SILVER OAK UNIVERSITY
EDUCATION TO INNOVATION

UNIT-5 CLASSIFICATION AND CLUSTER ANALYSIS: BASIC CONCEPTS

BRANCH: BACHELOR OF COMPUTER APPLICATION

SEMESTER:5TH

PREPARED BY: PROF. MANSI DAVE

OUTLINES

- Basic Concepts of Classification
- What is Classification?
- General approach to Classification
- Decision Tree Induction
- Bayes Classification method
- Basic Concepts of Clustering
- What is Cluster Analysis?
- Requirements of Cluster Analysis
- Overview of Basic Clustering Methods
- Partitioning Methods
- K-Means: A centroid based technique

BASIC CONCEPTS OF CLASSIFICATION

- It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.
- Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

BASIC CONCEPTS OF CLASSIFICATION

- Example: Before starting any project, we need to check its feasibility. In this case, a classifier is required to predict class labels such as 'Safe' and 'Risky' for adopting the Project and to further approve it. It is a two-step process such as:
- **Learning Step (Training Phase):** Construction of Classification Model
- Different Algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.
- **Classification Step:** Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.

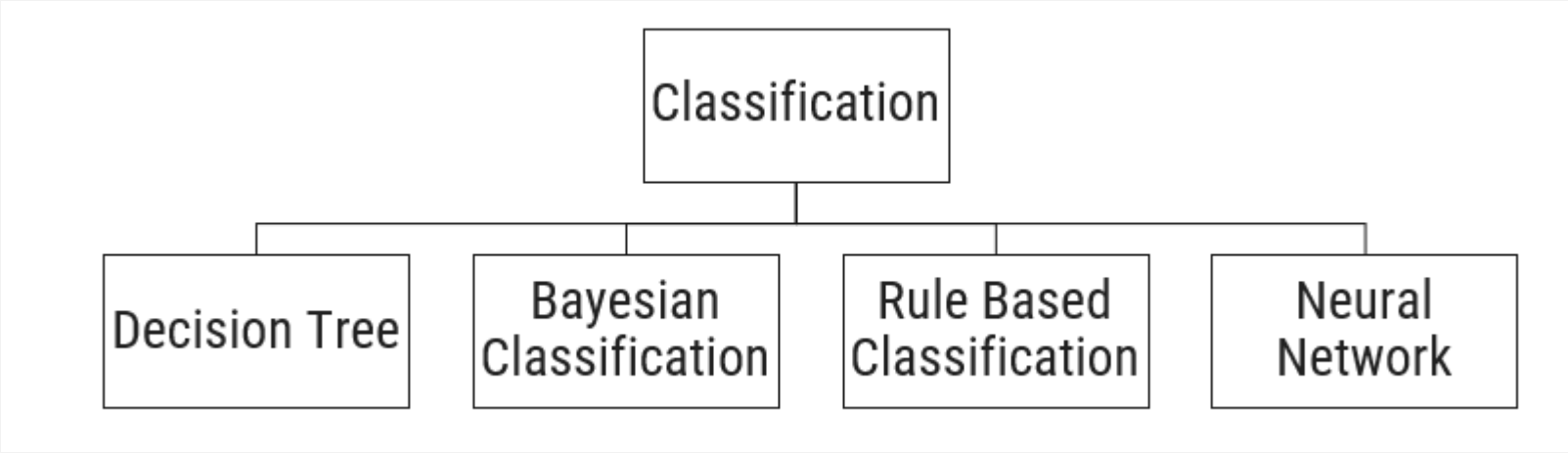
WHAT IS CLASSIFICATION?

- Classification is a task in data mining that involves assigning a class label to each instance in a dataset based on its features.
- The goal of classification is to build a model that accurately predicts the class labels of new instances based on their features.

GENERAL APPROACH TO CLASSIFICATION

- Classifiers can be categorized into two major types:
- **Discriminative:** It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the observed data, depends heavily on the quality of data rather than on distributions.
- Example: Logistic Regression
- **Generative:** It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data.
- Example: Naive Bayes Classifier

METHODS OF CLASSIFICATION

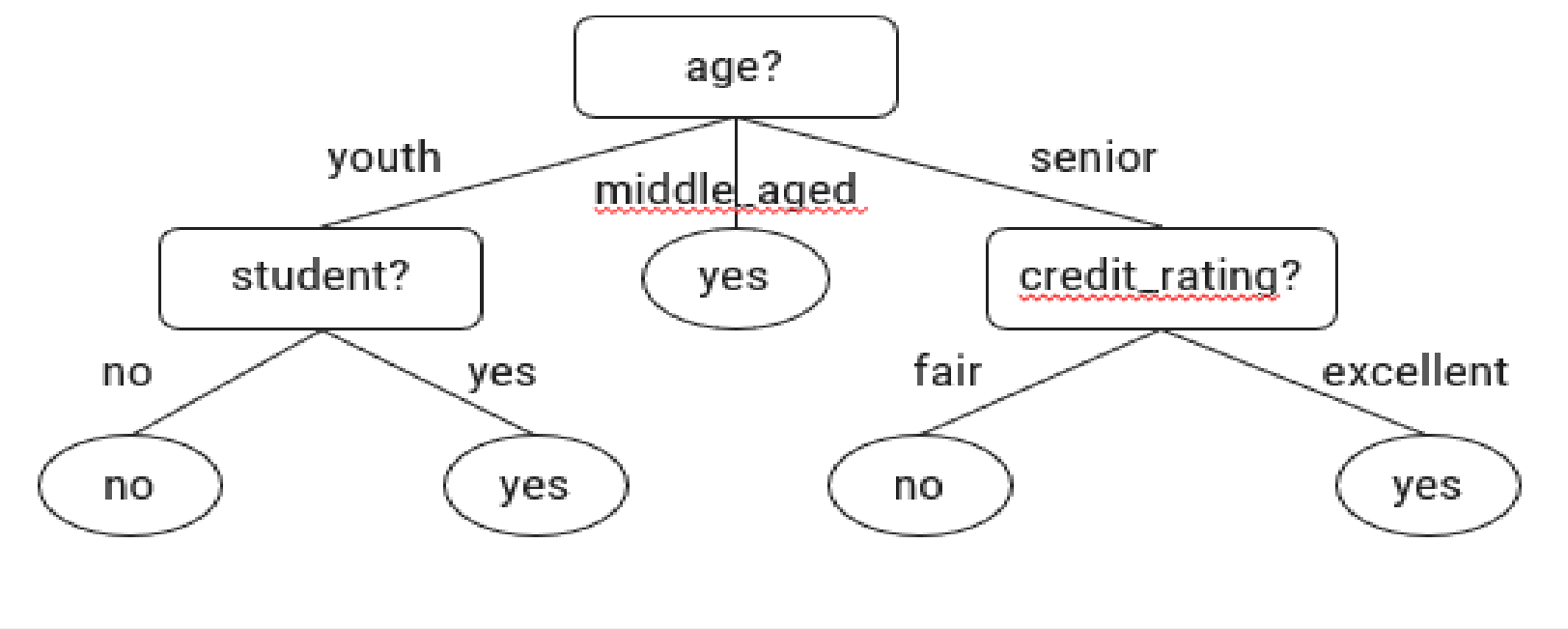


DECISION TREE

- Decision tree induction is the learning of decision trees from class-labeled training tuples.
- A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute.
- Each branch represents an outcome of the test.
- Each leaf node (or terminal node) holds a class label.
- The topmost node in a tree is the root node.

DECISION TREE

- Decision Tree represents the concept buys_computer, i.e. it predicts whether a customer at AllElectronics is likely to purchase a computer.



BAYES CLASSIFICATION METHOD

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

WHY IT IS CALLED NAÏVE BAYES?

- Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

WHY IT IS CALLED NAÏVE BAYES?

- Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

- Where,
- P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
- P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
- P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
- P(B) is Marginal Probability: Probability of Evidence.

WHY IT IS CALLED NAÏVE BAYES?

- Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

- Where,
- P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
- P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
- P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
- P(B) is Marginal Probability: Probability of Evidence.

WORKING OF NAÏVE WORKING OF NAÏVE BAYES' CLASSIFIER:

- Working of Naïve Bayes' Classifier can be understood with the help of the below example:
- Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:
- Convert the given dataset into frequency tables.
- Generate Likelihood table by finding the probabilities of given features.
- Now, use Bayes theorem to calculate the posterior probability.
- Problem: If the weather is sunny, then the Player should play or not?

WORKING OF NAÏVE WORKING OF NAÏVE BAYES' CLASSIFIER:

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

WORKING OF NAÏVE WORKING OF NAÏVE BAYES' CLASSIFIER:

Frequency table for the Weather Conditions:

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Likelihood table weather condition:

Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

WORKING OF NAÏVE WORKING OF NAÏVE BAYES' CLASSIFIER:

- Applying Bayes'theorem:
- P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)
- P(Sunny|Yes)= 3/10= 0.3
- P(Sunny)= 0.35
- P(Yes)=0.71
- So P(Yes|Sunny) = 0.3*0.71/0.35= 0.60
- P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)
- P(Sunny|NO)= 2/4=0.5
- P(No)= 0.29
- P(Sunny)= 0.35
- So P(No|Sunny)= 0.5*0.29/0.35 = 0.41
- So as we can see from the above calculation that P(Yes|Sunny)>P(No|Sunny)
- Hence on a Sunny day, Player can play the game.

ADVANTAGES AND DISADVANTAJES NAÏVE BAYES' CLASSIFIER:

- **Advantages of Naïve Bayes Classifier:**
- Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in Multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for text classification problems.
- **Disadvantages of Naïve Bayes Classifier:**
- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

WHAT IS CLUSTERING?

- Clustering is one of the most important research areas in the field of data mining.
- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
- It is an unsupervised learning technique.
- Data clustering is the subject of active research in several fields such as statistics, pattern recognition and machine learning.
- From a practical perspective clustering plays an outstanding role in data mining applications in many domains.
- The main advantage of clustering is that interesting patterns and structures can be found directly from very large data sets with little or none of the background knowledge.
- Clustering algorithms an be applied in many areas, like marketing, biology, libraries, insurance, city-planning, earthquake studies and www document classification.

REQUIREMENTS OF APPLICATIONS OF CLUSTER ANALYSIS

- **Marketing**
- Finding group of customers with similar behavior given a large data-base of customers.
- Data containing their properties and past buying records (Conceptual Clustering).
- **Biology**
- Classification of Plants and Animals Based on the properties under observation (Conceptual Clustering).
- **Insurance**
- Identifying groups of car insurance policy holders with a high average claim cost (Conceptual Clustering).
- **City-Planning**
- Groups of houses according to their house type, value and geographical location it can be both (Conceptual Clustering and Distance Based Clustering)
- **Libraries**
- It is used in clustering different books on the basis of topics and information.
- **Earthquake studies**
- By learning the earthquake-affected areas we can determine the dangerous zones.

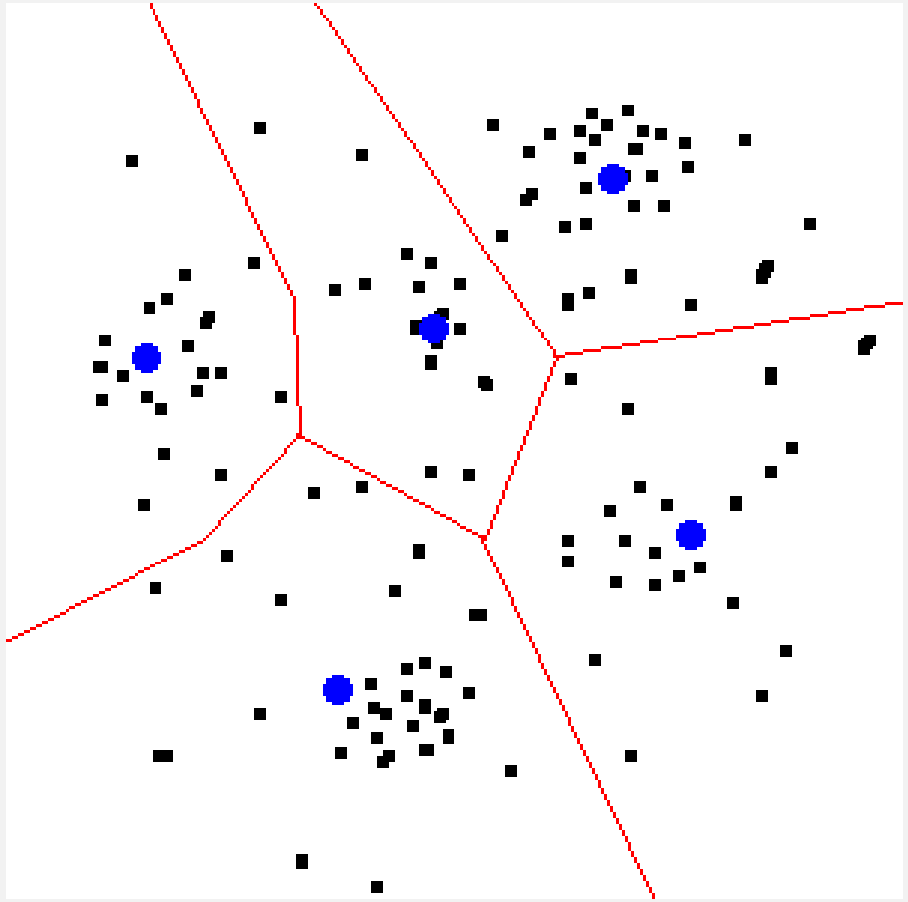
WHAT IS PARTITIONING?

- Clustering is a division of data into groups of similar objects.
- Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups.
- It represents many data objects by few clusters and hence, it models data by its clusters.
- A **cluster** is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

K-MEANS ALGORITHM

- K-Means is one of the simplest unsupervised learning algorithm that solve the well known clustering problem.
- The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (k-clusters).
- The main idea is to define k centroids, one for each cluster.
- A centroid is “the center of mass of a geometric object of uniform density”, though here, we'll consider mean vectors as centroids.
- It is a method of classifying/grouping items into k groups (where k is the number of pre-chosen groups).
- The grouping is done by minimizing the sum of squared distances between items or objects and the corresponding centroid.

K-MEANS ALGORITHM CONT..



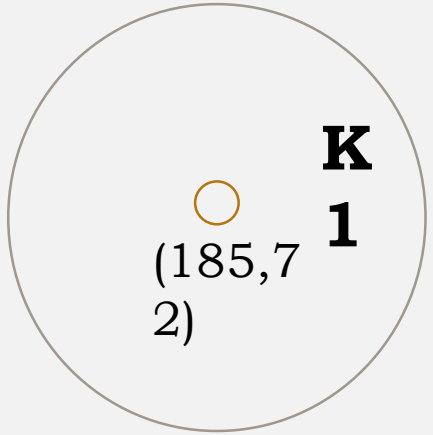
- A clustered scatter plot.
- The black dots are data points.
- The red lines illustrate the partitions created by the k-means algorithm.
- The blue dots represent the centroids which define the partitions.

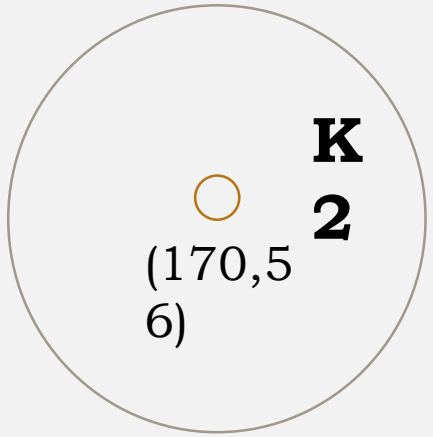
K-MEANS ALGORITHM CONT..

- The initial partitioning can be done in a variety of ways.
- **Dynamically Chosen**
  - This method is good when the amount of data is expected to grow.
  - The initial cluster means can simply be the first few items of data from the set.
  - For instance, if the data will be grouped into 3 clusters, then the initial cluster means will be the first 3 items of data.
- **Randomly Chosen**
  - Almost self-explanatory, the initial cluster means are randomly chosen values within the same range as the highest and lowest of the data values.
- **Choosing from Upper and Lower Bounds**
  - Depending on the types of data in the set, the highest and lowest of the data range are chosen as the initial cluster means.

K-MEANS ALGORITHM - EXAMPLE

Sr.	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

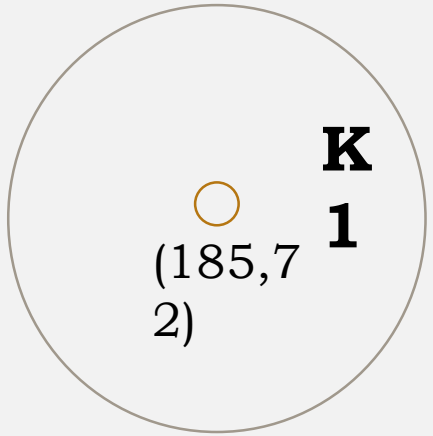


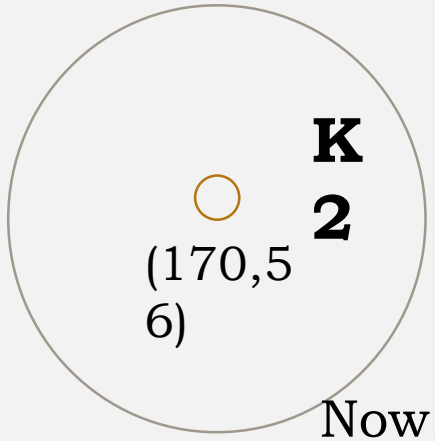


- First we take **K=2** So, two clusters or groups.
- We choose first (185,72) & second (170,56) row as centroid of each cluster or group.
- Now, we have to find Euclidean Distance,
  - $ED = \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$
- Where
  - X_0 & Y_o = Observed Value
  - X_c & Y_c = Centroid Value

K-MEANS ALGORITHM – EXAMPLE CONT..

Sr.	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76





- ED From **K1** ***to*** **(168, 60)**

$$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$$
$$= \sqrt{(168 - 185)^2 + (60 - 72)^2}$$
$$= \sqrt{(-17)^2 + (-12)^2}$$
$$= \sqrt{289 + 144}$$
$$= \sqrt{433}$$
$$= 20.80$$

- ED From **K2** ***to*** **(168, 60)**

$$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$$
$$= \sqrt{(168 - 170)^2 + (60 - 56)^2}$$
$$= \sqrt{(-2)^2 + (-4)^2}$$
$$= \sqrt{4 + 16}$$
$$= \sqrt{20}$$
$$= 4.48$$

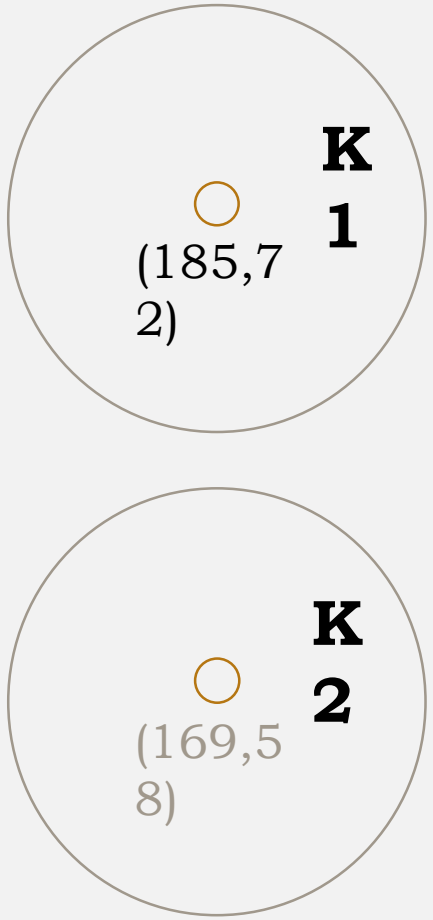
Now, data (168,60) nearer to K2, so it belongs to K2.

K1 = {1}

K2 = {2,3}

K-MEANS ALGORITHM – EXAMPLE CONT..

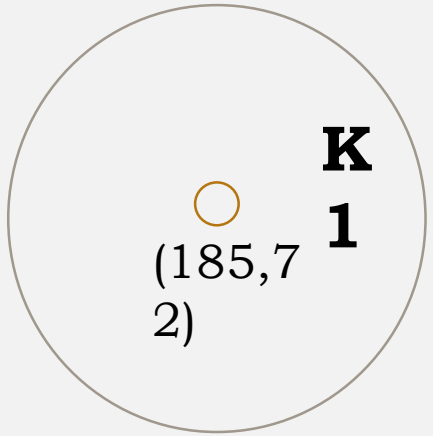
Sr.	Height	Weight
1	185	72
2	**170**	**56**
3	**168**	**60**
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

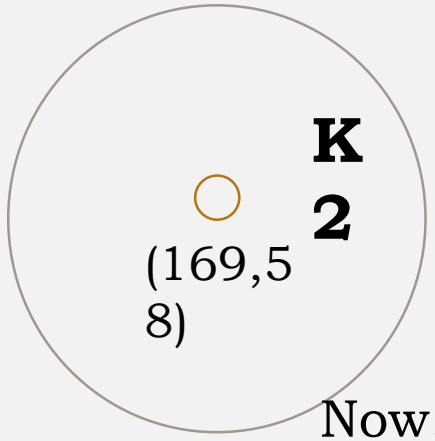


Now, New Centroid Calculation
For K2 = {2,3}
So, **K2** = {(170,56),(168,60)}
= 170+168/2 & 56+60/2
We get new centroid **C** = (169,58)

K-MEANS ALGORITHM – EXAMPLE CONT..

Sr.	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76





➥ ED From **K1** ***to*** **(179, 68)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(179 - 185)^2 + (68 - 72)^2}$

$= \sqrt{(-6)^2 + (-4)^2}$

$= \sqrt{36 + 16}$

$= \sqrt{52}$

$= 7.21$

➥ ED From **K2** ***to*** **(179, 68)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(179 - 169)^2 + (68 - 58)^2}$

$= \sqrt{(10)^2 + (10)^2}$

$= \sqrt{100 + 100}$

$= \sqrt{200}$

$= 14.14$

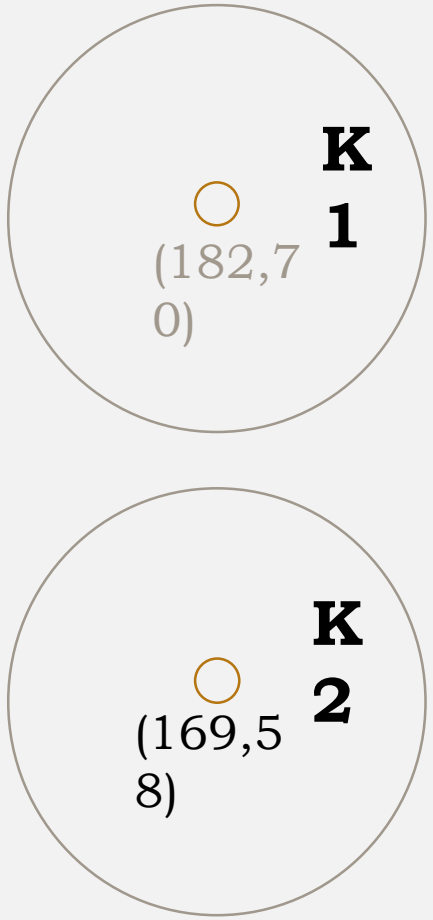
Now, data (179,68) nearer to K1, so it belongs to K1.

K1 = {1,4}

K2 = {2,3}

K-MEANS ALGORITHM – EXAMPLE CONT..

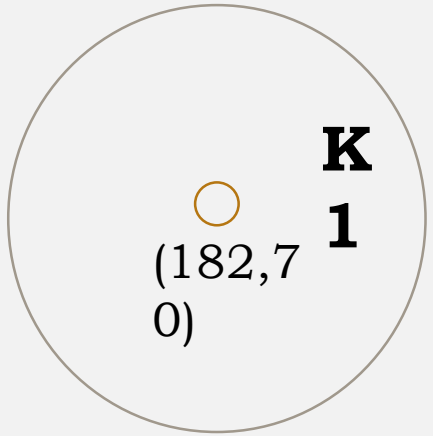
Sr.	Height	Weight
1	**185**	**72**
2	170	56
3	168	60
4	**179**	**68**
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

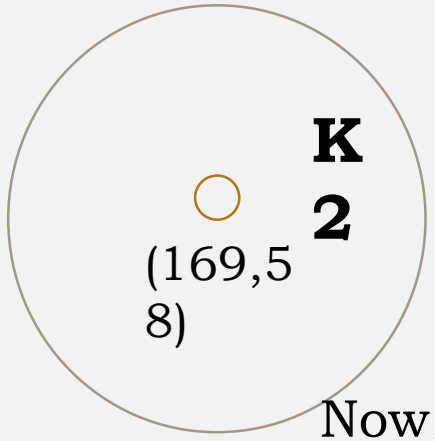


Now, New Centroid Calculation
For K1 = {1,4}
So, **K2** = {(185,72),(179,68)}
= 185+179/2 & 72+68/2
We get new centroid **C** = (182,70)

K-MEANS ALGORITHM – EXAMPLE CONT..

Sr.	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76





➥ ED From **K1** ***to*** **(182, 72)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(182 - 182)^2 + (72 - 70)^2}$

$= \sqrt{(0)^2 + (2)^2}$

$= \sqrt{0 + 4}$

$= \sqrt{4}$

$= 2$

➥ ED From **K2** ***to*** **(182, 72)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(182 - 169)^2 + (72 - 58)^2}$

$= \sqrt{(-13)^2 + (-14)^2}$

$= \sqrt{169 + 196}$

$= \sqrt{365}$

$= 19.10$

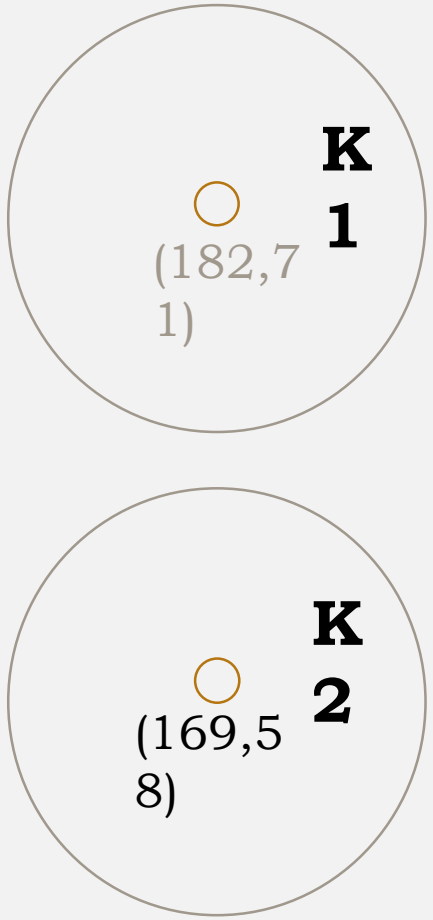
Now, data (182,72) nearer to K1, so it belongs to K1.

K1 = {1,4,5}

K2 = {2,3}

K-MEANS ALGORITHM – EXAMPLE CONT..

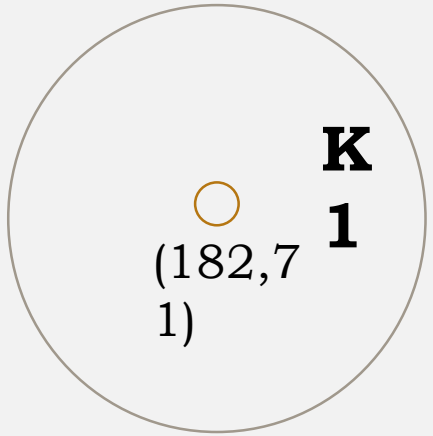
Sr.	Height	Weight
1	**185**	**72**
2	170	56
3	168	60
4	**179**	**68**
5	**182**	**72**
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

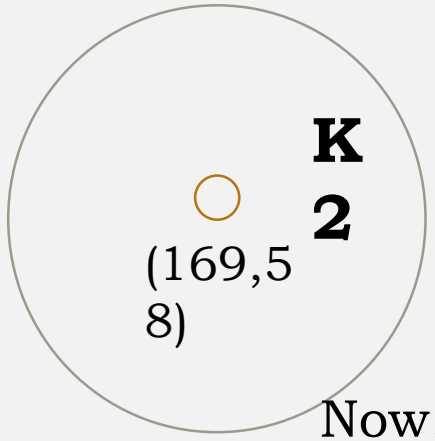


Now, New Centroid Calculation
For K1 = {1,4,5}
So, **K2** = {(185,72),(179,68),(182,72)}
= 185+179+182/3 & 72+68+72/3
We get new centroid **C** = (182,70.666) ~ (182,71)

K-MEANS ALGORITHM – EXAMPLE CONT..

Sr.	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	**188**	**77**
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76





➥ ED From **K1** ***to*** **(188, 77)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(188 - 182)^2 + (77 - 71)^2}$

$= \sqrt{(6)^2 + (6)^2}$

$= \sqrt{36 + 36}$

$= \sqrt{72}$

$= 8.48$

➥ ED From **K2** ***to*** **(188, 77)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(188 - 169)^2 + (77 - 58)^2}$

$= \sqrt{(19)^2 + (19)^2}$

$= \sqrt{361 + 361}$

$= \sqrt{722}$

$= 26.87$

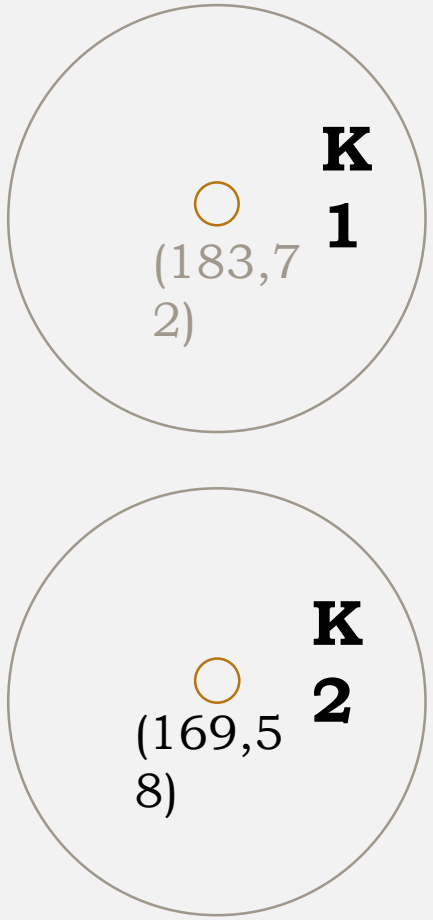
Now, data (188,77) nearer to K1, so it belongs to K1.

K1 = {1,4,5,6}

K2 = {2,3}

K-MEANS ALGORITHM – EXAMPLE CONT..

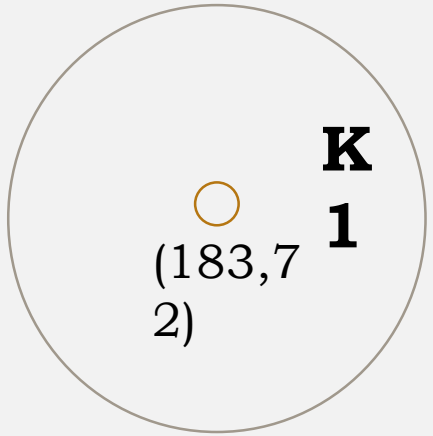
Sr.	Height	Weight
1	**185**	**72**
2	170	56
3	168	60
4	**179**	**68**
5	**182**	**72**
6	**188**	**77**
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

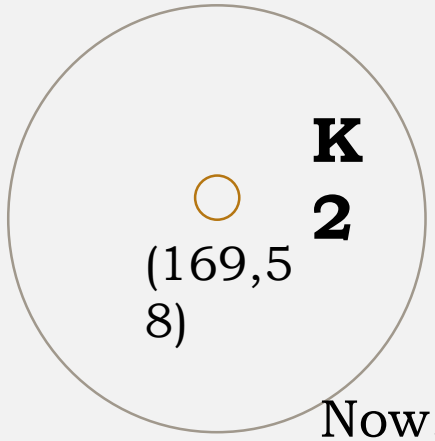


Now, New Centroid Calculation
For K1 = {1,4,5,6}
So, **K2** = {(185,72),(179,68),(182,72),(188,77)}
= 185+179+182+188/4 & 72+68+72+77/7
We get new centroid **C** = (183.50,72.25) ~ (183,72)

K-MEANS ALGORITHM – EXAMPLE CONT..

Sr.	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	**180**	**71**
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76





➥ ED From **K1** ***to*** **(180, 71)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(180 - 183)^2 + (71 - 72)^2}$

$= \sqrt{(-3)^2 + (-1)^2}$

$= \sqrt{9 + 1}$

$= \sqrt{10}$

$= 3.16$

➥ ED From **K2** ***to*** **(180, 71)**

$= \sqrt{(X_o - X_c)^2 + (Y_o - Y_c)^2}$

$= \sqrt{(180 - 169)^2 + (71 - 58)^2}$

$= \sqrt{(11)^2 + (13)^2}$

$= \sqrt{121 + 169}$

$= \sqrt{290}$

$= 17.02$

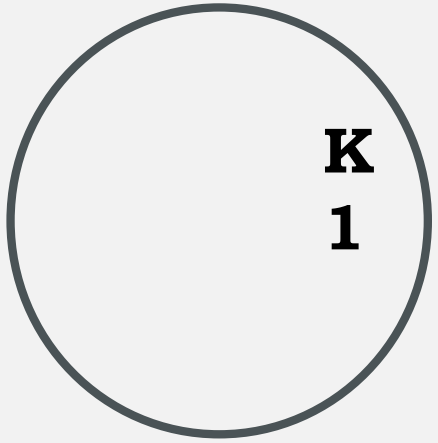
Now, data (180,71) nearer to K1, so it belongs to K1.

K1 = {1,4,5,6,7}

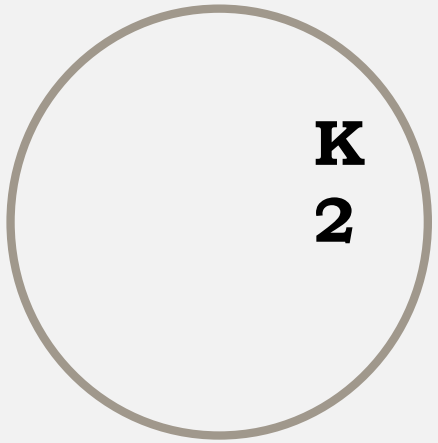
K2 = {2,3}

K-MEANS ALGORITHM – EXAMPLE CONT..

Sr.	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76



Cluster K1 = {1,4,5,6,7,8,9,10,11,12}



Cluster K2 = {2,3}

K-MEANS ALGORITHM CONT..

- Let us assume two clusters, and each individual's scores include two variables.
- **Step-1**
  - Choose the number of clusters.
- **Step-2**
  - Set the initial partition, and the initial mean vectors for each cluster.
- **Step-3**
  - For each remaining individual…
- **Step-4**
  - Get averages for comparison to the Cluster 1:
    - Add individual's A value to the sum of A values of the individuals in Cluster 1, then divide by the total number of scores that were summed.
    - Add individual's B value to the sum of B values of the individuals in Cluster 1, then divide by the total number of scores that were summed.

K-MEANS ALGORITHM CONT..

- **Step-5**
  - Get averages for comparison to the Cluster 2:
    - Add individual's A value to the sum of A values of the individuals in Cluster 2, then divide by the total number of scores that were summed.
    - Add individual's B value to the sum of B values of the individuals in Cluster 2, then divide by the total number of scores that were summed.
- **Step-6**
  - If the averages found in Step 4 are closer to the mean values of Cluster 1, then this individual belongs to Cluster 1, and the averages found now become the new mean vectors for Cluster 1.
  - If closer to Cluster 2, then it goes to Cluster 2, along with the averages as new mean vectors.
- **Step-7**
  - If there are more individual's to process, continue again with Step 4. Otherwise go to Step 8.
- **Step-8**

K-MEANS ALGORITHM CONT..

- **Step-9**
  - If any relocations occurred in Step 8, the algorithm must continue again with Step 3, using all individuals and the new mean vectors.
  - If no relocations occurred, stop. Clustering is complete.

THANK YOU