Name:

UTeid:

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

Notes: This is a closed book and closed notes exam. The maximal score on this exam is 100 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.

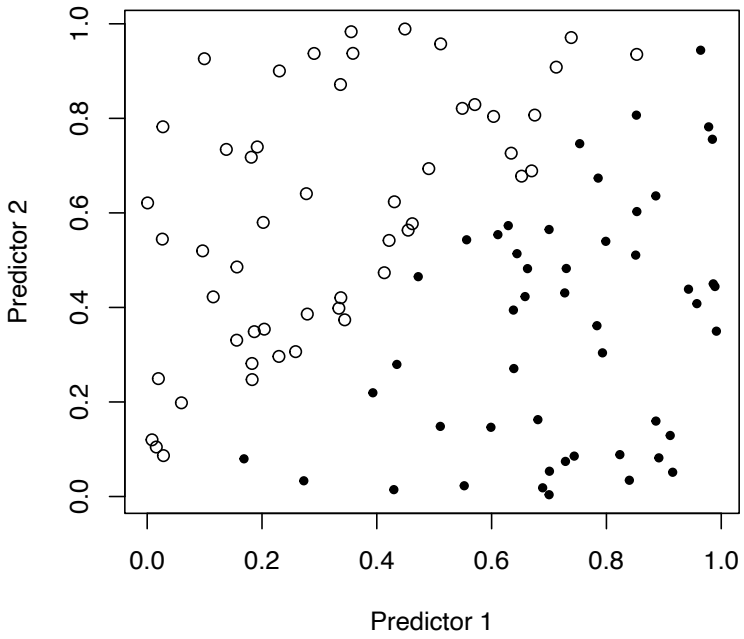
Problem 2.1. (10 points) What are the differences and similarities among bagging, random forests, and boosting in the context of decision trees?

Solution: Solutions will vary. The salient point of any response which is to earn credit must be that bagging is a bootstrap method with while boosting produces a sequence of trees where each new tree depends on the previous tree (the algorithm *slowly learns*).

Problem 2.2. (10 points) Please outline a set of assumptions under which QDA and naive Bayes coincide.

Solution: The distribution over predictors is Gaussian with a diagonal variance-covariance matrix.

Problem 2.3. (10 points) You encounter a data set with two predictors and a binary response. The following is a scatterplot of the training set with the two classes indicated by two types of points.



Would you choose linear discriminant analysis or a decision tree to classify these points? Why? Which other classification method do you think would be successful?

Solution: One would use linear discriminant analysis (recall this lecture: `https://mcudina.github.io/page/M339G/slides/ch8-single-tree-epilogue.pdf`). Other potentially successful classification methods would be logistic regression or KNN.

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 2.4. (15 points) The extremely popular *Fit Fathers* exercise club team enrolled in a fitness regimen. Here is a summary of their pre- and post-regimen weights (in lbs).

	mean	standard deviation
PRE	220	20
POST	185	10

Assume that the data are modeled as bivariate normal with the correlation coefficient equal to 0.8. Of the fathers who weighed **exactly** 260 pounds before the fitness regimen, about what percentage weighed above average after the regimen?

Solution: Let (U, V) be the random pair which stands for the fathers' pre- and post-regimen weights in lbs. Let (X, Y) be the random pair which stands for their weights in standard units.

Conditioning on U being **exactly** 260 is equivalent to conditioning on

$$X = \frac{260 - 220}{20} = 2$$

So, the probability that we are looking for is

$$\mathbb{P}[V > 185 \,|\, U = 260] = \mathbb{P}\left[\frac{V - 185}{10} > \frac{185 - 185}{10} \,|\, U = 260\right] = \mathbb{P}[Y > 0 \,|\, X = -2].$$

Recall that

$$Y \,|\, X = x \sim N(mean = \rho x, sd = \sqrt{1 - \rho^2})$$

. So, the probability we are seeking equals (with $x = -2$)

$$\mathbb{P}\left[Z > \frac{0 - \rho x}{\sqrt{1 - \rho^2}}\right] = \mathbb{P}\left[Z > \frac{0 - 0.8(2)}{\sqrt{1 - 0.64}} \approx -2.67\right] \approx \mathbb{P}[Z \leq 2.67] = \Phi(2.67) = 0.9963.$$

Problem 2.5. (20 points) Consider the following observations of (X, Y) with X being the predictor and Y being the response:

$$(0, 8), \quad (2, 6), \quad (4, 5), \quad (6, 7).$$

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 $(0, 8)$ is the sole element in R, the mean response for the remaining points is

$$\frac{6 + 5 + 7}{3} = \frac{18}{3} = 6.$$

The RSS is

$$(6 - 6)^2 + (5 - 6)^2 + (7 - 6)^2 = 2\,.$$

If $(0, 8)$ and $(2, 6)$ form the first region, the mean response in that region is $\frac{14}{2} = 7$. The remaining points $(4, 5)$ and $(6, 7)$ are in the other region and their mean response is $\frac{12}{2} = 6$. So, the RSS is

$$(8 - 7)^2 + (6 - 7)^2 + (5 - 6)^2 + (7 - 6)^2 = 4\,.$$

If $(0, 8), (2, 6)$, and $(4, 5)$ are in the first region and only $(6, 7)$ remains in the other region, then the average of the first region's values of the response variable is

$$\frac{8 + 6 + 5}{3} = \frac{19}{3}.$$

So, the RSS equals

$$\left(8 - \frac{19}{3}\right)^2 + \left(6 - \frac{19}{3}\right)^2 + \left(5 - \frac{19}{3}\right)^2 = \frac{14}{3}.$$

Overall, the smallest RSS corresponds to the first partition with $(0, 8)$ in its own region, and the remaining points in the other region.

Problem 2.6. (10 points)A classification tree is constructed to predict whether a student will pass *Predictive Analytics*. There are two categorical predictors:

- X_1 indicating whether the student passed *Linear Algebra* prior to enrolling in *Predictive Analytics*, and
- X_2 indicating whether the student passed *Mathematical Statistics* prior to enrolling in *Predictive Analytics*.

Here is the table from a data set of students:

X_1	X_2	outcome
0	0	5 passed, 20 didn't
1	0	10 passed, 10 didn't
0	1	15 passed, 20 didn't
1	1	15 passed, 5 didn't

What's the first split made using the Gini index? Be careful to calculate the **weighted** average.

Solution: Since both X_1 and X_2 are binary, we can make the first split in a unique way with respect to X_1 or with respect to X_2. If we split along X_1, we have that

$X_1 = 0$ 20 passed and 40 didn't;
$X_1 = 1$ 25 passed and 15 didn't.

So, the Gini index is

$$\frac{60}{100} \cdot 2 \cdot \left(\frac{20}{60}\left(1 - \frac{20}{60}\right)\right) + \frac{40}{100} \cdot 2 \cdot \left(\frac{25}{40}\left(1 - \frac{25}{40}\right)\right) = 0.4541667$$

If we split along X_2, we have that

$X_2 = 0$ 15 passed and 30 didn't;
$X_2 = 1$ 30 passed and 25 didn't.

So, the Gini index is

$$\frac{45}{100} \cdot 2 \cdot \left(\frac{15}{45}\left(1 - \frac{15}{45}\right)\right) + \frac{55}{100} \cdot 2 \cdot \left(\frac{30}{55}\left(1 - \frac{30}{55}\right)\right) = 0.4727273$$

Since the first Gini index is smaller, we should split along X_1.

2.3. **MULTIPLE CHOICE QUESTIONS.**

Problem 2.7. (5 points) You encounter a data set with two categorical predictors, two numerical predictors, and a categorical response. Which of the following methods would be applicable in this situation? **Circle ALL that apply!**

(a) $K-$means clustering.

(b) Regression trees.

(c) Multiclass logistic regression.

(d) Classification trees.

(e) None of the above.

Solution: (c) and (d)

Problem 2.8. (5 points)Which of the following statements about *boosting* is true?

I. Boosting always uses trees with either one or two splits only.

II. Boosting involves an incremental approach where the subsequent trees are based on the prior trees.

III. Boosting is a general approach that can be applied to many statistical learning methods for regression or classification.

(a) None.

(b) I and II only.

(c) I and III only.

(d) II and III only.

(e) The correct answer is not given above.

Solution: (d)
See subsection 8.2.3 in the textbook.

Problem 2.9. (5 points) Which of the following statements about *pruning* is true?

I. In *cost complexity pruning* the optimal tuning parameter λ is determined through cross-validation.

II. *Pruning* is a special case of bootstrap applied to decision trees.

III. *Pruning* always reduces the tree to at most 3 terminal nodes.

(a) None.

(b) I only.

(c) II only.

(d) III only.

(e) The correct answer is not given above.

Solution: (b)

Problem 2.10. (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	N	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 $R^c = \{X > 10\}$
III. $R = \{X \leq 14\}$ 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 I, there are 2 misclassifications total; for II, there is 1 misclassifiation; for III, there are 2 misclassifications.

Problem 2.11. (5 points) Consider the following statements involving classification trees.

I. We use cross-entropy to assess the fit only in the case of a binary response.
II. In a single terminal node with two possible classes A and B, cross-entropy can be expressed as a function of a single variable p denoting the proportion of points in class A at that node.
III. In a single terminal node with two possible classes A and B, cross-entropy is an increasing function of p described above.

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: (b)
Statement I is incorrect since cross-entropy is well-defined for response variables with multiple classes. Statement II is **correct** by the graph you were supposed to plot in a homework assignment. Statement III is incorrect by the graph you were supposed to plot in a homework assignment.