M339G Predictive Analytics
University of Texas at Austin
Practice Problems for In-Term Exam I
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:

1.1. DEFINITIONS.

Problem 1.1. (5 points) Provide the definition of *bias*.

Solution: See the solutions to the first homework assignment.

Problem 1.2. (5 points) Provide the definition of the *mean-squared error* in the context of parameter estimation.

Solution: See the solutions to the first homework assignment.

1.2. CONCEPTUAL QUESTIONS.

Problem 1.3. (10 points) Explain why cross-validation is not a suitable tool to assess the quality of an implementation of principal components analysis.

Solution: Solutions will vary. The salient point of any response which is to earn credit must be that cross-validation does not apply in an unsupervised-learning setting.

Problem 1.4. (10 points) Explain why standardization is an appropriate preliminary step before we employ the $K-$nearest neighbors procedure.

Solution: Solutions will vary. The salient point of any response which is to earn credit must be that $K-$nearest neighbors employs a distance which is heavily affected by the natural scale of various predictors.

1.3. 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 1.5. (10 points) Consider a simple linear regression fitted on 20 observations. In our usual notation, you are given the following:

- $\bar{x} = 5$
- $\sum x_i^2 = 500$
- $\bar{y} = 4$
- $\sum y_i^2 = 1024$
- $\sum x_i y_i = 500.$
- $RSS = 256$

Find the coefficient of determination.

Solution: The total sum of squares is

$$TSS = \sum (y - y_i)^2 = \sum y_i^2 - 20\bar{y} = 1024 - 20(4)^2 = 704.$$

Thus,

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{256}{704} = \frac{448}{704}\,.$$

Problem 1.6. (15 points) Consider the following multiple linear regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

You fit the above regression on 25 data points. In our usual notation, you are given that:

- $\sum (y_i - \hat{y}_i)^2 = 40$
- $\sum (y_i - \bar{y})^2 = 60$

Find the value of the F statistic.

Solution: We are given that

$$TSS = 60 \quad \text{and} \quad RSS = 40.$$

With p denoting the number of predictors, the F statistic is

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{(60-40)/3}{40/(25-3-1)} = \frac{20/3}{40/21} = \frac{7}{2}\,.$$

Problem 1.7. (10 points) You are using $K-$nearest neighbors in a classification problem with X as the explanatory variable and Y as the response. Here are the observed values:

x	1	2	5.5	6.5	9	13
y	1	2	1	1	2	2

Using $K = 3$, figure out the misclassification error.

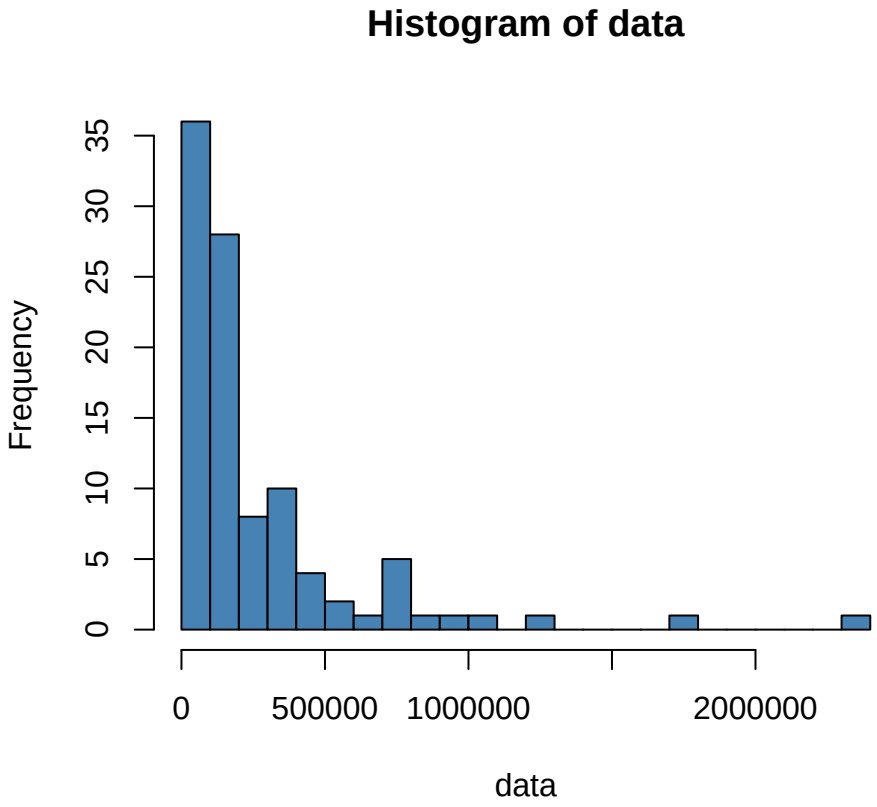
Solution: With the neighbourhood of size 3, we get the following predictions:

x	1	2	5.5	6.5	9	13
$\hat{y}$	1	1	1	1	2	2

We did fine: the misclassification error rate is 1/6.

1.4. **MULTIPLE CHOICE QUESTIONS.**

Problem 1.8. (5 points) You have a sample of size 12 from a distribution that you know from past experience looks like this:



Your task is to estimate its mean. Of the following, what is the **best** choice of a procedure in this case?

(a) A 95% bootstrap confidence interval using quantiles.

(b) Using the 't.test' command in **R**.

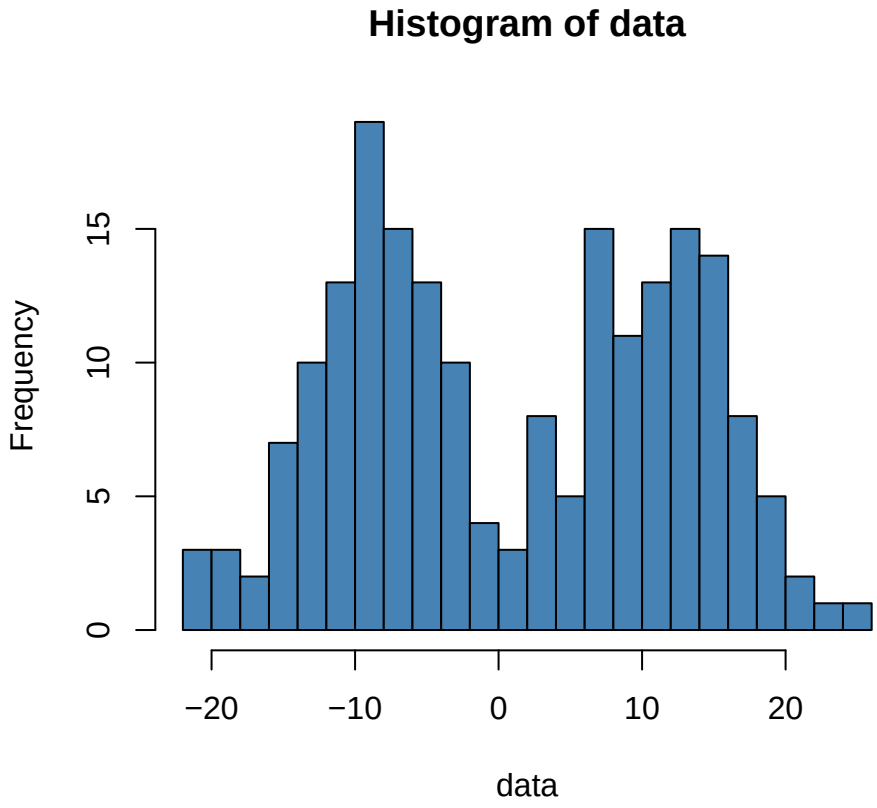
(c) A $2SE$ bootstrap confidence interval.

(d) The standard $z-$procedure 95%-confidence interval.

(e) An arithmetic average of the observations.

Solution: (a)

Problem 1.9. (5 points) You have a sample of size 20 from a distribution that you know from past experience looks like this:



Your task is to estimate its mean. Of the following, what is the **best** choice of a procedure in this case?

(a) The standard $z-$ procedure for confidence intervals.

(b) A bootstrap 95%-confidence interval.

(c) Using the 't.test' command in **R**.

(d) An arithmetic average of the observations.

(e) The median of the observations.

Solution: (b)

Problem 1.10. (5 points) Here is an example of a problem by the *Advanced Research Group* at UCLA.

People's occupational choices might be influenced by their parents' occupations and their own education level. We can study the relationship of one's occupation choice with education level and the parents' occupations. The occupational choices will be the outcome variable which consists of categories of occupations.

Which of the following procedures is the most applicable in this case?

(a) Multiclass logistic regression.

(b) Simple linear regression.

(c) Principal component analysis.

(d) $K-$means clustering.

(e) None of the above techniques apply in this case.

Solution: ()

Problem 1.11. (5 points) *Source: Sample SRM problems.*
An analyst is modeling the probability of a certain phenomenon occurring. The analyst has observed that the simple linear model currently in use results in predicted values less than zero and greater than one. Which of the following is the most appropriate to address the issue?

(a) Limit the data to observations that are expected to result in predicted values between 0 and 1.

(b) Consider predicted values below 0 as 0 and values above 1 as 1.

(c) Use a logit function to transform the linear model into only predicting values between 0 and 1.

(d) Apply the arctan function to transform the linear model into only predicting values between 0 and 1.

(e) None of the above.

Solution: (c)

Problem 1.12. (5 points) *Source: MAS exam Spring 2019.*
You are reviewing a data set with 100 observations in four variables: X_1, X_2, X_3, X_4. You analyze these data using two principal components:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \varphi_{31}X_3 + \varphi_{41}X_4$$
$$Z_2 = \varphi_{12}X_1 + \varphi_{22}X_2 + \varphi_{32}X_3 + \varphi_{42}X_4$$

Which of these statements is always **true**?

I $\sum_{i=1}^{100}\left(\sum_{j=1}^{4}\varphi_{j1}x_{ij}\right)^2 = \sum_{i=1}^{100}\left(\sum_{j=1}^{4}\varphi_{j2}x_{ij}\right)^2$

II $\sum_{j=1}^{4}\varphi_{j1}\varphi_{j2} = 0$

III $\sum_{j=1}^{4}\varphi_{j1}^2 + \sum_{j=1}^{4}\varphi_{j2}^2 = 1$

(a) I only

(b) II only

(c) III only

(d) I, II, and III

(e) None of the above.

Solution: (b)

1.5. **SUGGESTED TEXTBOOK PROBLEMS. 2.4.2 (p.52), 2.4.4 (p.53), 2.4.7 (p.54), 3.7.4 (p.122), 4.8.1 (p.189), 4.8.6, 4.8.8, 4.8.9, 4.8.12 (p.191), 5.4.1, 5.4.2 (a-g) (p.219), 5.4.4 (p.220), 6.6.1 (p.283), 7.9.3, 7.9.4 (p.323)**

Remark 1.1. The above problems are **in addition** to your past homework assignments. Do not forget to re-solve those!