**Group #\_\_\_\_\_\_\_**

**Names and IDs\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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:**

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:**

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

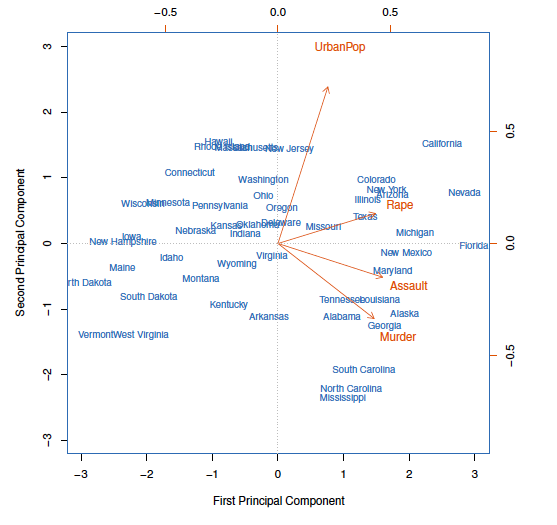
**Your answer:**

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:**

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



**Your answer:**

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

**Your answer:**

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

**Your answer:**

1. What are the assumptions for K-means?

**Your answer:**

1. What algorithm can we use to prevent local minima resulting from K-means?

**Your answer:**

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

**Your answer:**

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

**Your answer:**

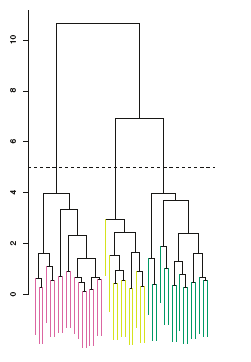
1. **(10 points) Write the pseudo code for agglomerative hierarchical clustering.**

**Your answer:**

1. **(5 points) What are the 3 dissimilarity measures in hierarchical clustering?**

**Your answer:**

1. (**3 points)** How many clusters do we have if we cut at a height of 5 in this Figure?



**Your answer:**

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:**