Stat 202 **Midterm exam**

SOLUTION

Your name:

You SUNetID (Stanford email handle):

You have 50 minutes to complete the exam. The exam is closed book and you are not allowed to use calculators, cell phones or any other help. Please show your work and justify your answers.

Scores:

1

2

3

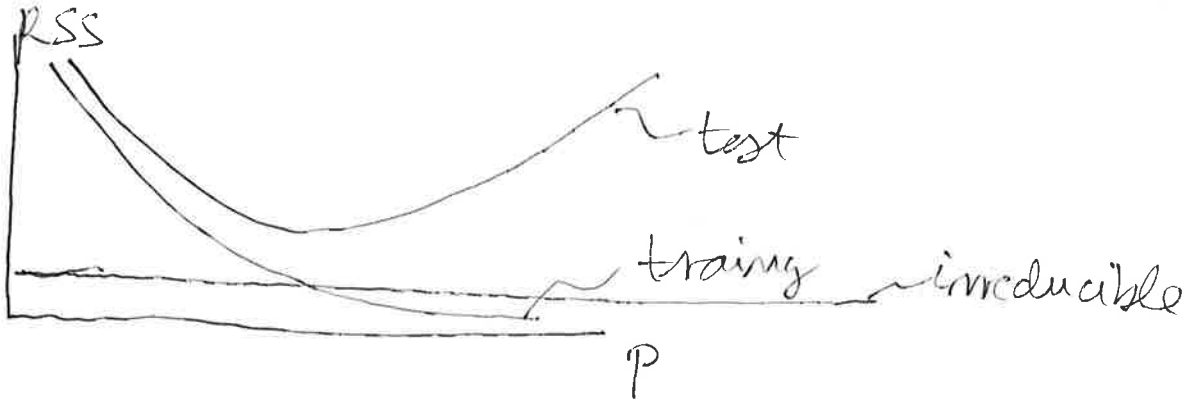
4

5

Total

1. (3+3+3 points)
You want to do a regression and there are 100 predictors available. You run best subset selection to find the best model with $p = 1$ predictor, with $p = 2$ predictors, etc.
(a) Provide a sketch of typical training error, test error, and irreducible error. The horizontal axis will represent p, and the vertical axis will represent RSS.
(b) Explain why the shape of the test error curve is different from that of the training error curve.
(c) Suppose you use forward selection instead of best subset selection. Do you expect that to materially change any of the answers in (a) or (b)? Explain.

a



b The test error curve first decreases as a larger p reduces bias, but then increases as an increasing variance dominates the reduction in bias. The training error keeps decreasing with p as a more flexible model allows a better fit.

c No. The bias-variance trade off will still apply even if we use a greedy optimizer like forward selection. We expect the same conclusion as in b for the training error as a larger p means a more flexible model.

2. (3+3 points)
For each of the following two classification problems, write down whether one should prefer LDA or logistic regression, and explain why:
(a) We want to predict gender based on height and weight. The training set consists of heights and weights for 80 men and 60 women.

(b) We want to predict gender based on annual income and weekly working hours. The training set consists of annual income and weekly working hours for 970 men and 810 women.

a LDA is preferable as heights and weights are normally distributed with the same correlation ~~for men~~ and SDs for men and women.

b logistic regression is preferable as income does not follow the normal curve but has a long right tail.

3. (3+3 points)
For each part (a) and (b), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than that of an inflexible method. Explain briefly.
(a) The sample size is extremely large and the number of predictors is small.
(b) The number of predictors is extemely large, and the number of observations is small.

a A flexible method should be better. It has a smaller bias than an inflexible method and since a large sample size a small number of predictors mean that the variance will be small, we expect a smaller test error (=sum of bias^2 +variance)

b The inflexible method will be better, as now the large variance part of the flexible method is expected to dominate the bias reduction.

4. (8 points)
A venture capital company wants to predict the likelihood that a startup will succeed as a function of the amount of funding it provides. To this end it fits a logistic regression model

$$P(\text{success} = \text{yes}) = \frac{e^{\beta_0+\beta_1 \text{funding}}}{1+e^{\beta_0+\beta_1 \text{funding}}}$$

They find that $\hat{\beta}_1 = 0.08$. Explain verbally how one can find the standard error for $\hat{\beta}_1$ with the bootstrap. (Please don't write code but give a clear explanation of the steps involved. Also give the formula that you would use in order to compute the standard error from the results of the resampling.)

Randomly sample n observations with replacement to obtain a bootstrap data set. Fit the logistic regression on the bootstrap data set to obtain the estimate $\hat{\beta}_1^{(1)}$. Repeat this $B=1000$ (say) times to obtain $\hat{\beta}_1^{(1)}, \ldots \hat{\beta}_1^{(B)}$. The bootstrap SE of $\hat{\beta}_1$ is

$$\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\beta}_1^{(i)} - \bar{\hat{\beta}}_1\right)^2}$$

where $\bar{\hat{\beta}}_1 = \frac{1}{B}\sum_{i=1}^{B}\hat{\beta}_1^{(i)}$

5. (6 points)
The unemployment rate is published monthly by the US Dept. of Labor and is based on survey data. You want to predict the unemployment rate based on other time series about the economy. To this end, you use 'Google Correlate': You feed the monthly unemployment rate for the last 5 years into Google Correlate, and Google Correlate provides a list of other times series that are highly correlated with the unemployment rate.
You take the first 100 of these time series. All of these have a correlation coefficient of more than 0.95 with the unemployment rate. Then you fit a linear model by regressing the unemployment rate on these 100 predictors.
Describe a major problem that will arise when fitting and using this regression model.

Collinearity will be a major problem:

Since all the predictors are highly correlated to the unemployment rate, the will also be highly correlated among themselves.

Collinearity will cause large standard errors for the regression coefficients as these are not uniquely defined.