STAT 6543 Sample Exam

July 03, 2024

Name: _____

ABC123: _____

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 answers for each problem. Any violation of Academic Conduct (including looking other people's work) will result a 0 on the exam and subsequential treatment. This exam will be due on July 11, 2024 by 11:59 pm central time, and no submission 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 the Canvas. The submission link is available in the Midterm folder. You will have had 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**. Each problem worth 2 points.

1. _____ In general the more flexible a method is, the lower its RMSE of the test data will be.

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

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

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

5. _____ The bias-variance trade-off means that as a method gets more flexible the bias will decrease and the variance will increase but expected RMSE of the testing data may go up or down.

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

7. _____ When the sample size n is extremely large, and the number of predictors p is small, we do not expect the performance of a flexible statistical learning method to be better than an inflexible method.

8. ____ 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.

9. ____ The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator. Each "bootstrap set" is created by sampling without replacement, and the size is smaller than our original dataset.

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

II Free Response Questions (40 Points)

Problem 1 (Total: 10 Points)
You think of some real-life applications for statistical learning and predictive modelling.

(a) Describe a real-life application in which classification might be useful. Describe the response, as well as the predictors. Is the goal of this application inference or prediction? Clearly explain your answer. (5 Points)

(b) Describe a real-life application in which regression might be useful. Describe the response, as well as the predictors. Is the goal of this application inference or prediction? Clearly explain your answer. (5 Points)

Problem 2 (Total: 10 Points)
During the class time, we learned k-fold cross-validation.

(a) (5 points) Explain how k-fold cross-validation is implemented.

(b) (5 points) What are the advantages and disadvantages of k-fold cross-validation relative to the bootstrap sample.

Problem 3 (Total: 10 Points)
What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Problem 4 (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 (40 Points)

Problem 5 (Total: 16 Points)
Suppose we are interested in examining the relationship between the response variable sales and the amount of money spent advertising on the TV, radio, and newspapers (i.e, there are three predictors: TV, radio, and newspapers). We fit a multiple linear regression with four predictors (TV, radio, newspaper, and the TV and radio interaction term, denoted by TV:radio) and obtain the following results:

```
> summary(fit)

Call:
lm(formula = sales ~ TV + radio + newspaper + radio * TV, data = dat

Residuals:
   Min     1Q Median     3Q    Max
-6.293 -0.398  0.181  0.596  1.501

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.73e+00   2.53e-01   26.56   <2e-16 ***
TV          1.91e-02   1.51e-03   12.63   <2e-16 ***
radio       2.80e-02   9.14e-03    3.06   0.0025 **
newspaper   1.44e-03   3.30e-03    0.44   0.6617
TV:radio    1.09e-03   5.26e-05   20.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.945 on 195 degrees of freedom
Multiple R-squared:  0.968,     Adjusted R-squared:  0.967
F-statistic: 1.47e+03 on 4 and 195 DF,  p-value: <2e-16
```

(a) (4 points) Provide an appropriate interpretation for the coefficient 1.91e-02.

(b) (4 points) True or false: Since the coefficient for the TV and radio interaction term "TV:radio" is quite small, there is very little evidence that this interaction term is important in predicting the response variable "sales". Justify your answer.

(c) (4 points) Suppose that the company has two options to split $\$12,000$ for the three types of advertising: (i) invest equally $\$4,000$ for each type of advertising, (ii)invest $\$6,000$ for TV, $\$3,000$ for radio, and $\$3,000$ for newspapers. Which option should be recommended for the company. Justify your answer.

(d) (4 points) Based on this model fit, which predictors are important in predicting the sales? In other words, explain what conclusions you can draw based on the p-values. Your explanation should be phrased in terms of sales, TV, radio, newspaper, and TV:radio, rather than in terms of the coefficients of the linear model.

Problem 6 (Total: 24 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%). [4 points]

(b) Fit a linear model using least squares on the training set, and report the test error obtained. [5 points]

(c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained. [5 points]

(d) Fit an ENET model on the training set with tuning parameters chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates. [5 points]

(e) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches? [5 points]