STAT 6543 Midterm Examination

Due at 11:59 pm Central Time on July 11 2024.

Name: ________________

Note: Please PRINT your name and ABC123 number above. You may use your computer/laptop to access the class notes and textbook, but you are NOT allowed to use the internet to Google and/or search for answers to each problem. Any violation of Academic Conduct (including looking at others' work) will result in a 0 on the exam and subsequent action. This exam must be submitted by July 11, 2024, at 11:59 PM Central Time. No submissions will be accepted after midnight. You need to submit a PDF, HTML, or Word file, and your R code or RMD in a separate file to Canvas. Other formats such as QMD may NOT be graded if I am unable to open the file or see your plots and results, which could support your claims. It is your responsibility to run your own code to obtain and provide detailed outputs and graphs if necessary. Resubmission is NOT allowed. The submission link is available in the Midterm folder. You will have three days to complete the exam. If you have any questions or concerns, please contact me as soon as possible. **Good luck!**

I True (T) or False (F). (20 Points)

For these problems, **if false, briefly justify your answer. You will receive partial credits if you fail to justify your answer**. Each problem worth 2 points.

1. _____ We can divide learning problems into supervised and unsupervised situations. Linear regression, Principal component regression, and Lasso regression are all supervised learning approaches.

2. _____ When we fit the linear regression model, the collinearity between predictors will improve the coefficient estimates.

3. _____ R^2 can go from -1 to 1, where -1 means no variation in response Y has been explained by X, and 1 means the variation has all been explained.

4. _____ In linear regression problems, there is always a unique least squares coefficient estimate for each predictor.

5. _____ Principal components analysis (PCA) is an unsupervised learning approach, because it does not make use of the response Y.

6. _____ All types of statistical models discussed in this course are beneficial from data pre-processing.

7. _____ One advantage of Principal component analysis (PCA) is that it is a data reduction technique which creates uncorrelated components.

8. _____ The trade-off between prediction accuracy and interpretability means that a predictive model that is most powerful is usually the least interpretable.

9. _____ Elastic net, OLS, Ridge regression, Lasso regression can all be used and implemented in situations where the number of predictors is larger than the sample size.

10. _____ The last name of the instructor of this course is Min.

II Free Response Questions (25 Points)

Problem 1 (Total: 15 Points)
Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry, and the CEO salary. We are interested in understanding which factors affect CEO salary. (5 Points)

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables. (5 Points)

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market. (5 Points)

Problem 2 (Total: 10 Points)
In this class, we discussed the bias-variance trade-off. Answer the following questions.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be four curves. Make sure to label each one.

(b) Briefly explain why each of the four curves has the shape displayed in part (a)

III Coding Questions (55 Points)

Problem 3 (Total: 18 Points - 3 points each)
This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data. We may start R and use these commands to load the data:

```
library(ISLR)
data("Auto")
Auto
```

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

(c) What is the mean and standard deviation of each quantitative predictor?

(d) Now remove the 20th through 80th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

Problem 4 (Total: 15 Points)
we will predict the number of applications received using the other variables in the College data set available in the R package **ISLR**, which can be accessed as follows.

```
library(ISLR)
data(College)
#data basic information
head(College)
dim(College)
# The column Apps is the response variable, and others may be treated as predictors.
#For instance, for linear regression model in R, you may use
lm(Apps~.,data=College)
```

(a) Appropriately split the data set into a training set (80%) and a test set (20%). [3 points]

(b) Fit a support vector machine (SVM) model with the radial basis function function model using least squares on the training set. Clearly report the test error obtained. [3 points]

(c) Fit a neural network model on the training set by creating an appropriate grid for tuning parameters. Clearly report the test error obtained. [3 points]

(d) Fit a Multivariate Adaptive Regression Splines (MARS) model on the training set with tuning parameters chosen by cross-validation. Clearly report the test error obtained, along with the importance of each predictor. estimates. [3 points]

(e) Comment on the results obtained. Is there much difference among the test errors resulting from these three nonparametric approaches? [3 points]

Problem 5 (Total: 22 Points)
A chemical manufacturing process for a pharmaceutical product was discussed in Sect. 1.4 of the textbook. In this problem, the objective is to understand the relationship between biological measurements of the raw materials (predictors), measurements of the manufacturing process (predictors), and the response of product yield. Biological predictors cannot be changed but can be used to assess the quality of the raw material before processing. On the other hand, manufacturing process predictors can be changed in the manufacturing process. Improving product yield by 1% will boost revenue by approximately one hundred thousand dollars per batch. We may start R and use these commands to load the data:

```
library(AppliedPredictiveModeling)
data("ChemicalManufacturingProcess")
```

The matrix processPredictors contains the 57 predictors (12 describing the input biological material and 45 describing the process predictors) for the 176 manufacturing runs. yield contains the percent yield for each run.

(a) A small percentage of cells in the predictor set contain missing values. Use an appropriate imputation function to fill in these missing values. [3 points]

(b) Split the data into a training and a test set, pre-process the data, and build at least four different models from Chapter 6. For those models with tuning parameters (e.g., ENET), what are the optimal values of the tuning parameter(s)? [8 points]

(c) Which model has the best predictive ability? Is any model significantly better or worse than the others? You need to conduct a hypothesis testing to justify your choice if necessary. [5 points]

(d) Which predictors are most important in the model you have trained? Do either the biological or process predictors dominate the list [3 points]

(e) Explore the relationships between each of the top predictors and the response. How could this information be helpful in improving yield in future runs of the manufacturing process? [3 points]