Name:

UTeid:

M339G Predictive Analytics
University of Texas at Austin
Mock In-Term Exam II
Instructor: Milica Čudina

Notes: This is a closed book and closed notes exam. The maximal score on this exam is points.

All written work handed in by the student is considered to be
their own work, prepared without unauthorized assistance.

The University Code of Conduct

"The core values of The University of Texas at Austin are learning, discovery, freedom, leadership, individual opportunity, and responsibility. Each member of the university is expected to uphold these values through integrity, honesty, trust, fairness, and respect toward peers and community. As a student of The University of Texas at Austin, I shall abide by the core values of the University and uphold academic integrity."

"I agree that I have complied with the UT Honor Code during my completion of this exam."

Signature:

2.1. CONCEPTUAL QUESTIONS.

Problem 2.1. (10 points) Describe the differences and the similarities between LDA and the logistic regression for a two-class classification problem.

Solution: Solutions will vary. The salient point of any response which is to earn credit must be similar to the last couple of slides here `https://mcudina.github.io/page/M339G/slides/ch4-lda.pdf`.

Problem 2.2. (10 points) Justify this statement from the textbook:

"Neither QDA nor naive Bayes is a special case of the other."

Solution: Solutions will vary. The salient point of any response which is to earn credit must be something along the lines of the second bullet point on page 160 from the textbook.

2.2. **FREE RESPONSE PROBLEMS.** Please, explain carefully all your statements and assumptions. Numerical results or single-word answers without an explanation (even if they're correct) are worth 0 points.

Problem 2.3. (10 points)*Source: Pitman's "Probability."*
Here is a summary of PSAT and SAT scores of a large group of students.

	mean	standard deviation
PSAT	1200	100
SAT	1300	90

Assume that the data are modeled as bivariate normal with the correlation coefficient equal to 0.6. Of the students who scored 1000 on the PSAT, about what percentage scored above average on the SAT?

Solution: Let (U, V) be the random pair which stands for the students' PSAT and SAT scores in real units. Let (X, Y) be the random pair which stands for the students' PSAT and SAT scores in standard units.

Conditioning on U being **exactly** 1000 is equivalent to conditioning on

$$X = \frac{1000 - 1200}{100} = -2$$

So, the probability that we are looking for is

$$\mathbb{P}[Y > 0 \,|\, X = -2].$$

Recall that $Y \,|\, X = x \sim N(\rho x, 1 - \rho^2)$. So, the probability we are seeking equals (with $x = -2$)

$$\mathbb{P}\left[Z > \frac{0 - \rho x}{\sqrt{1 - \rho^2}}\right] = \mathbb{P}\left[Z > -\frac{x\rho}{\sqrt{1 - \rho^2}}\right]$$

With the given correlation coefficient of 0.6, our answer is

$$\Phi\left(\frac{-2(0.6)}{\sqrt{1 - 0.36}}\right) = \Phi(-1.5) = 0.0668072.$$

Problem 2.4. (20 points) Consider the following observations of (X, Y) with X being the predictor and Y being the response:

$$(1, 4), \quad (2, 6), \quad (3, 5), \quad (6, 1).$$

After one iteration of recursive binary splitting, there are two groups of observations. Find the members of the two groups.

Solution: Remember that - in general - the criterion for choosing the splits is to minimize the residual sum of squares (RSS)

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

where $\hat{y}_{R_j}$ stands for the average of the response variable in region R_j for $j = 1, \ldots, J$. However, since this problem is computationally too complex, we resort to **recursive binary splitting**. Hence, as there is one predictor only in our current problem, we must make the split along its possible values. Every available split is **binary** and partitions the support of X into R and R^c. So, it creates an RSS with this structure

$$\sum_{i \in R} (y_i - \hat{y}_R)^2 + \sum_{i \in R^c} (y_i - \hat{y}_{R^c})^2.$$

In this problem, we can now proceed "by hand" from the lowest to the highest observed value of the predictor.

If $(1, 4)$ is the sole element in R, the mean response for the remaining points is

$$\frac{6 + 5 + 1}{3} = \frac{12}{3} = 4.$$

The RSS is

$$(6 - 4)^2 + (5 - 4)^2 + (1 - 4)^2 = 14\,.$$

If $(1, 4)$ and $(2, 6)$ form the first region, the mean response in that region is $\frac{10}{2} = 5$. The remaining points $(3, 5)$ and $(6, 1)$ are in the other region and their mean response is $\frac{6}{2} = 3$. So, the RSS is

$$(4 - 5)^2 + (6 - 5)^2 + (5 - 3)^2 + (1 - 3)^2 = 10\,.$$

If $(1, 4), (2, 6)$, and $(3, 5)$ are in the first region and only $(6, 1)$ remains in the other region, then the average of the first region's values of the response variable is

$$\frac{4 + 6 + 5}{3} = 5.$$

So, the RSS equals

$$(4 - 5)^2 + (6 - 5)^2 + (5 - 5)^2 = 2.$$

Overall, the smallest RSS corresponds to the third partition with $(6, 1)$ in its own region, and the remaining points in the other region.

2.3. MULTIPLE CHOICE QUESTIONS.

Problem 2.5. (5 points)*Source: MAS-II, Fall 2019.*
You are given the following three statements about tree-based methods for regression and classification:

I. The main difference between bagging and random forests is the number of predictors considered at each step in building individual trees.
II. Single decision tree models generally have a higher variance than random forest models.
III. Random forests provide an improvement over bagging because trees in a random forest are less correlated than those in bagged trees.

(a) I only.
(b) II only.
(c) III only.
(d) I, II and III.
(e) The correct answer is not given above.

Solution: (d)
All three statements are true.

Problem 2.6. (5 points) Consider the following data set with the explanatory random variable X and the categorical response Y:

X	1	2	6	8	12	16	17	20	22
Y	N	N	L	N	L	N	L	L	L

Determine which of these splits is/are the best using classification error as the criterion.

I. $R = \{X \leq 7\}$ and $R^c = \{X > 7\}$
II. $R = \{X \leq 10\}$ and $R^c = \{X > 10\}$
I. $R = \{X \leq 14\}$ and $R^c = \{X > 14\}$

(a) I only.
(b) II only.
(c) I and II only.
(d) I and III only.
(e) II and III only.

Solution: (b)
For both I and II, there are 3 misclassifications total whereas there are 2 for II.

Problem 2.7. (5 points) Consider the following statements involving classification trees.

I. The use of Gini index or cross-entropy may result in split nodes with the same predicted class.
II. Cross-validation can be used to prune the trees.
III. In a single node with two possible classes, the Gini index always exceeds the cross-entropy.

Which of the statements above are true?

(a) I only.
(b) II only.

(c) I and II only.

(d) I and III only.

(e) II and III only.

Solution: (c)
Statement III is incorrect by the graph you were supposed to plot in a homework assignment.