M339G Predictive Analytics
University of Texas at Austin
Practice Problems for 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. **There are many ways in which any single problem can be solved. The solutions herein are just one possible way to tackle the given problems.**

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 is the difference between bagging and random forests? Which one is preferable and why?

Problem 2.2. (10 points) What is the crucial difference in the assumptions between LDA and QDA? What consequence does that change have on the classification criterion?

Problem 2.3. (10 points) Explain how naive Bayes is different from LDA. Provide one advantage which stems from this different approach. Are there disadvantages?

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. (10 points) *Source: Pitman's "Probability."*
Heights and weights of a large group of people follow a bivariate normal distribution, with correlation 0.75. Of the people **at** the 90^{th} percentile of weights, about what percentage are **above** the 90^{th} percentile of heights?

Problem 2.5. (15 points) *Source: An old SRM manual.*
For a regression tree, two nodes in the tree - defining regions R_1 and R_2 - have the following values of the response variable (on the training set):

$$R_1 : 2, 3, 3, 4, 5$$
$$R_2 : 2, 4, 6$$

The tree is pruned using cost complexity pruning. The split creating R_4 and R_5 is the optimal one to prune. Determine the smallest value of α for which this pruning will occur.

Problem 2.6. (10 points) Consider the following data set with the explanatory random variable X and the categorical response Y:

X	1	2	2	3	3	4	6	10	12	12
Y	G	B	G	G	B	B	M	B	M	M

The data are split according to the partition by $X < 6$ and $X \geq 6$. What is the total (weighted) cross-entropy, Gini index, and classification error.

Problem 2.7. (5 points) Let L be a line in $\mathbb{R}^2$ through the point $\vec{p} = (2, 3)$ in the direction $\vec{v} = (-1, 2)$. Provide an example of a normal vector of this line, write down the normal equation, and the standard equation of the form $\beta_0 + \beta_1 x + \beta_2 y = 0$ for the line L. Give an example of a point which is on one side of the line and another which is on the other side of the line. Prove **algebraically** that they are, indeed, on opposite sides of the line.

Problem 2.8. (15 points) Find the minimum value of the function

$$f(x, y, z) = x^2 + 2y^2 + z^2$$

subject to constraints

$$x + 2y + 3z = 1$$
$$x - 2y + z = 5$$

2.3. MULTIPLE CHOICE QUESTIONS.

Problem 2.9. (5 points) *Source: SRM Sample Problem #29.*
Determine which of the following considerations may make decision trees preferable to other statistical methods.

I. Decision trees are easily interpretable.
II. Decision trees can be displayed graphically.
III. Decision trees are easier to explain than linear regression models.

(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.

Problem 2.10. (5 points) Which of the following statements about *boosting* is true?

I. Boosting has three tuning parameters.
II. Boosting involves creating multiple copies of the original training data set using the bootstrap.
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.

Remark 2.1. The above problems are **in addition** to your past homework assignments. Do not forget to re-solve those!