Stat 202 **Midterm exam**

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. (6 points)

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

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.

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.)

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.

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.