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:

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.

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

Your answer:

3. **(total 4 points)**
   1) Is the total variance of a dataset equal to the variance explained by components identified in PCA? **(1 point)**

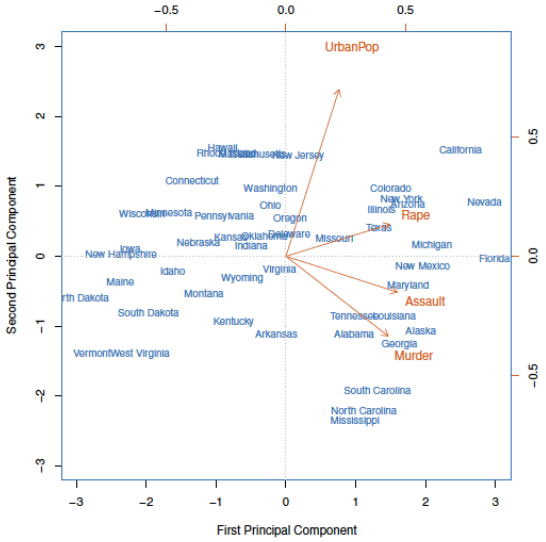
   Your answer:

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

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Your answer:

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



Your answer:

4. **(5 points; 1 point *5)**

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

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

3) What are the assumptions for K-means?
Your answer:

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

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

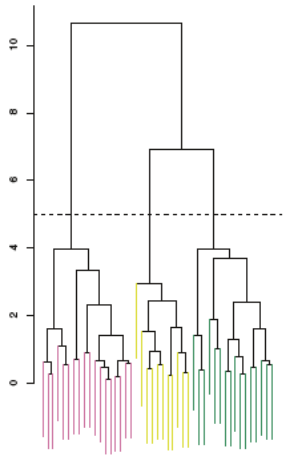
Your answer:

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

6. **(3 points) Write the pseudo code for agglomerative hierarchical clustering.**
   Your answer:

7. **(3 points) What are the 3 dissimilarity measures in hierarchical clustering?**
   Your answer:

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



Your answer:

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: