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?

**Solution:** Solutions will vary. The salient point of any response which is to earn credit must be that random forests always generate trees with a smaller randomly chosen number of predictors. Thus, correlation between trees is reduced.

**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?

**Solution:** Solutions will vary. The salient point of any response which is to earn credit must be that while LDA assumes the same standard deviation for each subpopulation, the QDA does not. Thus, a quadratic term shows up in the discriminant for the QDA case as opposed to the linear expressions in the predictors in the LDA case.

**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?

**Solution:** Solutions will vary. The salient point of any response which is to earn credit must be that naive Bayes **assumes** independence between predictors and **does not** in general assume a Gaussian distribution. In particular, naive Bayes can handle categorical predictors.

---

## 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?

**Solution:** Let $(U, V)$ be the random pair which stands for the people's weights and heights in real units. Let $(X, Y)$ be the random pair which stands for the people's weights and heights in standard units. Conditioning on $U$ is **at** the $90^{th}$ percentile is (close to) equivalent to conditioning on $\{X = 1.28\}$. So, the probability that we are looking for is

$$\mathbb{P}[Y > 1.28 \,|\, X = 1.28].$$

Recall that $Y \,|\, X = x \sim N(\rho x, 1 - \rho^2)$. So, the probability we are seeking equals (with $x = 1.28$)

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

With the given correlation coefficient of 0.75, our answer is

$$1 - \Phi\left(1.28\sqrt{\frac{1-0.75}{1+0.75}}\right) = 0.3142659$$

**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 $R_1$ and $R_2$ is the optimal one to prune. Determine the smallest value of $\alpha$ for which this pruning will occur.

**Solution:** The mean of the values in $R_1$ is 3.4 and the average of values in $R_2$ is 4. Without the pruning, the total contribution to the RSS from these two nodes is

$$(2-3.4)^2 + 2(3-3.4)^2 + (4-3.4)^2 + (5-3.4)^2 + (2-4)^2 + (6-4)^2 = 13.2.$$

If the pruning happens, all the response values will end up in the same terminal node whose mean will be 3.625. That node's contribution to the RSS is

$$2(2-3.625)^2 + 2(3-3.625)^2 + 2(4-3.625)^2 + (5-3.625)^2 + (6-3.625)^2 = 13.875.$$

For any $\alpha \geq 13.875 - 13.2 = 0.675$, the 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.

**Solution:** The region where $X < 6$ has weight 0.6. So, the region where $X \geq 6$ has the weight 0.4. For the Gini index, we get

$$0.6(2)(0.5)(0.5) + 0.4(2)(0.25)(0.75) = 0.45$$

For the cross-entropy, we get

$$-(0.6(2)(0.5\ln(0.5) + 0.4(0.25\ln(0.25) + 0.75\ln(0.75)) = 0.6408224$$

Classification error: $\frac{3+1}{10} = 0.4$.

**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.

**Solution:** An easily found normal vectors is $\vec{n} = (2, 1)$ as

$$\vec{n} \cdot \vec{v} = 2(-1) + 1(2) = 0.$$

Since we know that $\vec{p}$ is on the line $L$, we have the following normal equation

$$(2, 1) \cdot (x-2, y-3) = 0 \quad \Rightarrow \quad 2(x-2) + (y-3) = 0.$$

Finally, we get the equation $2x + y - 7 = 0$.

The origin $(0,0)$ is "below" the line as $2(0)+(0)-7=-7<0$. The point $(10,20)$ is "above" the line as $2(10)+(20)-7=33>0$.

**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

$$\begin{aligned} x+2y+3z&=1\\ x-2y+z&=5 \end{aligned}$$

**Solution:** The Lagrangian is

$$x^2+2y^2+z^2+\lambda_1(x+2y+3z-1)+\lambda_2(x-2y+z-5).$$

The partial derivatives are

$$\begin{aligned} \frac{\partial L}{\partial x}&=2x+\lambda_1+\lambda_2\\ \frac{\partial L}{\partial y}&=4y+2\lambda_1-2\lambda_2\\ \frac{\partial L}{\partial z}&=2z+3\lambda_1+\lambda_2\\ \frac{\partial L}{\partial \lambda_1}&=x+2y+3z-1\\ \frac{\partial L}{\partial \lambda_2}&=x-2y+z-5 \end{aligned}$$

Equating the above to zero, we obtain the following system of equations:

$$\begin{aligned} 2x+\lambda_1+\lambda_2&=0\\ 2y+\lambda_1-\lambda_2&=0\\ 2z+3\lambda_1+\lambda_2&=0\\ x+2y+3z&=1\\ x-2y+z&=5 \end{aligned}$$

From the first equation, we get

$$x=-\frac{\lambda_1+\lambda_2}{2}.$$

From the second equation, we obtain

$$y=\frac{\lambda_2-\lambda_1}{2}.$$

The third equation yields

$$z=-\frac{3\lambda_1+\lambda_2}{2}$$

If we substitute these expressions for $x, y$ and $z$ into the first constraint equation, we get

$$-\frac{\lambda_1+\lambda_2}{2}+(\lambda_2-\lambda_1)-\frac{9\lambda_1+3\lambda_2}{2}=1 \quad\Rightarrow\quad -6\lambda_1-\lambda_2=1.$$

Substituting the same expressions into the second constraint, we obtain

$$-\frac{\lambda_1+\lambda_2}{2}-(\lambda_2-\lambda_1)-\frac{3\lambda_1+\lambda_2}{2}=5 \quad \Rightarrow \quad -\lambda_1-2\lambda_2=5$$

Solving the latest system of two equations with two unknowns, we get

$$\lambda_1=\frac{3}{11} \quad \text{and} \quad \lambda_2=-\frac{29}{11}\,.$$

Now, it's easy to get

$$x=\frac{13}{11}, \quad y=-\frac{16}{11}, \quad z=\frac{10}{11}\,.$$

The function's minimal value is

$$f\left(\frac{13}{11},-\frac{16}{11},\frac{10}{11}\right)=\frac{71}{11}=6.454545.$$

---

## 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.

**Solution: (e)**
All three statements are true.

**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.

**Solution: (c)**
See subsection 8.2.3 in the textbook.

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