerous tasks or in hazardous environments
- Repetitive tasks
- High precision tasks or those requiring high quality
- Labor savings

Control technologies:

- Autonomous (self-controlled), tele-operated (remote control)

Industrial Robots

- Uses for robots in manufacturing:
  - Welding
  - Painting
  - Cutting
  - Dispensing
  - Assembly
  - Polishing/Finishing
  - Material Handling
    - Packaging, Palletizing
    - Machine loading



Space Robots

- Mars Rovers – Spirit and Opportunity
  - Autonomous navigation features with human remote control and oversight

Service Robots

- Many uses…
  - Cleaning & Housekeeping
  - Humanitarian Demining
  - Rehabilitation
  - Inspection
  - Agriculture & Harvesting
  - Lawn Mowers
  - Surveillance
  - Mining Applications
  - Construction
  - Automatic Refilling
  - Fire Fighters
  - Search & Rescue

iRobot Roomba vacuum cleaner robot

Issues in Machine Learning

- What algorithms can approximate functions well and when?
  - How does the number of training examples influence accuracy
- Problem representation / feature extraction
- Intention/independent learning
- Integrating learning with systems
- What are the theoretical limits of learnability
- Transfer learning
- Continuous learning

Scaling issues in ML

- Number of
  - Inputs
  - Outputs
  - Batch vs realtime
  - Training vs testing

Machine Learning VS Human Learning

- Some ML behavior can challenge the performance of human experts (e.g., playing chess)
- Although ML sometimes matches human learning capabilities, it is not able to learn as well as humans or in the same way that humans do
- There is no claim that machine learning can be applied in a truly creative way
- Formal theories of ML systems exist but are often lacking (why a method succeeds or fails is not clear)
- ML success is often attributed to manipulation of symbols (rather than mere numeric information)

END OF MODULE-1

MODULE-2

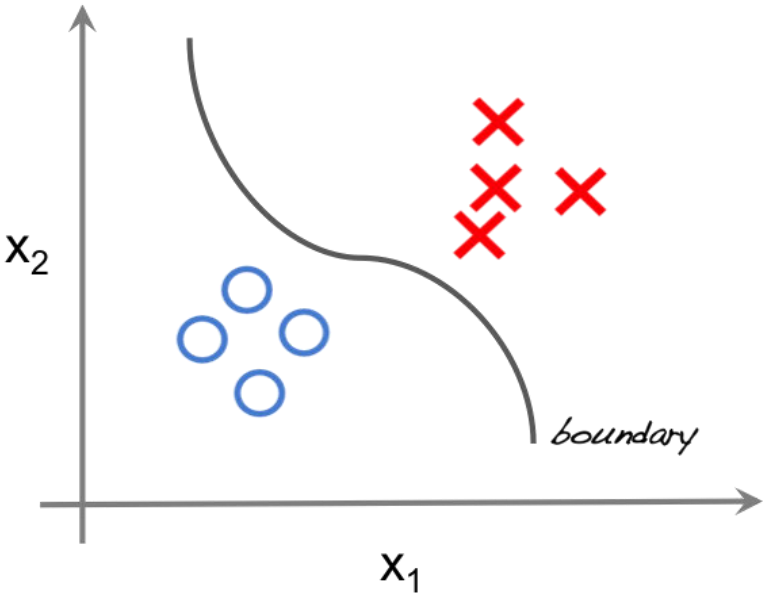
Supervised Learning

SUPERVISED LEARNING

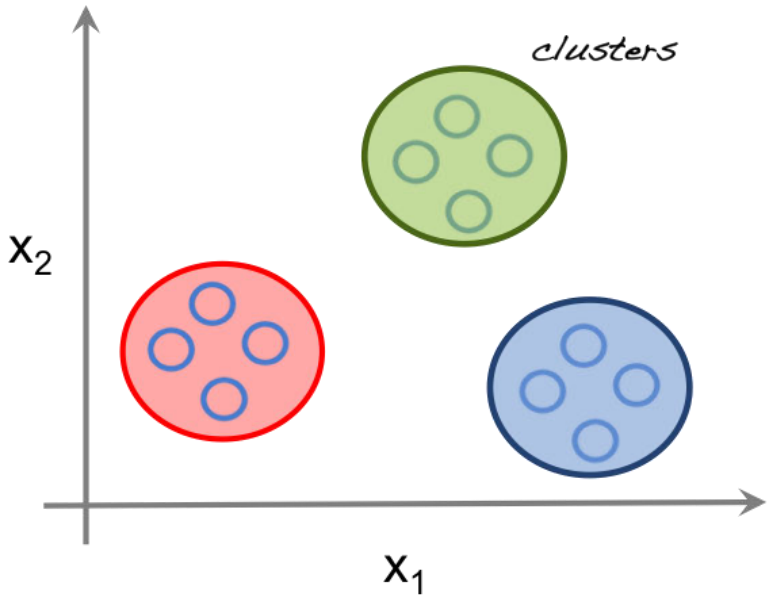
- In supervised learning, the labelled training data provide the basis for learning.
- The process of learning from the training data by a machine can be related to an expert supervising the learning process of a student.
- Here the expert is the training data.
- Training data is the past information with known value of class field or ‘label’.
- Unsupervised learning uses no labelled data.
- Semi-supervised learning uses a small amount of labelled data.

Supervised vs Unsupervised

Supervised learning



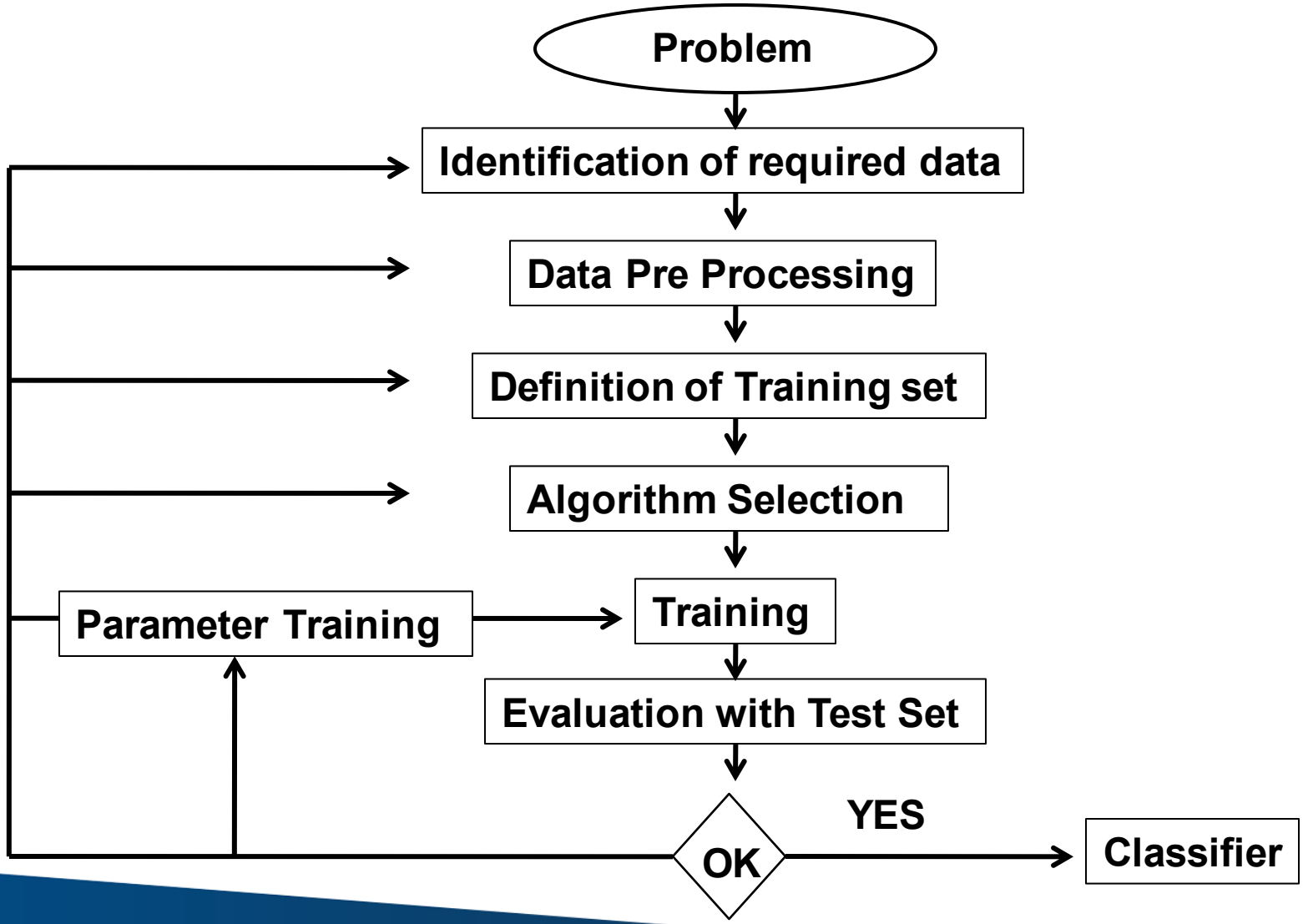
Unsupervised learning



Classification Model

- When we try to predict a categorical or nominal variable, the problem is known as a classification problem.
- Here, the problem centres around assigning a label or category or class to the test data on the basis of the label or category or class information imparted by training data.
- Classification is a type of supervised learning where a target feature, i.e. A categorical type, is predicted for test data on the basis of information obtained from training data.
- This categorical feature is known as class.

Classification Learning Steps



Common Classification Algorithms

1. k-Nearest Neighbour (kNN)
2. Decision tree
3. Random forest
4. Support Vector Machine (SVM)
5. Naive Bayes classifier

ORIGINS OF K-NN

- Nearest Neighbors have been used in statistical estimation and pattern recognition already in the beginning of 1970's (non-parametric techniques).

- The method prevailed in several disciplines and still it is one of the top 10 Data Mining algorithm.

IN A SENTENCE K-NN IS…..

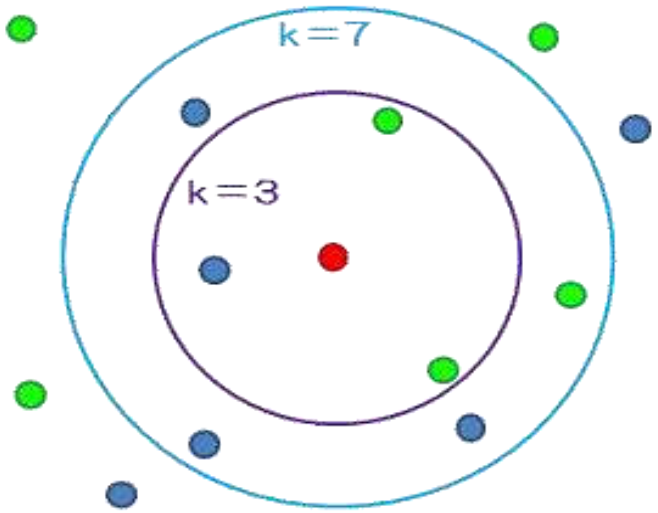
- It's how people judge by observing our peers.
- We tend to move with people of similar attributes so does data.

DEFINITION

- K-Nearest Neighbor is considered a lazy learning algorithm that classifies data sets based on their similarity with neighbors.

- "K" stands for number of data set items that are considered for the classification.

Ex: Image shows classification for different k-values.



TECHNICALLY..

- For the given attributes $A=\{X_1, X_2 \ldots\ldots X_D\}$ Where D is the dimension of the data, we need to predict the corresponding classification group $G=\{Y_1, Y_2 \ldots Y_n\}$ using the proximity metric over K items in D dimension that defines the closeness of association such that $X \in R^D$ and $Y_p \in G$.

THAT IS..

- Attribute A={Color, Outline, Dot}
- Classification Group, G={triangle, square}
- D=3, we are free to choose K value.

Attributes A

#	Attribute			Shape
	Color	Outline	Dot	
1	green	dashed	no	triange
2	green	dashed	yes	triange
3	yellow	dashed	no	square
4	red	dashed	no	square
5	red	solid	no	square
6	red	solid	yes	triange
7	green	solid	no	square
8	green	dashed	no	triange
9	yellow	solid	yes	square
10	red	solid	no	square
11	green	solid	yes	square
12	yellow	dashed	yes	square
13	yellow	solid	no	square
14	red	dashed	yes	triange

Classification Group

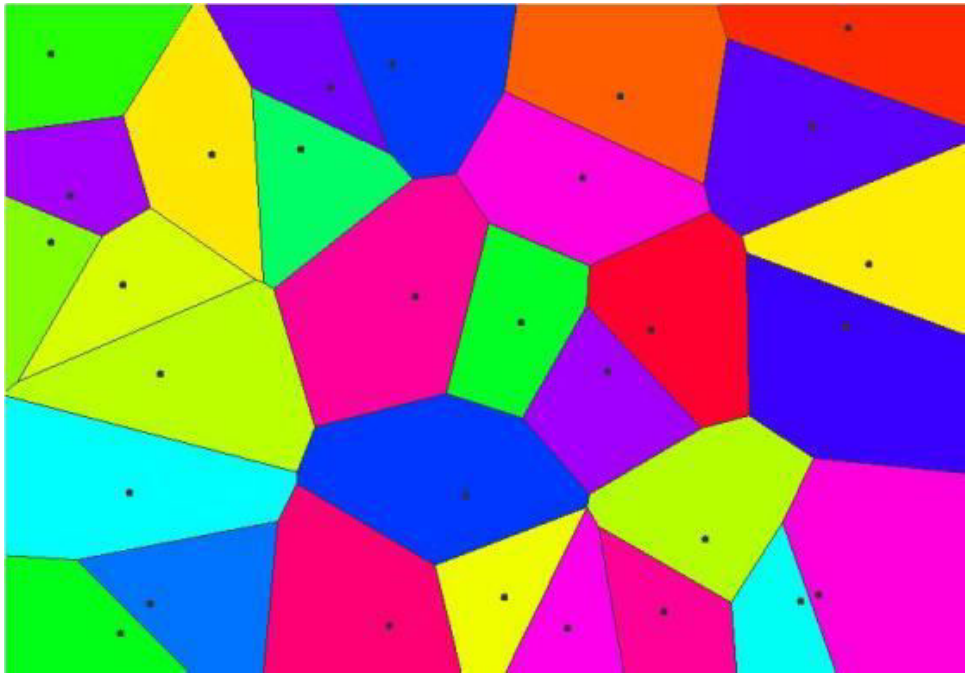
PROXIMITY METRIC

- Definition: Also termed as "Similarity Measure" quantifies the association among different items.
- Following is a table of measures for different data items:

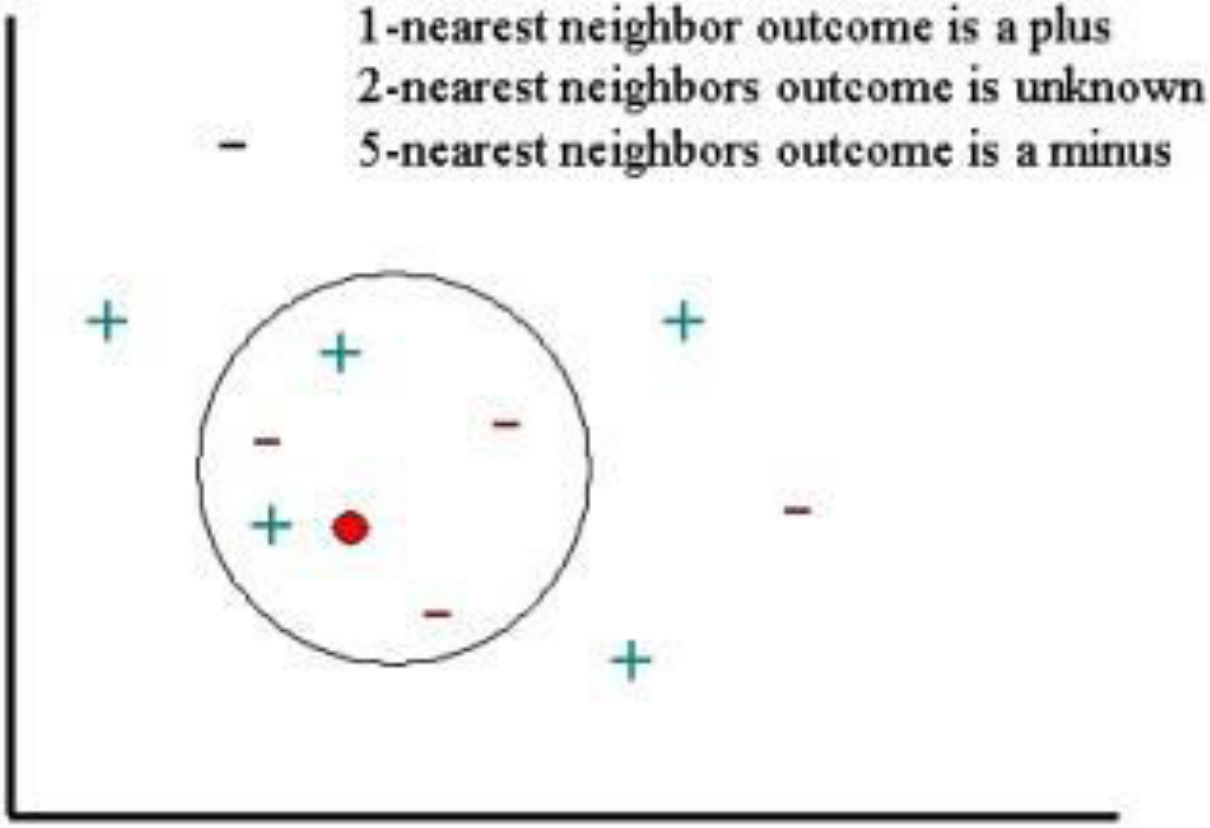
Similarity Measure	Data Format
Contingency Table, Jaccard coefficient, Distance Measure	Binary
Z-Score, Min-Max Normalization, Distance Measures	Numeric
Cosine Similarity, Dot Product	Vectors

Voronoi diagram

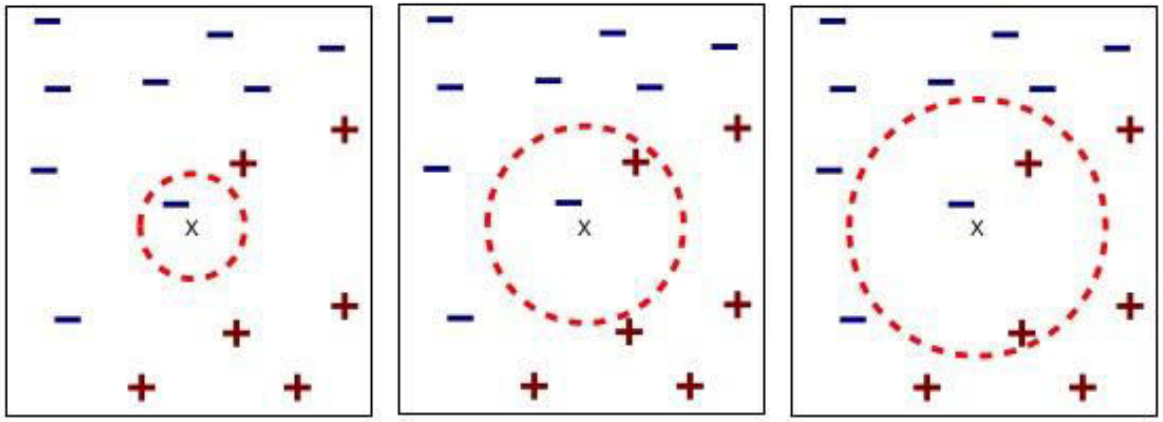
- A Voronoi diagram is a partitioning of a plane into regions based on distance to points in a specific subset of the plane.
- Here, k=1.



K-NN Example


1-nearest neighbor outcome is a plus
2-nearest neighbors outcome is unknown
5-nearest neighbors outcome is a minus

K-NN Example



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

PROXIMITY METRIC

- For the numeric data let us consider some distance measures:
  - Manhattan Distance:

$$X = \langle x_1, x_2, \cdots, x_n \rangle \quad Y = \langle y_1, y_2, \cdots, y_n \rangle$$

$$dist(X,Y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

  - Ex: Given X = {1,2} & Y = {2,5}
    Manhattan Distance = dist(X,Y) = |1-2|+|2-5|
    = 1+3
    = 4

PROXIMITY METRIC

- Euclidean Distance:

$$X = \langle x_1, x_2, \cdots, x_n \rangle \qquad Y = \langle y_1, y_2, \cdots, y_n \rangle$$

$$dist(X,Y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

- Ex: Given X = {-2,2} & Y = {2,5}
  Euclidean Distance

K-NN IN ACTION

- Consider the following data:
  A={weight,color}
  G={Apple(A), Banana(B)}

- We need to predict the type of a fruit with:
  weight = 378
  color = red

weight (g)	color	Type of fruit
303	3	Banana
370	1	Apple
298	3	Banana
277	3	Banana
377	4	Apple
299	3	Banana
382	1	Apple
374	4	Apple
303	4	Banana
309	3	Banana
359	1	Apple
366	1	Apple
311	3	Banana
302	3	Banana
373	4	Apple
305	3	Banana
371	3	Apple

SOME PROCESSING..

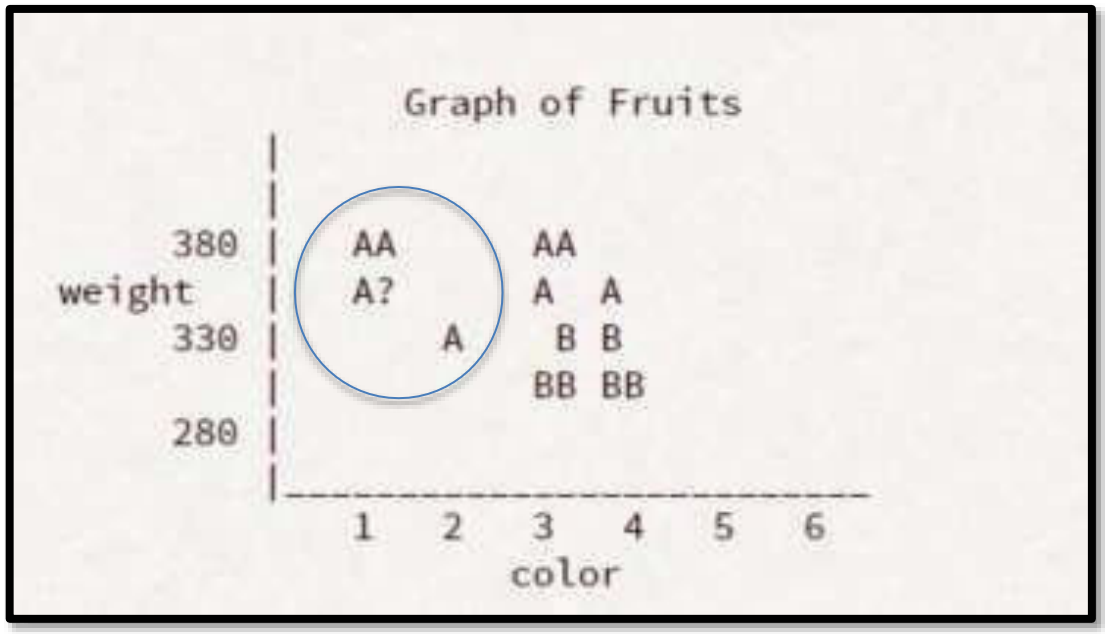
- Assign color codes to convert into numerical data:

red	1
orange	2
yellow	3
green	4
blue	5
purple	6

- Let's label Apple as "A" and Banana as "B"

PLOTTING

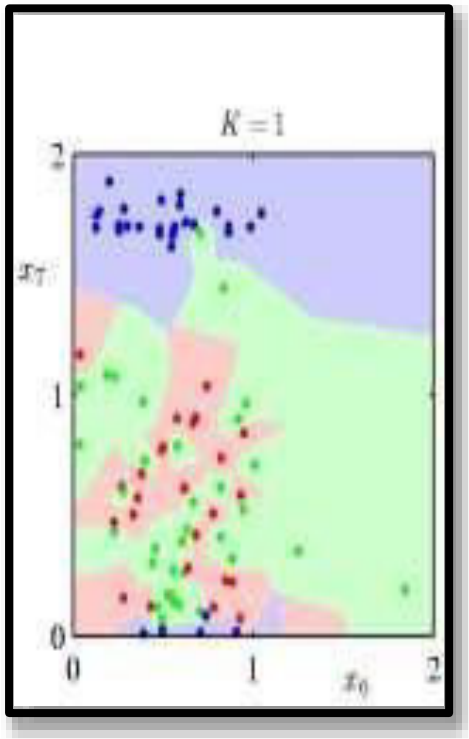
- Using K=3,
  Our result will be,

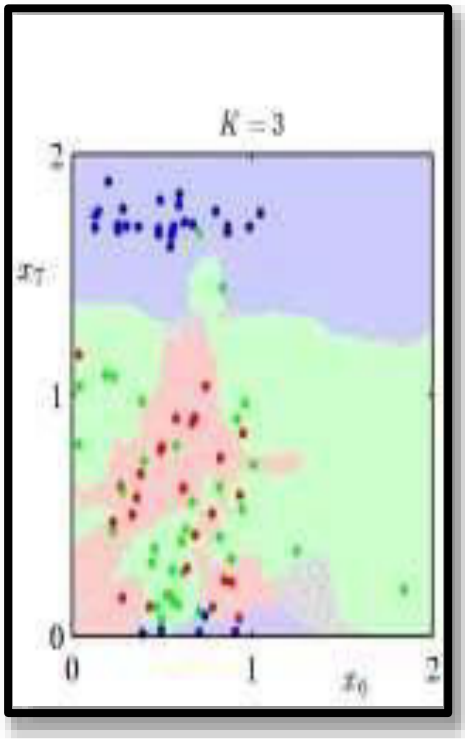


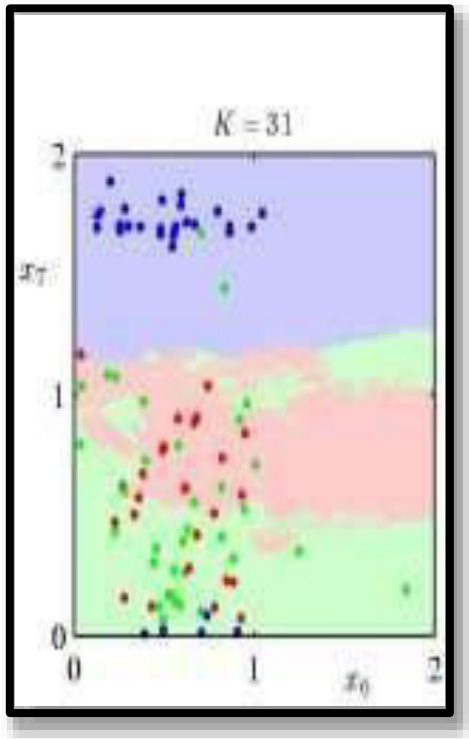
AS 'K' VARIES….

- Clearly, K has an impact on the classification.

Can you guess?







K-NN PROPERTIES

- K-NN is a lazy algorithm
- The processing defers with respect to K value.
- Result is generated after analysis of stored data.
- It neglects any intermediate values.

REMARKS: FIRST THE GOOD

Advantages

- Can be applied to the data from any distribution, for example, data does not have to be separable with a linear boundary
- Very simple and intuitive
- Good classification if the number of samples is large enough

NOW THE BAD….

Disadvantages

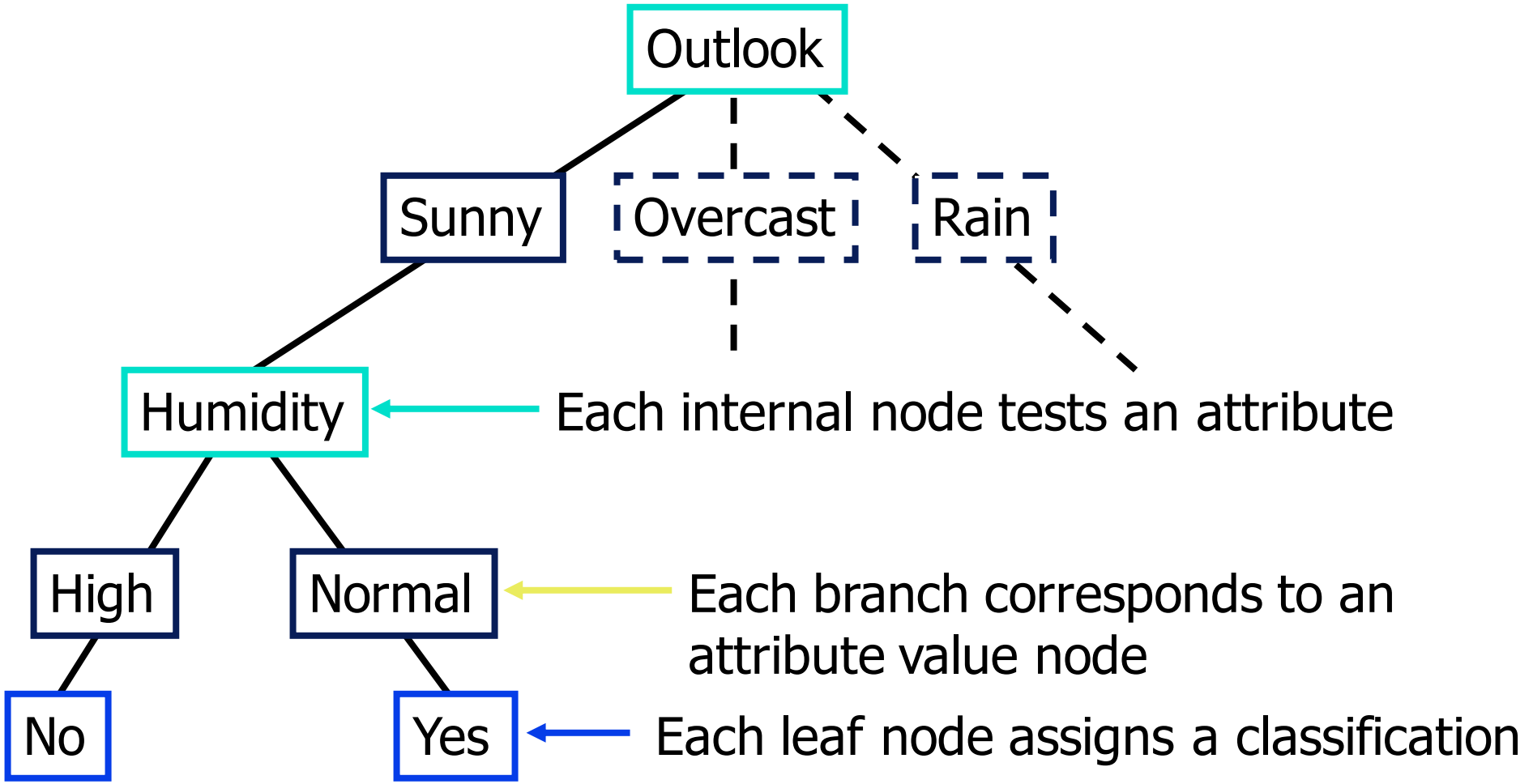
- Dependent on K Value
- Test stage is computationally expensive
- No training stage, all the work is done during the test stage
- This is actually the opposite of what we want. Usually we can afford training step to take a long time, but we want fast test step.
- Need large number of samples for accuracy

DECISION TREE

- This is one of the most adopted algorithms for classification.
- It builds a model in the form of a tree structure.
- A decision tree is used for multi-dimensional analysis with multiple classes and is characterized by ease of interpretation of rules and fast execution.
- The goal of decision tree learning is to create a model that predicts the value of the output variable based on the input variables in the feature vector.
- It contains a decision node and a leaf node.
- Each decision node corresponds to one of the feature vector.

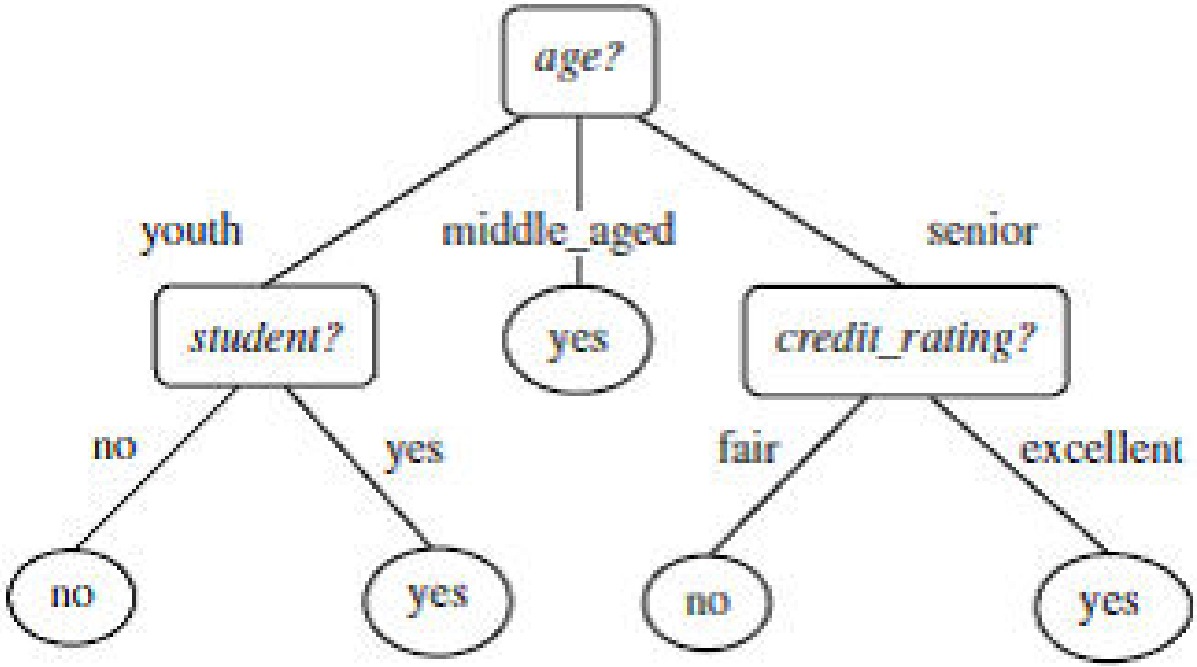
- From every node, there are edges to children, wherein there is an edge for each of the possible values of the feature associated with the node.
- The output variable is determined by following a path that starts at the root and is guided by the values of the input variables.
- Decision trees can be used for both classification and regression.

Decision Tree for PlayTennis



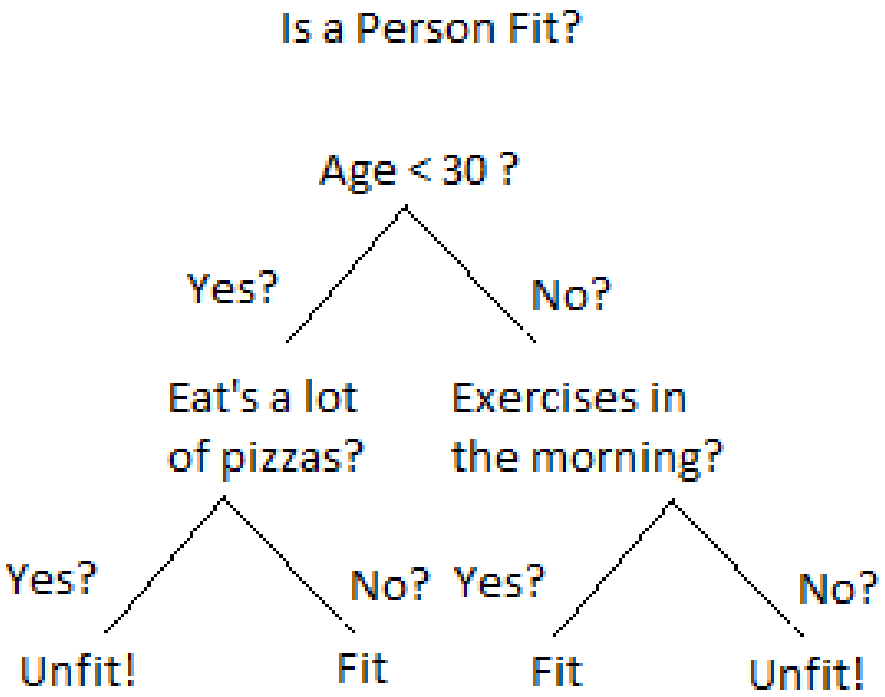
Example-1

- Will a person buy a computer?



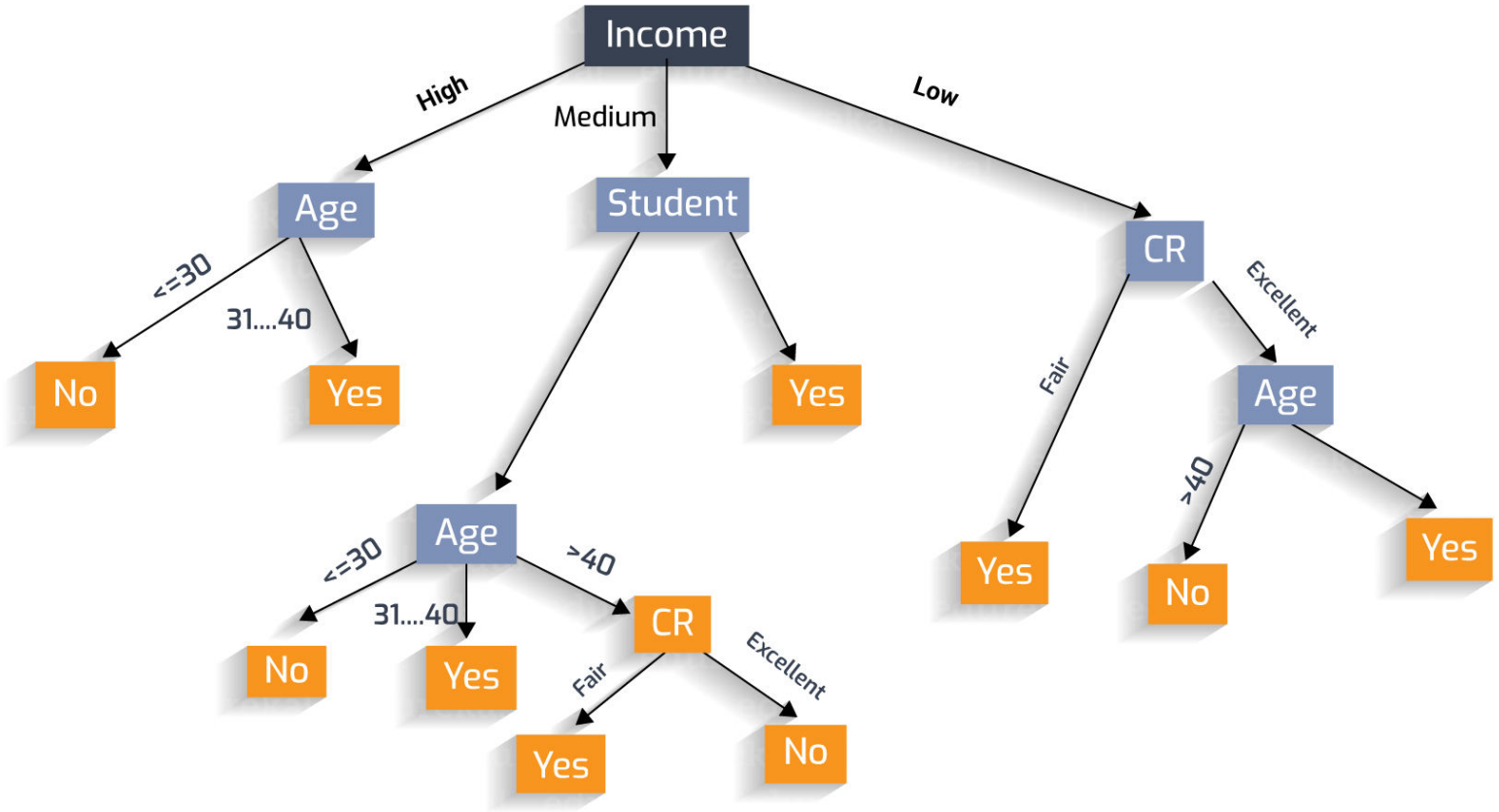
Example-2

- Is a person fit?



Example-3

- Should the LOAN be sanctioned?



Training Data for GTS recruitment

CGPA	Communication	Aptitude	Programming Skills	Job Offered?
High	Good	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	Low	Good	No
Low	Good	Low	Bad	No
High	Good	High	Bad	Yes
High	Good	High	Good	Yes
Medium	Bad	Low	Bad	No
Medium	Bad	Low	Good	No
High	Bad	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	High	Bad	No
Low	Bad	High	Bad	No
Medium	Good	High	Bad	Yes
Low	Good	Low	Good	No
High	Bad	Low	Bad	No
Medium	Bad	High	Good	No
High	Bad	Low	Bad	No
Medium	Good	High	Bad	Yes

Entropy of a decision tree

- Entropy, as it relates to machine learning, is a measure of the randomness in the information being processed.
- The higher the entropy, the harder it is to draw any conclusions from that information.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

- Ex: For class 'Job Offered?' we have two values: Yes and No.
- Pi values for Yes= 8/18 = 0.44 & No= 10/18= 0.56
- Entropy(S) = -0.44 log2(0.44) – 0.56 log2(0.56)

  = 0.99

Information gain of a decision tree

- The information gain is created on the basis of the decrease in entropy(S) after a data set is split according to a particular attribute(A).
- Constructing a decision tree is all about finding an attribute that returns the highest information gain.
- If information gain is 0, it means that there is no reduction in entropy due to split of the data set according to that particular feature.
- The maximum amount of information gain which may happen is the entropy of the data set before the split.

- Information gain for a particular feature A is calculated by the difference in entropy before a split(Sbs) with the entropy after the split(Sas).
- Information gain(S, A) = Entropy(Sbs) – Entropy(Sas)
- For weighted summation, the proportion of examples falling into each partition is used as weight.
- Entropy(Sas) = $\sum$ (i=1 to n) wi Entropy(pi)

a) Original data set

	Yes	No	Total
Count	8	10	18
pi	0.44	0.56	
-pi*LOG(pi)	0.52	0.47	0.99

Total Entropy = 0.99

b) Splitted data set(based on the CGPA)

CGPA = High				CGPA = Medium				CGPA = Low			
	Yes	No	Total		Yes	No	Total		Yes	No	Total
Count	4	2	6	Count	4	3	7	Count	0	5	5
pi	0.67	0.33		pi	0.57	0.43		pi	0	1	
-pi*LOG(pi)	0.39	0.53	0.92	-pi*LOG(pi)	0.46	0.52	0.99	-pi*LOG(pi)	0	0	0

Total Entropy = (6/18*0.92 + 7/18*0.99 + 5/18*0)
= 0.69

Information Gain
= 0.99-0.69 = 0.30

c) Splitted data set(based on ‘Communication’)

Communication = ‘Good’ Communication = ‘Bad’

Total Entropy = 0.63 Information Gain = 0.36

d) Splitted data set(based on ‘Aptitude’)

Aptitude = ‘High’ Aptitude = ‘Low’

Total Entropy = 0.52 Information Gain = 0.47(Entropy=0)

e) Splitted data set(based on ‘Programming Skills’)

Programming Skills = ‘Good’ Programming Skills = ‘Bad’

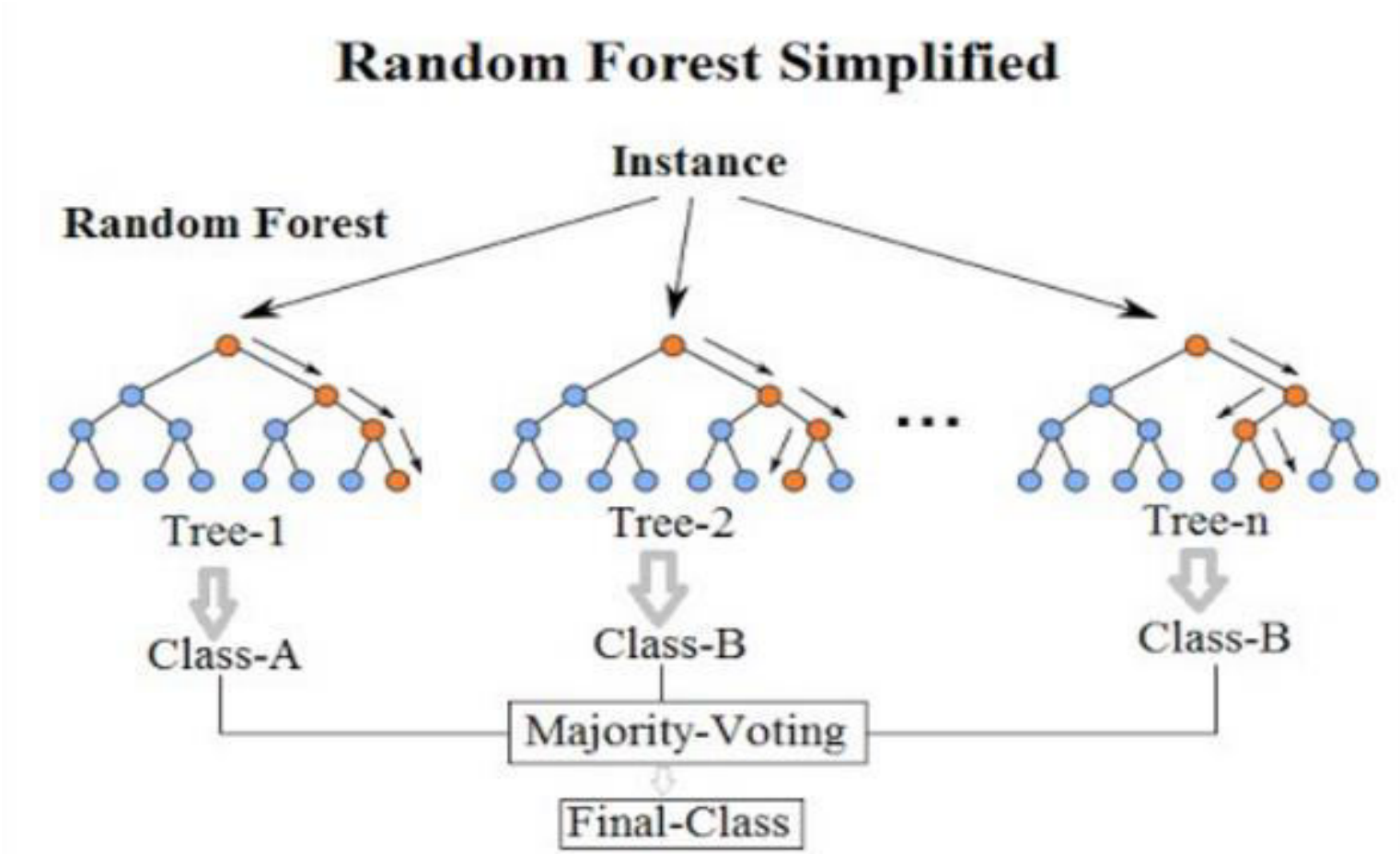
Total Entropy = 0.95 Information Gain = 0.04

Avoiding overfitting in decision tree- pruning

- The decision tree algorithm, unless a stopping criterion is applied, may keep growing indefinitely.
- To prevent a decision tree getting overfitted to the training data, pruning of the decision tree is essential.
- Pruning a decision tree reduces the size of the tree such that the model is more generalized and can classify unknown and unlabeled data in a better way.
- Pre-pruning: Stop growing the tree before it reaches perfection.
- Post-pruning: Allow the tree to grow entirely and then post-prune some of the branches from it.

Random Forest Model

- It is an ensemble classifier, i.e., a combining classifier that uses and combines many decision tree classifiers.
- Ensembling is usually done using the concept of bagging with different feature sets.
- The reason for using large number of trees in random forest is to train the trees enough such that contribution from each feature comes in a number of models.
- After the random forest is generated by combining the trees, majority vote is applied to combine the output of the different trees.
- Ensembled model yields better result than decision trees.


Random Forest Simplified
Instance
Random Forest
Tree-1
Tree-2
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class

Random forest algorithm

- The algorithm works as follows:

1. If there are N variables or features in the input data set, select a subset of ‘m’ (m<N) features at random out of the N features.
2. Use the best split principle on these ‘m’ features to calculate the number of nodes ‘d’.
3. Keep splitting the nodes to child nodes till the tree is grown to maximum possible extent.
4. Select a different subset of the training data ‘with replacement’ to train another DT with steps (1) to (3). Repeat this to build and train ‘n’ decision trees.
5. Final class assignment is done on the basis of the majority votes from the ‘n’ trees.

Strengths of RF

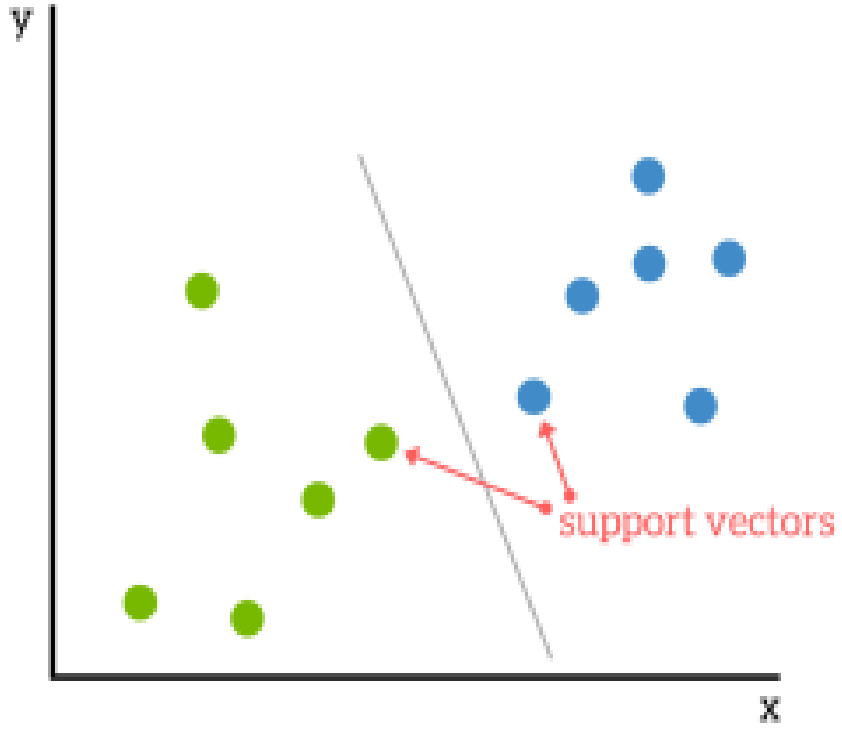
- It runs efficiently on large and expensive data sets.
- It has a robust method for estimating missing data and maintains precision when a large proportion of data is absent.
- It has powerful techniques for balancing errors in a class population of unbalanced data sets.
- It gives estimates about which features are the most important ones in the overall classification.

Drawback of RF

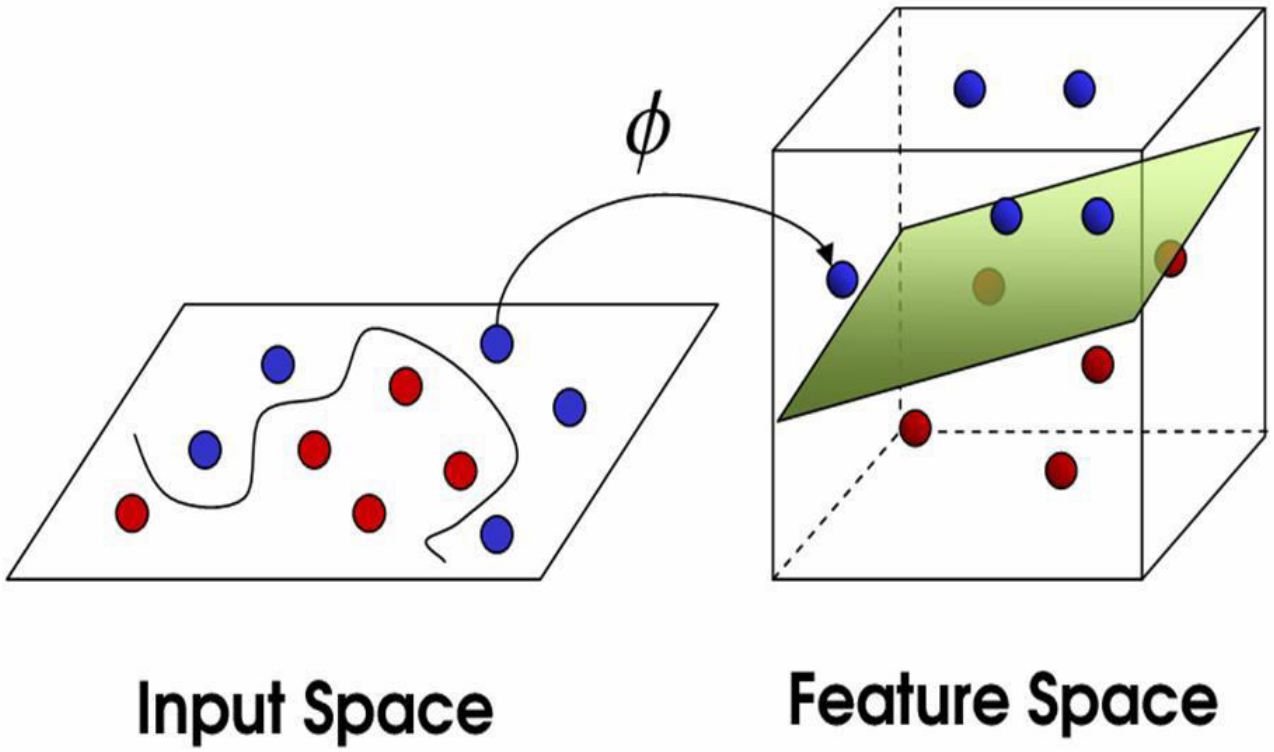
- As it combines many decision trees, it is not easy to understand as a decision tree model.
- Computationally, it is much more expensive than a simple decision tree.

Support Vector Machine

- SVM is a model which can perform linear classification as well as regression.
- It is based on the concept of a surface called hyperplane, which draws a boundary between data instances plotted on a multi-dimensional feature space.
- The output prediction is one of the two classes defined in the training data.

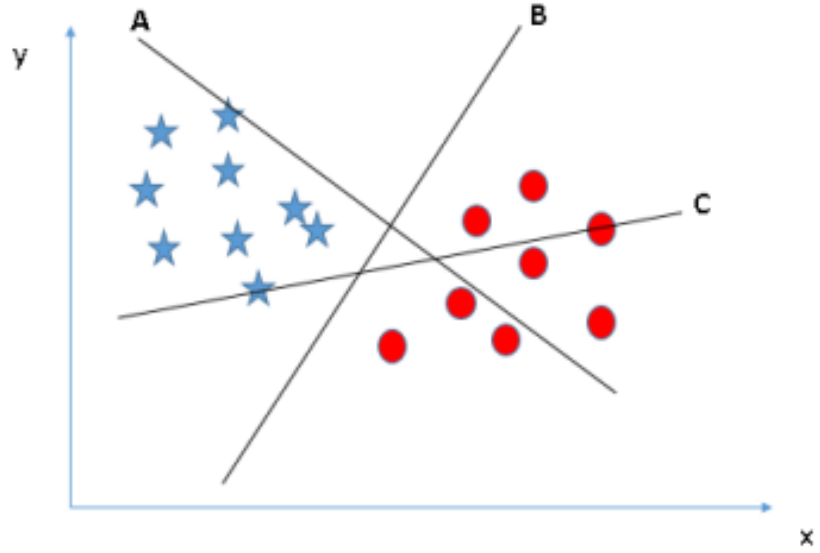

y
x
support vectors

Principle of Support Vector Machines (SVM)



Identify the right hyper-plane (Scenario-1)

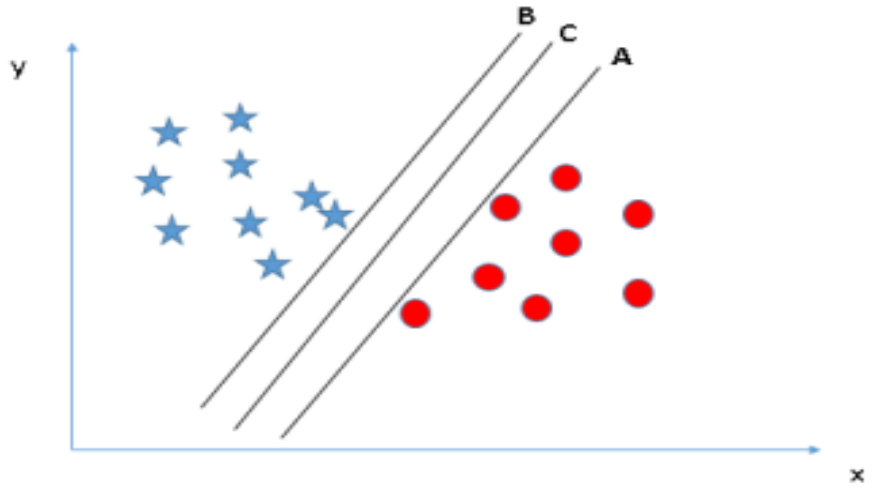
- Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.
- Select the hyper-plane which segregates the two classes better?



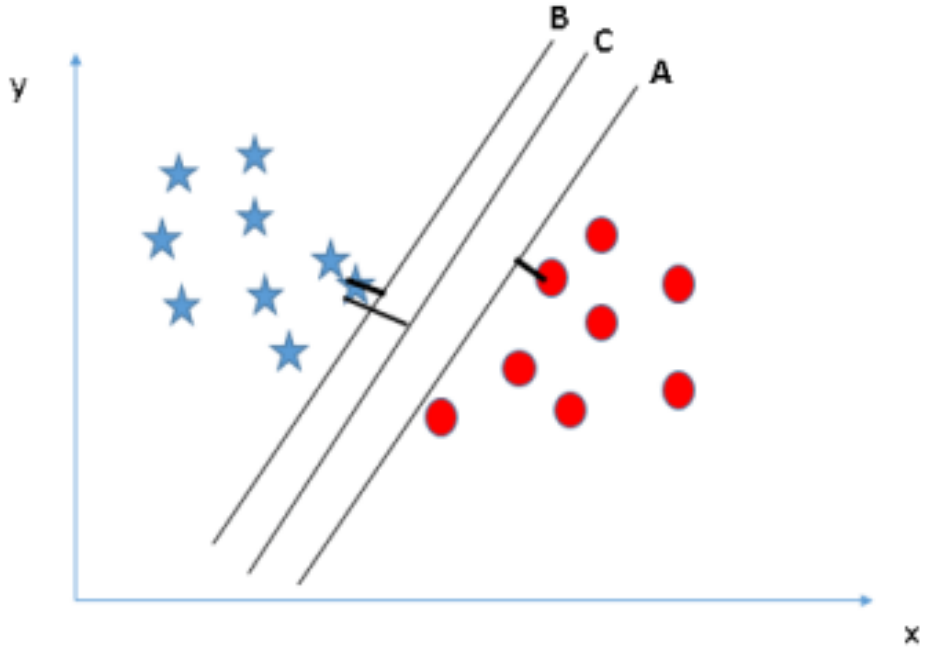
- In this scenario, hyper-plane “B” has excellently performed this job

Identify the right hyper-plane (Scenario-2)

- Here, we have three hyper-planes (A, B and C) and all are segregating the classes well. Now, How can we identify the right hyper-plane?

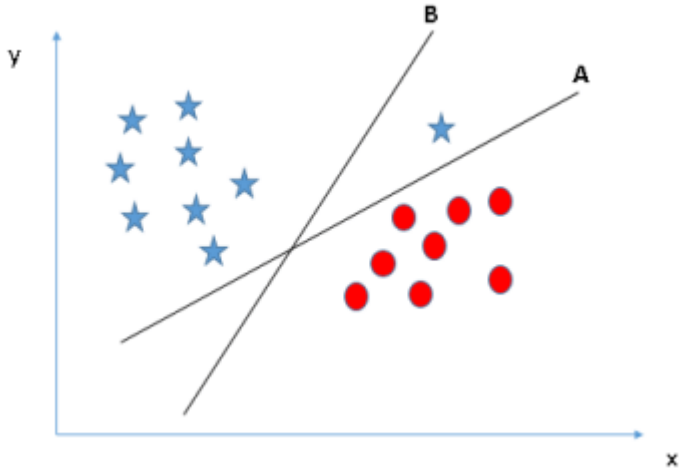


- Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin.



- We can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C.

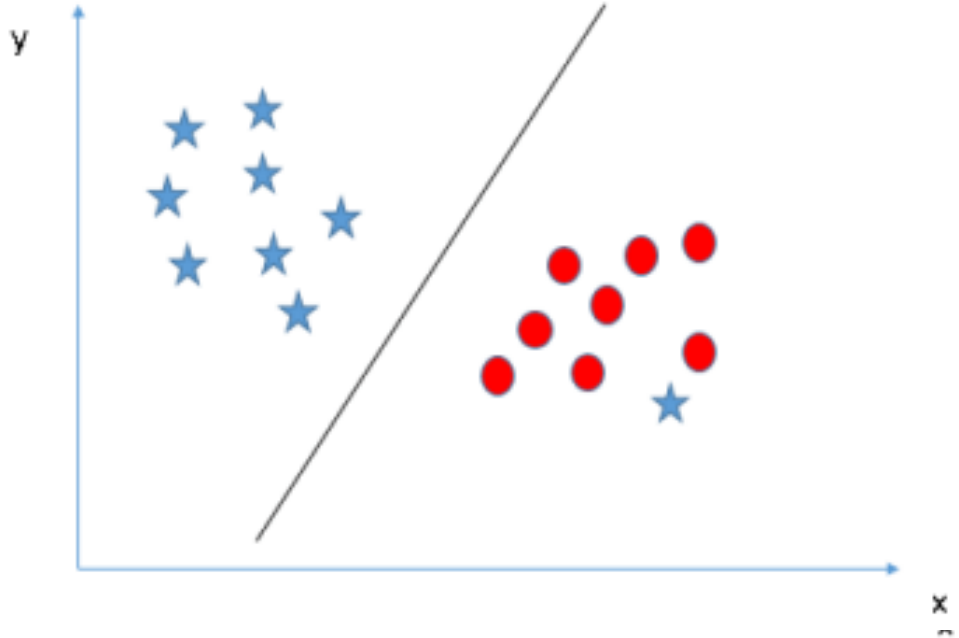
Scenario-3



- Here hyper-plane B as it has higher margin compared to A.
- SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin.
- Hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is A.

Scenario-4

- Here, we are unable to segregate the two classes using a straight line, as one of star lies in the territory of other(circle) class as an outlier.
- SVM has a feature to ignore outliers and find the hyper-plane that has maximum margin. SVM is robust to outliers.



Strengths of SVM

- SVM can be used for both classification & regression.
- It is robust, i.e. not much impacted by data with noise or outliers.
- The prediction results using this model are very strong.

Weakness of SVM

- SVM is applicable only for binary classification, i.e. when there are only two classes in the problem.
- While dealing with high dimensional data, it becomes very complex.
- It is slow for large dataset, i.e. a data set with more features or instances.
- It is memory-intensive(throughput is bounded by the device memory bandwidth).

Introduction to Regression

- Regression is a technique used to model and analyze the relationships between variables and often times how they contribute and are related to producing a particular outcome together.
- Here, dependent variable (Y) is the one whose value is to be predicted, ex.- the price quote of the real estate property.
- This variable is presumed to be functionally related to one (X) or more independent variables called predictors.
- $Y = f(X)$

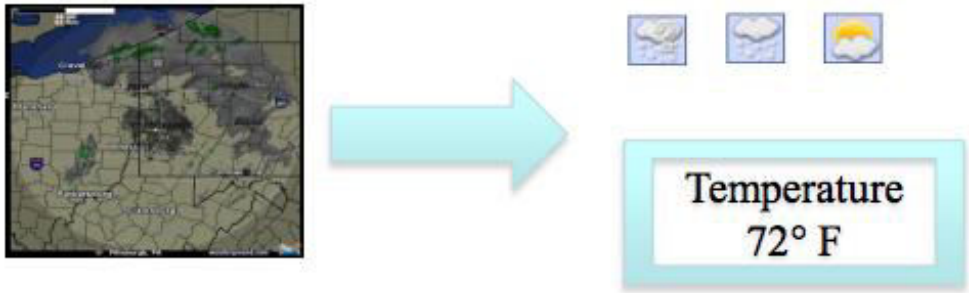
Common Regression Algorithms

- The common regression algorithms are:
  1. Simple linear regression
  2. Multiple linear regression
  3. Polynomial regression
  4. Multivariate adaptive regression splines
  5. Logistic regression
  6. Maximum likelihood estimation(least square)

Regression examples

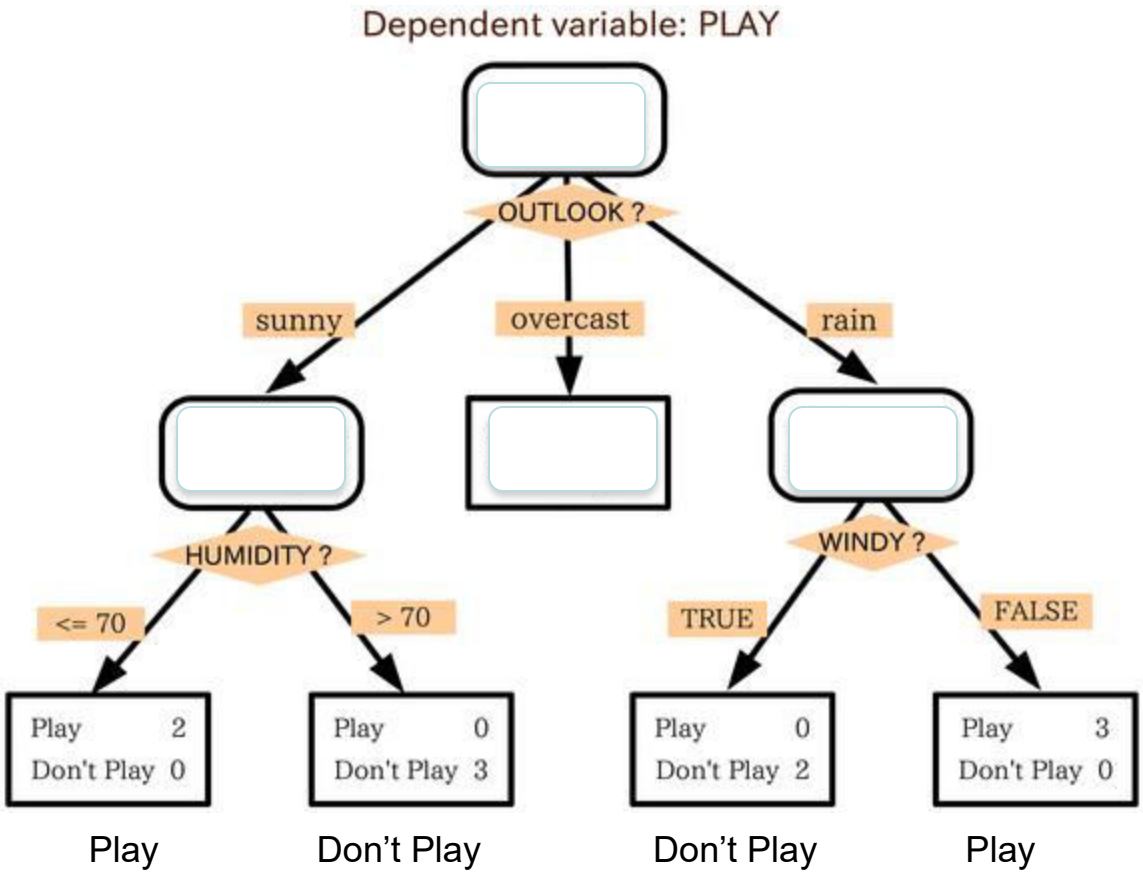
Stock market

Weather prediction

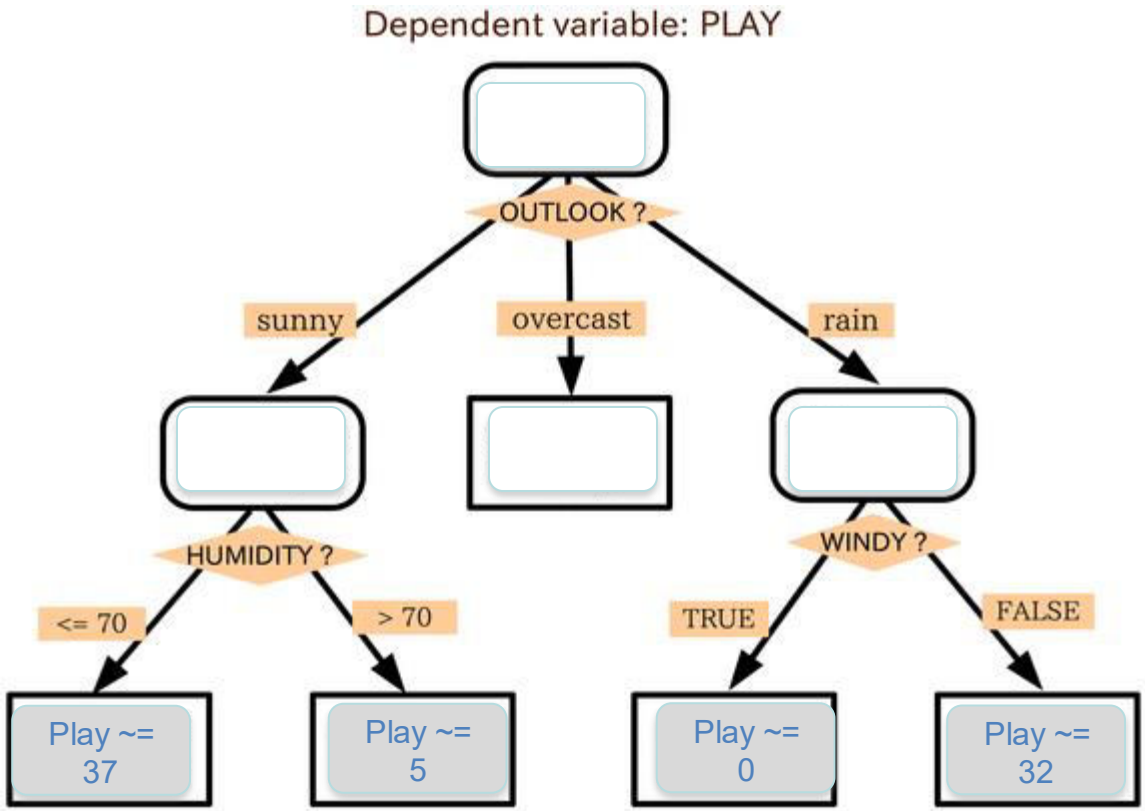


Predict the temperature at any given location

A decision tree: classification



A regression tree



Play = 30m, 45min

Play = 0m, 0m, 15m

Play = 0m, 0m

Play = 20m, 30m, 45m,

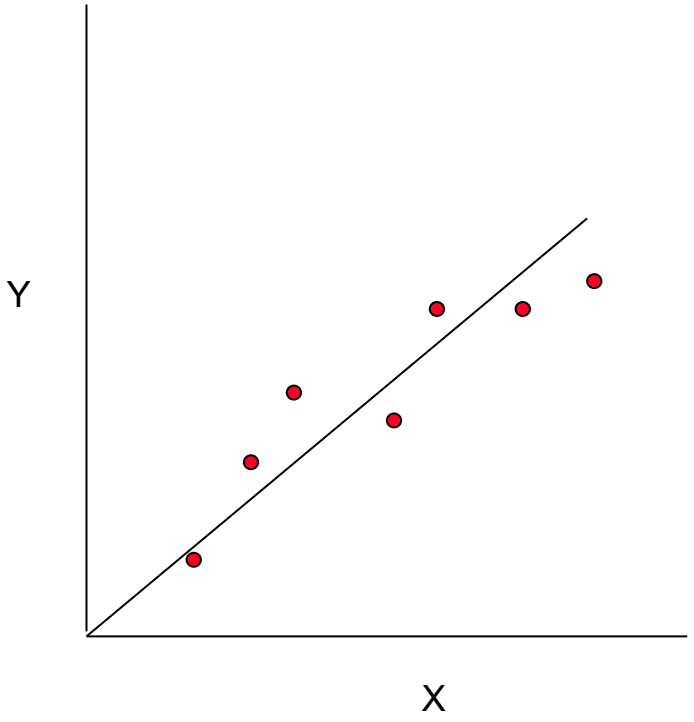
Linear regression

- Given an input x we would like to compute an output y
- For example:

  - Predict height from age

  - Predict Google's price from Yahoo's price

  - Predict distance from wall from sensors



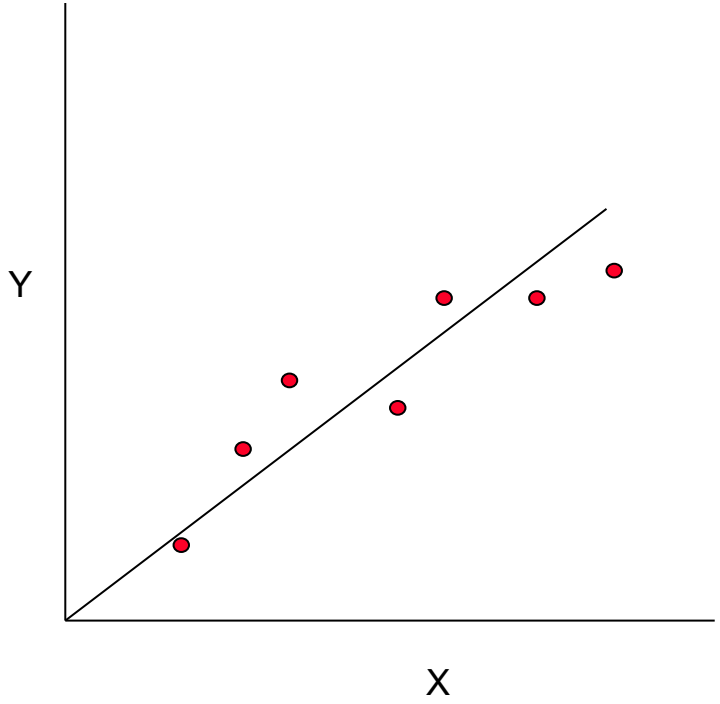
Linear regression

- Given an input x we would like to compute an output y
- In linear regression we assume that y and x are related with the following equation:

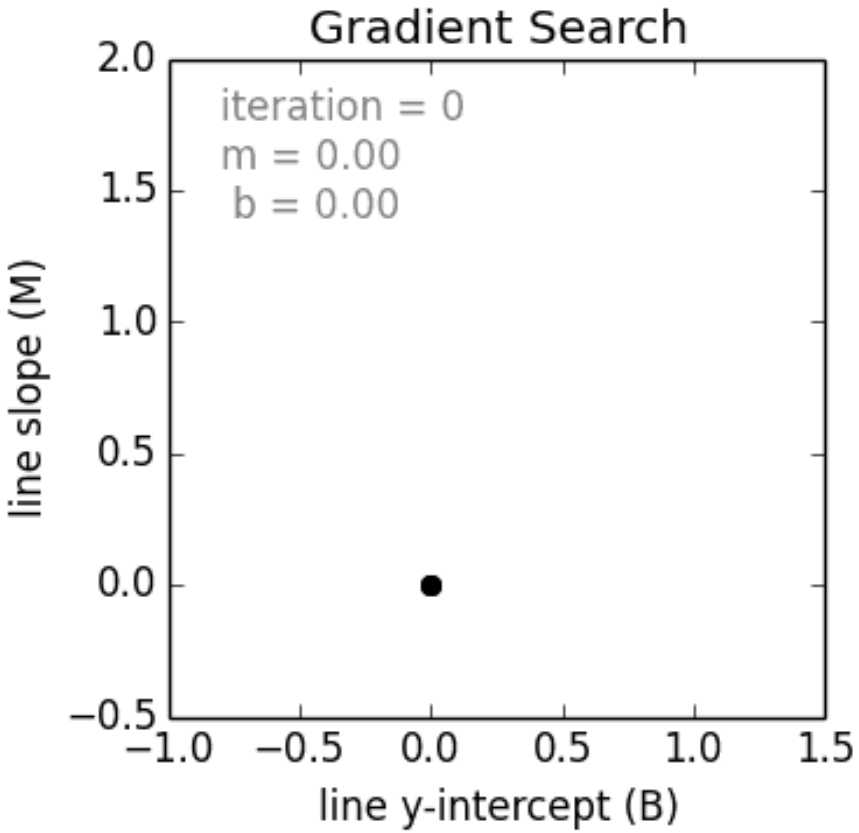
What we are trying to predict

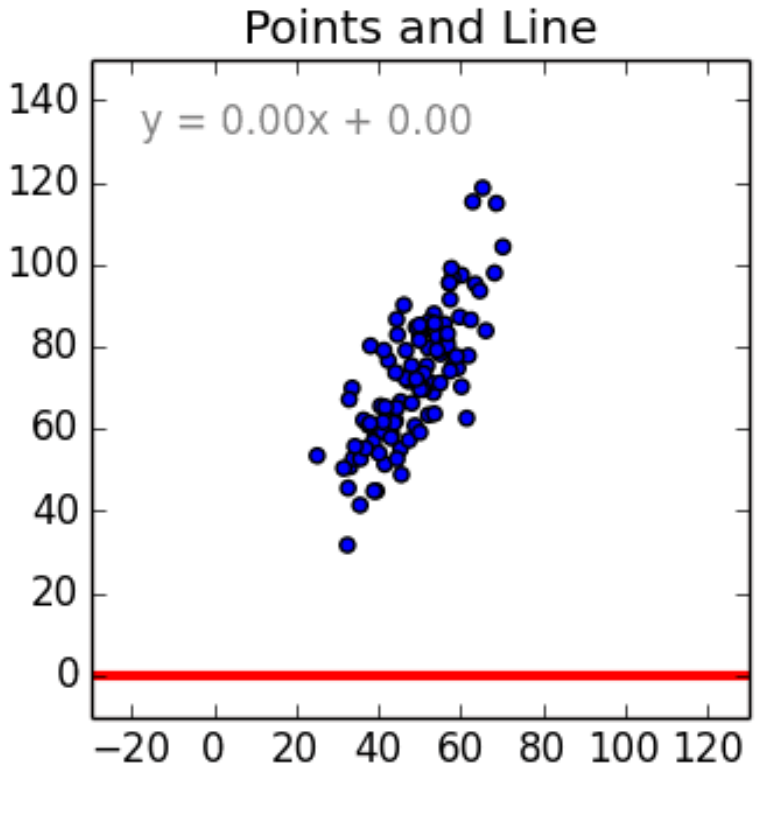
Observed values

$$y = wx+\varepsilon$$



where w is a parameter and ε represents measurement noise, model noise or other (data) noise


Gradient Search
iteration = 0
m = 0.00
b = 0.00
line slope (M)
line y-intercept (B)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Points and Line
y = 0.00x + 0.00
140
120
100
80
60
40
20
0
−20
0
20
40
60
80
100
120

Linear regression

• Our goal is to estimate *w* from a training data of $<x_i, y_i>$ pairs

• Optimization goal: minimize squared error (least squares):

$$\arg\min_w \sum_i (y_i - wx_i)^2$$

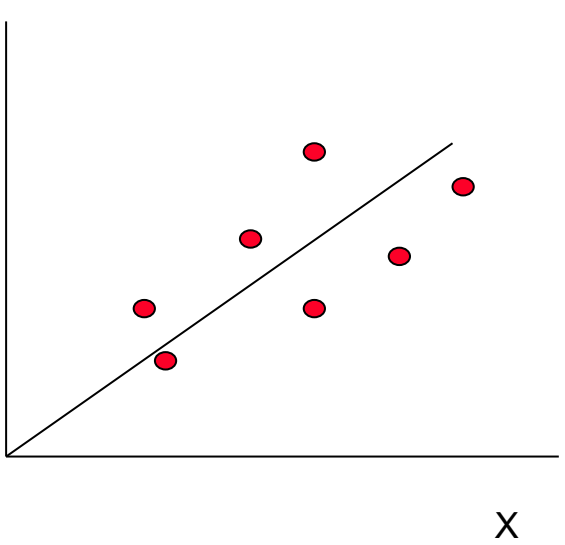
• Why least squares?

- minimizes squared distance between measurements and predicted line

- has a nice probabilistic interpretation (Gaussian Likelihood same as Mean Sq.)

- first degree polynomial model

$$y = wx +$$



Multivariate regression

- What if we have several inputs?

  - Stock prices for Yahoo, Microsoft and Ebay for the Google prediction task

- This becomes a multivariate regression problem

- Again, its easy to model:

$$y = w_0 + w_1 x_1 + \ldots + w_k x_k + \varepsilon$$

Google's stock price

Yahoo's stock price

Microsoft's stock price

- Price of property = f (Area, location, floor, ageing, amenities)
- Y = a + b1 X1 + b2 X2

Where Y is the three-dimensional space, X1 & X2 are the predictor variables, b1 & b2 are referred as partial regression coefficients.

Assumptions in Regression Analysis

1. The dependent variable(Y) can be calculated as a linear function of a specific set of independent variables(X) and an error term(ϵ).
2. The number of observations(n) is greater than the number of parameters(k) to be estimated, ie n>k.
3. Regression line can be valid only over a limited range of data.
4. Variance is the same for all values of X.
5. The error term(ϵ) is normally distributed.
6. The values of the error term(ϵ) are independent and are not related to any values of X.

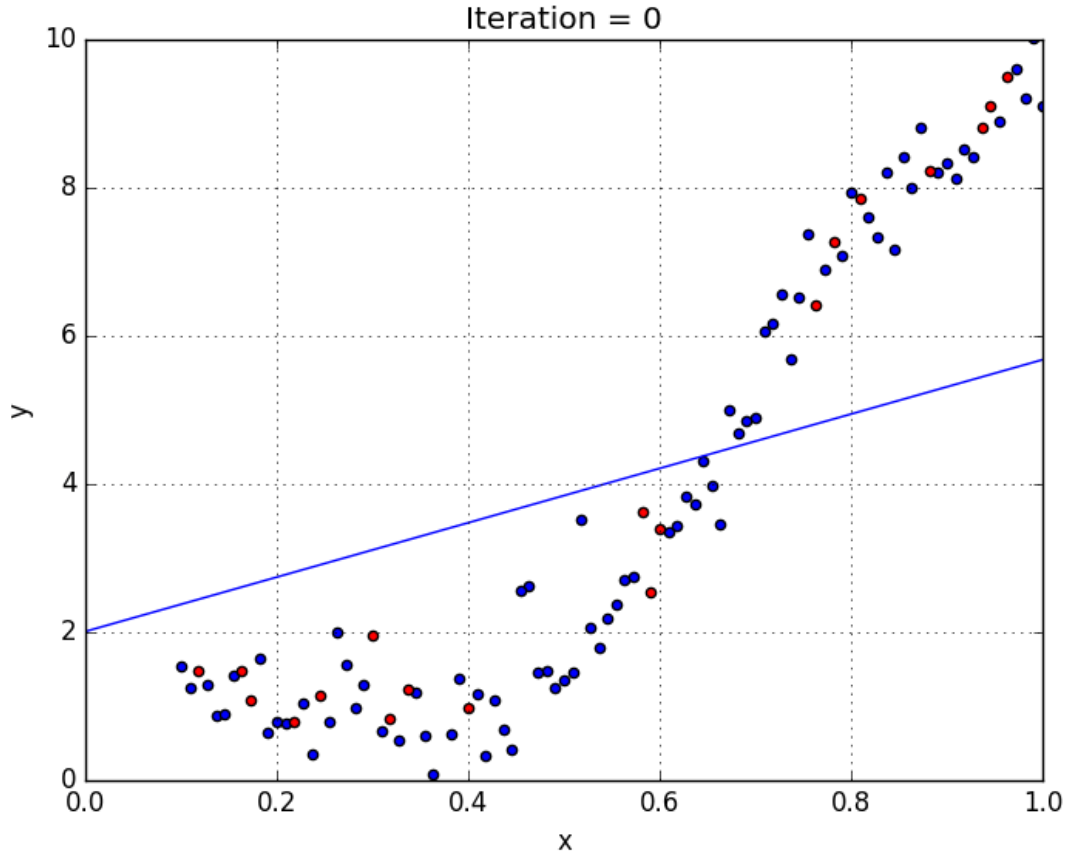
The OLS(Ordinary least square) estimator is the Best Linear Unbiased Estimator(BLUE) and this is called as Gauss-Markov Theorem.

Polynomial regression

Polynomial regression model is the extension of the simple linear model by adding extra predictors obtained by raising(squaring) each of the original predictors to a power.

If there are three variable, X, X^2, X^3 are used as predictors.

$$F(x) = C_0 + C_1 X^1 + C_2 X^2 + C_3 X^3$$


Iteration = 0
10
8
6
4
2
0
y
0.0
0.2
0.4
0.6
0.8
1.0
x

Let us use the below data set of (X, Y) for degree 3 polynomial

Internal Exam(X)	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam(X)	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

The regression line is slightly curved for degree = 3.

The regression line will curve further if we increase the polynomial degree.

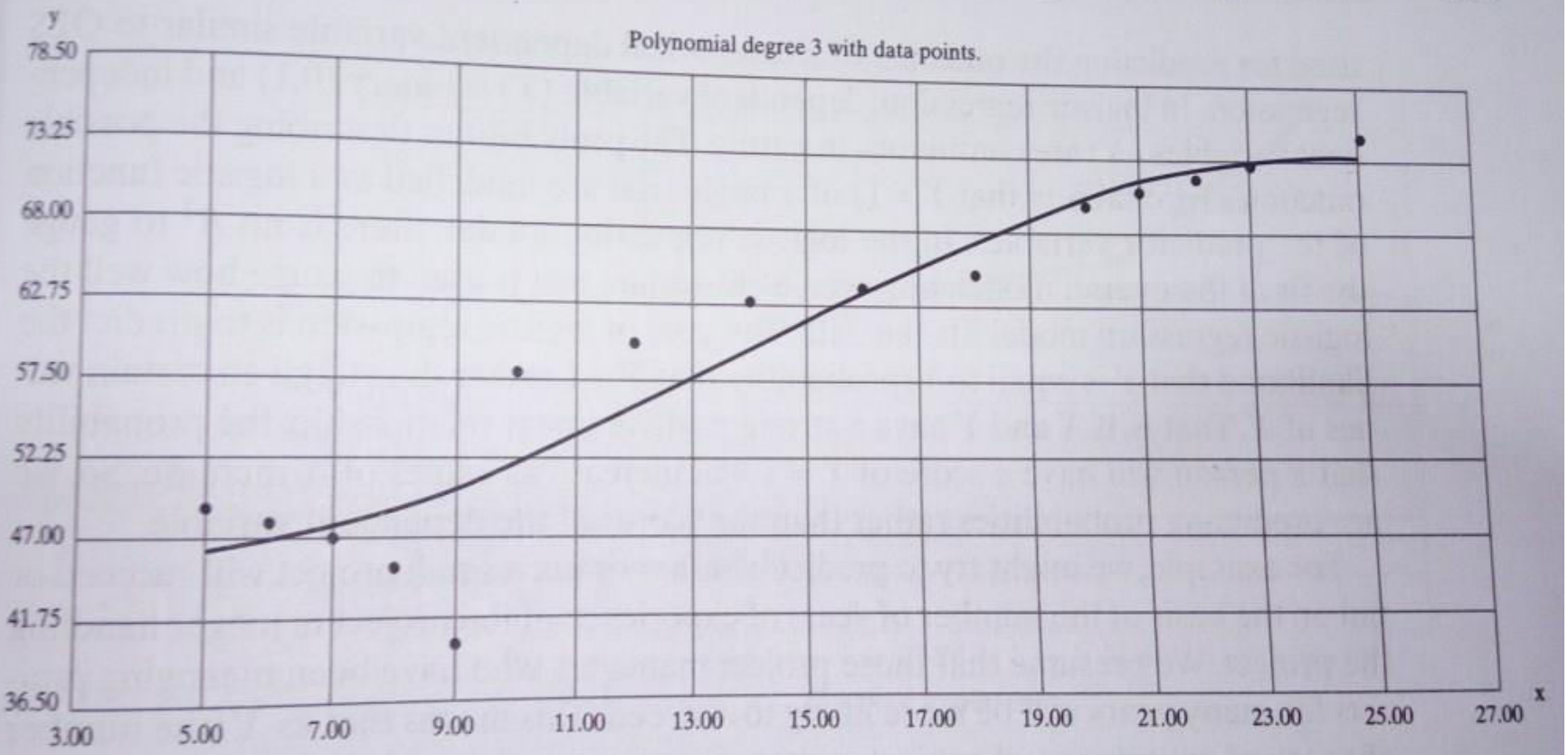


FIG. 8.16
Polynomial regression degree 3

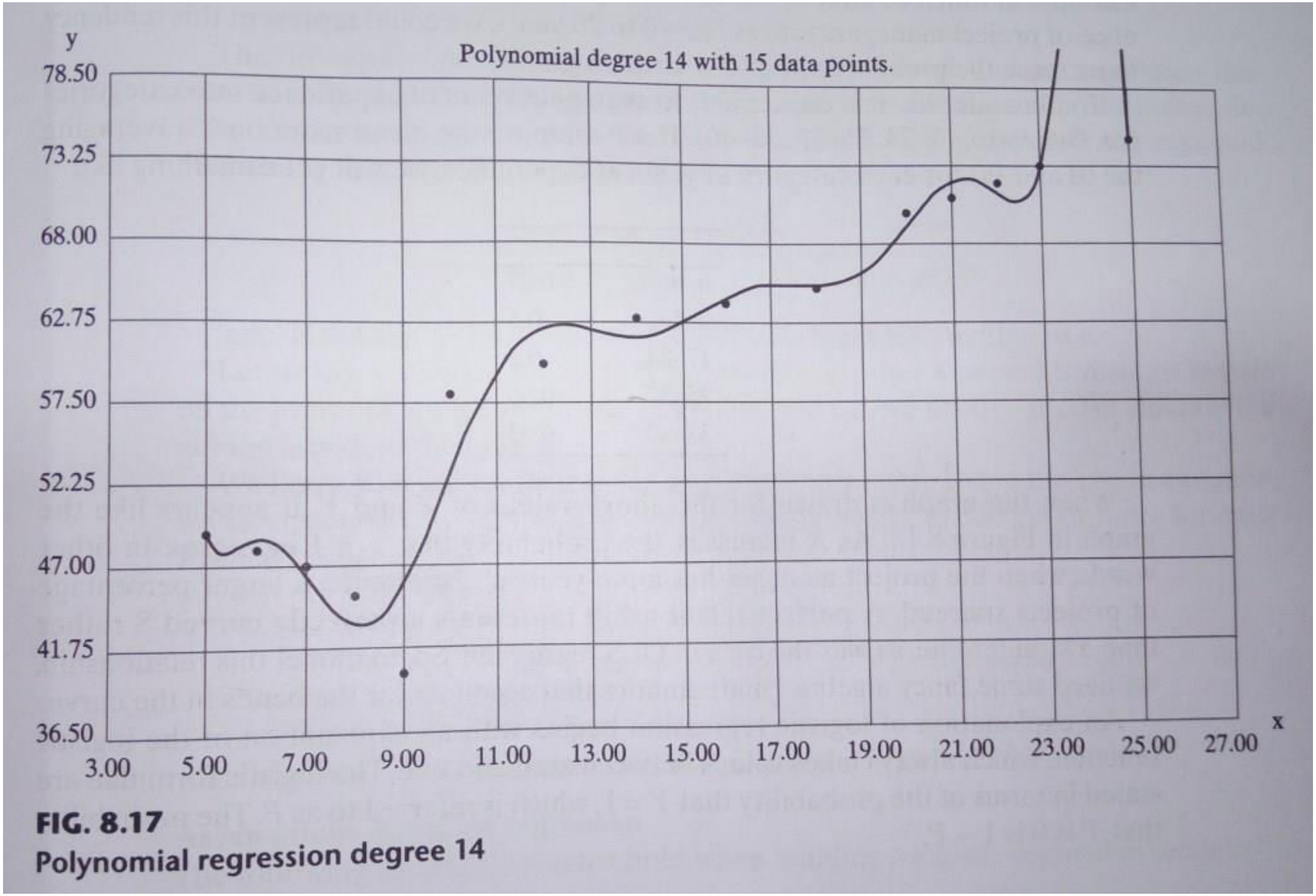


FIG. 8.17
Polynomial regression degree 14

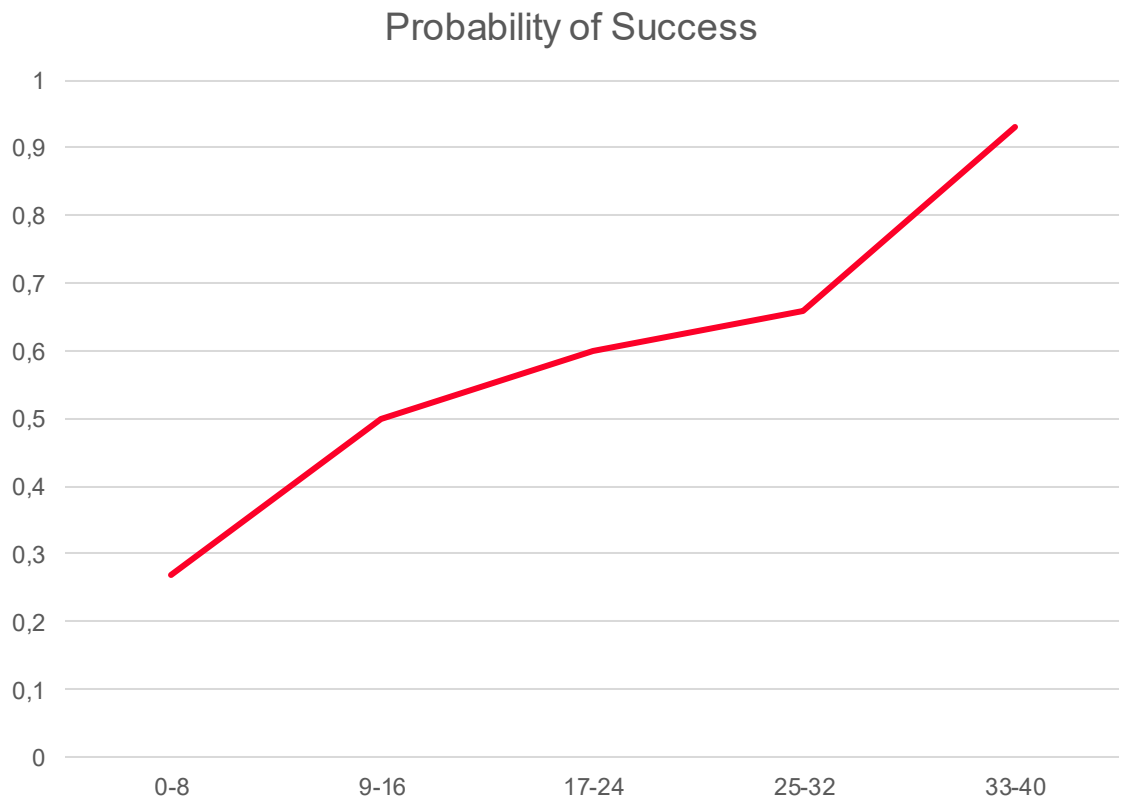
Logistic Regression

- The **logistic model** is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.
- Logistic regression(LR) is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.
- It can be used for both Classification and Regression based on the given problem.
- The goal of LR is to predict the likelihood that Y is equal to 1(probability that Y = 1 rather than 0) given certain values of X.

Example

- We may predict the success or failure of a small project on the basis of the number of years of experience of the project manager handling the project.
- This means that as X(the number of years of experience of manager) increases, the probability that Y will be equal to 1(success of project) will tend to increase.

Years of Experience	Probability of Success
0-8	0.27
9-16	0.5
17-24	0.6
25-32	0.66
33-40	0.93



END OF MODULE-2

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**

Linear regression

- **Method of least squares**: estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line.

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1\bar{x}$$

- *D:* a training set
- *x*: values of predictor variable
- *y*: values of response variable
- |*D*| : data points of the form($x1$, $y1$), ($x2$, $y2$),…, ($x|D|$, $y|D|$).
- $\bar{x}$ *:* the mean value of $x1$, $x2$, … , $x|D|$
- $\bar{y}$ *:* the mean value of $y1$, $y2$, … , $y|D|$

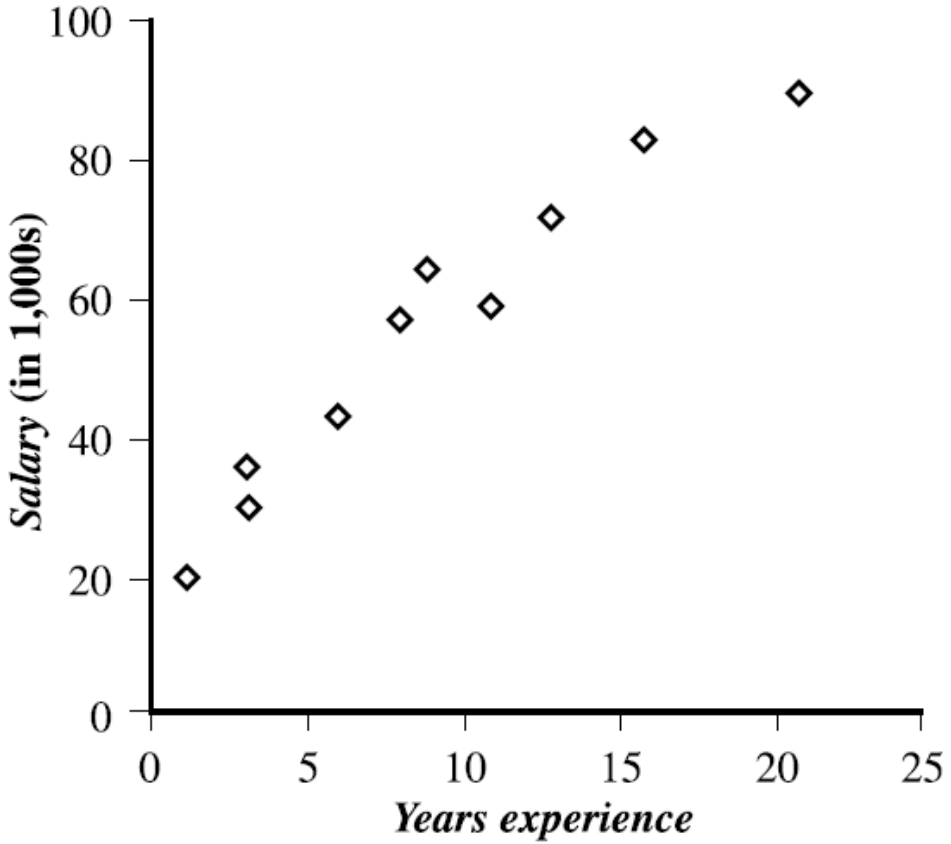
Example: Salary problem

- The table shows a set of paired data where *x* is the number of years of work experience of a college graduate and *y* is the corresponding salary of the graduate.

x years experience	*y* salary (in $1000s)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Linear Regression

- The 2-D data can be graphed on a **scatter plot**.
- The plot suggests a linear relationship between the two variables, *x and y.*



Example: Salary data

- Given the above data, we compute

$$\bar{x} = 9.1 \text{ and } \bar{y} = 55.4$$

- we get

$$w_1 = \frac{(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\cdots+(16-9.1)(83-55.4)}{(3-9.1)^2+(8-9.1)^2+\cdots+(16-9.1)^2} = 3.5$$

$$w_0 = 55.4-(3.5)(9.1) = 23.6$$

- The equation of the least squares line is estimated by

$$y = 23.6 + 3.5x$$

Example: Salary data