Classification

1) Consider the training examples shown in Table 1 for a binary classification problem. (Choose (a) or (b))

   a) Calculate the information gained when splitting on A and B. Which attribute would the decision tree induction algorithm chooses. (Hint: Information gain is the gain with the entropy impurity measure

Tad	A	B	Target Class
1	T	F	+
2	T	T	+
3	T	T	+
4	T	F	-
5	T	T	+
6	F	F	-
7	F	F	-
8	F	F	-
9	T	T	-
10	T	F	+

Answer

$$Entropy = -p_+ log_2 p_+ - p_- log_2 p_-$$

$$Entropy = -\frac{5}{10} log_2(\frac{5}{10}) - \frac{5}{10} log_2(\frac{5}{10}) = 1$$

A	Target Class
T	+
T	+
T	+
T	-
T	+
T	-
T	+

A	Target Class
F	-
F	-
F	-

$$information\ gain(A) = E(S) - \sum_{i=1}^{k} p_1\, E(S,Q)$$

$$information\ gain(A) = 1 - \frac{7}{10}\left(-\frac{5}{7} log_2(\frac{5}{7}) - \frac{2}{7} log_2(\frac{2}{7})\right) - \frac{3}{10}\left(-\frac{0}{3} log_2(\frac{0}{3}) - \frac{3}{3} log_2(\frac{3}{3})\right) = 0.396$$

B	Target Class
T	+
T	+
T	+
T	-

B	Target Class
F	+
F	-
F	-
F	-
F	-
F	+

$$information\ gain(B) = E(S) - \sum_{i=1}^{k} p_1\, E(S,Q)$$

$$information\ gain(B) = 1 - \frac{4}{10}\left(-\frac{3}{4} log_2(\frac{3}{4}) - \frac{1}{4} log_2(\frac{1}{4})\right) - \frac{6}{10}\left(-\frac{2}{6} log_2(\frac{2}{6}) - \frac{4}{6} log_2(\frac{4}{6})\right) = 0.126$$

The best start node is (A)

   b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

Answer

$$Gain\ index\ (A) = 1 - C * \sum_{c=1}^{i} (p_i)^2$$

$$Gini\ index\ (A) = \left(\frac{7}{10}\right) * \left(1 - \left(\left(\frac{5}{7}\right)^2 + \left(\frac{2}{7}\right)^2\right)\right) + \left(\frac{3}{10}\right) * \left(1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right)\right)$$

$$= 0.4 + 0.6 = 1$$

$$Gini\ index\ (B)\ =\ 1\ -\ C\ *\sum_{c=1}^{\iota}(p_i)^2$$

$$Gini\ index(B) = \left(\frac{4}{10}\right)*\left(1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right)+\left(\frac{6}{10}\right)\left(1-\left(\left(\frac{2}{6}\right)^2+\left(\frac{4}{6}\right)^2\right)\right)$$

$$= 0.2 + 0.8 = 1$$

Choose A or (B) to start because its equal

2) In the basic decision tree construction procedure, what happen if.

a) There are no records associated with the generated child.

***Answer*:** the node is declared a leaf node with the same class label as the majority class of training records associated with its parent node.

b) The records associated with the generated child have identical attribute values (except for the class label)?

***Answer**:* the node is declared a leaf node with the same class label as the majority class of training records associated with this node

3) Given the attribute shirt size = {small, medium, large, Extra Large}

a) Is it binary, nominal, or ordinal attribute?

Answer: *ordinal attribute*

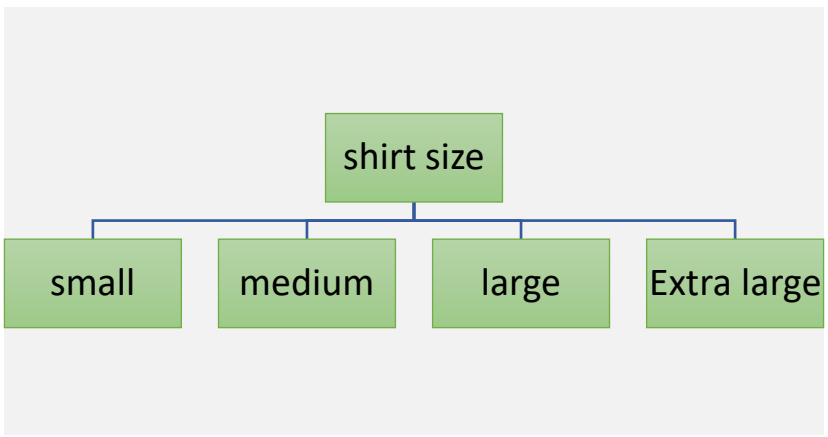
b) How many **binary splits** are possible for this attribute?

***Answer**: binary splits* $= (2^{k-1} - 1) = (2^{4-1} - 1) = 2^3 - 1 = 7$

- {small}, {medium, large, Extra-large}
- {medium}, {small, large, Extra Large}
- {large}, {small, medium, Extra Large}
- {Extra-large}, {small, medium, large}
- {small, medium}, {Large, Extra-large}
- {small, large}, {medium, Extra-large}
- {small, Extra-large}, {medium, large}

c) How many children can be obtained using this attribute if multiway split is considered?

Answer: The number of children = 4



4) Given the attribute Marital Status = {Single, Divorced, Married}

a) Is it binary, nominal, or ordinal attribute?

Answer: nominal

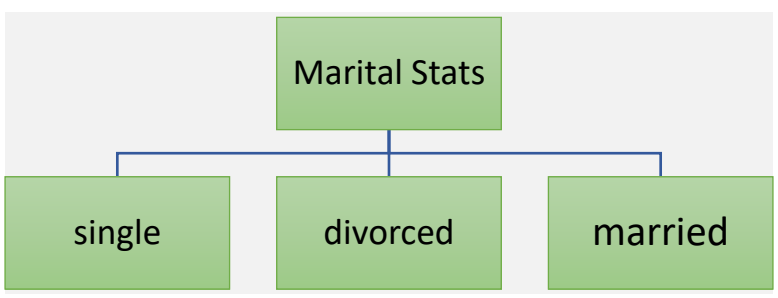
b) How many binary splits are possible for this attribute?

Answer: binary splits $= (2^{k-1} - 1) = (2^{3-1} - 1) = 2^2 - 1 = 3$

- {single}, {divorced, married}
- {divorced}, {single, married}
- {married}, {single, divorced}

c) How many children can be obtained using this attribute if multiway split is considered?

Answer: The number of children = 3



5) Define classification. How to calculate the accuracy of a classification model, i.e., a classified

Answer:

Classification is the task of assigning objects to one of several predefined classes.

- **To calculate accuracy:**

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

6) Contrast eager learners (such as decision tree) with lazy ones (such as Nearest-Neighbor classifiers).

Answer:

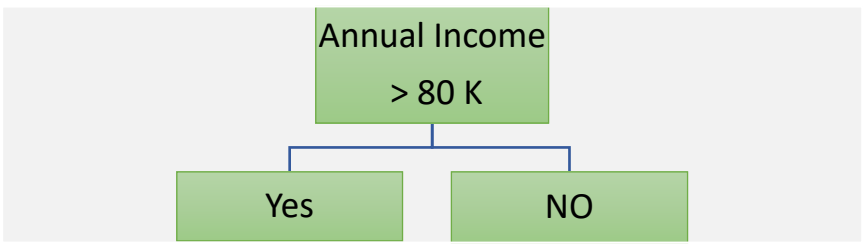
Eager learners are machine learning algorithms that first build a model from the training dataset before making any prediction on future datasets. Decision trees are an example of eager learners.

Lazy learners, on the other hand, delay abstracting from the data until it is asked to make a prediction2. Nearest-Neighbor classifiers are an example of lazy learners

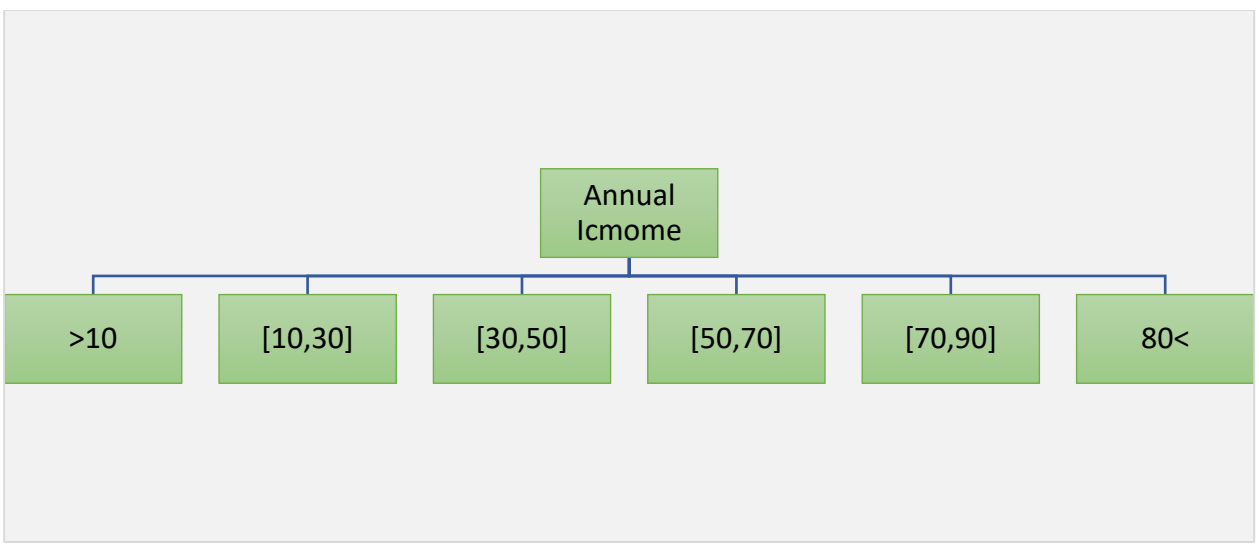
7) نفس السؤال مكرار وتم حله في رقم 3

8) Consider the continuous attribute Annual Income = {60k, 70k, 100k, 120k}. Show how to perform binary split and multiway split by giving an example.

- Binary split



- Multiway



9) What is the difference between classification algorithm and classifier?

Answer:

Classification algorithm is a routine which takes training data (data objects with known labels) as an input and deduces a model called **classifier** which fits the training data very well and is able to predict the class label of new data objects.

10) Define: Classification, Redundant attribute

Answer:

Classification: is the task of assigning objects to one of the several predefined categories, e.g., detecting spam email messages based upon the message header and contents.

Redundant attribute: attribute is redundant I it's strongly correlated with another attribute I the data. The presence of redundant attributes does not affect the accuracy of the decision tree.

11) Mention two problems facing any decision tree construction algorithm. (Data fragmentation, replication)

Answer:

Data Fragmentation: Since most of decision tree algorithms apply a top-down recursive partitioning approaches, the number of records become smaller as we travers down the tree. At the leaf nodes, the number of records may be too small to make statistically significant decision about the class representation of the nodes.

Replication: subtree can be replicated multiple times in decision tree. this makes the D.T more complex and more difficult to interpret.

Most DT use the divide and conquer strategy. The same test condition applied to different parts of attribute leading to subtree replication problem.

12) Consider the training examples shown in Table 2 for a binary classification problem.

Cid	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C1
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a) What is the information gain of Gender and Car type relative to these training examples?
(Hint: Information gain is the gain with the entropy impurity measure.)

Answer

$$Entropy = -p_+ log_2 p_+ - p_- log_2 p_-$$

$$Entropy = -\frac{10}{20} log_2(\frac{10}{20}) - \frac{10}{20} log_2(\frac{10}{20}) = 1$$

Information Gain (Gender)

Gender	Class
M	C0
M	C0
M	C0
M	C0
M	C0
M	C0
M	C1
M	C1
M	C1
M	C1

Gender	Class
F	C0
F	C0
F	C0
F	C0
F	C1
F	C1
F	C1
F	C1
F	C1
F	C1

$$information\ gain(Gender) = E(S) - \sum_{i=1}^{k} p_1\, E(S,Q)$$

$$information\ gain(Gender) = 1 - \frac{10}{20}\left(-\frac{6}{10} log_2(\frac{6}{10}) - \frac{4}{10} log_2(\frac{4}{10})\right) - \frac{10}{20}\left(-\frac{4}{10} log_2(\frac{4}{10}) - \frac{6}{10} log_2(\frac{6}{10})\right) = 0.03$$

Information Gain (Car type)

Car type	Class
Family	C0
Family	C1
Family	C1
Family	C1

Car type	Class
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0

Car type	Class
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0
Sport	C0

$$information\ gain(car\ type) = E(S) - \sum_{i=1}^{\kappa} p_1\ E(S,Q)$$

$$information\ gain(car\ type) = 1 - \frac{4}{20}\left(-\frac{1}{4}log_2(\frac{1}{4}) - \frac{3}{4}log_2(\frac{3}{4})\right) - \frac{8}{10}\left(-\frac{8}{8}log_2(\frac{8}{8}) - \frac{0}{8}log_2(\frac{0}{8})\right) - \frac{8}{10}\left(-\frac{0}{8}log_2(\frac{0}{8}) - \frac{8}{8}log_2(\frac{8}{8})\right) = 0.6$$

b) What is the best split (between a1 and a2) according to the information gain?

Answer: The best split of information gain Is (car type)

c) What is the best split (between a1 and a2) according to the Gini index?

Answer

$$Gain\ index\ (Gender) = 1 - C * \sum_{c=1}^{i} (p_i)^2$$

$$Gain\ index\ (Gender) = \left(\frac{10}{20}\right) * \left(1 - \left(\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right)\right) + \left(\frac{10}{20}\right) * \left(1 - \left(\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right)\right) = 0.24 + 0.24 = 0.48$$

$$Gain\ index\ (Gender) = 1 - C * \sum_{c=1}^{i} (p_i)^2$$

$$Gain\ index(Car\ tyep) = \left(\frac{4}{20}\right) * \left(1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right)\right) + \left(\frac{8}{10}\right) * \left(1 - \left(\left(\frac{8}{8}\right)^2 + \left(\frac{0}{8}\right)^2\right)\right) + \left(\frac{8}{10}\right) * \left(1 - \left(\left(\frac{0}{8}\right)^2 + \left(\frac{8}{8}\right)^2\right)\right) = 0.075 + 0 + 0 = 0.075$$

Answer: The best split of Gini index Is (car type)

13) Consider the training examples shown in Table 3 for a binary classification problem.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

a) What is the information gain of a_1 and a_2 relative to these training examples?

(Hint: Information gain is the gain with the **entropy** impurity measure.

$$Entropy = -p_+ log_2 p_+ - p_- log_2 p_-$$

$$Entropy = -\frac{4}{9}log_2(\frac{4}{9}) - \frac{5}{9}log_2(\frac{5}{9}) = 0.991$$

Information Gain(a1)

A	Target Class
T	+
T	+
T	-
T	+

A	Target Class
F	+
F	-
F	-
F	-
F	-

$$information\ gain(A) = E(S) - \sum_{i=1}^{k} p_1\, E(S,Q)$$

$$information\ gain(A) = 0.991 - \frac{4}{9}\left(-\frac{3}{4}log_2(\frac{3}{4}) - \frac{1}{4}log_2(\frac{1}{4})\right) - \frac{5}{9}\left(-\frac{1}{5}log_2(\frac{1}{5}) - \frac{4}{5}log_2(\frac{4}{5})\right)$$
$$= 0.23$$

Information Gain(a2)

B	Target Class
T	+
T	+
T	-
T	-
T	-

B	Target Class
F	-
F	+
F	-
F	+

$$information\ gain(B) = E(S) - \sum_{i=1}^{k} p_1\, E(S,Q)$$

$$information\ gain(B) = 0.991 - \frac{5}{9}\left(-\frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5})\right) - \frac{4}{9}\left(-\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4})\right) = 0.008$$

b) What is the best split (between a1 and a2) according to the information gain?

Answer: The best split of information gain Is (a1)

c) What is the best split (between a1 and a2) according to the Gini index?

Answer

$$Gain\ index\ (A) = 1 - C * \sum_{c=1}^{i} (p_i)^2$$

$$Gini\ index\ (A) = \left(\frac{4}{9}\right) * \left(1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right)\right) + \left(\frac{5}{9}\right) * \left(1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)\right)$$
$$= 0.17 + 0.27 = 0.44$$

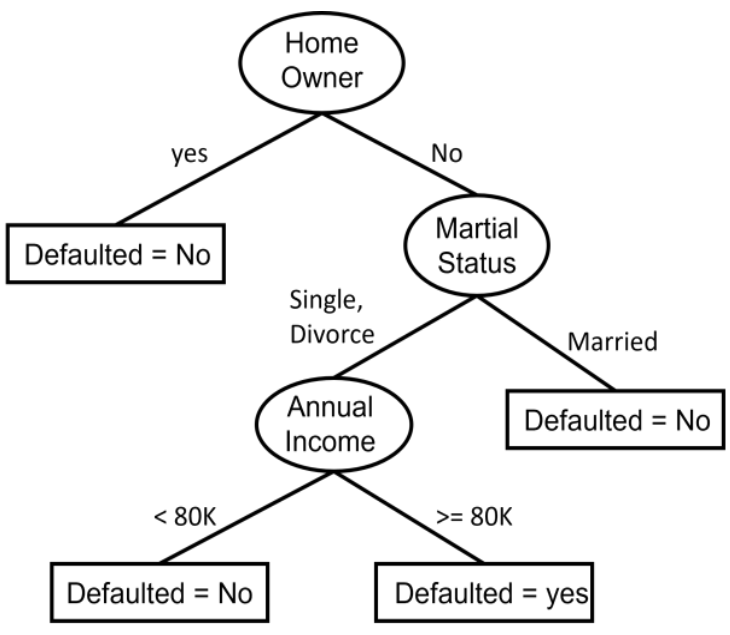
$$Gain\ index\ (B) = 1 - C * \sum_{c=1}^{i} (p_i)^2$$

$$Gini\ index(B) = \left(\frac{5}{9}\right) * \left(1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)\right) + \left(\frac{4}{9}\right) * \left(1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right)\right)$$
$$= 0.27 + 0.22 = 0.49$$

Answer: The best split of Gini index Is (a2)

14) Find the accuracy of the decision tree given in Figure 1 on the test examples given in Table 4

Tad	Homeowner	Marital status	Annual Income	Default Borrower
1	Yes	Single	100k	? (Yes)
2	Yes	Divorced	80k	? (No)
3	No	Married	125k	? (Yes)
4	No	Single	90k	? (Yes)



Answer

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2}{4} = 0.5 = 50\%$$

15) Answer each of the following:

a) In a decision tree, each leaf node is assigned class label and each non-terminal node contains test node.

16) State whether the following are true or false.

a) In classification, the class label must be a discrete attribute (T).

b) In regression, the class label must be a continuous attribute (T).

c) In decision tree construction, the attribute chosen as the test attribute may vary depending on the choice of the impurity measure (T).

d) In decision tree construction, the record id can be used as a splitting attribute (F).

e) The strategy used to prune the decision tree has a greater impact on the final tree than the choice of impurity measure (F)

17) Choose the correct answer for each of the following.

a) Classification algorithms seek models that attain .....when applied to the test set.

I. the highest accuracy

II. the highest error rates.

III. both (I) and (ii)

b) A classifier (classification model) fits the relationship between the .....and.....

I. attribute set...class attribute

II. subset of the attribute set....class attribute

c) The class label of the test records is .....

I. given

II. unknown

III. masked

Clustering

1) What is cluster analysis. Mention two utilities for which cluster analysis is used.

Answer

Cluster analysis is a technique used to group similar objects into sets or clusters. It is used in a variety of fields such as biology, marketing, and social sciences.

Two utilities for which cluster analysis is used are:

a) Market segmentation: Cluster analysis can be used to segment customers into groups based on their purchasing behavior. This can help businesses tailor their marketing strategies to specific groups of customers.
b) Image segmentation: Cluster analysis can be used to segment images into regions based on their color or texture. This can help in image processing tasks such as object recognition.

2) Contrast well-separated clusters with prototype-based and contiguity-based ones. What type of cluster K-means algorithm produces. Is K-means able to handle all different types of clusters?

Answer

- **Well Separated**: A cluster is a set of objects in which each object is closer (or more similar) to every other object in the cluster than to any object not in the cluster.
- **Prototype-based**: A cluster is a set of objects in which each object is closer (or more similar) to the prototype that defines the cluster than to the prototype of any other cluster. The prototype of a cluster is often a centroid.
- **Contiguity-based**: where two objects are connected only if they are within a specified distance of each other. This implies that each object in a contiguity-based cluster is closer to some other object in the cluster than to any point in a different cluster.

➔ K-means is a prototype-based clustering.

➔ K-Means cannot handle non-globular data of different sizes and densities. (Size of cluster is number of k).

3) Explain agglomerative clustering and write down its pseudo-code. Show that the time complexity of agglomerative clustering is O (n2 log n). Mention the several distance measures that can be used to compute the distance between any two clusters.

Agglomerative clustering: Start with the points as individual clusters and, at each step, merge the closest pair of clusters. This requires defining a notion of cluster proximity.

- Initialize each data point as its own cluster.
- Compute the distance between each pair of clusters.
- Merge the two closest clusters.
- Repeat steps 2-3 until only one cluster remains.

4) Consider the sample dataset that consists of 6 two-dimensional points shown in Table 5. The Euclidean distances between these points are shown in Table 7.

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

distance		result
Distance (p1, p1)	$\sqrt{(0.40-0.40)^2+(0.53-0.53)^2}$	0
Distance (p1, p2)	$\sqrt{(0.40-0.22)^2+(0.53-0.38)^2}$	0.24
Distance (p1, p3)	$\sqrt{(0.40-0.35)^2+(0.53-0.32)^2}$	0.22
Distance (p1, p4)	$\sqrt{(0.40-0.26)^2+(0.53-0.19)^2}$	0.37
Distance (p1, p5)	$\sqrt{(0.40-0.08)^2+(0.53-0.41)^2}$	0.34
Distance (p1, p6)	$\sqrt{(0.40-0.45)^2+(0.53-0.30)^2}$	0.23
Distance (p2, p2)	$\sqrt{(0.22-0.22)^2+(0.38-0.38)^2}$	0
Distance (p2, p3)	$\sqrt{(0.22-0.35)^2+(0.38-0.32)^2}$	0.15
Distance (p2, p4)	$\sqrt{(0.22-0.26)^2+(0.38-0.19)^2}$	0.20
Distance (p2, p5)	$\sqrt{(0.22-0.08)^2+(0.38-0.41)^2}$	0.14
Distance (p2, p6)	$\sqrt{(0.22-0.45)^2+(0.38-0.30)^2}$	0.25
Distance (p3, p3)	$\sqrt{(0.35-0.35)^2+(0.32-0.32)^2}$	0
Distance (p3, p4)	$\sqrt{(0.35-0.26)^2+(0.32-0.19)^2}$	0.15
Distance (p3, p5)	$\sqrt{(0.35-0.08)^2+(0.32-0.41)^2}$	0.28
Distance (p3, p6)	$\sqrt{(0.35-0.45)^2+(0.32-0.30)^2}$	0.11
Distance (p4, p4)	$\sqrt{(0.26-0.26)^2+(0.19-0.19)^2}$	0
Distance (p4, p5)	$\sqrt{(0.26-0.08)^2+(0.19-0.41)^2}$	0.29
Distance (p4, p6)	$\sqrt{(0.26-0.45)^2+(0.19-0.30)^2}$	0.22
Distance (p5, p5)	$\sqrt{(0.08-0.08)^2+(0.41-0.10)^2}$	0
Distance (p5, p6)	$\sqrt{(0.08-0.45)^2+(0.41-0.30)^2}$	0.39
Distance (p6, p6)	$\sqrt{(0.45-0.45)^2+(0.30-0.30)^2}$	0

The end table.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

a) Apply agglomerative clustering to the sample dataset using **MIN** as cluster proximity. Show the result clustering using the dendrogram.

- MIN ((p3, p6), p1) = MIN (distance (p3, p1), distance (p6, p1)) = MIN (0.22,0.23) =0.22
- MIN ((p3, p6), p2) = MIN (distance (p3, p2), distance (p6, p2)) = MIN (0.15,0.25) =0.15
- MIN ((p3, p6), p4) = MIN (distance (p3, p4), distance (p6, p4)) = MIN (0.15,0.22) =0.15
- MIN ((p3, p6), p5) = MIN (distance (p3, p5), distance (p6, p5)) = MIN (0.28,0.39) =0.28

	p1	p2	**p3, p6**	p4	p5
p1	0.00	0.24	**0.22**	0.37	0.34
p2	0.24	0.00	**0.15**	0.20	0.14
p3, p6	**0.22**	**0.15**	**0.00**	**0.15**	**0.28**
p4	0.37	0.20	**0.15**	0.00	0.29
p5	0.34	0.14	**0.28**	0.29	0.00

- MIN ((p2, p5), p1) = MIN (distance (p2, p1), distance (p5, p1)) = MIN (0.24,0.34) =0.24
- MIN ((p2, p5), (p3, p6)) = MIN (distance (p2, (p3, p6)), distance (p5, (p3, p6)) = MIN (0.15,0.28) =0.15
- MIN ((p2, p5), p4) = MIN (distance (p2, p4), distance (p5, p4)) = MIN (0.20,0.29) =0.20
- MIN ((p2, p5), p4) = MIN (distance (p2, p4), distance (p5, p4)) = MIN (0.20,0.29) =0.20

	p1	**p2, p5**	p3, p6	p4
p1	0.00	**0.24**	0.22	0.37
p2, p5	**0.24**	**0.00**	**0.15**	**0.20**
p3, p6	0.22	**0.15**	0.00	0.15
p4	0.37	**0.20**	0.15	0.00

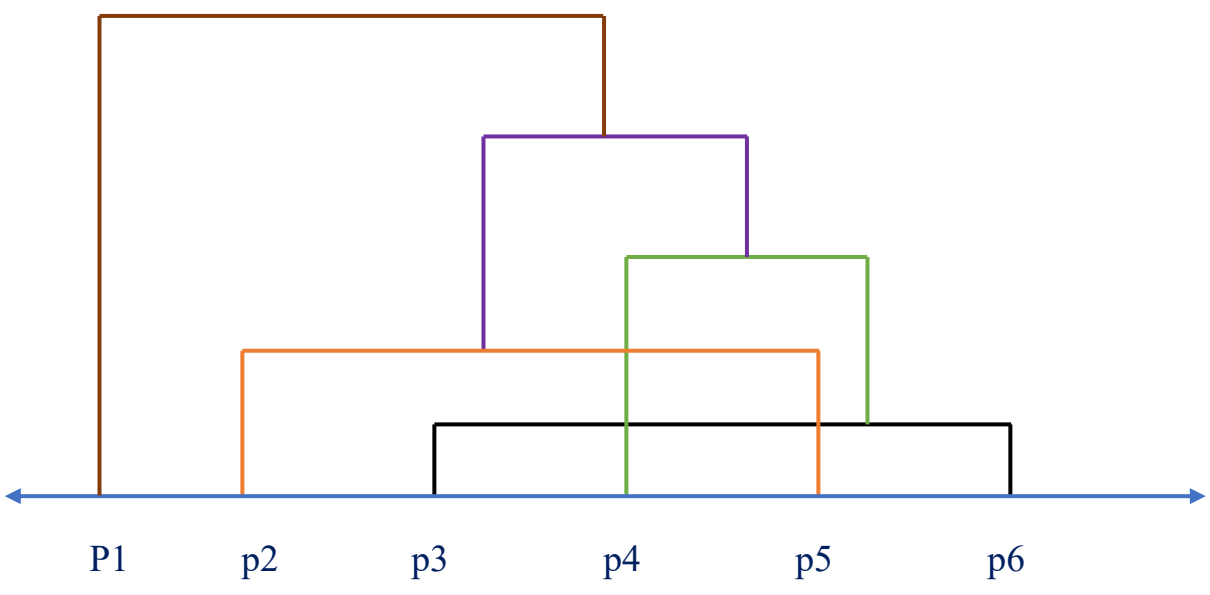
- MIN (((p3, p6), p4), p1) = MIN (distance ((p3, p6), p1), distance (p4, p1)) = MIN (0.22,0.37) =0.22
- MIN (((p3, p6), p4), (p2, p5)) = MIN (distance ((p3, p6), (p2,5)), distance (p4, (p2, p5)) = MIN (0.15,0.15) =0.15

	p1	p2, p5	**P3, p6, p4**
p1	0.00	0.24	**0.22**
p2, p5	0.24	0.00	**0.15**
P3, p6, p4	**0.22**	**0.15**	**0.00**

- MIN ((p2, p5, p3, p6, p4), p4) = MIN (distance ((p2, p5, p3, p6, p4), p4)) = (0.24, 0.22) = 0.22
- MIN ((p2, p5, p3, p6, p4), p4) = MIN (distance ((p2, p5, p3, p6, p4), p4)) = (0.24, 0.22) = 0.22

	p1	**p2, p5, p3, p6, p4**
p1	0.00	**0.22**
p2, p5, p3, p6, p4	**0.22**	**0.00**

the dendrogram



b) Apply agglomerative clustering to the sample dataset using **MAX** as cluster proximity. Show the result clustering using the dendrogram.

- MAX ((p3, p6), p1) = Max (distance (p3, p1), distance (p6, p1)) = Max (0.22,0.23) =0.23
- MAX ((p3, p6), p2) = Max (distance (p3, p2), distance (p6, p2)) = Max (0.15,0.25) =0.25
- MAX ((p3, p6), p4) = Max (distance (p3, p4), distance (p6, p4)) = Max (0.15,0.22) =0.22
- MAX ((p3, p6), p5) = Max (distance (p3, p5), distance (p6, p5)) = Max (0.28,0.39) =0.39

	p1	p2	**p3, p6**	p4	p5
p1	0.00	0.24	**0.23**	0.37	0.34
p2	0.24	0.00	**0.25**	0.20	0.14
p3, p6	**0.23**	**0.25**	**0.00**	**0.22**	**0.39**
p4	0.37	0.20	**0.22**	0.00	0.29
p5	0.34	0.14	**0.39**	0.29	0.00

- MAX ((p2, p5), p1) = Max (distance (p2, p1), distance (p5, p1)) = Max (0.24,0.34) =0.34
- MAX ((p2, p5), (p3, p6)) = MAX (distance (p2, (p3, p6)), distance (p5, (p3, p6)) = MAX (0.25,0.39) =0.39
- MAX ((p2, p5), p4) = Max (distance (p2, p4), distance (p5, p4)) = Max (0.20,0.29) =0.29

	p1	**p2, p5**	p3, p6	p4
p1	0.00	**0.34**	0.23	0.37
p2, p5	**0.34**	**0.00**	**0.39**	**0.29**
p3, p6	0.23	**0.39**	0.00	0.22
p4	0.37	**0.29**	0.22	0.00

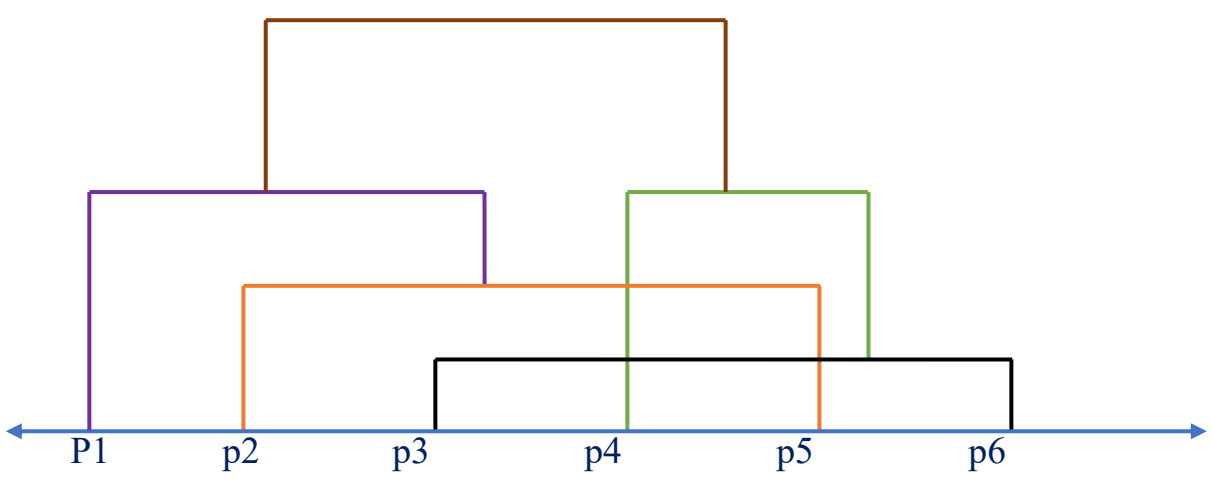
- MAX (((p3, p6), p4), p1) = MAX (distance ((p3, p6), p1), distance (p4, p1)) = MAX (0.23,0.37) =0.37
- MAX (((p3, p6), p4), (p2, p5)) = MAX (distance ((p3, p6), (p2,5)), distance (p4, (p2, p5)) = MAX (0.39,0.22) =0.39

	p1	p2, p5	**P3, p6, p4**
p1	0.00	0.34	**0.37**
p2, p5	0.34	0.00	**0.39**
P3, p6, p4	**0.37**	**0.39**	**0.00**

- MAX ((p1, p2, p5), (p3, p6, p4)) = MAX (distance (((p1, p2, p5), (p3, p6, p4)) = MAX (0.34, 0.37) = 0.37

	p1, p2, p5	p3, p6, p4
p1, p2, p5	**0.00**	**0.39**
p3, p6, p4	**0.39**	0.00

the dendrogram



c) Apply agglomerative clustering to the sample dataset using **group average** as cluster proximity. Show the result clustering using the dendrogram.

- AVG ((p3, p6), p1) = AVG (distance (p3, p1), distance (p6, p1)) = AVG (0.22,0.23) = $\frac{0.22+0.23}{2} = 0.23$
- AVG ((p3, p6), p2) = AVG (distance (p3, p2), distance (p6, p2)) = AVG (0.15,0.25) = $\frac{0.15+0.25}{2} = 0.2$
- AVG ((p3, p6), p4) = AVG (distance (p3, p4), distance (p6, p4)) = AVG (0.15,0.22) = $\frac{0.15+0.22}{2} = 0.19$
- AVG ((p3, p6), p5) = AVG (distance (p3, p5), distance (p6, p5)) = AVG (0.28,0.39) = $\frac{0.28+0.39}{2} = 0.34$

	p1	p2	**p3, p6**	p4	p5
p1	0.00	0.24	**0.23**	0.37	0.34
p2	0.24	0.00	**0.2**	0.20	0.14
p3, p6	**0.23**	**0.2**	**0.00**	**0.19**	**0.34**
p4	0.37	0.20	**0.19**	0.00	0.29
p5	0.34	0.14	**0.34**	0.29	0.00

- AVG ((p2, p5), p1) = AVG (distance (p2, p1), distance (p5, p1)) = AVG (0.24,0.34) =$\frac{0.24+0.34}{2} = 0.29$
- AVG ((p2, p5), (p3, p6)) = AVG (distance (p2, (p3, p6)), distance (p5, (p3, p6)) = AVG (0.2,0.34) = $\frac{0.2+0.34}{2} = 0.27$
- AVG ((p2, p5), p4) = AVG (distance (p2, p4), distance (p5, p4)) = AVG (0.20,0.29) = $\frac{0.20+0.29}{2} = 0.25$

	p1	**p2, p5**	p3, p6	p4
p1	0.00	**0.29**	0.23	0.37
p2, p5	**0.29**	**0.00**	**0.27**	**0.25**
p3, p6	0.23	**0.27**	0.00	0.22
p4	0.37	**0.25**	0.22	0.00

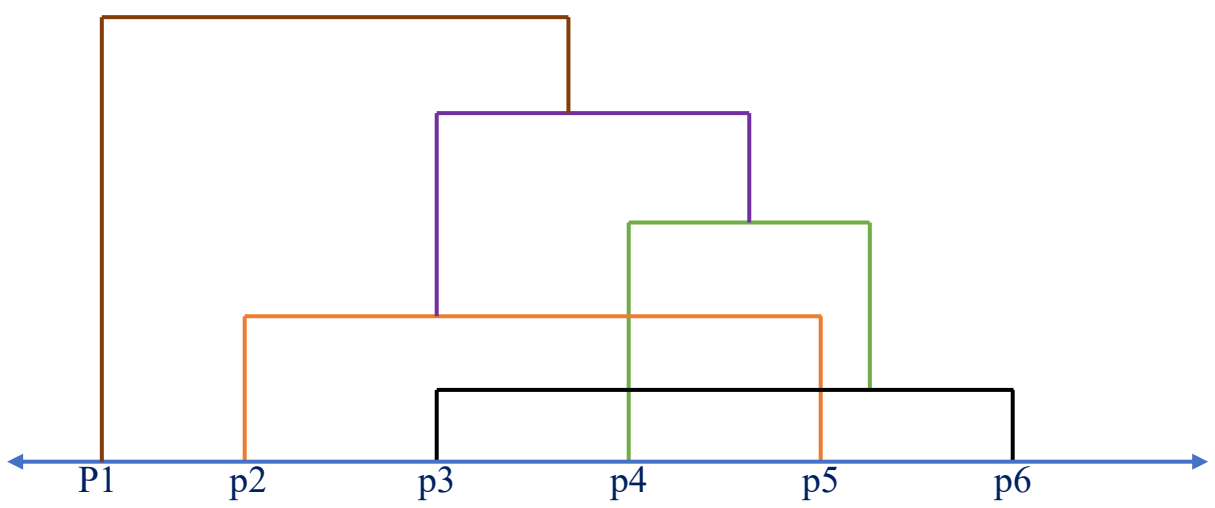
- AVG (((p3, p6), p4), p1) = AVG (distance ((p3, p6), p1), distance (p4, p1)) = AVG (0.23,0.37) =$\frac{0.23+0.37}{2} = 0.3$
- AVG (((p3, p6), p4), (p2, p5)) = AVG (distance ((p3, p6), (p2,5)), distance (p4, (p2, p5)) = AVG (0.27,0.25) =0.26

	p1	p2, p5	**P3, p6, p4**
p1	0.00	0.34	**0.3**
p2, p5	0.29	0.00	**0.26**
P3, p6, p4	**0.3**	**0.26**	**0.00**

- AVG ((p2, p5, p3, p6, p4), p1) = AVG (distance ((p2, p5, p3, p6, p4), p1) = AVG (0.29,0.3) =$\frac{0.29+0.3}{2} = 0.3$

	p1	**P2, p5, p3, p6, p4**
p1	0.00	**0.39**
P2, p5, p3, p6, p4	**0.39**	**0.00**

The diagrams



5) Use the similarity matrix in Table to perform MIN and MAX hierarchical clustering. Shows your result by drawing a dendrogram.

	p_1	p_2	p_3	p_4	p_5
p_1	0.00	0.10	0.41	0.55	0.35
p_2	0.10	0.00	0.64	0.47	0.98
p_3	0.41	0.64	0.00	0.44	0.85
p_4	0.55	0.47	0.44	0.00	0.76
p_5	0.35	0.98	0.85	0.76	0.00

MIN as cluster

	p1, p2	p3	p4	p5
p1, p2	**0.00**	**0.41**	**0.47**	**0.35**
p3	**0.41**	0.00	0.44	0.85
p4	**0.47**	0.44	0.00	0.76
p5	**0.35**	0.85	0.76	0.00

- Min ((p1, p2), p3) = Min (distance (p1, p3), distance (p2, p3)) = Min (0.41,0.64) =0.41
- Min ((p1, p2), p4) = Min (distance (p1, p4), distance (p2, p4)) = Min (0.55,0.47) =0.47
- Min ((p1, p2), p5) = Min (distance (p1, p5), distance (p2, p5)) = Min (0.35,0.98) =0.35

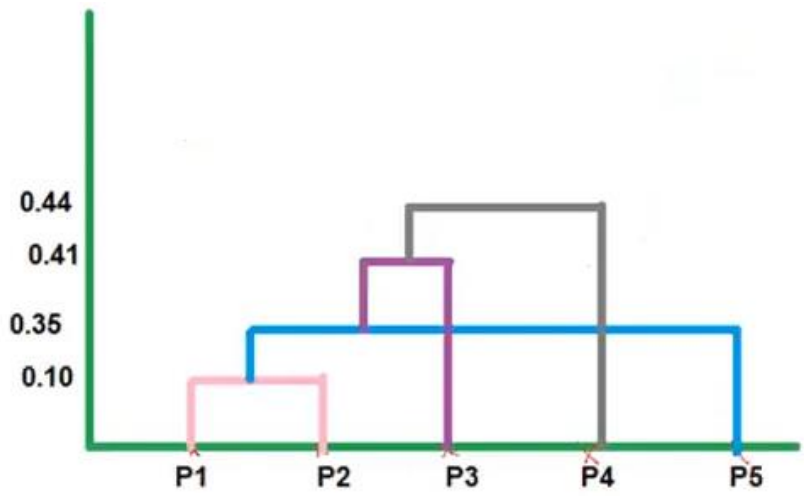
	p1, p2, p5	p3	p4
p1, p2, p5	0.00	0.35	0.47
p3	0.41	0.00	0.44
p4	0.47	0.44	0.00

- Min (((p1, p2), p5)), p3) = Min (distance ((p1, p2), p3), distance ((p5, p3) = Min (0.41,0.85) =0.41
- Min (((p1, p2), p5)), p4) = Min (distance ((p1, p2), p4), distance ((p5, p4) = Min (0.47,0.76) =0.47

The end table.

	p1, p2, p5, p3	p4
p1, p2, p5, p3	**0.00**	**0.44**
p4	0.44	0.00

the dendrogram



MAX as cluster

	p1, p2	p3	p4	p5
p1, p2	**0.00**	**0.64**	**0.55**	**0.98**
p3	**0.64**	0.00	0.44	0.85
p4	**0.55**	0.44	0.00	0.76
p5	**0.98**	0.85	0.76	0.00

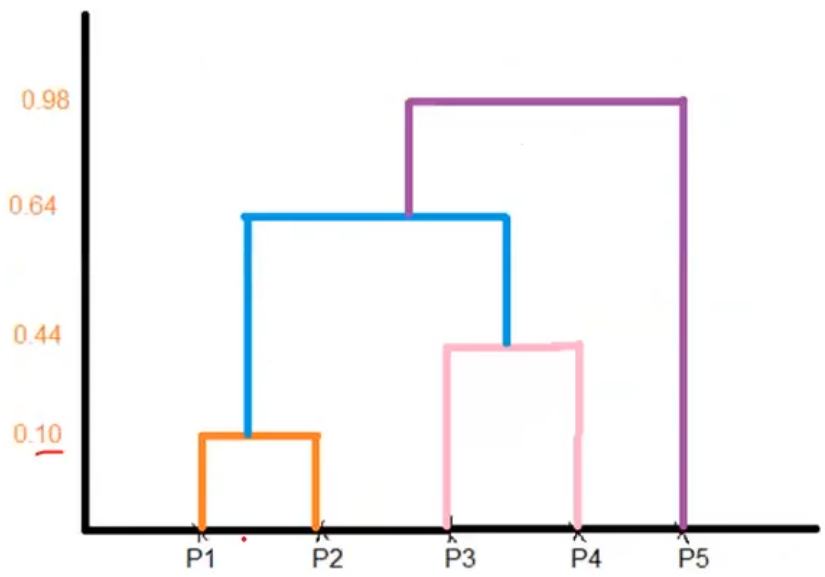
- Max ((p1, p2), p3) = Max (distance (p1, p3), distance (p2, p3)) = Max (0.41,0.64) =0.64
- Max ((p1, p2), p4) = Max (distance (p1, p4), distance (p2, p4)) = Max (0.55,0.47) =0.55
- Max ((p1, p2), p5) = Max (distance (p1, p5), distance (p2, p5)) = Max (0.35,0.98) =0.98

	p1, p2	**p3, p4**	p5
p1, p2	0.00	**0.64**	0.98
p3, p4	**0.64**	**0.00**	**0.85**
p5	0.98	**0.85**	0.00

- Max ((p3, p4), (p1, p2)) = Max (distance (p3, (p1, p2)), distance (p5, (p1, p2)) = Max (0.64,0.55) =0.64
- Max ((p3, p4), p5) = Max (distance (p3, p5), distance (p4, p5) = Max (0.85,0.76) =0.85

The end table.

	p1, p2, p3, p4	p5
p1, p2, p3, p4	**0.00**	**0.98**
p5	**0.98**	0.00



6) The following is a set of one-dimensional points: {6, 12, 18, 24, 30, 42, 48}

a) For each of the following set of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroid.

I. {18 ,45}

m1=18

C1 = {6, 12, 18, 24,30}

The new centroids are:

$$m1 = \frac{6 + 12 + 18 + 24 + 30}{5} = 18$$

C1 = {6, 10, 12, 18, 20, 24,30}

C1 and C1 are the same.

m2 =35

C2 = {42,48}

The new centroids are:

$$m2 = \frac{42 + 48}{2} = 45$$

C2 = {42,48}

C2 and C2 are the same.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon c1} dist(c_i, x)^2$$

$$SSE = (18-6)^2 + (18-12)^2 + (18-18)^2 + (18-24)^2 + (18-30)^2 + (45-42)^2 + (45-48)^2$$
$$= 144 + 36 + 0 + 36 + 144 + 9 + 9 = 378$$

II. {15,40}

m1=15

C1 = {6, 12, 18, 24}

The new centroids are:

$$m1 = \frac{6 + 12 + 18 + 24}{4} = 15$$

C1 = {6, 10, 12, 18, 20, 24}

C1 and C1 are the same.

m2 =40

C2 = {30,42,48}

The new centroids are:

$$m2 = \frac{30 + 42 + 48}{3} = 40$$

C2 = {30,42,48}

C2 and C2 are the same.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon c1} dist(c_i, x)^2$$

$$SSE = (15-6)^2 + (15-12)^2 + (15-18)^2 + (15-24)^2 + (40-30)^2 + (40-42)^2 + (40-48)^2 = 81 + 9 + 9 + 81 + 100 + 4 + 64 = 348$$

b) What are the two clusters produced by Agglomerative clustering using MIN for cluster proximity?

. I do not know the answer.

7) State briefly the main steps of K-means algorithm. Contrast the variant methods that are used for choosing initial centroids in K-means.

The K-means algorithm is a clustering algorithm that partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean

The main steps of K-means algorithm are:

- Choose the number of clusters k and randomly initialize k centroids.
- Assign each observation to the nearest centroid.
- Recalculate the centroid of each cluster as the meaning of all observations assigned to it.
- Repeat steps 2-3 until convergence.

8) State briefly the main steps of K-means algorithm. Contrast the variant methods that are used for choosing initial centroids in K-means.

The main steps of K-means algorithm are:

- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. assignment of data points to clusters isn't changing.
- Compute the sum of the squared distance between data points and all centroids.

There are several methods for choosing initial centroids in K-means such as:

- ✓ Random initialization
- ✓ K-means++ initialization
- ✓ Initialization based on hierarchical clustering.

9) Define cluster prototype and show how cluster prototypes are used for data compression.

A cluster prototype is a representative point that summarizes a dataset into "representative points" or "cluster prototypes."

Cluster prototypes can be used for data compression by replacing each data point with its nearest prototype. This reduces the number of data points and can make it easier to analyze the data.

10) Answer each of the following:

a) Considering data whose proximity measure is Euclidean distance, the objective function, which measures the quality of a clustering is sum squared error.

b) The centroid of a cluster containing the three two-dimensional points (1,1), (2, 3), and (6, 2) is (3,2)

11) State whether the following are true or false.

a) The goal of cluster analysis is to group the data such that the objects in different groups are similar to one another and different from the object in its group. (false)

b) In cluster analysis, the greater the similarity within a group and the greater the difference between groups, the better or more distinct the clustering. (true)

c) in K-means clustering, the number of clusters produced is not known. (false)

d) Bisecting K-means is less susceptible to initialization problem. (true)

e) Cluster analysis is the techniques that are used to divide data objects into groups. (true)

f) Partitional **clustering** is a division of data objects into non-overlapping clusters. (true)

g) A **hierarchical clustering** is a division of data objects into nested clusters. (true)

h) Well-separated clusters do not need to be globular but can have any shape. (true)

i) A medoid must be an actual point (data object). (true)

j) When random initialization of centroids is used, different runs of K-means typically produce different clustering's (total SSEs). (true)

k) In agglomerative hierarchical clustering, we start with one cluster including all the points, and at each step split a cluster until only singleton clusters of individual points remain. (false)

l) In agglomerative hierarchical clustering, we start with the points as individual clusters, and at each step, we merge the closest pair of clusters until one cluster remains. (true)

12) Choose the correct answer for each of the following.

a) In ....., each data object is assigned to a single cluster.

I. Fuzzy clustering
II. Exclusive clustering
III. non-exclusive clustering.

b) In ....., each data object is assigned to more than one cluster.

I. Fuzzy clustering
II. Exclusive clustering
III. non-exclusive clustering

c) In ....., each data object is assigned to every cluster with a membership weight.

I. Fuzzy clustering
II. Exclusive clustering
III. non-exclusive clustering

d) .....tend to be globular.

I. Center-based cluster
II. Well-separated clusters
III. Contiguity-based

e) ……. almost never corresponds to an actual data point.

I. Centroid
II. Medoid

f) Bisecting K-means is an approach introduced for…….

I. Speeding up k-means
II. avoiding the initialization problem of K-means
III. both of (I) and (II).

g) K-means cannot handle .....

I. Non-globular clusters
II. clusters of different sizes and densities
III. both of (I) and (II).

h) The goal of K-means clustering is to .....the squared distance of each point to its closest centroid.

I. Minimize
II. maximize

Association

1) Define association rules, their support and confidence?

Rule generation: object to find all high confidence rules from frequent itemset.

Support: the support value of X with respect to this defined as the proportion of transactions in the database which contains the item-set X

Confidence: the confidence value of a rule X => Y, with respect with a set of transactions T, is the proportion the transaction that contains X witch also contains Y
Confidence of defined as:
Confidence (X => Y) = support (X *U* Y) / support(X)

2) State whether the following are true or false.

a) In any dataset, the number of maximal item sets is greater than the number of closed item sets. (False)
b) The frequency of item sets is a monotonic function. (True)
c) Apriori algorithm follows a depth first exploration of the itemset search space. (False)

3) Consider the example market basket transaction given in Table 8.

Tad	Items
1	{Bread. Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Tad	Items
1	{a, b, c}
2	{a, b, c, d}
3	{b, c, e}
4	{a, c, d, e}
5	{d, e}

item	Support
Bread	4
Milk	4
Diapers	4
Beer	3
Eggs	1
Cola	2

Item	Support
Bread, Milk	3
Bread, Diapers	3
Bread, Beer	2
Bread, Eggs	1
Bread, Cola	1
Milk, Diapers	3
Milk, Beer	2
Milk, Eggs	0
Milk, Cola	2
Diapers, Beer	3
Diapers, Eggs	1
Diapers, Cola	2
Beer, Eggs	1
Beer, Cola	1
Eggs, Cola	0

Item	Support
Bread, Milk, Diapers	2
Bread, Milk, Beer	1
Bread, Milk, Eggs	0
Bread, Milk, Cola	1
Bread, Diapers, Beer	2
Bread, Diapers, Eggs	1
Bread, Diapers, Cola	1
Bread, Beer, Eggs	1
Bread, Beer, Cola	0
Bread, Eggs, Cola	0
Milk, Diapers, Beer	2
Milk, Diapers, Eggs	0
Milk, Diapers, Cola	2
Milk, Beer, Eggs	0
Milk, Beer, Cola	1
Diapers, Eggs, Cola	0
Diapers, Eggs, beer	1
Diapers, cola, Beer	1
Diapers, Eggs, Cola	0
Beer, Eggs, Cola	0

Item	Support
Bread, Milk, Diapers, Beer	1
Bread, Milk, Diapers, Cola	1
Milk, Diapers, Beer, Cola	1

a) How many distinct items are in the dataset?

Answer: 6

b) How many item sets of length 3 are in the datasets?

Answer:20

c) if min sup = 40%, how many frequent item sets of length 2?

Answer: 8

d) How many times does the apriori algorithm scans the dataset if *min sup* = 40%?

Answer: 4items

e) How many association rules of confidence greater than 50%?

Each frequent k-itemset can produce $(2^k - 2)$ association rule.

$$(2^2 - 2) * 8 + (2^3 - 2) * 4 = 40$$

f) Define the apriori principle. How many item sets can be pruned by the apriori algorithm following this principle?

apriori principle: any subset of a frequent itemset is also frequent itemset

4 items

g) Show how the apriori algorithm generates frequent item sets from the dataset given in Table 8. (*Min sup* = 40%)

تم شرح السوال في الجداول الموجوده في الصفحة السابقة

h) How many (subset) checks apriori algorithm performs while collecting item sets frequencies, if (1) using the hash tree to store candidates, (2) store the candidates in a list? (*Min sup* = 40%)

I do not know the answer.

i) How many maximal item sets found? (*Min sup* = 40%)

Answer: 4

j) How many closed item sets found? (*Min sup* = 40%)

Answer: 10

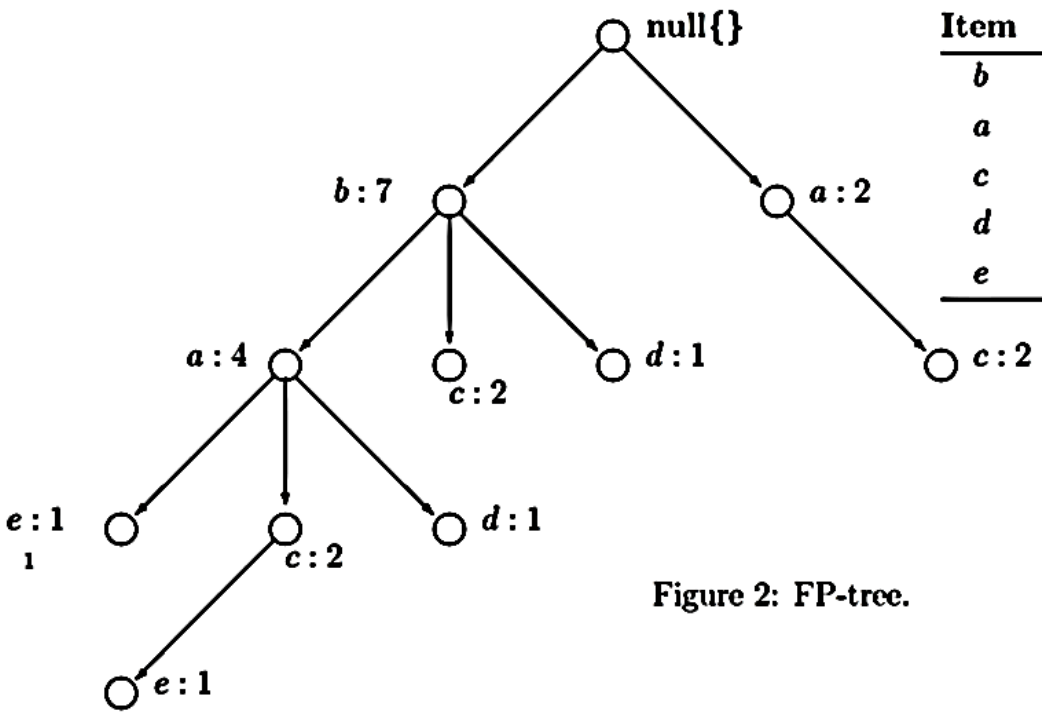
k) Convert the dataset into the vertical format?

Answer

Vertical format					
bread	Milk	Diapers	Beer	Eggs	Cola
1	1	2	2	2	3
2	3	3	3		5
4	4	4	4		
5	5	5			

Binary format						
Num	bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	0	1	1	1	1	0
3	1	0	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

4) Recover the original dataset that corresponds the FP-tree in Figure 2



Item	Supp	Node-link
b	7	
a	6	
c	6	
d	2	
e	2	

Figure 2: FP-tree.

Answer

Number	Transaction
1	**b a e**
2	**b a c e**
3	**b a c**
4	**b a d**
5	**b c**
6	**b c**
7	**b d**
8	**a c**
9	**a c**

لا تنسونا من الدعاء

Mohamed Agami