**Group # 1**

**Names and IDs: Taha Bin Amer 01972684, Pranav Vinod 01984464\_\_**

**Sectional Written Homework #2**: (**75 points**):

1. **(10 points)** Given the observed data below,



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:**

**P (Class = Mammal|X) α p (X | Class = mammal) P (Class = mammal )**

**P( X | Class = mammal ) = P( GiveBirth = No | Class = mammal) x P(CanFly=Yes| Class = mammal ) x P( LiveinWater=Yes| Class = mammal ) x P( HaveLegs = No| Class = mammal )**

**P (Class = mammal) = N\_mammal/N**

**P (Class = Non-Mammal|X) α p (X | Class = Non-Mammal) P (Class = Non-Mammal)**

**P(X | Class = Non-Mammal) = P( GiveBirth = No | Class = Non-Mammal) x P(CanFly=Yes| Class = Non-Mammal) x P( LiveinWater=Yes| Class = Non-Mammal) x P( HaveLegs = No| Class = Non-Mammal)**

**P (Class = Non-Mammal) = N\_ Non-Mammal /N**

**P(X|M) = 1/7 \* 1/7 \* 2/7 \* 2/7 = 0.0016**

**P(X|N) = 12/13 \* 3/13 \* 3/13 \* 4/13 = 0.0151**

**P(X|M)P(M) = 0.0016 \* 7/20 = 0.00056**

**P(X|N)P(N) = 0.0151\*13/20 = 0.00982**

**Since P (Class = Non-Mammal|X) > P(Class = Mammal|X), this new case is classified as Non-Mammal.**

1. **(10 points)** Given the observed data and the reference table below,





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.



**Your answer: To classify the new customer, we must calculate and compare two quantities.**

**P (Yes | X) = P (Refund=Yes | Class=Yes) x P (Marital Status=Single | Class=Yes) x P (Taxable Income=200k | Class=Yes) x P (Yes)**

**P (Yes | X) = 0 \* 2/7 \* 6.3 \* 10-107 \* 3/10 = 0**

**P (Yes | X) = 0**

**P (No | X) = P (Refund=Yes | Class=No) x P (Marital Status=Single | Class=No) x P (Taxable Income=200k | Class=No) x P (No)**

**P (No | X) = 3/7 \* 2/7 \* 0.0073 \* 0.256 \* 7/10**

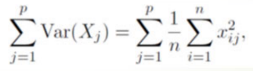
**P (No | X) = 0.00016**

**Since P (No | X) > P (Yes | X), therefore the new customer will be classified as No evade.**

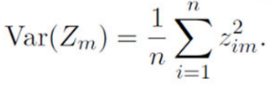
1. **(15 points; 5 points \*3)**
2. Is the total variance of a dataset equal to the variance explained by components identified in PCA?

**Your answer:**

**Yes. The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as:**

****

**And the variance explained by the principal component is:**



**** **It can be shown that with M = min (n − 1, p).**

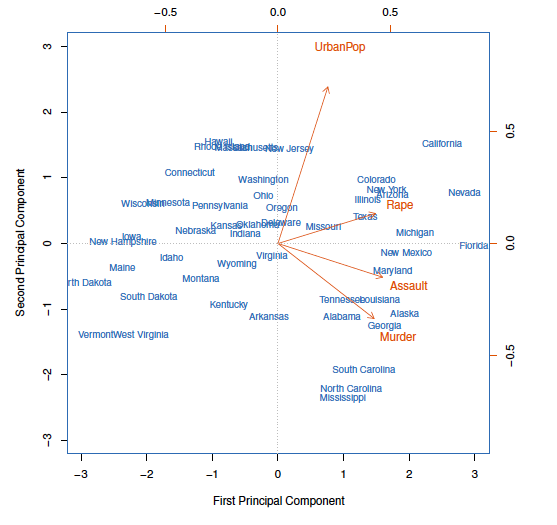
**This tells us the variance in data is fully explained by the identified principal components.**

1. Based on the loading matrix from the USarrests data, which variables will be counted into PC1 and which one will be counted into PC2?



**Your answer: Based on the loading matrix, the variables Murder, Assault and Rape will be counted into PC1 because they have similar and larger weights than UrbanPop. Whereas UrbanPop has a higher weight placed on by PC2 and will be counted into PC2.**

1. What are the principal components scores shown on this bi-plot using USarrest data? What do the arrows indicate?



**Your answer:**

**The PCA biplot displays both the PC scores and the PC loadings. We notice that the crime-related variables (Murder, Assault, and Rape) are close to each other which indicates high co-relation between them. The UrbanPop variable is far away from the other three meaning that it’s less co-related with the crime-related variables. The blue state names are the principal component scores, and the orange arrows indicate the first two PC loading vectors for example, the loading for rape on the 1st PC is 0.54, and on the second is 0.17.**

1. **(10 points; 2 points \*5)**
2. How to deal with random initialization issues in K-means?

**Your answer: There are three ways to deal with issues rising from bad initializations:**

1. **Trying multiple initializations and choosing the best result.**
2. **Choosing the first centroid as one of the examples, the second one which is the farthest from the first and the third one which is farthest from both and so on.**
3. **Lastly, we could look at the K-means++ algorithm to make smarter initializations.**
4. What algorithm can be used to deal with outliers, if k-means is sensitive to outliers?

**Your answer: K-Medians can be used to deal with outliers if K-Means is sensitive to outliers.**

**This is because K-means uses mean to calculate new centroids. Mean as a statistic is sensitive to outliers and can be skewed in the presence of an outliers. Using medians instead of mean to find new centroids makes the algorithm more robust to the presence of outliers.**

1. What are the assumptions for K-means?

**Your answer: There are 2 main assumptions of the K-means clustering algorithm:**

1. **The clusters are spherical or all data points in a cluster are centered around the centroid of that cluster.**
2. **The spread/variance of the clusters are similar. This means that each data point belongs to the closest cluster.**
3. What algorithm can we use to prevent local minima resulting from K-means?

**Your answer: We can use the K-means++ algorithm to prevent local minima. It randomly chooses the first center. Then the next one such that the probability of that point being chosen is proportional to the contribution of that point to the total error. We repeat this until we have K centers.**

1. How to choose the optimal number of K clusters?

**Your answer: There are 2 primary ways to choose the optimal number of clusters K:**

1. **Plot the K-means objective versus K for different values of K and check for the elbow point.**
2. **Perform Silhouette Analysis and look for clusters with silhouette score close to 1.**
3. **(10 points) Write the K-means pseudo code for choosing 2-clusters for a sample of 100 cases with 2 attributes.**

**Your answer:**

**1. Choose the number of clusters(K) and obtain the data points in this case K=2.**

**2. Place the centroids c\_1, c\_2 randomly**

**3. Repeat steps 4 and 5 until convergence or until the end of a fixed number of iterations**

**4. for each data point x\_i:**

**- find the nearest centroid (c\_1, c\_2)**

**- assign the point to that cluster**

**5. for each cluster j = 1..2**

**- new centroid = mean of all points assigned to that cluster**

**6. End**

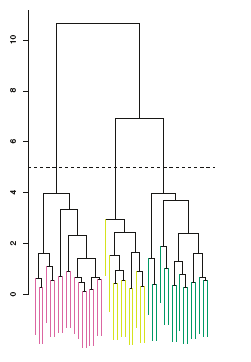
1. **(10 points) Write the pseudo code for agglomerative hierarchical clustering.**

**Your answer: Pseudo code for agglomerative hierarchical clustering is:**

1. **Start by considering each point as its own cluster**
2. **Choose appropriate dissimilarity measure and create a distance matrix.**
3. **Put the two points with the minimum dissimilarity into the same cluster.**
4. **Choose appropriate linkage and create an updated distance matrix.**
5. **Repeat steps 3 and 4 until all points have been assigned a cluster and a dendrogram is created.**
6. **Use gap statistic to find optimal number of clusters and cut dendrogram at that point.**
7. **(5 points) What are the 3 dissimilarity measures in hierarchical clustering?**

**Your answer: There are 3 main dissimilarity measures in hierarchical clustering:**

1. **Min-link or single link – This type of linkage chooses the minimum distance between data points to be the dissimilarity between the clusters. This sort of linkage results in chaining and clusters can get very large.**
2. **Max-link or complete link – This type of linkage chooses the maximum distance between data points to be the dissimilarity between the clusters. Clusters created are small and round.**
3. **Average-link – This type of linkage chooses the average distance between data points in each cluster to be the dissimilarity between the clusters. It is a compromise between min-link and max-link.**
4. **There is also a 4th type of dissimilarity measure: Ward Method, which minimizes the total in-cluster variance.**
5. (**3 points) How many clusters do we have if we cut at a height of 5 in this Figure?**



**Your answer: 3**

1. **(2 Points)**Gap statistic and silhouette plots can be used to select the optimal number of clusters in hierarchical clustering? True or False

**Your answer: (True)**