Stat 202 **Midterm exam**

Your name: SOLUTION
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. (6 points)

In a regression setting, explain what 'collinearity' means and why it is a problem.

Collinearity refers to the situation where two or more predictors are highly correlated.

It is a problem because it causes large standard errors for parameter estimates and some computations, such as inverting matrices, become difficult or impossible.

It also makes it difficult to separate out the individual effects of collinear predictors on the response.

2. (5+5 points)

True or False, and explain briefly:
(a) Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary.

(b) If the Bayes decision boundary for a given problem is nonlinear, then we will achieve a superior test error rate using QDA rather than LDA.

a: False: QDA is more flexible than LDA and hence has more variance, which can result in a larger test error rate (e.g. in high-dimensional space)

b False: While QDA may have a smaller bias than LDA, its variance is larger and it's not clear that the bias-variance trade off goes in its favor. E.g. in high dimensional space the variance penalty may be larger than the gain in bias.

3. (2+8 points)

(a) What is the purpose of applying principal components analysis to a data set?

(b) The book discusses the USArrests data set, which for each of the 50 states contains the number of arrests per 100,000 residents for each of three crimes: assault, murder, and rape. In this data set, the proportion of variance explained by the first two principal components is 87%. Explain how one can find the standard error for this number with the bootstrap. (That is, list the various steps that one has to do in the bootstrap procedure in this case.)

a When there is a large set of correlated variables, then PCA allows to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set.

b Randomly sample 50 states with replacement from the list of 50 states. Apply PCA to the sample and compute the proportion of variance explained. Repeat all of this many times (say $B = 1{,}000$ times) to get B such proportions, call them $\hat{p}^{(1)}, \ldots, \hat{p}^{(B)}$. The bootstrap SE is

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

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

4. (5+5 points)

You plan to use K-nearest neighbors procedure for classification. You don't know what the right choice of K might be, and you decide to apply the method for a range of $K = 1, 2, \ldots$

(a) Describe the shape of the curve that you get when you plot the training error vs. K. Explain why one gets this shape.

(b) Describe the shape of the curve that you get when you plot the test error vs. K. Explain why one gets this shape.

a Increasing k decreases the flexibility of the method, hence ~~the~~ the training error increases with k as the training error decreases ~~when~~ when the flexibility increases.

b U-shaped, because the test error is ~~made out~~ the sum of $(\text{bias})^2$ and variance. As k increas, the bias ~~decreases~~ increases and the variance decreases, as the flexibility of K-nearest neighbors decreases.

5. (4+6 points)
The Advertising data set in the book consists of the sales of a product in 200 markets, along with the advertising budget in each market for three different media: TV, radio and print. You want to use k-nearest neighbor regression to predict the sales as a function of the spending on advertising on these three media.

(a) For a fixed value of k, explain how k-nearest neighbor regression predicts the sales number, given the advertising spending on TV, radio and print.

(b) How would you choose k to get a good prediction? Name a method for doing this and briefly explain the method.

a For the given numbers (TV, radio, print), find the K nearest ~~neighbori~~ observations (using Euclidean distance in 3-D space (TV, radio, print)). Then compute the average of the K corresponding Y-values (i.e. sales numbers) as the predictor.

b Using cross-validation. For example, leave one out cross-validation predicts each of the 200 sales values in turn from its K nearest neighbors, as desribed in a. Then we estimate the test error by averaging the 200 squared ~~distan~~ differences between predicted sale and actual sale. Then we choose the K that minimizes this estimated test error.