Statistics 151B Midterm Thu 17 Mar 2011

Name: _____

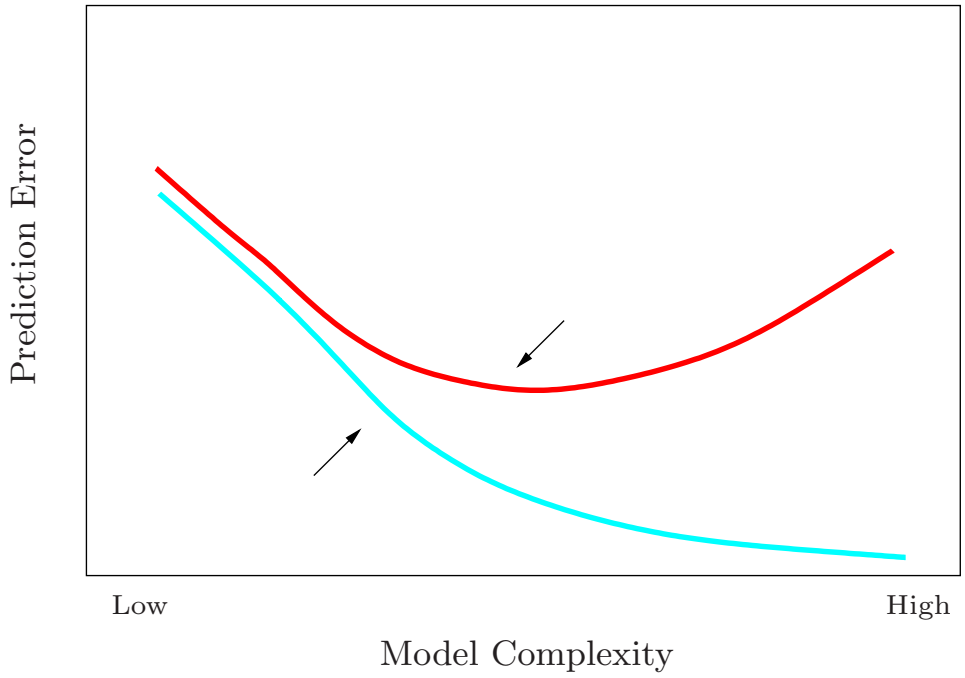
Student ID: _____

There are five questions on the exam. Each question is worth 20 points total. **CHOOSE FOUR OUT OF THE FIVE QUESTIONS, AND SOLVE ONLY THOSE FOUR**. The exam is geared to take you around an hour to complete. Write your solutions in the accompanying packet of answer sheets. If you need extra paper, ask the proctor. Hand in both your answer sheets and this copy of the exam questions.

Question 1

Part A. Explain what is meant by the training error and test error of a prediction rule.

Part B. Refer to the following figure.



Label one of the curves as "training" and the other as "test". Explain your reasoning.

Part C. Which of these choices best describes the 1-nearest-neighbor prediction rule? Why?

(I) low bias, low variance
(II) low bias, high variance
(III) high bias, low variance
(IV) high bias, high variance

Now answer the same question, but for a linear prediction rule based on a single predictor variable.

Question 2

Having fit a regression tree to a training set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, we can write its prediction on the vector x_{test} as

$$\hat{f}_{\text{tree}}(x_{\text{test}}) = \sum_{m=1}^{M} \hat{c}_m \mathbb{1}[x_{\text{test}} \in R_m]. \tag{1}$$

Here M is the number of leaves in the tree, R_m is the (hyper-)rectangle in predictor space corresponding to the mth leaf, and $\hat{c}_m$ is the constant predicted value for the mth leaf:

$$\hat{c}_m := \frac{1}{N_m} \sum_{\{i\,:\,x_i \in R_m\}} y_i, \tag{2}$$

with N_m denoting the number of training points x_i which belong to rectangle R_m.

Part A. Show that the tree's prediction can also be written in the form

$$\hat{f}_{\text{tree}}(x_{\text{test}}) = \sum_{i=1}^{N} w(x_{\text{test}}, x_i) y_i, \tag{3}$$

where the weight function $w(x_{\text{test}}, x_i)$ has these properties: for a fixed value of x_{test},

$$w(x_{\text{test}}, x_i) \geq 0 \text{ for all } i, \tag{4}$$

and (again for a fixed value of x_{test}),

$$\sum_{i=1}^{N} w(x_{\text{test}}, x_i) = 1. \tag{5}$$

Part B. The weight function from Part A has particular characteristics which lead us to describe CART as an "adaptive nearest-neighbor" prediction rule. Describe these characteristics and explain the relationship to nearest-neighbors prediction.

Question 3

Part A. Suppose X is an $N \times p$ design matrix built from p real-valued predictor variables. Why do the columns of X need to be standardized (i.e., centered, then scaled to unit variance) prior to fitting a ridge regression or the lasso? What would happen if you tried to put the intercept term (a column of 1's) into X before standardizing?

Part B. Suppose we have data $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, where each $x_i = (x_i^1, \ldots, x_i^p)$ is a vector of p real values. We standardize the design matrix, then fit a ridge regression with complexity parameter λ, obtaining regression coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$. Give a formula for $\hat{f}_{\mathrm{RR}}(x_{\mathrm{test}})$, the value of the ridge-regression prediction rule at the input x_{test}. (Your formula can use the $\hat{\beta}_j$'s; you do not need to write them in terms of the data.)

Part C. Why does the lasso usually set some of its fitted regression coefficients $\hat{\beta}_j$ to exactly zero, whereas ridge regression essentially never does? Draw a picture if it helps you explain.

Part D. Can putting additional predictor variables into an OLS regression cause the residual sum of squares on the training set to increase? Why or why not? What implications does this have for the variable-subset selection problem?

Question 4

Part A. You and Joe are discussing your favorite topic: cross-validation. Joe says, "When you do cross-validation, you don't need a test set. Whichever complexity setting $\hat{\lambda}$ you finally choose, you can just use the cross-validated error estimate of $\hat{f}_{\hat{\lambda}}(x)$ to assess its future prediction error." Do you agree or disagree with Joe? Explain.

Part B. (HTF 7.10.2) Consider a classification problem with a large number of predictors, as may arise, for example, in genomic or proteomic applications. A typical strategy for analysis might be as follows:

1. Screen the predictors: find a subset of "good" predictors that show fairly strong (univariate) correlation with the class labels;

2. using just this subset of predictors, build a multivariate classifier;

3. use cross-validation to estimate the unknown model complexity parameters.

Is this a correct application of cross-validation? Why or why not?

Question 5

Part A. You and Joe are getting into it again; this time the subject is local versus global prediction rules. Joe: "It makes no sense to build a prediction rule with OLS regression. The assumption that $\mathbb{E}[Y|x]$ is linear in x over all of predictor space will never be true in practice. I'm always going to use k-nearest neighbors for my prediction problems." Explain to Joe a problem with k-NN that should make him reconsider.

Part B. Joe thinks that least-squares fitting always means we are fitting a prediction rule which is linear in x. Name and describe to him a general method by which OLS can be used to produce nonlinear (in x) prediction rules. Give details.