Problem Set 2

STAT 380
UAEU

This is a Student's Activity Task containing a few multiple type questions and a number of descriptive type problems. The activity is not a graded component of the course. Its objective is to encourage students learning while solving problems and also to prepare for the upcoming exam.

Part I: Short answer type questions.

Must Review: Review All the Quiz Problems. Also Review all the problems in the problem set that we have circulated before the midterm.

1. If we consider a Dataset that has 100 observations and 170 covariates (independent variables) to model a continuous response variable. Identify whether the following statement is True or False.
Statement: If we consider a principal component analysis of the data, no more than 100 eigenvalues of the corresponding variance covariance matrix can be positive.

   Ans: TURE FALSE

2. **Statement**: 'single-linkage' is a criteria used to identify the distance between two clusters based on the distances of each pair of points that are located withing different clusters.

   Ans: TURE FALSE

3. **Statement:** To start a K-Means clustering algorithm we need to specify the number of clusters the data may have.

   Ans: TURE FALSE

4. Under which of the following conditions is k-fold cross-validation the same as leave-one-out cross-validation?

   The training set and test-set have the same number of examples

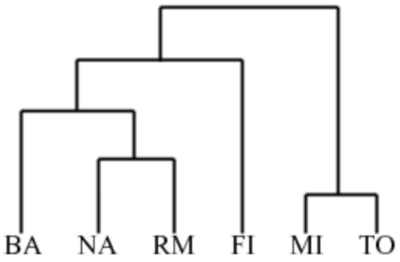
   k=1 k=n None of the above

5. Which one of the following is the main reason for pruning a Decision Tree?

Ans:

- T o save computing time during testing
- To save space for storing the Decision Tree
- To make the training set error smaller
- To avoid over fitting the training set

6. Consider the dendrogram:



Using this dendrogram to create 3 clusters, what would the clusters be?

- {BA,NA},{RM,FI},{MI,TO}
- {NA,RM},{BA,FI},{MI,TO}
- {BA,NA,RM,FI},{MI},{TO}
- {BA,NA,RM},{FI},{MI, TO}

7. In order to model a multi-category response variable, a classification tree is utilized. Let the response variable can take only 4 categories. During the modeling, a specific terminal node of the tree has the corresponding proportion/ probabilities to be $\hat{\pi} = (0.1, 0.8, 0.04, 0.06)^T$ to the different response categories for the points assigned to it. Calculate the Gini's Index for the obtained probability vector.

Ans:

8. In order to model a multi-category response variable, a classification tree is utilized. Let the response variable can take only 4 categories. During the modeling, a specific terminal node of the tree has the corresponding proportion/ probabilities to be $\hat{\pi} = (0.08, 0.81, 0.05, 0.06)^T$ to the different response categories for the points assigned to it. Calculate the Cross-Entropy Index for the obtained probability vector.

Ans:

9. In order to model a multi-category response variable, a classification tree is utilized. Let the response variable can take only 4 categories. During the modeling, a specific terminal node of the tree has the corresponding proportion/ probabilities to be $\hat{\pi} = (0.15, 0.10.65, 0.1)^T$ to the different response categories for the points assigned to it. Calculate the mis-classification rate for the obtained probability vector.

Ans:

Part II: Descriptive Problems.

1. Must Review: Review All the Quiz Problems. Also Review all the problems in the problem set that we have circulated before the midterm.

2. Consider a clustering problem of bi-variate data with the following distance matrix:

	A	B	C	D	E
A	0	6.5	2.4	4.2	5.9
B	6.5	0	4.7	4.5	2.8
C	2.4	4.7	0	2.4	4.1
D	4.2	4.5	2.4	0	1.7
E	5.9	2.8	4.1	1.7	0

Point names are denoted by the Letters A, B, C, D, E.

   (a) Construct a Dendogram for hierarchical agglomerative clustering using the **complete-linkage** procedure.

   (b) Identify the cluatering assignment of the points if one decides to split the data into two clusters.

Ans:

3.

Consider a clustering problem of bi-variate data

$$A = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, B = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, C = \begin{bmatrix} 2 \\ 6 \end{bmatrix}, D = \begin{bmatrix} 5 \\ 0 \end{bmatrix}, E = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, F = \begin{bmatrix} 4 \\ 0 \end{bmatrix},$$

where the points are labeled as A, B, C, D, E, F. The Euclidean distance between the points are provided in the distance matrix below:

	A	B	C	D	E	F
A	0	2.24	3.61	3.16	4	3
B	2.24	0	4	3.61	2.24	2.83
C	3.61	4	0	6.71	3.61	6.32
D	3.16	3.61	6.71	0	5.83	1
E	4	2.24	3.61	5.83	0	5
F	3	2.83	6.32	1	5	0

During an iteration of the K-means algorithm with 2 clusters, the points $\{A, D, F\}$ are assigned to Cluster1 and the rest of the points to the Cluster2.

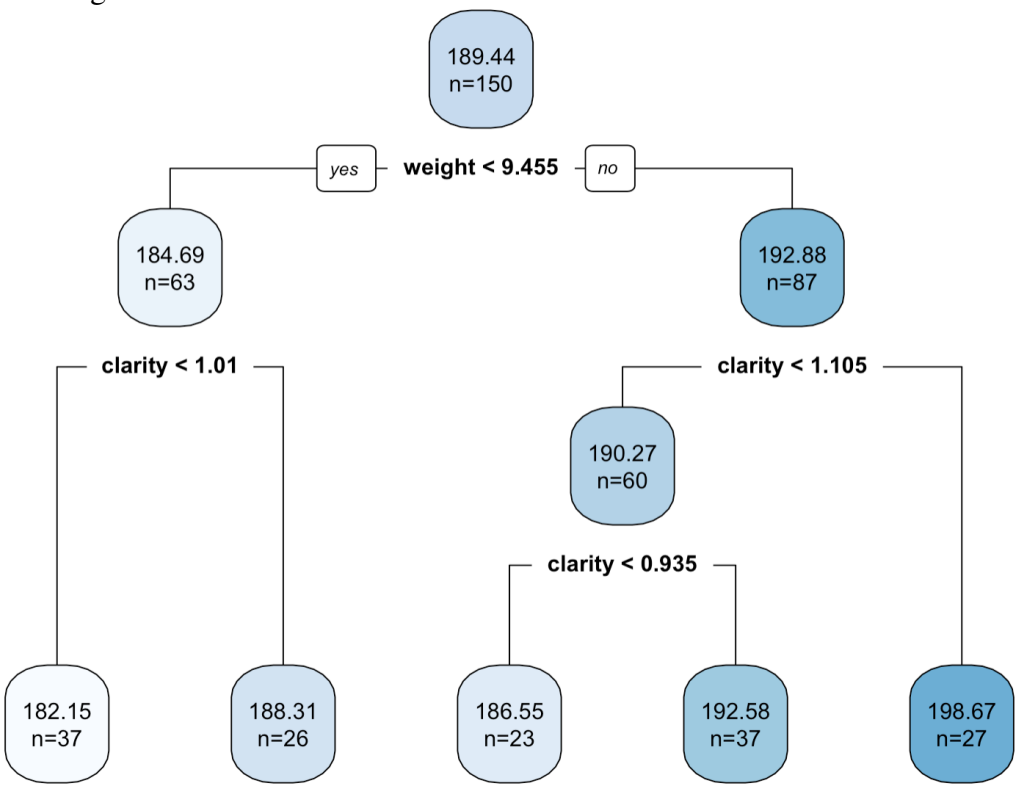
(a) Identify the resulting cluster centers for both the groups at the beginning of the next iteration.

(b) Compute the Withing Group Sum of Squares for clusters provided in part(a). Note that, you may take help from the provided distance matrix.

4.

Consider a dataset that includes data on 150 diamonds sold at an auction. We applied a regression-tree model to build a model to predict the price of a diamond based on the 'weight', 'clarity', and 'color' of a diamond. The variables that are present in the data are the following:
"value": The price at which the diamond is sold. (in 10^5 USD)
"weight": Weight of the diamond. (grams)
"clarity": Clarity measure of the diamond.(A score between 0 to 2)
"color" : Color of the diamond. (A score between 1 to 10)

The following is the diagram of the obtained tree that we denote here by T^* when modeling the continuous variable 'value':



The numbers provided at the terminal node, denote the the corresponding predicted response ans the number of observations are allocated to the specific terminal node. Answer the following questions based on the estimated tree given in this problem.

(a) Let $\alpha(T)$ denote the usual complexity measure obtained by the number of the terminal node of the tree or the number of decision regions that it corresponds. What is the value of $\alpha(T^*)$?

(b) Write down/Describe the different decision regions that the tree corresponds to.

(c) What would be predicted response of a point that has the following values for the corresponding covariates

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
0.17	12.5	7.87	0	0.53	6.01	85.9	6.5	5	311	15.2	17.1	?

5. Consider a small data set of 6 observations:

DataId	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	6.5	3	5.2	2	virginica
2	5.1	3.5	1.4	0.2	setosa
3	7	3.2	4.7	1.4	versicolor
4	4.9	3	1.4	0.2	setosa
5	5.8	2.7	5.1	1.9	virginica
6	4.6	3.4	1.4	0.3	setosa

The objective of this problem is to obtain a bootstrap sample of the above data. As we have only 6 observations here, we may assume that a random data row can be selected by rolling a dice. In particular, we can select the specific DataId that matches with the number that appears as a result of the throw of the dice. The following is the results of 10 independent throws of a dice:

$$\{5,3,4,1,2,4,5,1,6,1\}$$

(a) Obtain a Bootstrap sample of the above data. Clearly indicate which DataId's are you selecting.

6.

Let $\hat{\Sigma}$ be an estimated variance covariance matrix from a data set. As the variance covariance matrix is non negative definite matrix, a spectral decomposition of the matrix is possible to be done. Based on a computation procedure to applied to $\hat{\Sigma}$, the following spectral decomposition is obtained:

$$\hat{\Sigma} := \Gamma\Lambda\Gamma^T, \text{ where}$$

$$\Gamma = \begin{bmatrix} 0.78 & -0.06 & -0.04 & -0.38 & -0.26 & -0.13 & -0.32 & 0.25 \\ 0.24 & -0.33 & 0.78 & 0.03 & 0.27 & 0.25 & -0.03 & -0.29 \\ 0.37 & 0.66 & -0.14 & 0.04 & 0.45 & 0.41 & 0.18 & -0.06 \\ 0.06 & 0.22 & 0.35 & 0.57 & -0.42 & 0.17 & 0.05 & 0.54 \\ -0.32 & 0.21 & 0.19 & -0.61 & -0.42 & 0.52 & 0 & -0.01 \\ 0.3 & -0.15 & -0.21 & 0.21 & -0.52 & 0.14 & 0.46 & -0.55 \\ -0.06 & 0.37 & 0.05 & 0.25 & -0.19 & -0.08 & -0.72 & -0.49 \\ -0.01 & -0.46 & -0.4 & 0.21 & 0.1 & 0.65 & -0.37 & 0.13 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.001 \end{bmatrix}.$$

Note that Γ is an Orthogonal-Matrix i.e. $\Gamma^T\Gamma = \Gamma\Gamma^T = \mathrm{I}_{8\times 8}$ while, Λ is a diagonal matrix with non-negative diagonal elements. Also, note that the columns of the matrix Γ comprises of the eigen-vectors of the matrix $\hat{\Sigma}$ while the diagonal elements of the matrix Λ refers to the corresponding eigen-values of $\hat{\Sigma}$. Answer the following questions based on the results provided above.

(a) Derive the equation for the first three principal components of the observed data. i.e. What is the equation for the first three principal components, PC_1, PC_2, and PC_3, of the observed data?

(b) What is the loading of PC_1 on the different variables?

(c) Derive the variance of PC_1.

(d) Derive the co-variance between PC_1 and PC_2.

(e) Draw a scree plot of the importance factors for the principal components. Based on the scree-plot, identify the optimum number of components (say m) that would be adequate to capture the variability present in the observed data.

(f) What is the percentage of total variability that is captured by the first m principal components of the data.

(g) If a specific observed data point is $X_1 = 2.44, X_2 = 4.82, X_3 = 2.43, X_4 = -0.2, X_5 = -0.29, X_6 = -3.32, X_7 = 1.77, X_8 = -7.65$ What is the score of the plot along the first two principal components?

Let $\hat{\Sigma}$ be an estimated variance covariance matrix from a data set. As the variance covariance matrix is non negative definite matrix, a spectral decomposition of the matrix is possible to be done. Based on a computation procedure to applied to $\hat{\Sigma}$, the following spectral decomposition is obtained:

$$\hat{\Sigma} := \Gamma\Lambda\Gamma^T, \text{ where}$$

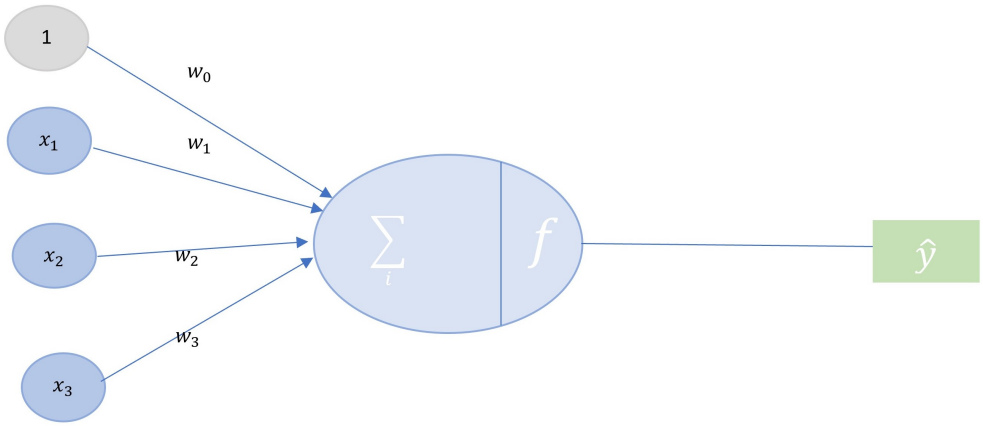
$$\Gamma = \begin{bmatrix} 0.15 & 0.93 & 0.11 & 0.28 & 0.11 \\ 0.58 & 0.03 & 0.22 & -0.21 & -0.76 \\ -0.06 & 0.28 & -0.8 & -0.52 & -0.13 \\ -0.05 & 0.14 & 0.51 & -0.78 & 0.33 \\ 0.8 & -0.17 & -0.22 & 0.01 & 0.53 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 12 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \end{bmatrix}.$$

7. Note that Γ is an Orthogonal-Matrix i.e. $\Gamma^T\Gamma = \Gamma\Gamma^T = \mathrm{I}_{5\times5}$ while, Λ is a diagonal matrix with non-negative diagonal elements. Also, note that the columns of the matrix Γ comprises of the eigen-vectors of the matrix $\hat{\Sigma}$ while the diagonal elements of the matrix Λ refers to the corresponding eigen-values of $\hat{\Sigma}$. Answer the following questions based on the results provided above.
   (a) Derive the equation for the first three principal components of the observed data. i.e. What is the equation for the first three principal components, PC_1, PC_2, and PC_3, of the observed data?
   (b) What is the loading of PC_1 on the different variables?
   (c) Derive the variance of PC_1.
   (d) Derive the co-variance between PC_1 and PC_2.
   (e) Draw a scree plot of the importance factors for the principal components. Based on the scree-plot, identify the optimum number of components (say m) that would be adequate to capture the variability present in the observed data.
   (f) What is the percentage of total variability that is captured by the first m principal components of the data.
   (g) If a specific observed data point is $X_1 = 2.44, X_2 = 4.82, X_3 = 2.43, X_4 = -0.2, X_5 = -0.29$ What is the score of the plot along the first two principal components?

8.

Consider the following diagram of a Perceptron:



Let the weights are given to be $\underset{\sim}{\mathbf{w}}_0 = 5$, $\underset{\sim}{\mathbf{w}}_1 = 2$, $\underset{\sim}{\mathbf{w}}_2 = -3$, $\underset{\sim}{\mathbf{w}}_3 = 1$. The observed covariates/imputs are given as $x_1 = 4, x_2 = 1, X_3 = -1$.

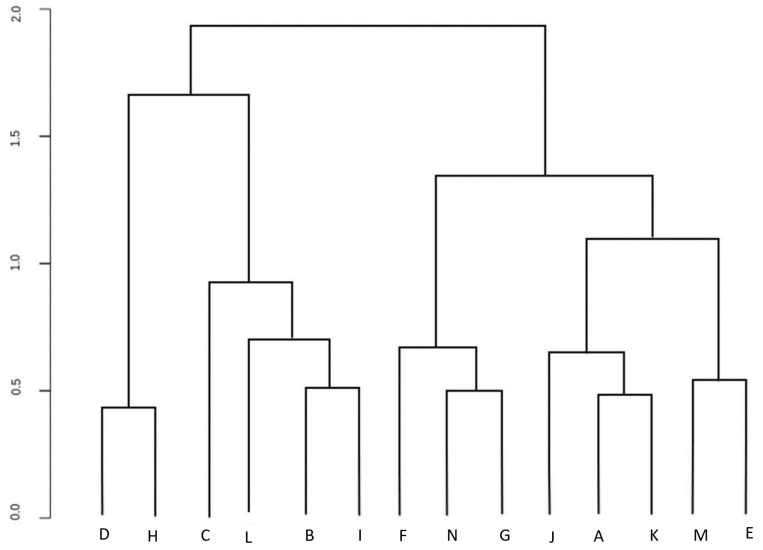
(a) Write down the mathemetical formulation of the above Perceptron. Keep the nonlinear function to be the generic f in your expression.

(b) If we consider the nonlinear function f to be the Sigmoid function, then what would be the output from the above Perceptron?

(c) If we consider the nonlinear function f to be the ReLU function, then what would be the output from the above Perceptron?

9.

Consider the dendrogram:



Using this dendrogram to create 4 clusters, what would the clusters be?

(a) Identify the clusters if it is given that there is only three clusters.
(b) Identity the two points which are closest to each other.