Midterm Exam

IE 8623 Fall 2021

Instructions (Read carefully before you start)

- You can only communicate about the exam problems with your instructor. **Don't discuss about the questions with anyone else.**
- There are three parts in the exam. The respective point distribution has been provided for each question.
- You are expected to use R markdown or R notebook to generate a pdf file with all your answers with your codes nested as chucks inside the file. Failing to use the expected format will result in penalty.
- All your codes should be available upon request.
- Unless otherwise noted, your exam should be submitted by 11:59pm Oct 17, 2021. **No late exams will be accepted.** Please plan ahead and submit your exam through canvas on time.

PART 1: REGRESSION (15 PTS)

The code to retrieve the dataset is as follows,

```
data("state")

statdata<-data.frame(state.x77,row.names=state.abb)

summary(statdata)
```

The data were collected from US Bureas of the Census on the 50 states from the 1970s. Use “statdata” to perform regression analysis. More information about the variables are available from the help file of R. We will take **life expectancy (Life.Exp)** as the response and the remaining variables as predictors. Use all the rows except for the row of MS as our training set, and the row of MS as our testing set.

1.1 (3 pts) Fit a linear regression model using the training set based on all the predictors. Interpret all the coefficients you estimated.

1.2 (1 pts) What is the R^2 value of the model 1.1? How to interpret this value?

1.3 (3 pts) Calculate the point estimate and 95% confidence interval (CI) for the mean life expectancy for MS.

1.4 (8 pts) Look at the model in 1.1, list at least THREE different models that can potentially improve the performance. Explain your rationality behind your selection, and train those three models. Do they actually perform better than the model obtained in 1.1? Explain possible reasons.

PART 2: CLASSIFICATION I (7 PTS)

The code to retrieve the data is as follows.

```
install.packages("mlbench")

library(mlbench)

data("PimaIndiansDiabetes2")
```

Take diabetes as the response, and all the other variables as predictors. (To simply the problem, you may remove the rows with NA using the function “na.omit”). Split the dataset to 80% training and 20% testing sets (set the seed as 100).

2.1 (3 pts) Fit a linear SVM model based on the training set (use function tune.svm to find the best C parameter). Evaluate its classification performance using the testing set. List the Type I and Type II errors, respectively.

2.2 (4 pts) Fit a nonlinear SVM model based on the training set (use function tune.svm to find the best C parameter). Compare the Type I and Type II errors with the ones you obtained from 2.1. Which one is better? Why?

PART 3. CLASSIFICATION II (8 PTS + BONUS 4 PTS)

3.1 (2 pts) Use mvrnorm function to generate the training data set below (use seed(100)):

a) Generate three groups of two-dimensional data (50 rows in each group) with their mean as $\mu_1 = [0,0]^T$, $\mu_2 = [0,3]^T$, $\mu_3 = [1.5,1.5]^T$, and the common covariance matrix $\sum = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

b) X is obtained by concatenating the three groups together (150 samples of two-dimensional input)

c) Y is a 150*1 vector with the first 50 elements equal to 1, the second 50 elements equal to 0, and the third 50 elements equal to 1.

3.2 (2 pts) Fit a logistic regression model based on the training data generated in 3.1.

3.3 (2 pts) Fit a linear discriminant analysis model based on the training data generated in 3.1.

3.4 (2 pts) If we know the input variables come from a normal distribution, theoretically which model should perform better, logistic regression or LDA? Why?

3.4 (Bonus 4 pts) Derive all the decision boundaries for the models you fit in 3.1 and 3.2, respectively.