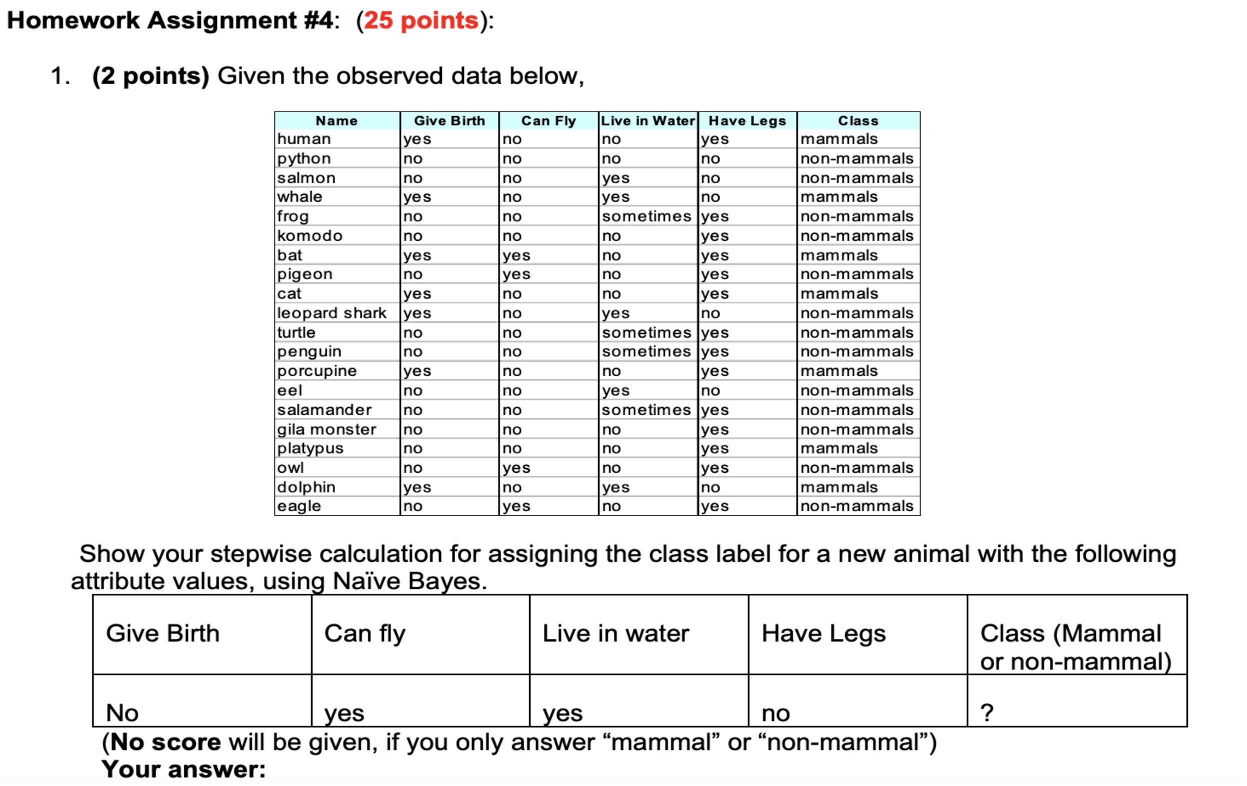
Group #________

Group Leader and ID________

Member Names and IDs____________________



Homework Assignment #4: (**25 points**):

1. **(2 points)** Given the observed data below,

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Show your stepwise calculation for assigning the class label for a new animal with the following attribute values, using Naïve Bayes.

Give Birth	Can fly	Live in water	Have Legs	Class (Mammal or non-mammal)
No	yes	yes	no	?

(**No score** will be given, if you only answer "mammal" or "non-mammal")
Your answer:

To assign the class label for the new animal using Naive Bayes, we need to calculate the probability of the animal being a mammal and the probability of the animal being a non-mammal, given the provided attribute values. We will then choose the class with the higher probability.

The attributes for the new animal are:

- Give Birth: No

- Can Fly: Yes

- Live in Water: Yes

- Have Legs: No

Step 1: Calculate the probability of the animal being a mammal.

P(mammal) = 6/16 = 0.375 (prior probability)

P(no give birth | mammal) = 1/6 = 0.167

P(yes fly | mammal) = 0/6 = 0 (assuming no mammals can fly)

P(yes water | mammal) = 2/6 = 0.333

P(no legs | mammal) = 2/6 = 0.333

P(attributes | mammal) = P(no give birth | mammal) * P(yes fly | mammal) * P(yes water | mammal) * P(no legs | mammal)

P(attributes | mammal) = 0.167 * 0 * 0.333 * 0.333 = 0

P(mammal | attributes) = (P(attributes | mammal) * P(mammal)) / P(attributes)

P(mammal | attributes) = 0 / (0 + 0.625) = 0

Step 2: Calculate the probability of the animal being a non-mammal.

P(non-mammal) = 10/16 = 0.625 (prior probability)

P(no give birth | non-mammal) = 7/10 = 0.7

P(yes fly | non-mammal) = 3/10 = 0.3

P(yes water | non-mammal) = 5/10 = 0.5

P(no legs | non-mammal) = 3/10 = 0.3

P(attributes | non-mammal) = P(no give birth | non-mammal) * P(yes fly | non-mammal) * P(yes water | non-mammal) * P(no legs | non-mammal)

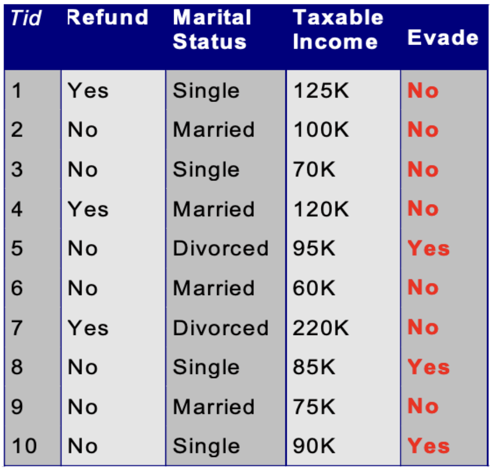
P(attributes | non-mammal) = 0.7 * 0.3 * 0.5 * 0.3 = 0.0315

P(non-mammal | attributes) = (P(attributes | non-mammal) * P(non-mammal)) / P(attributes)

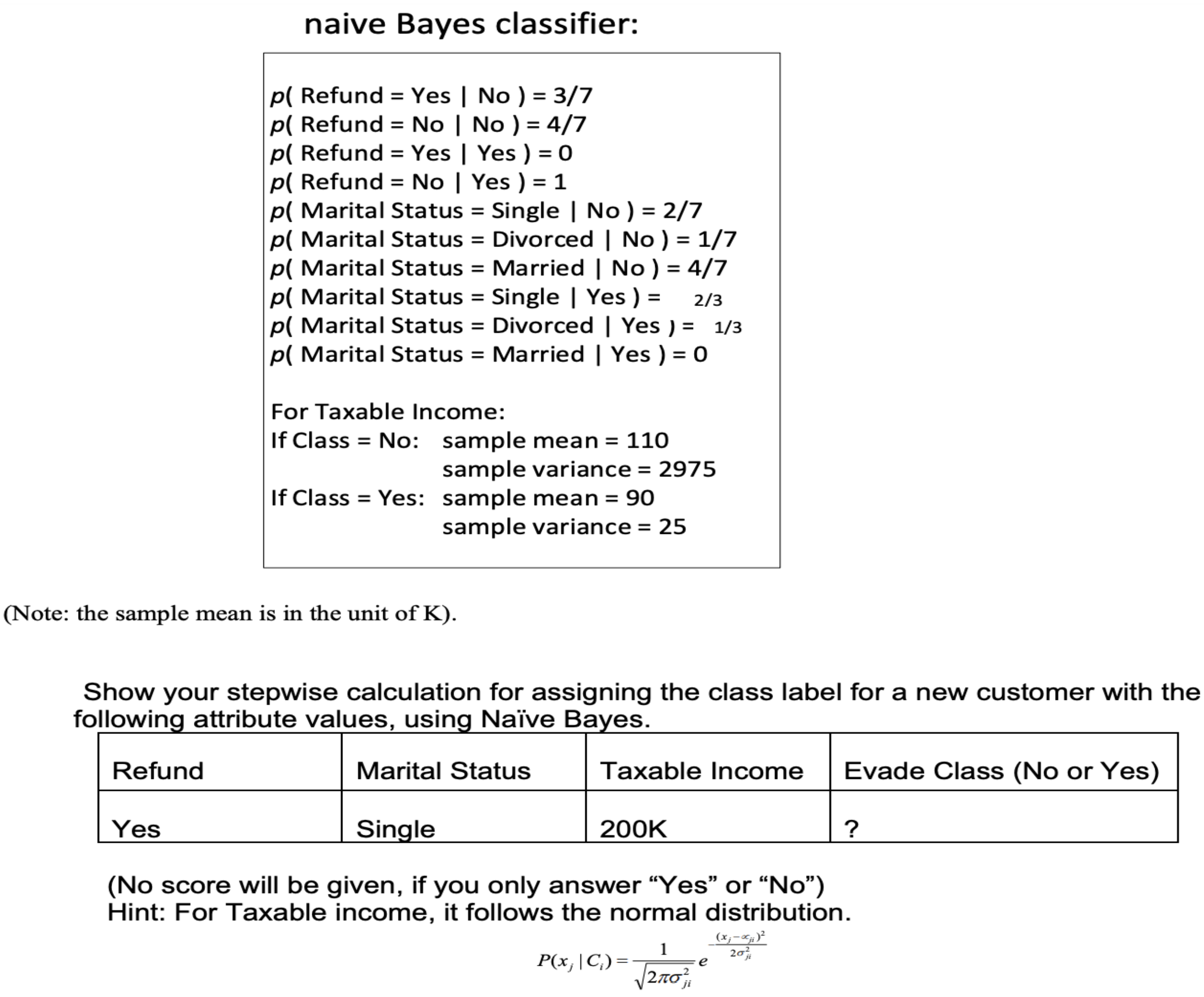
P(non-mammal | attributes) = 0.0315 / (0 + 0.0315) = 1

Since the probability of the animal being a non-mammal (1) is higher than the probability of the animal being a mammal (0), the class label for the new animal is “non-mammal”.

2. **(2 points)** Given the observed data and the reference table below,



Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



naive Bayes classifier:

p(Refund = Yes | No) = 3/7
p(Refund = No | No) = 4/7
p(Refund = Yes | Yes) = 0
p(Refund = No | Yes) = 1
p(Marital Status = Single | No) = 2/7
p(Marital Status = Divorced | No) = 1/7
p(Marital Status = Married | No) = 4/7
p(Marital Status = Single | Yes) = 2/3
p(Marital Status = Divorced | Yes) = 1/3
p(Marital Status = Married | Yes) = 0

For Taxable Income:

If Class = No:	sample mean = 110
	sample variance = 2975
If Class = Yes:	sample mean = 90
	sample variance = 25

(Note: the sample mean is in the unit of K).

Show your stepwise calculation for assigning the class label for a new customer with the following attribute values, using Naïve Bayes.

Refund	Marital Status	Taxable Income	Evade Class (No or Yes)
Yes	Single	200K	?

(No score will be given, if you only answer “Yes” or “No”)
Hint: For Taxable income, it follows the normal distribution.

$$P(x_j \mid C_i) = \frac{1}{\sqrt{2\pi\sigma_{ji}^2}} e^{-\frac{(x_j - \propto_{ji})^2}{2\sigma_{ji}^2}}$$

Ans:

To assign the class label for the new customer using Naive Bayes, we need to calculate the probability of the customer evading taxes (Class = Yes) and the probability of the customer not evading taxes (Class = No), given the provided attribute values. We will then choose the class with the higher probability.

The attributes for the new customer are:

- Refund: Yes

- Marital Status: Single

- Taxable Income: 200K

Step 1: Calculate the probability of the customer evading taxes (Class = Yes).

P(Class = Yes) = 3/10 = 0.3 (prior probability)

P(Refund = Yes | Class = Yes) = 0 (from the given probabilities)

P(Marital Status = Single | Class = Yes) = 2/7 = 0.2857

P(Taxable Income = 200K | Class = Yes) = (1 / (sqrt(2 * pi * 25))) * exp(-((200 - 90)^2) / (2 * 25)) = 0.0057

P(attributes | Class = Yes) = P(Refund = Yes | Class = Yes) * P(Marital Status = Single | Class = Yes) * P(Taxable Income = 200K | Class = Yes)

P(attributes | Class = Yes) = 0 * 0.2857 * 0.0057 = 0

P(Class = Yes | attributes) = (P(attributes | Class = Yes) * P(Class = Yes)) / P(attributes)

P(Class = Yes | attributes) = 0 / (0 + 0.1696) = 0

Step 2: Calculate the probability of the customer not evading taxes (Class = No).

P(Class = No) = 7/10 = 0.7 (prior probability)

P(Refund = Yes | Class = No) = 3/7 = 0.4286

P(Marital Status = Single | Class = No) = 2/7 = 0.2857

P(Taxable Income = 200K | Class = No) = (1 / (sqrt(2 * pi * 2975))) * exp(-((200 - 110)^2) / (2 * 2975)) = 0.0702

P(attributes | Class = No) = P(Refund = Yes | Class = No) * P(Marital Status = Single | Class = No) * P(Taxable Income = 200K | Class = No)

P(attributes | Class = No) = 0.4286 * 0.2857 * 0.0702 = 0.0086

P(Class = No | attributes) = (P(attributes | Class = No) * P(Class = No)) / P(attributes)

P(Class = No | attributes) = 0.0086 / (0.0086 + 0) = 1

Since the probability of the customer not evading taxes (1) is higher than the probability of the customer evading taxes (0), the class label for the new customer is **No**.

3. (total 4 points)

1) Is the total variance of a dataset equal to the variance explained by components identified in PCA? (1 point)

Your answer:

The total variance of a dataset does not always match the variance explained by the components found in Principal Component Analysis (PCA). PCA extracts components to maximize explained variance, but the total variance of the dataset is usually greater than the variance explained by the identified principal components.

2) Based on the loading matrix from the US arrests data, which variables will be counted into PC1, and which one will be counted into PC2? (1 point)

Ans:

1. Based on the provided loading matrix:

PC1:

- Murder (0.5358995)
- Assault (0.5831836)
- Rape (0.5434321)

PC1 will be influenced more by the variables Murder, Assault, and Rape, as they have higher loadings on PC1.

PC2:

- UrbanPop (0.8728062)

PC2 will be influenced more by the variable UrbanPop, as it has the highest loading on PC2.

3) What are the principal components scores shown on this bi-plot fusing USarrest data? What do the arrows indicate? (2 points)

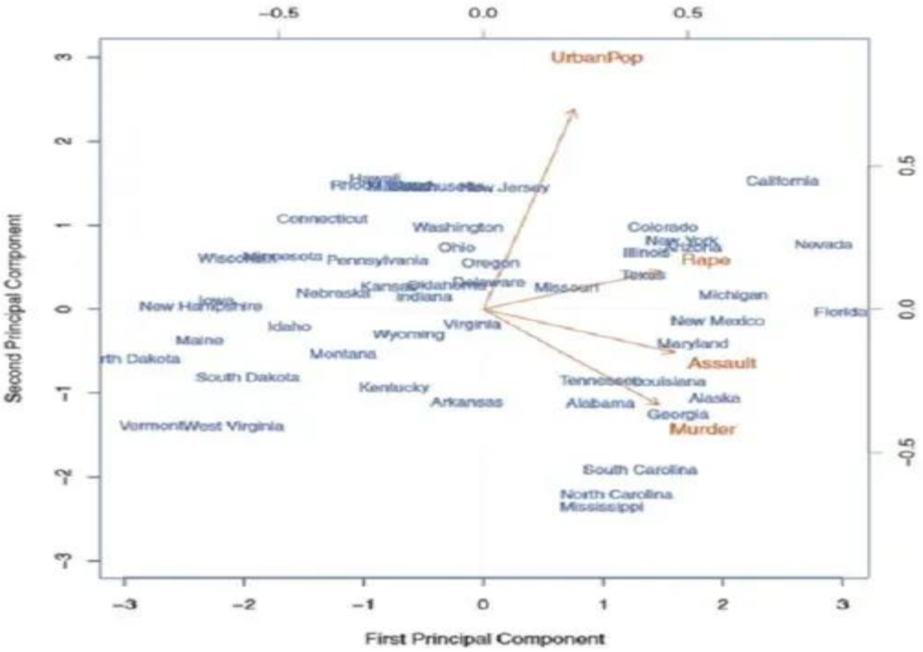


Fig3

Ans:

The bi-plot displays the principal component scores for each state or entity, showcasing the values of PC1 and PC2. The arrows in the bi-plot indicate the direction and strength of the original variables (Murder, Assault, UrbanPop, Rape) on the principal component axes.

The length of the arrows reflects the contribution of each variable to the principal components, while the direction indicates the correlation between the variables and the principal components. In particular, the lengthy arrow for UrbanPop signifies its significant impact on PC2, whereas the arrows for Murder, Assault, and Rape have a more pronounced influence on PC1. The positions of the state/entity labels in the bi-plot show how they are represented by the first two principal components.

4. (5 points; 1 point *5)

1) How to deal with random initialization issues in K-means?

Your answer:

Dealing with random initialization problems in K-means can be approached in several ways: - Running K-means multiple times with different initializations and choosing the solution with the lowest within-cluster sum of squares (WCSS). - Utilizing a more reliable initialization technique like the K-means++ algorithm, which selects initial centroids to prevent getting stuck in poor local optimums. - Experimenting with different numbers of clusters (K) and selecting the most effective one, possibly through methods like the elbow method or silhouette analysis.

2) What algorithm can be used to deal with outliers, if k-means is sensitive to outliers? Your answer:

If K-means is affected by outliers, you might want to consider using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm instead. DBSCAN is a clustering algorithm based on density and capable of effectively handling outliers. Unlike K-means, DBSCAN does not need the number of clusters as an input and can detect clusters of various shapes and sizes.

3) What are the assumptions for K-means?

Your answer:

The main assumptions for K-means clustering are that the data is numerical and continuous, the clusters have convex shapes and similar sizes, the clusters are well-separated and distinct, and the data points within a cluster are more like each other than to data points in other clusters.

4) What algorithm can we use to prevent local minima resulting from K-means?

Your answer:

To avoid getting stuck in local minimum points when using K-means, you can try the following methods: 1. Hierarchical clustering: Begin by treating each data point as its own cluster and gradually combine the nearest clusters until you reach the desired number of clusters. This strategy can help steer clear of local minimums. 2. Genetic algorithms: Implement a technique that utilizes a population-based approach to better navigate the solution space and pinpoint the overall best solution.

5) How to choose the optimal number of K clusters?

Your Answer:

When determining the ideal number of clusters (K) for K-means, there are multiple methods available:

1. Elbow method: Plot the within-cluster sum of squares (WCSS) across different K values and identify the "elbow" point where the WCSS levels off, indicating the optimal number of clusters.

2. Silhouette analysis: Evaluate the silhouette score for each data point to assess its fit within its cluster. Opt for the K value that maximizes the average silhouette score.

3. Information criteria (e.g., AIC, BIC): Employ information criteria to strike a balance between model effectiveness and complexity. Choose the K value that minimizes the specific information criterion selected.

5. (3 points) Write the K-means pseudo code for choosing 2-clusters for a sample of 100 cases with 2 attributes.

Your answer:

To choose 2 clusters for a sample of 100 cases with 2 attributes using K-means, follow these steps:

> 1. Begin by initializing the centroids of the 2 clusters randomly.
>
> 2. Repeat the following steps until convergence:
>
> > a. Assign each of the 100 data points (with 2 attributes) to the nearest centroid.
> >
> > b. Update the centroids by calculating the mean of all data points assigned to each cluster.
>
> 3. Output the final cluster assignments and centroids. The main steps involved are as follows:
>
> > 1. Start with random initial centroids for the 2 clusters.
> >
> > 2. Assign each of the 100 data points to the closest centroid and update the centroids to be the mean of the data points in each cluster iteratively.
> >
> > 3. Repeat step 2 until the cluster assignments stop changing (convergence).
> >
> > 4. Output the final cluster assignments and the coordinates of the 2 centroids.

This allows us to partition the 100 data points (each with 2 attributes) into 2 clusters based on their proximity to the learned centroids.

6. (3 points) Write the pseudo code for agglomerative hierarchical clustering.

Your answer:

Hierarchical clustering is a process that starts with each data point as its own cluster and gradually merges similar clusters until only one cluster remains.

Initially, there are n clusters, with each of the 100 data points being a separate cluster. The dissimilarity matrix is then computed to capture the differences between all pairs of data points. The main steps involved in agglomerative hierarchical clustering are as follows:

1. Start with n clusters, where each of the 100 data points is considered a separate cluster.

2. Calculate the dissimilarity matrix, which includes the dissimilarities between all 100 data points.

3. Repeat the following steps until only one cluster is left:

   a. Identify the two closest clusters.

   b. Merge these two clusters into a single cluster.

   c. Update the dissimilarity matrix to account for the newly merged cluster.

4. Finally, produce the hierarchical clustering dendrogram to visualize the clustering process. By following these steps, the algorithm iteratively clusters

- Initialize each data point as its own cluster, resulting in n total clusters.
- Compute the dissimilarity matrix, which stores the pairwise dissimilarity between each pair of data points.
- Repeatedly find the two most similar clusters, merge them into a single cluster, and update the dissimilarity matrix accordingly. This agglomerative process continues until there is only 1 cluster left.
- Output the final hierarchical clustering dendrogram, which shows the merging of clusters at different levels of similarity.

This step-by-step hierarchical clustering approach allows us to build the cluster hierarchy and visually inspect the relationships between the 100 data points.

7. (3 points) What are the 3 dissimilarity measures in hierarchical clustering?

Your answer:

In hierarchical clustering, there are three common dissimilarity measures:

1. Euclidean distance: - This is the straight-line distance between two data points.

   - It is calculated as the square root of the sum of squared differences between the corresponding feature values.

2. Manhattan (city block) distance: - This is the sum of the absolute differences between the corresponding feature values of two data points.

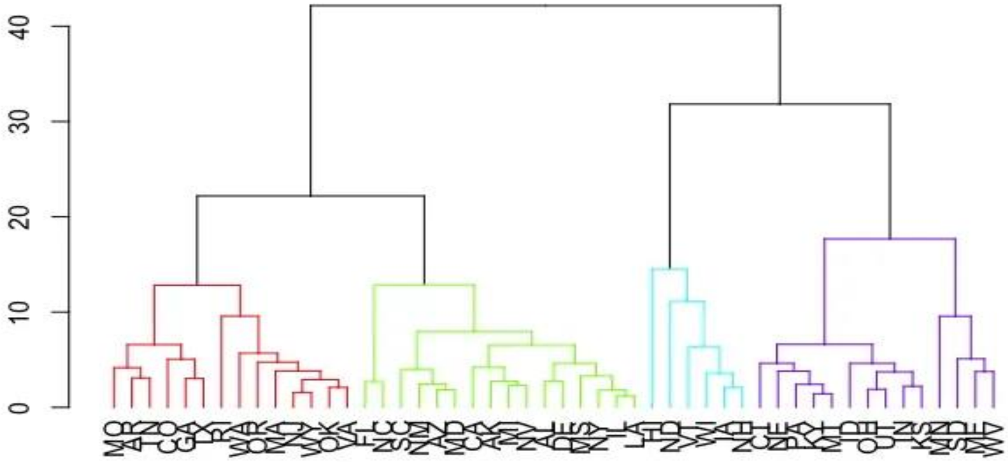
- It is also known as the L1 norm or taxicab geometry distance.

3. Cosine distance: - This is calculated as one minus the cosine of the angle between two data points (treated as vectors).

- It measures the cosine similarity between the two data points.
- The range of cosine distance is from 0 (identical) to 1 (completely dissimilar).

8. (2 points) How many clusters do we have if we cut at a height of 5 in this Figure?

Your answer:



The image shows a hierarchical clustering dendrogram. In a hierarchical clustering dendrogram, the y-axis represents the dissimilarity or distance between clusters, while the x-axis shows the individual data points or clusters.

If we draw a horizontal line at a height of 5 on the y-axis, this line will intersect the dendrogram at 3 distinct locations. This means that if we were to cut the dendrogram at a height of 5, we would end up with 3 separate clusters of data points.

The number of clusters is determined by the number of vertical lines that the horizontal cut-off line intersects. In this case, the cut-off line at height 5 intersects the dendrogram at 3 distinct locations, so the final number of clusters would be 3.

So, in summary, by cutting the hierarchical clustering dendrogram at a height of 5 on the y-axis, the analysis would result in 3 distinct clusters of data points.

9. (1 Point) Gap statistic and silhouette plots can be used to select the optimal number of clusters in hierarchical clustering? True or False

Your answer:

True. Gap statistic and silhouette plots can be used to select the optimal number of clusters in hierarchical clustering.

The gap statistic compares the within-cluster dispersion of the dataset to the expected within-cluster dispersion under a reference null distribution, and the optimal number of clusters is the value of k that maximizes the gap statistic.

The silhouette plot measures how well each data point fits into its assigned cluster, and the optimal number of clusters is the value of k that maximizes the average silhouette coefficient.

Both the gap statistic and silhouette plots are useful techniques for determining the appropriate number of clusters in hierarchical clustering, as they provide objective criteria for selecting the optimal clustering solution.