United Arab Emirates University

STAT 380

Final Exam

Name:

ID:

- There are a total of 105 points in this Question Paper. Answer as much as you can. If your acquired score is greater than equal to 100 it will be counted as 100%.
- The Exam is scheduled for 120 minutes
- Students who are late by 15 minutes or more from the comencement of the exam are not be allowed to enter the room.
- A student leaving the exam hall for any reason is not allawed to return.
- Students are not allowed to leave the room before 45 minutes from the commencement of the exam.
- Students are required to carry university ID, calculator and pen/pencil to the desk.
- Electronic gadgets such as a Laptop, mobile phone, smart watch, etc. are not allowed.
- This is a closed book, closed notes exam. However, you may take help from the "Exam Assistance Note" provided along with the exam paper.

For instructor's use only

Problem Number	Obtained Score	Total Score
Problem 1		25
Problem 2		10
Problem 3		10
Problem 4		10
Problem 5		25
Problem 6		10
Problem 7		10
Problem 8		5
TOTAL		105
TOTAL(out of 100)		100

You May Use This Page for Rough Work

Part-I

Multiple Choice or Short Answer Type Problems

1. (a) In order to model a multi-category response variable, a classification tree is utilized where the response variable can of 4 different categories. During the modeling, the proportion/probabilities of different response categories in a specific terminal node appears to be $\hat{\pi} = \begin{bmatrix} 0.15 & 0.08 & 0.65 & 0.12 \end{bmatrix}^T$. What is the **mis-classification rate** for the obtained probability vector $\hat{\pi}$.

Score: _____ Total Score: 5

Answer:

- [ ] 0
- [ ] 0.08
- [ ] 0.65
- [ ] 0.35

(b) Under which of the following conditions is k-fold cross-validation the same as leave-one-out cross-validation? Consider that there are n number of observation in the data.

Score: _____ Total Score: 5

Answer:

- [ ] The training set and test-set have the same number of examples
- [ ] k=1
- [ ] k=n
- [ ] None of the above

(c) In the context of Standard Linear Regression to model a continuous response variable, how do we check if there is 'Influential' point in the dataset.)

Score: _____ Total Score: 5

Answer:

- [ ] It is a Influential point if the corresponding Cook's Distance is more than 0.5.
- [ ] It is a Influential point if the corresponding Cook's Distance is less than 0.5.
- [ ] It is a Influential point if the corresponding Cook's Distance is less than 0.05.

(d) Consider a principal component analysis, where the estimated first two principal components are PC_1 and PC_2. Identify whether the following statement is TRUE or FALSE.
Statement: The correlation between PC_1 and PC_2 is 0.

Score: _____ Total Score: 5

Answer:

- [ ] TURE
- [ ] FALSE

(e) Which one of the following is the main reason for pruning a Decision Tree?

Score:
Total Score: 5

Answer:

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

2. (a) Consider modelling a multi-category response variable using a classification tree. Assume that the response variable can be of 3 different categories. During the modeling, a specific terminal node, the proportion/ probabilities of different response categories are estimated to be $\hat{\pi} = \begin{bmatrix} 0.35 & 0.55 & 0.10 \end{bmatrix}^T$. Calculate the **Gini's Index** for the obtained probability vector.

Score:
Total Score: 5

(b) What is the value of Logit(0.65)?

Score:
Total Score: 5

Part-II

Answer the following short type questions. Show your steps to get full credit.

3. Consider the output from a logistic regression model applied to the Credit Risk dataset. There are three variables in the dataset. The variable named **"loanStatus"** indicates whether there is default (loanStatus=1) of the loan or not (loanStatus=0 for not default). The covariates are **"intRate"** and **"age"** that represents the corresponding **interest rate** of the loan and the **age** of the corresponding individual. We fit a following logistic regression model

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\text{Logit}(\pi_i) := \beta_0 + \beta_1 \times \text{intRate}_i + \beta_2 \times \text{age}_i.$$

Here the response variable $Y_i = 1$ if there is a loan default (i.e. if $\text{loanStatus}_i = 1$) corresponding to the i^{th} data point. Answer the parts of this questions based on the following output from the R Statistical Software that is provided below:

```
Call:
glm(formula = loanStatus ~ intRate + age, data = dd)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.025   0.0109583  -2.505   0.01227 *
intRate      0.015   0.0005969  24.717   < 2e-16 ***
age          -0.019   0.006212 -3.042   0.00235 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Interpret the estimated value of regression coefficient corresponding to the variable 'intRate'.

Score:
Total Score: 3

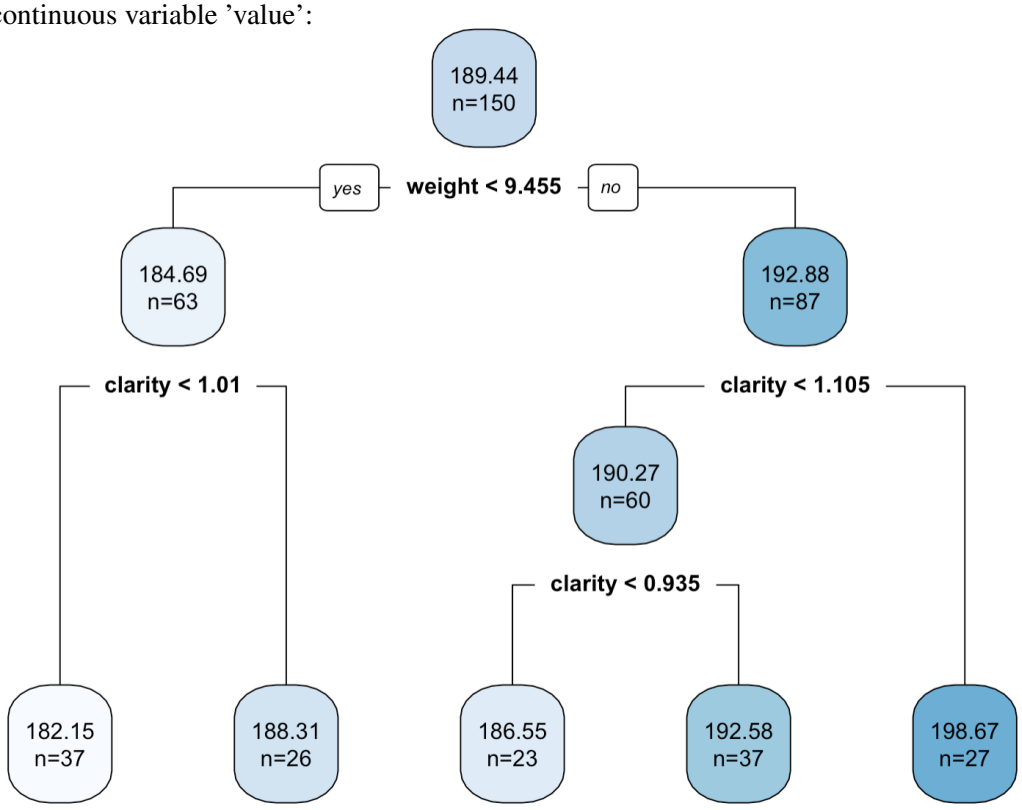
(b) Interpret the estimated value of regression coefficient corresponding to the variable 'age'.

Score:
Total Score: 3

(c) A new loan application has been filed, where the corresponding values of the variables are **age=50, intRate=15**. If the applied loan is approved, then based on the provided model, **predict the probability of the default for the loan.**

Score:
Total Score: 4

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 $\alpha(\mathrm{T})$ 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 $\alpha(\mathbf{T}^*)$?**

Score:
Total Score: 2

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

Score:
Total Score: 5

(c) Based on the given output, what is the predicted response of a point that has the following values for the corresponding covariates

weight	clarity	color
12	1.0	4.0

?

Score:
Total Score: 3

Part-III

Answer the following descriptive type questions. Show your steps to get full credit.

5. Let $\hat{\Sigma}$ be an estimated variance covariance matrix from a data set with 7 variables denoted as $X_1, \ldots, X_7$. 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:

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

$$\Gamma = \begin{bmatrix} -0.92 & 0 & 0.39 & 0 & 0 & 0 & 0 \\ 0 & 0.98 & 0 & -0.04 & 0 & 0 & -0.17 \\ 0.39 & 0 & 0.92 & 0 & 0 & 0 & 0 \\ 0 & 0.05 & 0 & -0.87 & 0 & 0 & 0.49 \\ 0 & 0 & 0 & 0 & -0.73 & 0.68 & 0 \\ 0 & 0 & 0 & 0 & 0.68 & 0.73 & 0 \\ 0 & -0.17 & 0 & -0.49 & 0 & 0 & -0.85 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \end{bmatrix}.$$

Note that Γ is an Orthogonal-Matrix i.e. $\Gamma^T\Gamma = \Gamma\Gamma^T = \mathrm{I}_{7\times 7}$, the identity matrix of dimension 7×7 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) What are the equations for the first **two principal components**, PC_1, and PC_2 of the observed data?

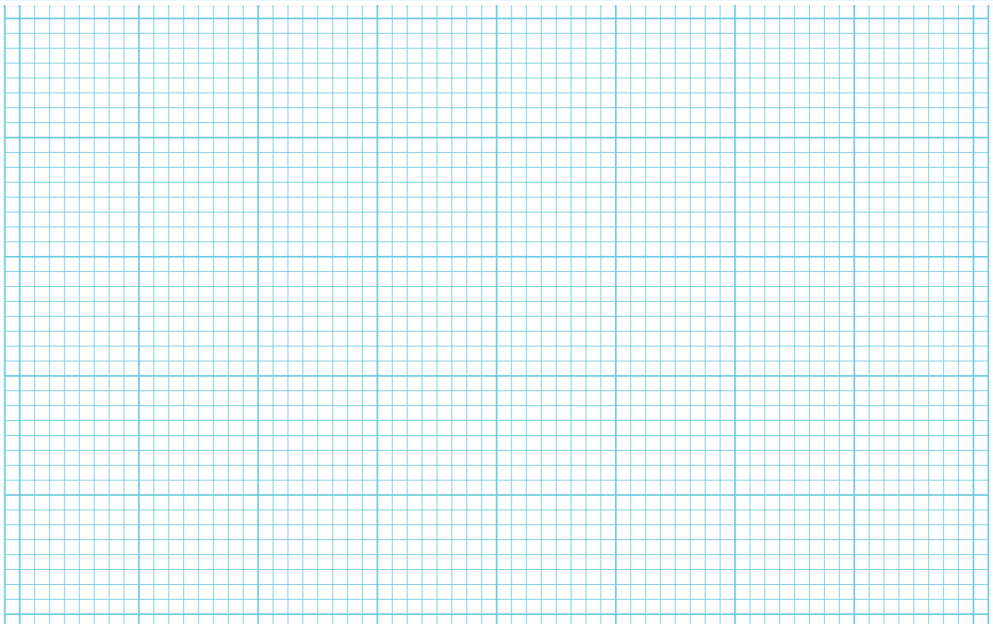
Score:
Total Score: 2+2

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

Score:
Total Score: 3

(c) 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.

Score:
Total Score: 4+1



(d) What is the percentage of total variability that is captured by the first m principal components of the data. (Here m is the optimum number of components that you have decided in the previous part of the problem.)

Score:
Total Score: 5

(e) If a specific observed data point is:
$X_1 = 3.0 \quad X_2 = 4.0 \quad X_3 = 2.0 \quad X_4 = 0 \quad X_5 = 0 \quad 6 = 0 \quad X_7 = 0$
What is the score of the plot along the **first** principal components?

Score:
Total Score: 3

(f) Derive the variance of PC_1.

Score:
Total Score: 5

6. Consider a clustering problem of bivariate data with the following distance matrix:

	A	B	C	D
A	0	7	4	6
B	7	0	4.5	3
C	4	4.5	0	1.5
D	6	3	1.5	0

(a) Complete the tables in the next page and construct a Dendogram for hierarchical agglomerative clustering using the **complete-linkage** procedure.

Score:
Total Score: 4+3

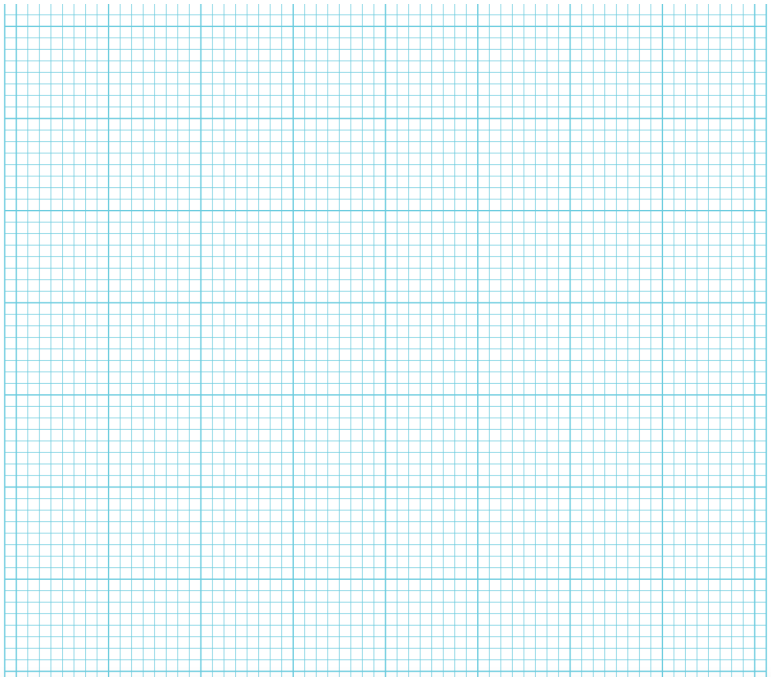
Answer:

Step1:

	A	B	C	D
A	0	7	4	6
B	7	0	4.5	3
C	4	4.5	0	1.5
D	6	3	1.5	0

Step2:

Step3:



(b) Based on the Dendogram that you have obtained, identify the clustering assignments of the points if there is only two clusters in the data.

Score:
Total Score: 3

7. Consider a clustering problem of bivariate data: $A = \begin{bmatrix} 6 \\ 2 \end{bmatrix}, B = \begin{bmatrix} 7 \\ 3 \end{bmatrix}, C = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, D = \begin{bmatrix} 5 \\ 4 \end{bmatrix}, E = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ 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
A	0	1.41	3	2.24	5.10
B	1.41	0	4.12	2.24	6.32
C	3	4.12	0	2.83	2.24
D	2.24	2.24	2.83	0	5
E	5.10	6.32	2.24	5	0

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

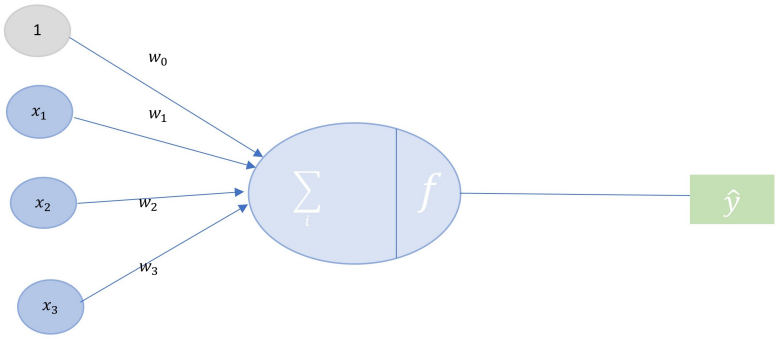
(a) Identify the resulting cluster centers for both the groups at the beginning of the next iteration.

Score:
Total Score: 5

(b) Compute the Withing Group Sum of Squares for clusters provided in part(a).

Score:
Total Score: 5

8. Consider the following diagram of a Perceptron:



Let the weights are given to be $\underset{\sim}{\mathbf{w}}_0 = 5$, $\underset{\sim}{\mathbf{w}}_1 = 2$, $\underset{\sim}{\mathbf{w}}_2 = -3$, $\underset{\sim}{\mathbf{w}}_3 = 1$.

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

Score:
Total Score: 3

(b) If we consider the nonlinear function f to be the **"ReLU"** (Rectified Linear Unit) function, then what would be the output from the above Perceptron? Assume, the observed covariates/imputs are given as $x_1 = 0, x_2 = 1, x_3 = 0$. Note that $\text{ReLU}(x) = \text{Max}(x, 0)$.

Score:
Total Score: 2