

School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS **(23BCAD4C01) MODULE 4: Machine Learning and** **Model Building**

Dr. Ananta Charan Ojha, Professor

Session -1

2

- Machine Learning
- Introduction

A Simple Example of an ML Problem

- ❑ A credit card company receives thousands of applications for new cards. Each application contains information about an applicant's
 - Age, Marital status, dependents, House, Vehicle, Annual income, Outstanding debts, Credit rating, etc.
- ❑ **Problem:** to decide whether an application should be **approved** or **not approved**
- ❑ **More Motivating Examples**
 - 1) **Google** gives accurate autocomplete prompts in **search box** as if it is reading your mind about search context.
 - 2) **Google Maps** predicts which route will be faster and shows an accurate ETA.
 - 3) **Gmail filters** spam mails to the spam folder in your Gmail account.
 - 4) **Amazon's** and **Netflix** recommending what you buy and watch- **production recommendations**.
 - 5) **Facebook** and **Twitter** monitor and check hate messages / fake news.
 - 6) **Airlines** keep track of buying trends and offer competitive pricing – **Dynamic pricing**.

What is Learning?

Learning makes someone Intelligent, perform a task better

- Learning is the **process of acquiring new understanding**, knowledge, behaviors, or skills **through study, experience** or being taught.
- The ability to learn is possessed by humans, birds and animals, even certain plants – **Natural Learning**.
- Today, computers/ machines are able to learn- **Artificial Learning/ Machine Learning**.
- **Computers can learn and act like humans do**, and improve their learning over time in autonomous fashion, **by taking input data** in the form of observations and real-world interactions.

Why Machines need to Learn?

5



- ❑ Learning is essential for unknown environments, dynamic environments
 - ❑ i.e., when system designer/ programmer lacks omniscience (every knowledge)
 - ❑ **First**, the designers cannot anticipate all possible situations that the computer program might find itself in.
 - ❑ E.g., a robot designed to navigate must learn the layout of the area it encounters.
 - ❑ **Second**, the designers cannot anticipate all changes over time;
 - ❑ E.g., a intelligent system designed to predict tomorrow's stock market prices must learn to adapt when conditions change from boom to bust.
 - ❑ **Third**, sometimes designers have no idea how to program a solution themselves.
 - ❑ E.g., most people are good at recognizing the faces of family members, but even the best programmers are unable to program a computer to accomplish that task, except by using learning algorithms.
- ❑ Learning is useful as a system construction method,
 - ❑ i.e., expose the computer program to reality rather than trying to write down everything. Let the computer create rules/ finds steps to solve a problem. Create learning algorithms that solve different problems.
- ❑ Learning modifies the computer program's decision mechanisms and improve performance
 - ❑ Dynamic and improved decisions

Machine Learning- Broad Timeline

6

- 1950, **Turing Test** (to determine whether a machine possesses Artificial Intelligence or not), at its simplest, the test requires a machine to carry on a conversation via text with a human being. If after five minutes the human is convinced that they're talking to another human, the machine is said to have passed.
- 1952, **Arthur Samuel** first came up with the phrase "**Machine Learning**". Arthur Samuel of IBM developed a computer program for playing checkers in the 1950s.
- 1967, the **nearest neighbor algorithm** was conceived, which was used for mapping routes and finding the most efficient route.
- 1990s, Work on machine learning intensified, developments in **neural network such as Backpropagation network**, Recurrent Neural Network, Reinforcement Learning got prominence. Commercialization of Machine Learning on Personal Computers
- 1997, IBM **Deep Blue**, a chess game program beats Kasparov, the world champion in Chess.
- 2002, Torch Machine Learning Library was released. Today, **PyTorch ML framework** is very popular.
- 2006, **Facial Recognition** Becomes a Reality, National Institute of Standards and Technology in its Face Recognition Grand Challenges programme found that some of the algorithms were able to **outperform human participants** in recognizing faces and could uniquely identify identical twins.
- 2010, **Kaggle**, a website that serves as a platform for machine learning competitions, is launched.
- 2011, IBM's **Watson beats two human champions in a Jeopardy** (US television quiz show) competition. It used a combination of machine learning, natural language processing and information retrieval techniques, .
- 2012, Google's X Lab developed **Google Brain**, an ML algorithm that can autonomously browse and find YouTube videos containing cats.
- 2014, Facebook developed **DeepFace**, an algorithm capable of recognizing or verifying individuals in photographs with the same accuracy as humans (97.35%).
- 2017, **Waymo** introduced completely **autonomous taxis** in the city of Phoenix, USA

Defining Machine Learning

7

- ❑ Machine learning (ML) is a technique by which a computer program performs a task better as it learns from the data given to it.
- ❑ Machine learning focuses on developing computer algorithms that can learn from data and improve their performance in performing a task with experience.
- ✓ Tom Mitchell defines machine learning which is formal and widely used.

A computer program is said to learn from **experience E** with respect to some **task T** and **performance measure P**, if its performance at task **T**, as measured by **P**, improves with experience **E**.

- ❑ The definition serves as a template to analyse machine learning problems with less ambiguity.
- ❑ It could be used to design machine learning software systems by clearly identifying three features:
 - ❑ **what task or decision the software needs to make (T), what data inputs the software is provided with (E), and how to evaluate its results (P).**
- ❖ E.g., a computer program that checks spam emails might improve its performance of correctly classifying an email as spam and putting it in the correct folder.
 - In this case, classifying an email as spam or not-spam is the task (T),
 - a set of example emails consisting of spam and not-spam is the experience or data inputs (E),
 - and classification accuracy as a percentage is the performance measure (P).

Identify T, E, and P in the following cases:

- 1) A program that approves a bank loan application.
- 2) A program that checks Covid-19 in chest X-rays image.
- 3) A program that predicts stock price for next one week.
- 4) A program that predicts if a student will graduate on time.
- 5) A program that finds if a person has sleeping disorder.
- 6) A program that finds diseases in a plant leaf.

THANK YOU

Any questions...?



School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS **(23BCAD4C01) MODULE 4: Machine Learning and** **Model Building**

Dr. Ananta Charan Ojha, Professor

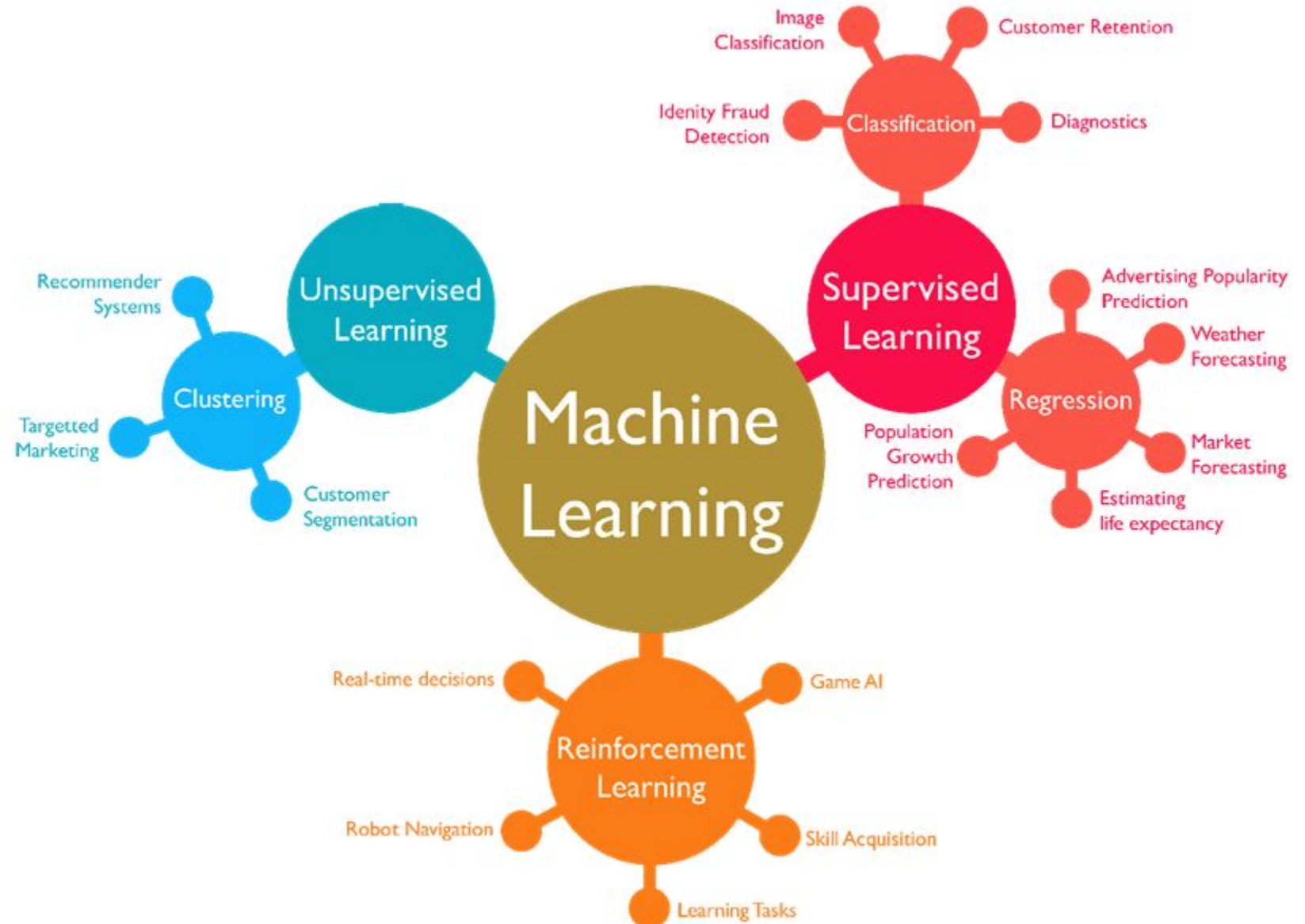
Session -2

2

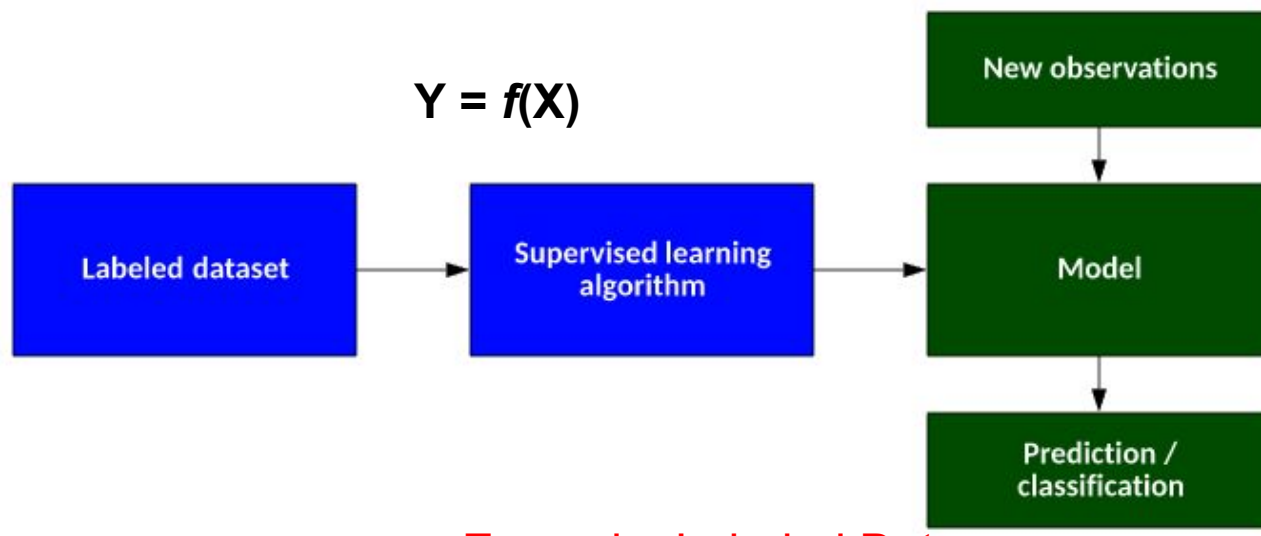
- Machine Learning
 - Types of ML
 - Modelling Process

Types of Machine Learning

- Supervised
- Unsupervised
- Reinforcement



$$Y = f(X)$$



Example: Labeled Data

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

Supervised (Learning-from-Examples)

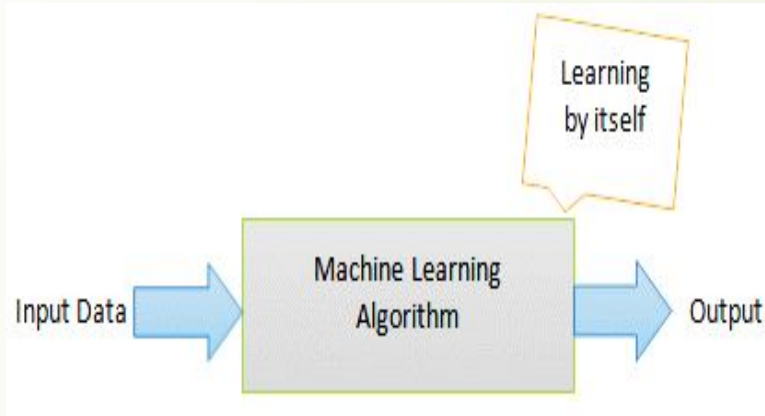
- The computer program is provided with a set of labeled data called a training set.
- In the training set, each input data is **labeled** with an output value as per the prior knowledge, **known as ground truth**.
- The goal of the learning algorithm is to **approximate a function** that maps each input to the desired output as observed in the sample data.
- Once learned, the mapping function, called model is used to predict a label for unseen input data.

Examples: Classification, regression

Nominal Attributes

- outlook {sunny, overcast, rainy}
- temperature {hot, mild, cool}
- humidity {high, normal}
- windy {TRUE, FALSE}
- play {yes, no}

rainy	mild	high	TRUE	?
-------	------	------	------	---



Example: Unlabeled Data

```

1,0,0,0,0,0,0,0
1,1,1,0,0,0,1,0
1,0,0,0,0,0,0,0
1,1,1,1,0,0,1,1
1,0,1,1,1,0,1,1
1,1,1,0,1,0,0,0
1,0,1,0,0,0,1,1
1,0,1,0,1,0,0,0
1,1,1,0,1,0,1,0
1,0,1,1,1,1,1,1
1,0,1,1,1,1,1,0
1,0,1,1,0,1,0,0
1,0,1,1,0,0,1,1
1,1,1,0,0,1,1,0
  
```

Unsupervised (Learning-from-Observation)

- There are many application areas of machine learning where the **ground truth** or prior knowledge on data is **not available** or difficult to obtain.
- In those situations, the available data is unlabeled i.e., **not in the form of input-output pairs**.
- So, no training data is provided to the computer program and the learning algorithm is **made to learn by itself** through searching hidden patterns in the input data.
- It then draws inferences that describe the hidden patterns.

❖ Example: Clustering

Example: BMW Customer Dataset

Dealership website:

0 = did not visit to browse;
1 = did visit to browser

Showroom:

0 = did not come into the showroom;
1 = did come inside showroom

Computer Search:

0 = Did not look at the cars online;
1 = Did look at the cars online

M5:

0 = Did not show interest in the M5;
1 = Did show interest in the M5

Z4:

0 = Did not show interest in the Z4;
1 = Did show interest in the Z4

3-series:

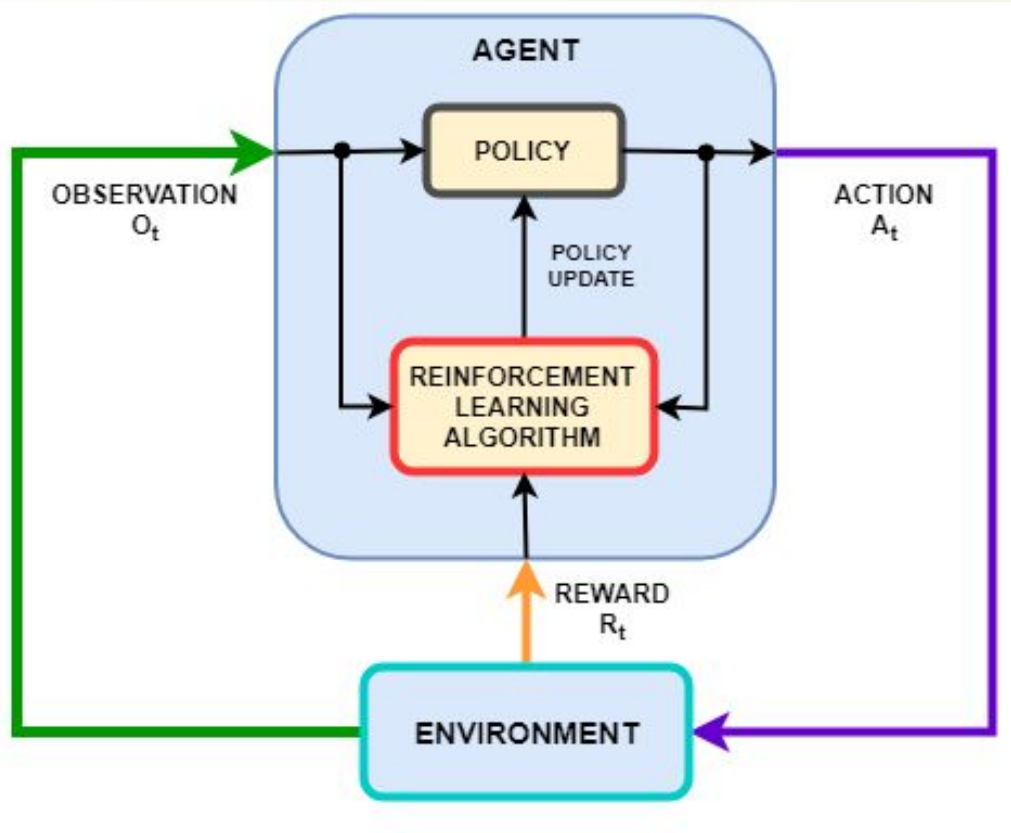
0 = Did not show interest in the 3 series;
1 = Did show interest in the 3 series

Financing:

0 = Did not qualify for financing;
1 = Did qualify for financing

Purchase:

0 = Did not purchase a car;
1 = Did purchase or lease a car

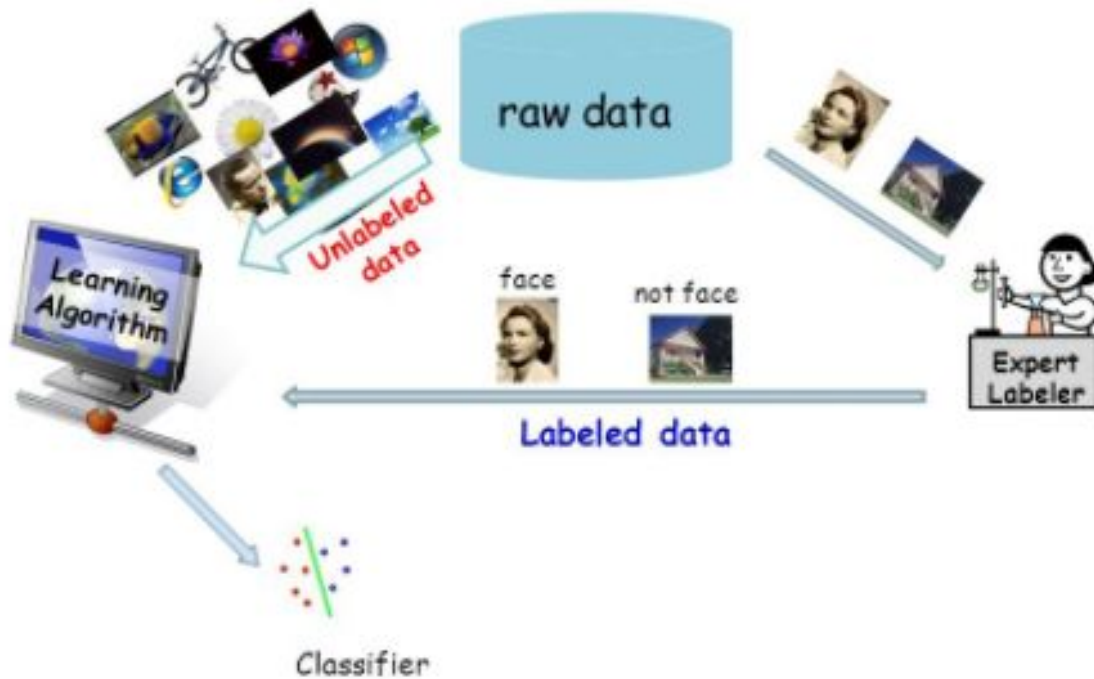


❑ Reinforcement Learning (Learning-from-Actions)

- ❑ In reinforcement learning, there is **no training data** provided to the computer program, rather the learning algorithm termed as an agent is **put to interact with a work environment** that employs a reward system.
- ❑ The agent learns to carry out its intended task by performing actions in the environment and **looking at the outcomes**.
- ❑ The **reward system interprets** the outcome of an action and decides if it is favourable or not.
- ❑ If the **outcome is favourable**, the reward system assigns a **positive reward score** and the action is encouraged or reinforced.
- ❑ If the **outcome is not favourable**, the reward system assigns a **negative reward score**, and the agent iterates for a better action that provides a reward score.
- ❑ The **agent learns through** experience following a **trial-and-error** approach to maximize the expected cumulative score.
- ❑ Since the agent gathers examples in form of good-action and bad-action, reinforcement learning is called “learning from actions”.
- ❑ Primarily, reinforcement learning is used in **autonomous robots, self-driving cars, and gaming applications**.
- ❑ It can also be used in many other domains such as **recommender systems, real-time bidding, stock trading using bots, chatbots, and portfolio management**.

Semi Supervised Learning

Semi-Supervised Learning



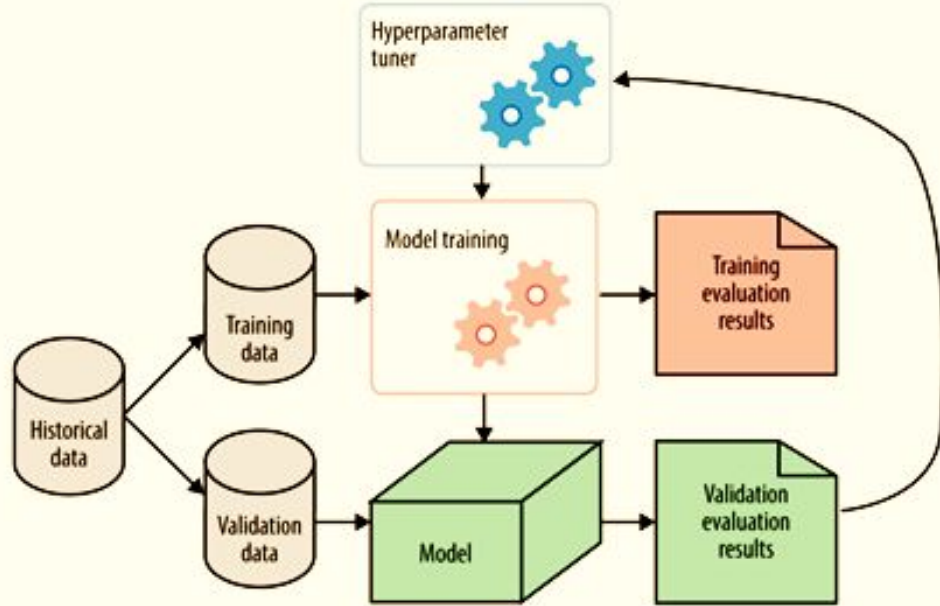
❖ When adequate amount of Labeled data is not available to train the ML algorithms, or no labeled data is available, semi supervised algorithm is used to label the unlabeled data, and then you train the ML algorithm.

□ Self-Training

- Train supervised model on the few of the available labeled data L
- Test on unlabeled data U
- Add the most confidently classified instances of U to L
- Repeat

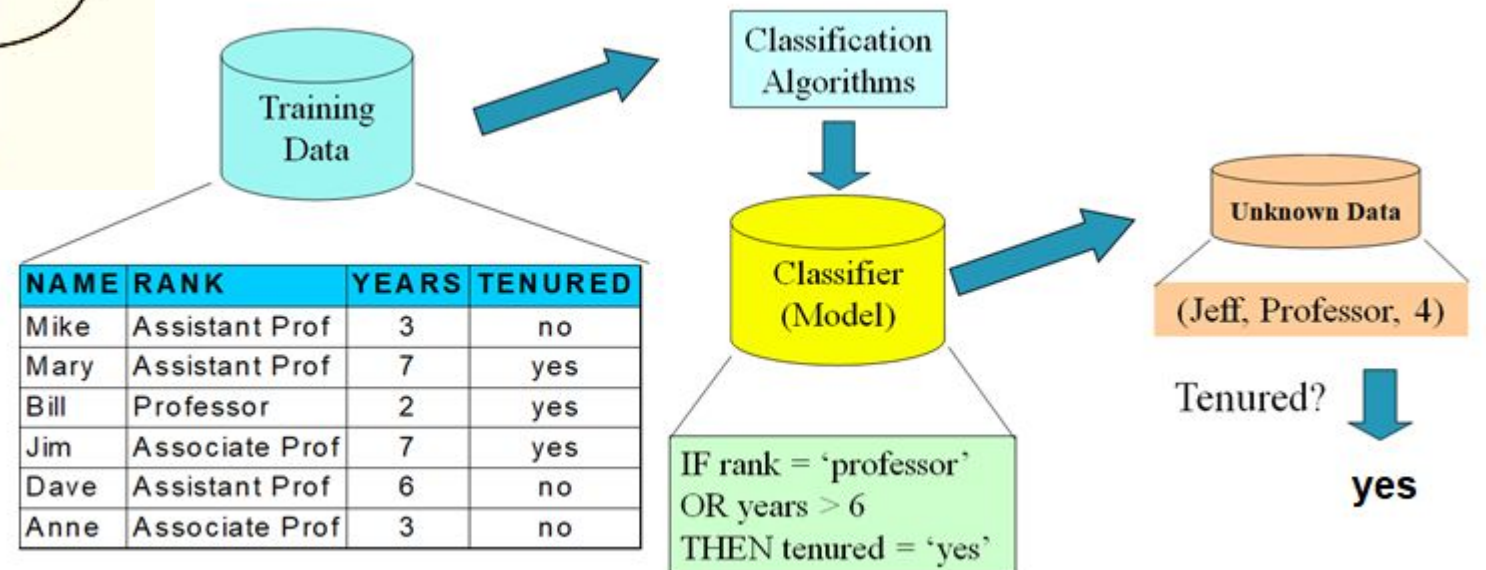
Modelling Process

8



□ The modeling phase consists of four steps:

- 1 Feature engineering
- 2 Training the model
- 3 Model validation and selection
- 4 Applying the trained model to unseen data



THANK YOU

Any questions...?



School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS **(23BCAD4C01) MODULE 4: Machine Learning and** **Model Building**

Dr. Ananta Charan Ojha, Professor

Session -3

2

- Machine Learning
 - Supervised Learning
 - Regression

Regression Example

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
...				

- The Advertising data set consists of the sales of a product in 200 different markets (200 rows), along with advertising budgets for the product in each of those markets for three different media: TV, Radio and Newspaper.
- Input Variables: X_1, X_2, X_3, \dots {TV/Radio/Newspaper budget}
 - Predictors, independent variables, factors
- Output Variable: Y {Sales}
 - Response, dependent variables
- Questions we might ask:
 - Is there a relationship between advertisement budget and sales?
 - How strong is the relationship between advertisement budget and sales?
 - Is the relationship linear?
 - How can we accurately predict the future sales of the product?
 - Which Ad type contributes to the sales the most?
 - What is the effect of each Ad type on sales?
 - Is there any synergy / interaction among different advertisement media?
Does one variable influence the predictive power of another variable?
- Answer: Regression Analysis

Regression

4

- Regression in general is a method of modelling a dependent variable based on independent variables known as predictors. The predicted **output is continuous**. It is a statistical method that is **used for predictive analysis**.
- Regression techniques mostly **differ** based **on the number** of independent variables and the type of **relationship** between the independent and dependent variables. (**simple, multiple, linear, non-linear**).
- Simple linear regression is a type of regression analysis where the **number of independent variables is one** and there is a **linear relationship** between the independent(x) and dependent(y) variables.
- Linear regression is a **linear model** that assumes a linear relationship between the input variables (X) and the single output variable (Y).

$Y = f(X) + \epsilon$ where f is a fixed but unknown function, ϵ is the error term

- More specifically, that Y can be calculated from a linear combination of the input variables (X).

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.

X:

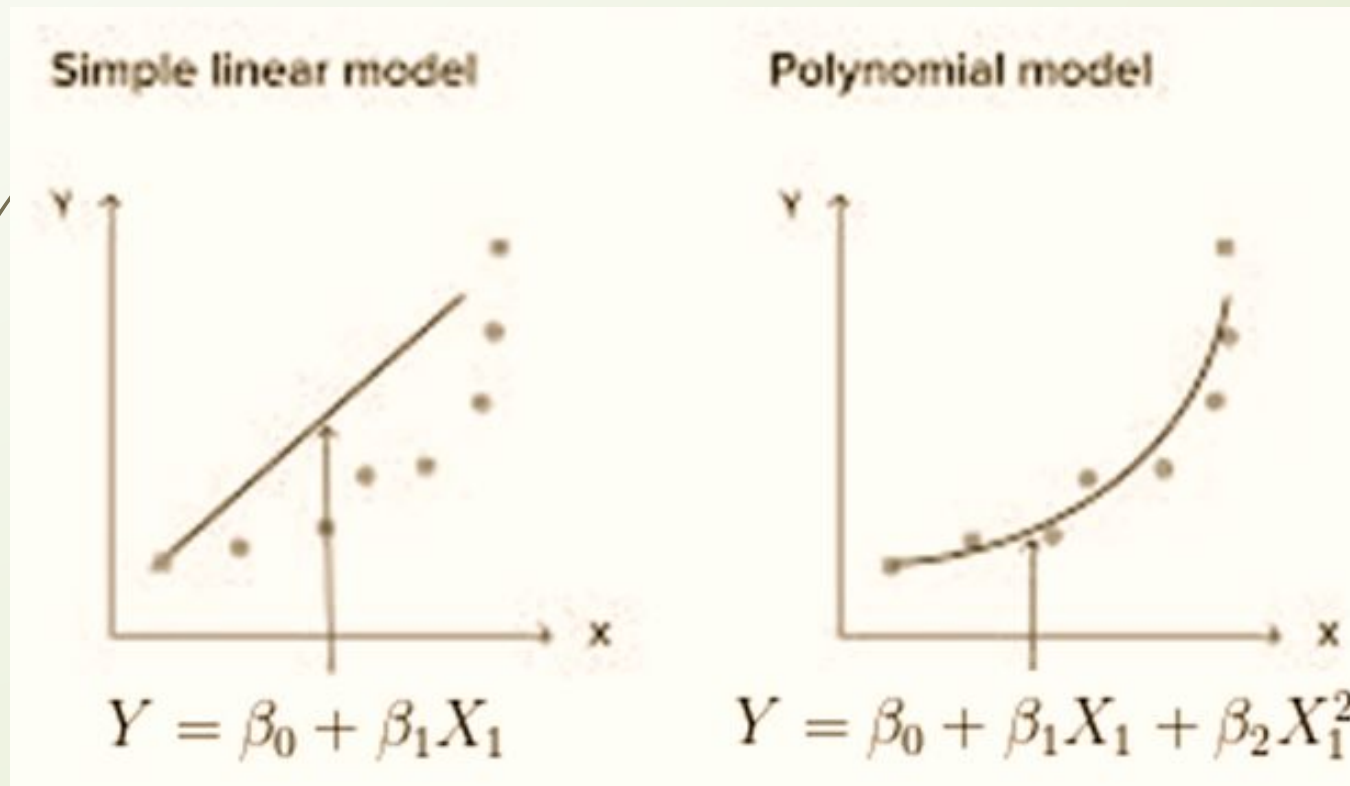
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

This is a *linear model*,

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon.$$

Relationship between Independent and Dependent variables

- **Linear regression** assumes the relationship between the independent variables and the dependent variable is linear, meaning the **change in the dependent variable is constant for the same change in the independent variable**. The model fits a **straight line** to the data.
- **Non-linear regression** models a non-linear relationship, meaning the **change in the dependent variable is not constant for the same change in the independent variable**. The data fits a **curved line** instead of a straight one.



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

Estimating the Coefficients of simple Regression

6

- To estimate these coefficients, we need to use training data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- The training observations, in this case pairs of X and Y measurements.
- The goal is to use these measurements to estimate the coefficients, such that the linear model fits our data as close as possible.
- Measuring *closeness* can be done by multiple methods, but **Least Squares** method is the most popular.

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i th value of X . Then $e_i = y_i - \hat{y}_i$ represents the i th *residual*
- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \dots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

Estimating the Regression Coefficients Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \dots, \beta_p$ in the multiple regression formula are unknown, and must be estimated.

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- The parameters are **estimated using the same least squares method** that we saw in the context of simple linear regression. We choose $\beta_0, \beta_1, \dots, \beta_p$ to minimize the sum of squared residuals (RSS).

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

- Minimizing these coefficients is more complicated than the simple linear regression setting, and is best represented using linear algebra.

Weka: Regression

- Dataset: cpu.arff

- Attributes:

1. MYCT: machine cycle time in nanoseconds (integer)
2. MMIN: minimum main memory in kilobytes (integer)
3. MMAX: maximum main memory in kilobytes (integer)
4. CACH: cache memory in kilobytes (integer)
5. CHMIN: minimum channels in units (integer)
6. CHMAX: maximum channels in units (integer)
7. class : cpu performance predicted (integer)

- Model Prediction

- Prepare a test set for prediction (we can use Linear Regression)

Time Series Forecasting

- **Time-series data** is a time-dependent series of data points collected over equal spaced time intervals.
- Time series forecasting is the process of using a model to generate predictions (forecasts) for future events based on known past events.
- ❖ Dataset: Weka's airline.arff data
 - Use different algorithms (Linear Regression, SMOreg, RandomForest etc)

THANK YOU

Any questions...?



School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS **(23BCAD4C01) MODULE 4: Machine Learning and** **Model Building**

Dr. Ananta Charan Ojha, Professor

Session -4

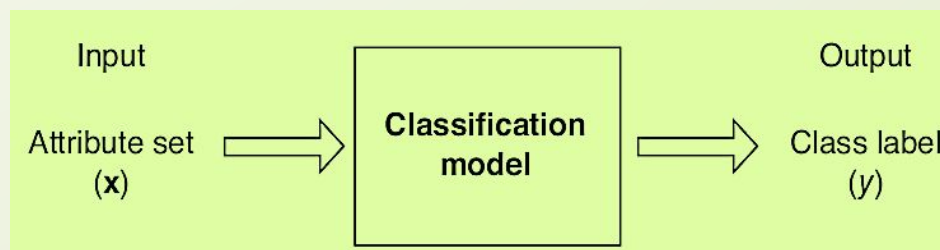
2

- Machine Learning
 - Supervised Learning
 - Classification

What is Classification?

3

- Classification is a type of supervised learning.
 - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
- **Training Data**: A set of data records (also called examples, instances or cases) described by
 - A set of k attributes / input variables / predictor variables: A_1, A_2, \dots, A_k .
 - A class variable / output variable (which is a categorical variable): Each example is labeled with a predefined class.
- **Goal**: To learn a classification model from the training data that can be used to predict the classes of new cases / instances (called test data) .
- In other words, the **training data set has n examples of input–output pairs** $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where each y_i is generated by an unknown function $y_i = f(x_i)$, each x_i has k attributes
- New data is classified based on the training set, where y_i is predicted when x_i is given
- **Task Examples**:
 - Given an email, classify if it is spam or ham
 - Given a plant leaf, identify the what disease it has.
 - Given a transaction, classify if it is fraudulent or not
 - Given a handwritten character, classify it as one of the known characters
 - Given a bank loan application, find if it will be granted, not granted
 - Given a student, identify if he will pass or fail in an examination



Training Data

Student#	IDA	DM	BDP	Result
1	32	30	28	Pass
2	12	15	17	Fail
3	20	24	18	Pass

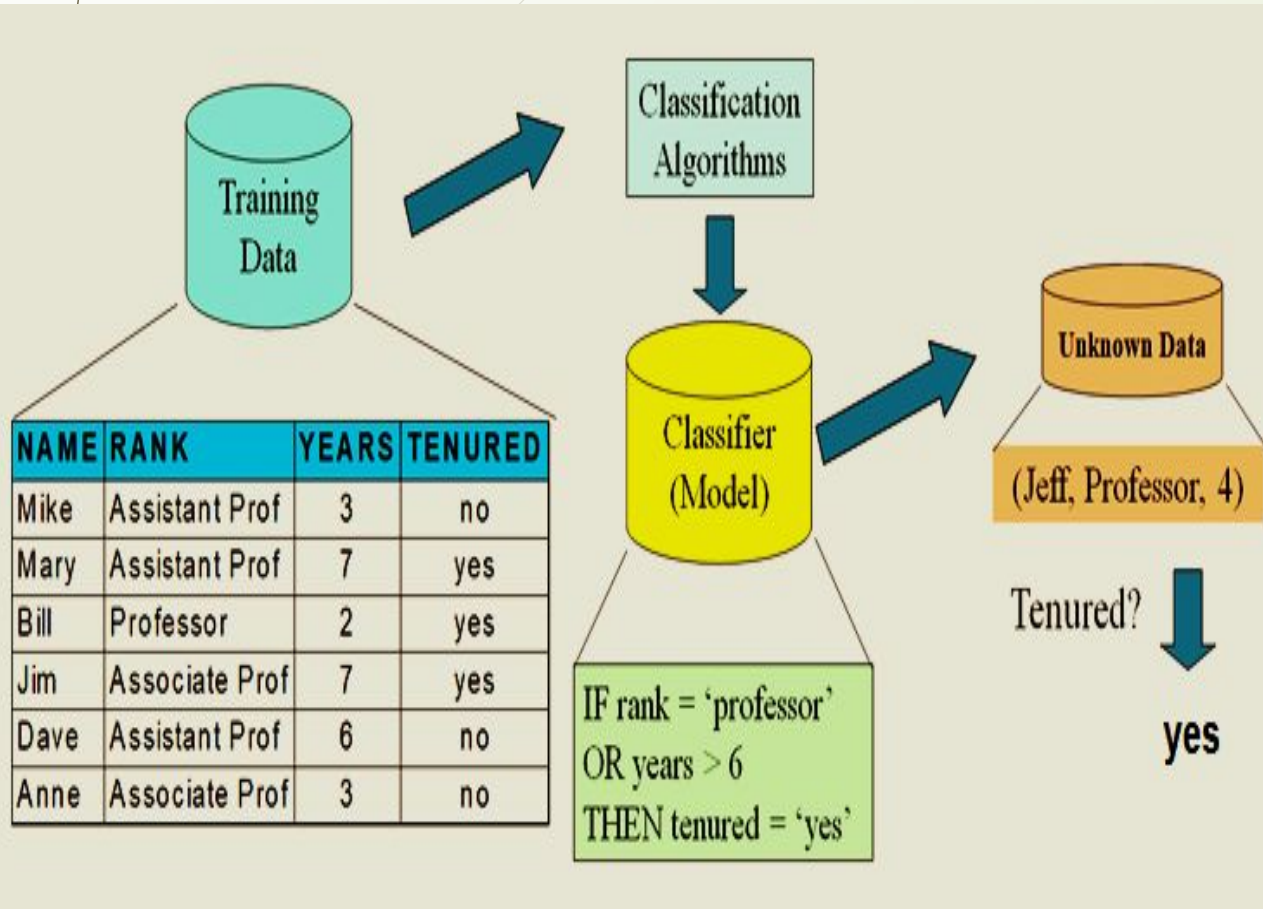
Test Data

Student#	IDA	DM	BDP	Result
1	19	30	09	?
2	21	15	27	?

Type of Classification

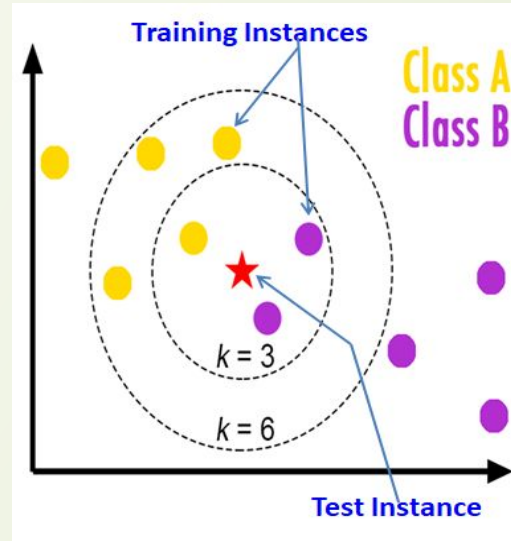
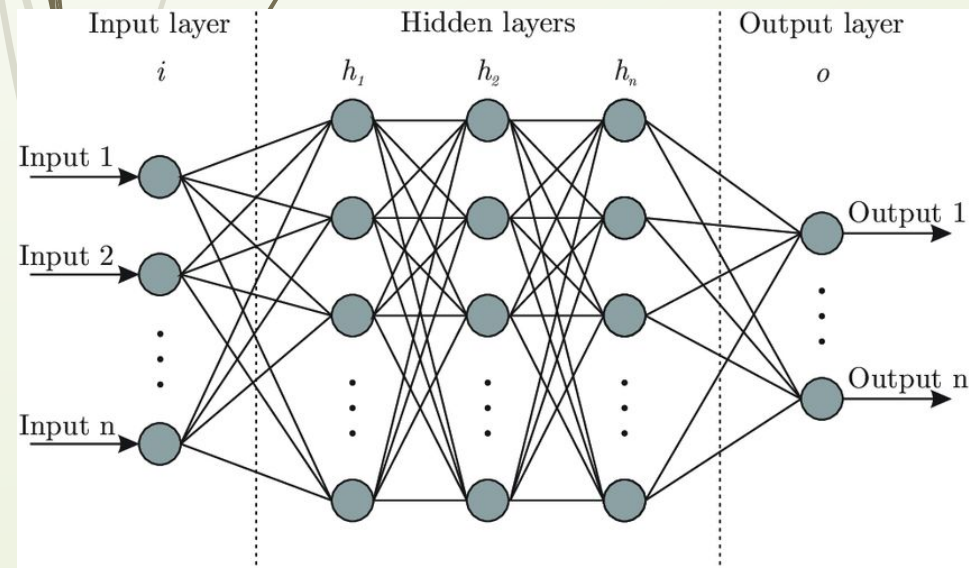
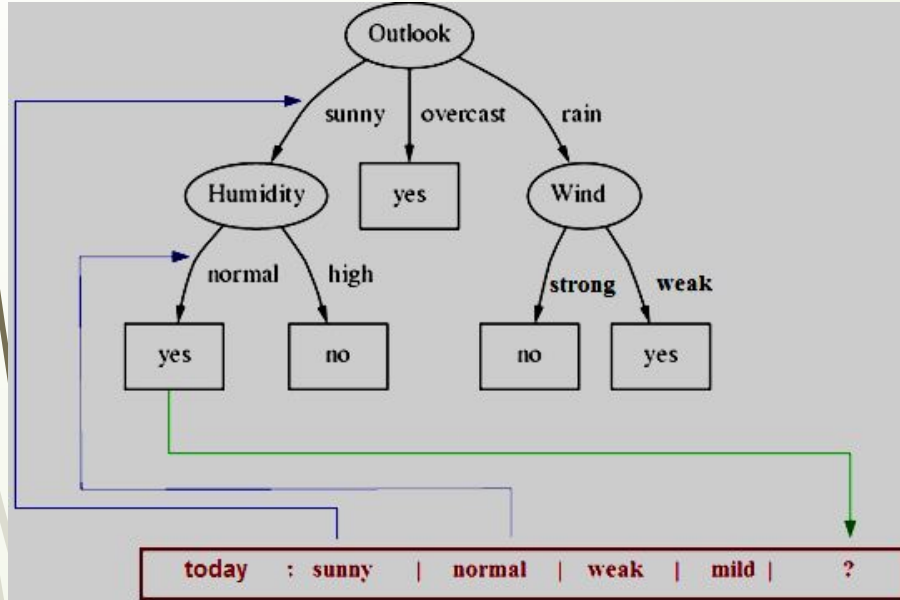
- ❑ There are 2 types of Classification: Binary Classification, Multi-Class Classification
- ❑ **Binary classification** refers to those classification tasks that **have two class labels**.
- ❑ Examples:
 - Email spam detection (spam or not).
 - Churn prediction (churn or not).
 - Bank Loan Application (approve or not).
- ❑ **Multi-class classification** refers to those classification tasks that **have more than two class labels**.
- ❑ Examples:
 - Face classification.
 - Plant disease classification.
 - Optical character recognition.

General Approach to Classification

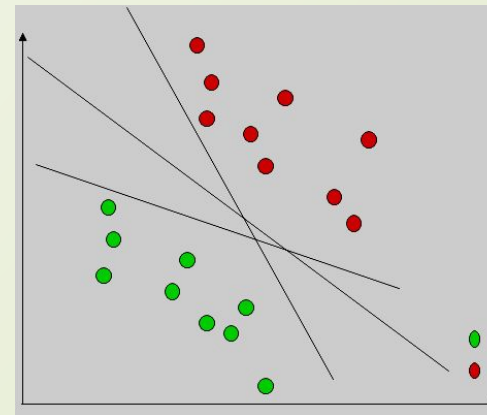


- **Classification is a two-step process:** consisting of a *learning step* (where a classification model is constructed) and a *classification step* (where the model is used to predict class labels for given data).
- **Learning Step/Model Construction:**
 - A classification model is constructed using a set of Examples (i.e. Training Set)
 - The model is represented as classification rules, decision trees, or mathematical formulae
- **Classification step/ Model Usage:**
 - The model is used to predict class labels for unseen/unknown data
 - Test data independent of training data is used for evaluation.

Some Popular Classifiers



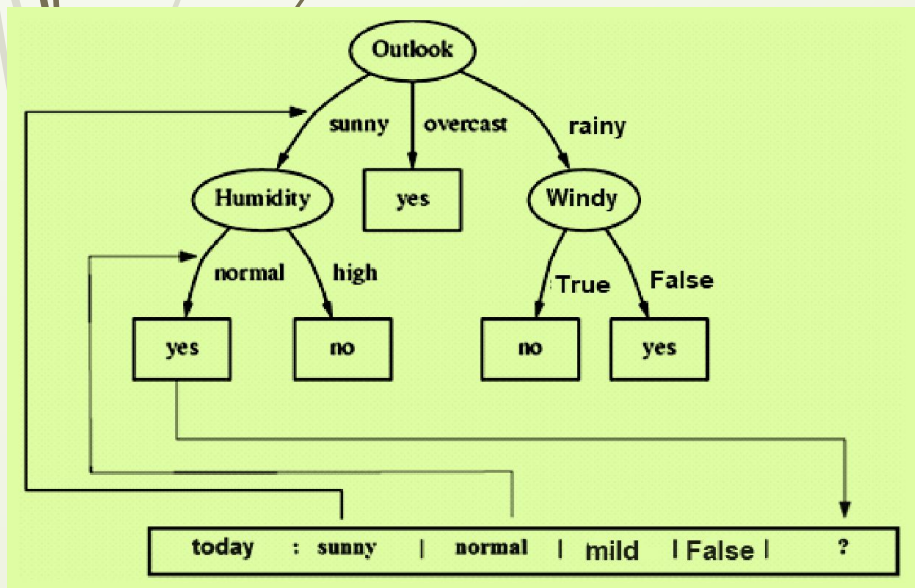
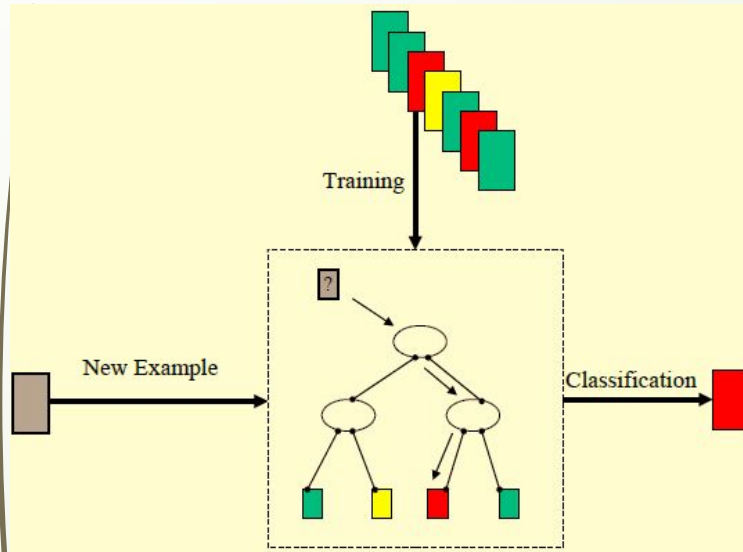
1. Decision Tree
2. K-Nearest Neighbor
3. Neural Network
4. Support Vector Machine
5. Naïve Bayes
6. Bayesian Belief Network
7. Random Forest
8. Rule-Based Classifier (e.g. ZeroR)



$$p(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

IF *condition* THEN *conclusion*

Decision Tree



- Decision tree learning is one of the simplest and yet the most successful algorithms of machine learning.
- The learning algorithm operates on the training data and develops a decision tree model.
- A decision tree represents an object or situation described by a set of input attributes, and outputs a “decision” - a single output value.
- A decision tree reaches its decision by performing a sequence of tests.
- **Decision Tree Representation:**
 - Each internal node of the tree tests the value of an input attribute
 - Branches from the node correspond to possible values of the attribute
 - Each leaf node assigns a classification (or predict the outcome)
- **Tree Interpretation:**
 - ❖ If Outlook=sunny and Humidity=normal then Play=yes

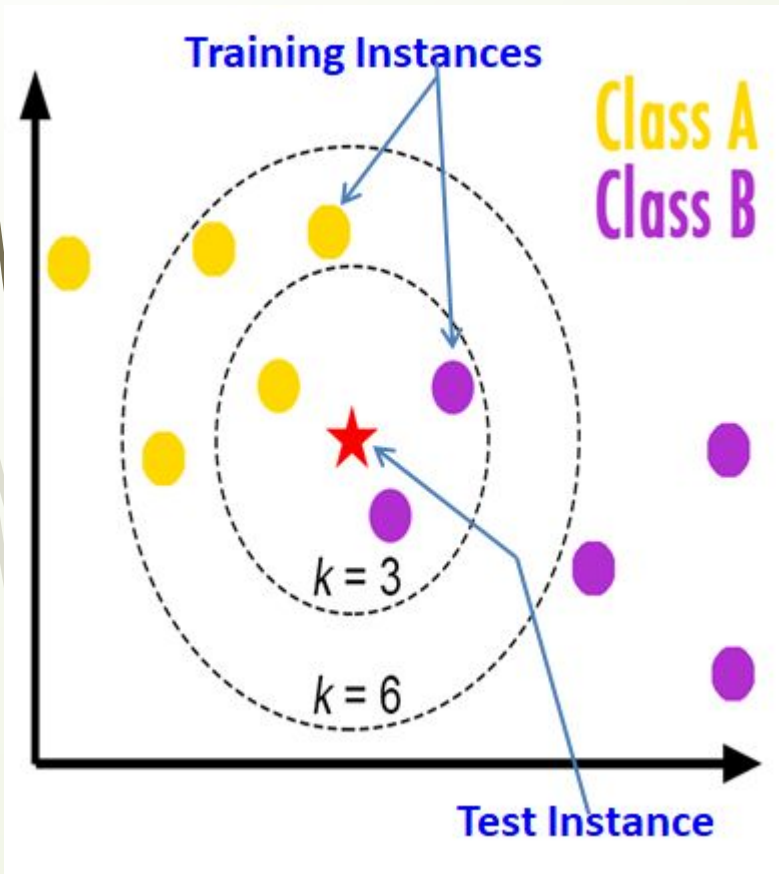
Example: Play Tennis

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

Decision Tree Learning

- The Decision-tree-learning algorithm adopts a **greedy divide-and-conquer strategy**: always test the most important attribute first.
- This test divides the problem up into smaller sub-problems that can then be solved recursively.
- **Basic Divide-And-Conquer Algorithm:**
 1. Select a test for root node
Create branch for each possible outcome of the test
 2. Split instances into subsets
One for each branch extending from the node
 3. Repeat recursively for each branch, using only instances that reach the branch
 4. Stop recursion for a branch if all its instances have the same class

K-Nearest Neighbor



- Unlike Decision Tree, K-Nearest Neighbor does not develop a model by learning from the training data to classify/ predict the test data.
- Rather, it uses all the instances to compute the classification/ prediction.
- It stores all the instances in the memory.
- All instances correspond to some points in an n -dimensional (feature vector) space.
- Classification is done by comparing feature vectors of different points with the test data based on a similarity measure like Euclidian distance, Manhattan distance etc.

Algorithm

- To determine the class of a test instance \mathbf{x}_t :
- Calculate the distance (i.e. similarity measure) between \mathbf{x}_t and all instances in the training set.
- Select k -nearest instances in the training set to \mathbf{x}_t .
- Assign \mathbf{x}_t to the majority class among its k -nearest neighbors.

Example: k-Nearest Neighbor

10

Using Euclidean

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Consider a simple loan application if it will be a defaulter.

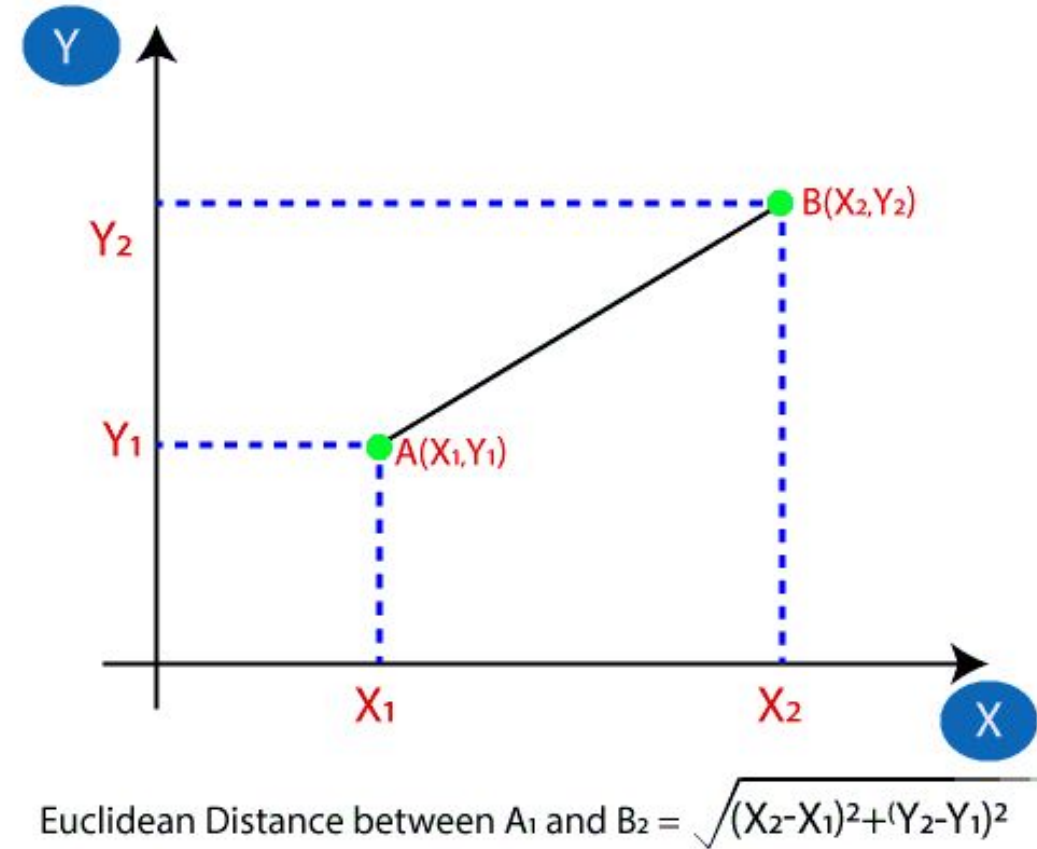
Age	Loan	Defaulter	Distance
25	40,000	N	102000
35	60,000	N	82000
45	80,000	N	62000
20	20,000	N	122000
35	120,000	N	22000
52	18,000	N	124000
23	95,000	Y	47000
40	62,000	Y	80000
60	100,000	Y	42000
48	220,000	Y	78000
33	150,000	Y	8000
48	142,000	?	

□ $k=3$;

□ An odd value of k facilitates Majority Vote

□ Larger k increases confidence in prediction.

□ Note that if k is too large, decision may be skewed.



THANK YOU

Any questions...?



School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS **(23BCAD4C01) MODULE 4: Machine Learning and** **Model Building**

Dr. Ananta Charan Ojha, Professor

Session -5

2

- Machine Learning
 - Unsupervised Learning
 - Clustering

What is Cluster Analysis?

Dataset

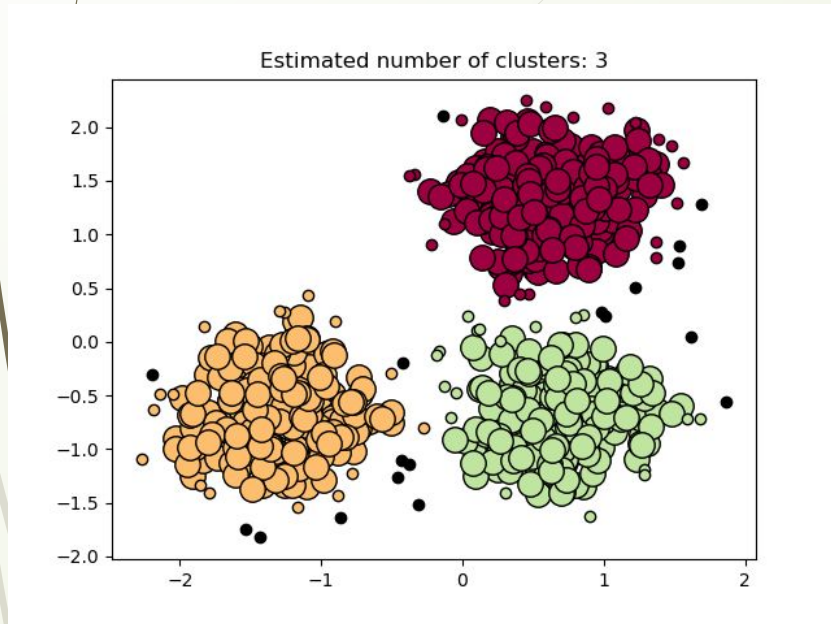
Student#	IDA	DM	BDP
1	32	30	28
2	12	15	17
3	20	24	18
4	19	30	09
5	21	15	27

- Clustering is an **unsupervised** learning because the class label information is not present in the dataset. For this reason, clustering is a form of *learning by observation*, rather than *learning by examples*.
- **Cluster**: A collection of data objects similar (or related) to one another within the same group but dissimilar (or unrelated) to the objects in other groups.
- **Cluster analysis** (or *clustering*):
 - The process of finding similarities between data according to the characteristics found in the data and grouping similar data objects into subsets called clusters.
 - Different clustering methods may generate different clusters on the same data.
 - Hence, clustering is useful in that it can lead to the discovery of previously unknown groups within the data.

Clustering Applications

- ❑ Cluster analysis is used to understand the data distribution and get insights on the data and the hidden pattern.
- ❑ **Search Engines:**
 - ❑ Search engines try to group similar user-queries in one cluster and the dissimilar user-queries far from each other.
 - ❑ An incoming user query is mapped to find the nearest cluster and provides the search result based on that cluster.
 - ❑ Better the clustering algorithm used, better are the chances of getting the required result on the front page.
- ❑ **Academics:**
 - ❑ Clustering algorithm can be used to monitor the students' academic performance.
 - ❑ Based on the students' score they are grouped into different clusters (using k-means), where each clusters denoting different level of performance.
 - ❑ By knowing the number of students' in each cluster we can know the average performance of a class as a whole.
- ❑ **Biology:** Taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- ❑ **Information retrieval:** Document clustering
- ❑ **Land use:** Identification of areas of similar land use in an earth observation database
- ❑ **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs (i.e. [customer segmentation and target marketing](#))
- ❑ **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- ❑ **Earth-quake studies:** Observed earth quake epicenters should be clustered along continent faults
- ❑ **Climate:** Understanding earth climate, find patterns of atmosphere and ocean
- ❑ **Economic Science:** Market research (finding different markets)

Quality: What Is Good Clustering?



- A good clustering method will produce high quality clusters
 - high intra-class similarity: **cohesive** within clusters
 - low inter-class similarity: **distinctive** between clusters
- The quality of a clustering method depends on
 - the similarity measure used by the method
 - its implementation, and
 - its ability to discover some or all of the hidden patterns
- **Silhouette Coefficient** or silhouette score is a metric used to calculate the goodness of a clustering technique.
- Its value ranges from -1 to 1.
 - 1: Means clusters are well apart from each other and clearly distinguished.
 - 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
 - -1: Means clusters are assigned in the wrong way.

Major Clustering Approaches

- Partitioning approach:

- Construct various partitions based on similarity and then evaluate them by some criterion, e.g., minimizing the sum of square errors
- Example Algorithms: **k-means**, k-medoids, CLARANS

- Hierarchical approach:

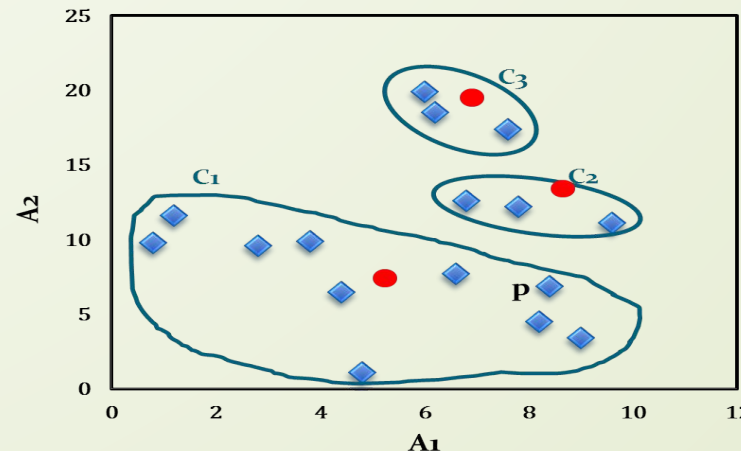
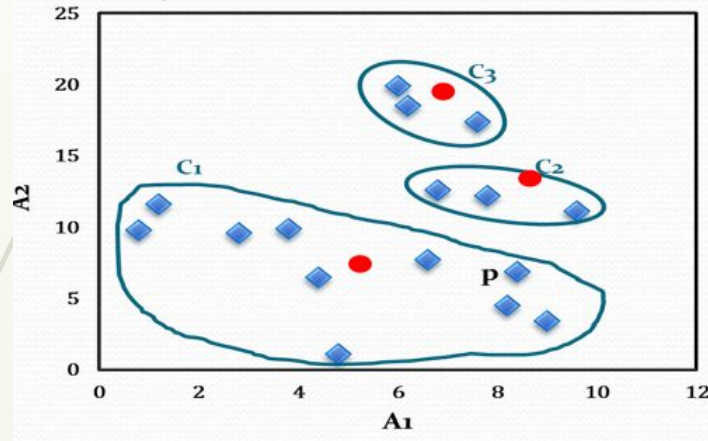
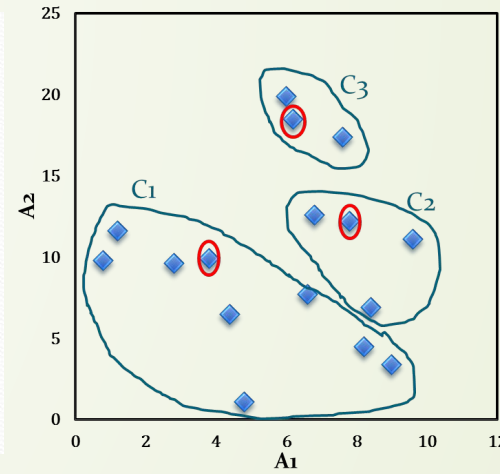
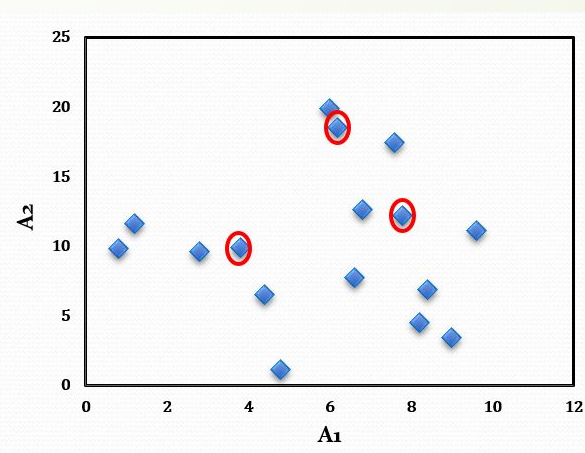
- Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Example Algorithms : Diana, Agnes, BIRCH, CAMELEON

- Density-based approach:

- Based on connectivity and density functions
- Example Algorithms : DBSACN, OPTICS, DenClue

- Model-based:

- A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
- Example Algorithms : EM, SOM, COBWEB



k-Means Algorithm

- Given a set of n distinct objects, the k-Means clustering algorithm partitions the objects into k number of clusters such that intra-cluster similarity is high but the inter-cluster similarity is low.

The algorithm can be stated as follows.

- First it selects k number of objects at random from the set of n objects. These k objects are treated as the centroids or center of gravities of k clusters.
- For each of the remaining objects, it is assigned to one of the closest centroid. Thus, it forms a collection of objects assigned to each centroid and is called a cluster.
- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).
- The assignment and update procedure is repeated until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no movement of objects among the clusters, etc.)

Determining the Number of Clusters

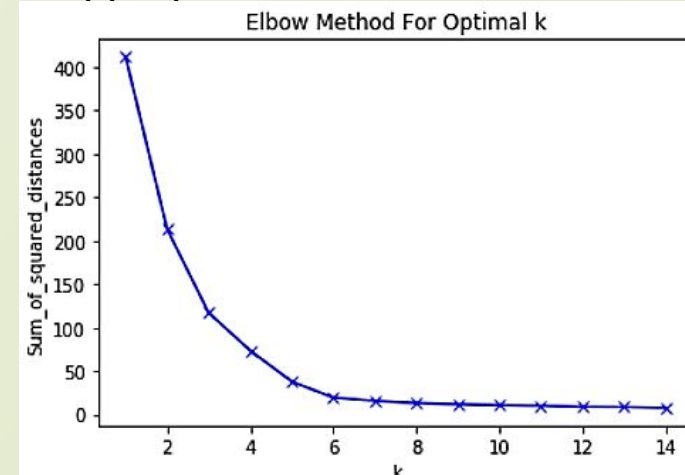
8

$$\sum_{k=1}^K \sum_{i \in S_k} \sum_{j=1}^P (x_{ij} - \bar{x}_{kj})^2$$

where S_k is the set of observations in the k th cluster
and \bar{x}_{kj} is the j th variable of the cluster center for the k th cluster

- Determining the optimal number of clusters is very important but far from easy.
- **Empirical method**
 - # of clusters $\approx \sqrt{(n/2)}$ for a dataset of n data points, and each cluster has $\sqrt{(2n)}$ data points.
- **Elbow method**: shows the turning point in the curve of sum of within cluster variance w.r.t the # of clusters.
 - **Steps:**
 1. Compute clustering algorithm (e.g., k-means clustering) for different values of k . For instance, by varying k from 2 to 14 clusters.
 2. For each k , calculate the total **within-cluster sum of square distance** (wss - Distance of each point from the centroid)
 3. Plot the curve of wss according to the number of clusters k .
 4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.
- ❖ In the plot, the elbow is at $k=5$ indicating the optimal k

Weka terms it within cluster sum of squared errors, it can be used to plot the graph and find the elbow. Smaller value of error gives cohesive clusters.



Case Study-1: BMW Dealership

9

- Dataset: bmw-browsers.arff
- A BMW dealership has customers who come to browse. Some will wander around the lot but not come into the show room. Some will come into the show room and ask about a particular car such as the 3-series, Z4 or the M5.
- Some will ultimately buy a car, some will lease a car, and others are just browsing.
- The dealership would like to have a means of identifying whether a customer is going to be a “buyer” or a “browser” and whether it is best to try to interest the customer in the 3-series, Z4 or M5 series.

Attributes

Dealership:

0 = did not visit to browse;
1 = did visit to browser

Showroom:

0 = did not come into the showroom;
1 = did come inside showroom

Computer Search:

0 = Did not look at the cars online;
1 = Did look at the cars online

M5:

0 = Did not show interest in the M5;
1 = Did show interest in the M5

Z4:

0 = Did not show interest in the Z4;
1 = Did show interest in the Z4

3-series:

0 = Did not show interest in the 3 series;
1 = Did show interest in the 3 series

Financing:

0 = Did not qualify for financing;
1 = Did qualify for financing

Purchase:

0 = Did not purchase a car;
1 = Did purchase or lease a car

```
1 @relation car-browsers
2
3 @attribute Dealership numeric
4 @attribute Showroom numeric
5 @attribute ComputerSearch numeric
6 @attribute M5 numeric
7 @attribute 3Series numeric
8 @attribute Z4 numeric
9 @attribute Financing numeric
10 @attribute Purchase numeric
11
12 @data
13
14 1,0,0,0,0,0,0,0
15 1,1,1,0,0,0,1,0
16 1,0,0,0,0,0,0,0
17 1,1,1,1,0,0,1,1
18 1,0,1,1,1,0,1,1
19 1,1,1,0,1,0,0,0
20 1,0,1,0,0,0,1,1
21 1,0,1,0,1,0,0,0
22 1,1,1,0,1,0,1,0
23 1,0,1,1,1,1,1,1
24 1,0,1,1,1,1,1,0
25 1,0,1,1,0,1,0,0
26 1,0,1,1,0,0,1,1
27 1,1,1,0,0,1,1,0
28 1,0,1,1,1,1,0,0
29 1,1,1,1,1,0,1,1
30 1,0,1,0,0,0,1,1
31 1,0,1,0,0,0,1,0
32 1,1,0,1,1,0,0,0
33 1,0,0,1,0,0,0,0
34 1,1,1,0,0,1,1,1
35 1,0,1,0,0,1,1,1
36 1,1,0,0,0,0,1,0
37 1,1,0,1,0,0,0,0
38 1,1,0,1,1,1,1,0
39 1,1,1,0,1,1,0,0
40 1,1,0,1,1,1,1,0
41 1,1,0,1,1,0,1,0
42 1,1,1,1,1,1,0,0
43 1,1,0,1,1,1,1,0
44 1,1,0,1,0,1,1,0
```

Cluster output

Attribute	Cluster#					
	Full Data (100)	0 (26)	1 (27)	2 (5)	3 (14)	4 (28)
=====						
Dealership	0.6	0.9615	0.6667	1	0.8571	0
Showroom	0.72	0.6923	0.6667	0	0.5714	1
ComputerSearch	0.43	0.6538	0	1	0.8571	0.3214
M5	0.53	0.4615	0.963	1	0.7143	0
3Series	0.55	0.3846	0.4444	0.8	0.0714	1
Z4	0.45	0.5385	0	0.8	0.5714	0.6786
Financing	0.61	0.4615	0.6296	0.8	1	0.5
Purchase	0.39	0	0.5185	0.4	1	0.3214

Clustered Instances

0	26 (26%)
1	27 (27%)
2	5 (5%)
3	14 (14%)
4	28 (28%)

- **Cluster 0**— We can call this group the “**BMW Dreamers**,” as they appear to wander around the dealership, looking at cars parked outside on the lots, but trail off when it comes to coming into the dealership, and worst of all, they don’t purchase anything.
- **Cluster 1**— We’ll call this group the “**M5 Lovers**” because they tend to walk straight to the M5s, ignoring the 3-series cars and the Z4. However, they don’t have a high purchase rate — only 52 percent. This is a potential problem and could be a focus for improvement for the dealership, perhaps by sending more salespeople to the M5 section.
- **Cluster 2**— This group is so small; we can call them the “**Throw-Aways**” because they aren’t relevant.
- **Cluster 3**— We’ll call this group the “**BMW Babies**” because they always end up purchasing a car and always end up financing it.
- **Cluster 4**— This group, we’ll call the “**BMW starters**” because they always look at the 3-series and never look at the much more expensive M5. The dealership could possibly increase sales to this group by relaxing their financing standards or by reducing the 3-series prices.

THANK YOU

Any questions...?



School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS **(22BCAD4C01) MODULE 4: Machine Learning and** **Model Building**

Dr. Ananta Charan Ojha, Professor

Session -6

2

□ Machine Learning

□ Association Rule Mining

Association Rule Mining

3

- **Association rule mining** is a **rule-based machine learning** method for discovering interesting relations between variables in large databases.
- Given a set of transactions, it finds rules that will predict the occurrence of an item based on the occurrences of other items in the **transaction**.
- **Transaction** data can be a set of *itemsets* containing bag of words, a set of genes in gene expression dataset, a set of patterns etc.

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules









$\{\text{Diaper}\} \rightarrow \{\text{Beer}\},$
 $\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\},$
 $\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\},$

In general: $X \Rightarrow Y$

if basket contains X, then it also contains Y

ID	Expressed Genes in Sample
1	GENE1, GENE2, GENE 5
2	GENE1, GENE3, GENE 5
3	GENE2
4	GENE8, GENE9
5	GENE8, GENE9, GENE10
6	GENE2, GENE8
7	GENE9, GENE 10
8	GENE2
9	GENE11

doc1:	Student, Teach, School
doc2:	Student, School
doc3:	Teach, School, City, Game
doc4:	Baseball, Basketball
doc5:	Basketball, Player, Spectator
doc6:	Baseball, Coach, Game, Team
doc7:	Basketball, Team, City, Game

1	 A	 B
2	 A	 C
3	 D	 C
4	 A	 A



- Association rules do not represent any sort of **causality** or **correlation** between the two itemsets.

- $X \Rightarrow Y$ does not mean X causes Y , so no Causality
- $X \Rightarrow Y$ can be different from $Y \Rightarrow X$, unlike correlation

□ Association rule types:

- **Actionable Rules** – contain high-quality, actionable information (interesting/**surprising** rules)
- **Trivial Rules** – information already well-known by those familiar with the domain
- **Inexplicable Rules** – no explanation, incomprehensible and do not suggest action

- Trivial and Inexplicable Rules occur most often, and not of much use.

□ Example of Actionable Rules:

- In a large transaction dataset of patient symptoms and interventions (including drugs taken):

{warfarin, levofloxacin} → {nose bleeds}

- Then a dangerous drug interaction is discovered. Both **warfarin** and **levofloxacin** are useful drugs by themselves, but together they are *dangerous*.
Both **warfarin** and **levofloxacin** are useful drugs by themselves, but together they are *dangerous*.
Both **warfarin** and **levofloxacin** are useful drugs by themselves, but together they are *dangerous*.
warfarin is a blood clotting, makes blood flow; used to treating heart disease
levofloxacin is an antibiotic medication to treat bacterial infection

Example Applications

❑ Retail Business

- Store layout or arrangement in a supermarket
- Determine what items are purchased together, purchased sequentially, and purchased by season
- Customer profiling and target marketing

❑ Telecommunication

- Analyzing customer churn (why customers leave the company?)

❑ Banking and Insurance

- Credit card purchase and usage profile of customers
- Insurance claims profiling

❑ Medical Diagnosis

- Adding new symptoms types for the given disease, and by defining new relations between these symptoms
- Analysis can be used for comorbid conditions and symptom analysis
- Finding disease and medication with effects and side-effects
- To reveal biologically relevant associations between different genes or between environmental effects and gene expression.

❑ Web Design

- Based on browsing behavior/pattern of visitors, website navigation can be designed

Basic Concepts

Itemset

- A collection of one or more items
 - Example: {Milk, Bread, Diaper}
- k-itemset
 - An itemset that contains **k** items

Support count (σ)

- Frequency of occurrence of an itemset
- E.g. $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$

Support (range from 0 to 1)

- Fraction of transactions that contain an itemset
- E.g. $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$

Frequent Itemset

- An itemset whose support is greater than or equal to a *minsup* threshold

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Beer, Diaper
5	Bread, Milk, Diaper, Coke

Association Rule

- An implication expression of the form
 $X \rightarrow Y$, where X and Y are disjoint itemsets,
- Example:
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$
- An association rule is a pattern that states when X occurs, Y occurs with certain probability.

Important Note

- Association rules do *not* consider order. So...
- $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$
- and
- $\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$
- .. are the same rule

- **Rule Evaluation Metrics**

- **Support** (s)
 - Fraction of transactions that contain both X and Y
- **Confidence** (c)
 - Measures how often item Y appear in transactions that contain X

$$support = \frac{(X \cup Y).count}{\# \text{ of transactions}}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

- **Why measure support?**
 - Very low support rules can happen by chance
 - Even if true rules, **low support rules are often not actionable**
- **Why measure confidence?**
 - Very low confidence rules are **not reliable**

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	<u>Milk, Diaper</u> , Beer, Coke
4	Bread, <u>Milk, Diaper</u> ,
5	Bread, <u>Milk, Diaper</u> , Coke

Example:

$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = 0.67$$

Goal: Find all rules that satisfy the user-specified *minimum support* (minsup) and *minimum confidence* (minconf).

Mining Association Rules

- Two-step approach:

1. **Frequent Itemset Generation**

- Generate all itemsets that have **minimum support**

2. **Rule Generation**

- Generate **high confidence** rules from each frequent itemset

- Several algorithms exist, use different strategies and data structures.

- **The Apriori Algorithm**

- **Probably the best-known algorithm**, proposed by **Agrawal et al in 1993**.

- Initially used for **Market Basket Analysis** to find how items purchased by customers are related.

- It is an important data mining model studied extensively by the database and data mining community.

- **The algorithm assumes all data are categorical.**

- It does not work for numeric data.

Weka Demonstration

- ❑ Weather data (nominal)
- ❑ **Strategy:**
 - ❑ Specify minimum confidence value and seek rules with maximum support and maximum confidence

THANK YOU

Any questions...?

