**Group # 2**

**Names and IDs -> Anubhav Shankar (01951462)**

**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:**

**Important ->** Naïve Bayes assumes independence among attributes xj when given a class.

Mathematically,

p( x1, x2, … , xn | Ci ) = p( x1 | Ci ) p( x2 | Ci ) … p( xn | Ci )

Here, X: attributes -> Give Birth, Can fly, Live in water, Have Legs.

Two classes -> M = Mammal, N = Non-Mammal

P(M) = 7/20 -(1)

P(N) = 13/20 –(2)

Two cases:

1. P (Class = Mammal|X) α p( x | Class = mammal ) p(Class = mammal )
2. P (Class = Non-Mammal|X) α p( x | Class = Non-Mammal ) p(Class = Non-Mammal )

**Case 1 -> P (Class = Mammal|X) α p( x | Class = mammal ) p(Class = mammal )**

*p( x | Class = mammal ) = p( GiveBirth = No | Class = mammal) \* p(CanFly= Yes | Class = mammal ) \*p( LiveinWater=Yes| Class = mammal ) \* p( HaveLegs = No| Class = mammal )*  - (3)

From (1)

n(Mammals) = 7

p(X|M) = 1/7 \* 1/7 \* 2/7 \* 2/7 = 0.0016 -(4)

p(X|M) p(M) = 0.0016 \* 7/20 = 5.83 \* 10-4  - (5)

-------------------------------------------------------------------------------------------------------------------

**Case 2 -> 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) \* p(CanFly= Yes | Class = Non-mammal ) \* p( LiveinWater=Yes| Class = Non-mammal ) \* p( HaveLegs = No| Class = Non-mammal )*  - (6)

From (2)

n(Non-Mammals) = 13

p(X|N) = 12/13 \* 3/13 \* 3/13 \* 4/13 = 0.0151 - (7)

p(X|N) p(N) = 0.0151 \* 13/20 = 0.0098 - (8)

From (5) and (8) ->

**P(Class = M | X) < P(Class = N | X) => This case is a 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:**

**Assumption ->** Income follows a Gaussian/Normal Distribution.

**New case, x = (Refund = Yes, Status = Single, Taxable Income = 200K)**

P(Yes) = 0.3 -(1)

P(No) = 0.7 -(2)

*p( x | Class = No ) = p(Refund = Yes | Class = No) \* p(Single | Class = No) \* p(Income = 200K | Class = No )*

* p( x | Class = No ) = 3/7 \* 2/7 \* p(Income = 200K | Class = No ) - (3)
* If class = No, then µji = 110; Sample variance = 2975
* P(Income = 200K | Class = No) = 1/(√2π \* 2975) \* e(200-110/(2 \* 2975)) = 0.0085 -(4)
* Substituting (4) in (3) -> p(x | Class = No) = 0.001 -(5)

*p( x | Class = Yes ) = p(Refund = Yes | Class = Yes) \* p(Single | Class = Yes) \* p(Income = 200K | Class = Yes)*

* P(x | Class = Yes) = 0 \* 2/7 \* p(Income = 200K | Class = Yes) – (6)
* If class = Yes, then µji = 90; Sample variance = 25
* P(Income = 200K | Class = No) = 1/(√2π \* 25) \* e(200-90/(2 \* 25)) = 0.482 -(7)
* Substituting (7) in (6) -> p(x | Class = Yes) = 0 -(8)

From (1), (2), (5), and (8) ->

**p(x | Class = No) \* P(No) > p(x | Class = Yes) \* P(Yes) = 0.3**

**Hence, P(No | x) > P(Yes | x), and this case 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 of the dataset is explained by the principal components identified. To understand the strength of each component, we compute the Proportion of Variance (PVE) explained by each component.

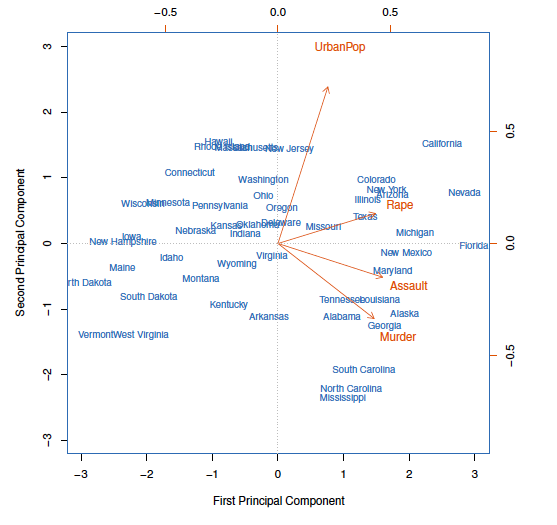
1. Based on the loading matrix from the USarrests data, which variables will be counted into PC1 and which one will be calculated into PC2?



**Your answer:**

Based on the loading matrix, the variables Murder, Assault, and Rape will be counted in PC1 and thus indicate the measure of overall rates of serious crimes. A high correlation between the variables is observed. The variable, UrbanPop, will be counted in PC2 and thus measure the level of urbanization in the state.

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



**Your answer:**

The arrows in the biplot indicate the first two principal components’ loading vectors. The PC loading scores for each variable is -> Murder(0.52, -0.41), Assault(0.58, -0.18), UrbanPop(0.27,0.87), Rape(0.54, 0.16). The negative sign indicates the direction.

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

**Your answer:** Trying multiple initializations and choosing the best results. We can use other robust algorithms like K-means ++.

1. What algorithm can deal with outliers if k-means is sensitive to outliers?

**Your answer:** Use the K-medians algorithm.

1. What are the assumptions for K-means?

**Your answer:** Following are the assumptions of K – Means:

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

**Your answer: K-Means ++**

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

**Your answer: Choose the elbow point from the scree-plot.**

1. **(10 points) Write the K-means pseudo-code for choosing 2-clusters for a sample of 100 cases with two attributes.**

**Your answer: Pseudocode ->**

**Select 2 points as the initial centroids**

**Repeat:**

**Form 2 clusters by assigning all 100 points to the closest centroid**

**Recompute the centroid of each cluster**

**Compare each individual’s distance to its updated cluster mean and to that of the other cluster**

**Until:**

**The centroids don’t change**

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

**Your answer: Pseudocode ->**

**Begin with ‘n’ observations and a measure of all the (n \* (n – 1))/2 pairwise dissimilarities. Treat each observation as its cluster.**

**For (i = n, n-1, …., 2) :**

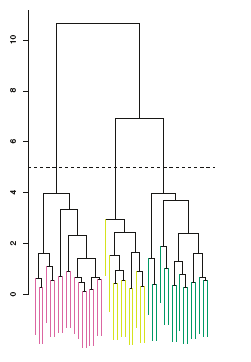
**Examine all pairwise inter-cluster dissimilarities among the clusters and identify the pair of clusters that are least dissimilar. Fuse the two clusters.**

**Compute the new pairwise inter-cluster dissimilarities among the remaining clusters.**

1. **(5 points) What are the three dissimilarity measures in hierarchical clustering?**

**Your answer:** The three dissimilarity measures in hierarchical clustering are ->

1. Min- link -> The minimum distance between the data points of each cluster.
2. Max – link -> The maximum distance between the data points of each cluster
3. Average link -> The mean distance between the data points of each cluster
4. (**3 points)** How many clusters do we have if we cut at the height of 5 in this Figure?



**Your answer: Three (3)**

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

**Your answer: True**