Final Exam

BioStat 527, Spring 2019

Due Wednesday, June 12, 2019, 11:59pm

Note: Whatever you hand in has to be **your own work**. "Group work" is not allowed.

Suggestion: When you start using a complex package like `rpart`, first try it on a simple data set (maybe simulated data or the `iris` data) where you know roughly what the results should be.

Problem 1: (70 points) Consider a classification problem with a binary class label Y and a single continuous feature X that takes values in $(-4,-2) \cup (2,4)$. Suppose (X,Y) is generated by choosing Y at random with $P(Y = 1) = P(Y = 2) = 1/2$, and then drawing X conditional on Y according to uniform distributions. Specifically, assume that the class-conditional densities for X are

$$\begin{aligned} p(x \,|\, Y = 1) &= \frac{1}{2} \cdot \mathbf{1}_{(-4,-2)}(x) \quad \text{and} \\ p(x \,|\, Y = 2) &= \frac{1}{2} \cdot \mathbf{1}_{(2,4)}(x)\,. \end{aligned}$$

In the below we consider 0-1 loss, that is, the risk of a classifier is the probability of an error.

(a) (10 points) What is the marginal distribution of X? What is the conditional distribution of Y given X?

(b) (10 points) What is the Bayes rule $f_B(x)$ and its risk $P(Y \neq f_B(x)$? Explain!

(c) (20 points) Let $\hat{f}_1(x; S)$ be the 1-nearest neighbor classifier based on a training sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of i.i.d observations of (X, Y). What is the risk $\Pr(Y \neq \hat{f}_1(X; S)$? Explain. (Here, the risk is computed by integrating over training data and a new independent pair (X, Y)).

(d) (20 points) Under the same scenario calculate the risk of the 3-nearest neighbor classifier.

(e) (10 points) Which method, 1-nearest neighbor or 3-nearest neighbor, has smaller risk in this problem?

Problem 2: (60 points) In this problem you will do spam classification. Consider the email spam dataset (in "spam.data"). This consists of 4601 email messages, from which 57 features (or covariates) have been extracted. These are as follows:

- 48 features, in $[0, 100]$, giving the percentage of words in a given message which match a given word on the list. The list contains words such as "business", ' 'free", "george", etc.
- 6 features, in $[0, 100]$, giving the percentage of characters in the email that match a given character on the list. The characters are ; ([! $ #
- Feature 55: The average length of an uninterrupted sequence of capital letters
- Feature 56: The length of the longest uninterrupted sequence of capital letters
- Feature 57: The sum of the lengths of uninterrupted sequence of capital letters

More detail about the data can be found in the file "spam.info".

Load the data from "spam.data", in which Columns 1-57 are the features and Column 58 is our response variable, the indicator of spam emails (1 is spam and 0 is non-spam). Divide the dataset into a training set (of size 3065) and a test set (of size 1536) by the indicator in the file "spam.traintest" (1 is test set and 0 is training set).

(a) (20 points) Fit a classification tree to the training set. Plot the fitted un-pruned tree. Make binary predictions and report the error rate on the training and test sets.

(b) (40 points) Prune the tree and select the optimal tuning parameter λ by minimizing the 10-fold cross validation error with the one standard error rule (1-SE Rule). The 1-SE Rule means that we should choose a simpler model if its penalized RSS is less than 1 SE worse than the next complex model. Plot the fitted pruned tree. Make binary predictions and report the error rate on the training and test sets.

Hint: You may find the "rpart" package in R helpful. The examples in the introduction below are also helpful.
http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Problem 3: (60 points) In this problem you will try binary classification using ridge regression. Here is the idea: You have training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, with $y_i \in \{0, 1\}$. Fit a linear model using ridge regression and obtain fitted values $\hat{y_1}, \ldots, \hat{y}_n$. For a given threshold c, predict $y = 1$ for all training observations with $\hat{y} > c$, and predict $y = 0$ otherwise. Choose c to minimize the number of misclassifications. (This rule could be modified to account for different priors / losses.) To predict y for a new observation with predictor vector $\mathbf{x}$ calculate the value $\hat{y}$ predicted by the ridge regression model. Predict $y = 1$ if $\hat{y} > c$.

Apply this method to the SPAM training set (from Problem 2). Use `glmnet` to estimate the optimal penalty parameter λ and the coefficients of the corresponding model. Calculate the resubstitution error rate and the test set error rate.

How does the performance of ridge regression compare to the performance of least squares?

Problem 4: (60 points) This problem uses the `College` data set which is part of the ISLR package.

(a) (20 points) Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

(b) (20 points) Fit an additive model on the training data, using out-of-state tuition as the response, and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

(c) (20 points) Evaluate the model obtained on the test set, and explain the results obtained.

Problem 5: (50 points) Suppose that a curve $\hat{g}$ is defined by

$$\hat{g} = \arg\min_{g} \left(\sum_{i=1}^{N} (y_i - g(x_i))^2 + \lambda \int \left[g^{(m)}(x) \right]^2 \right),$$

where $g^{(m)}$ represents the mth derivative of g, and $g^{(0)} = g$. Provide example sketches of $\hat{g}$ in each of the following scenarios:

1. $\lambda = \infty, m = 0$.
2. $\lambda = \infty, m = 1$.
3. $\lambda = \infty, m = 2$.
4. $\lambda = \infty, m = 3$.
5. $\lambda = 0, m = 3$.

Problem 6: (70 points) In this problem you will apply a k-nearest neighbor classifier to the handwritten digit data. The data are divided into a training set `zip-train.dat` and test set `zip-test.dat`. Both data sets have lines of length 257. The first entry in each line is the digit that was written $(0, \ldots 9)$ and the remaining 256 entries are grey levels pixels in a 16×16 bitmap.

(a) (30 points) Write an R function

```
knn.classifier(X.train, y.train, X.test, k.try = 1, pi = rep(1/K,
K), CV = F)
```

where X.train, y.train, X.test have the obvious meanings; k.try is a vector of neighborhood sizes pi is a vector of prior probabilities CV = T if cross-validation is to be used.

CV = T only makes sense if X.train = X.test

The function should return a (n.test x length(k.try)) matrix of predicted class identities for the n.test test observations and the different values of k provided in k.try.

(b) (10 points) Run the function on the Iris data with k = 5 and with both choices for CV. Print the respective number of misclassifications.

(c) (20 points) Run the function on the hand-written digit training data with k.try = c(1, 3, 7, 11, 15, 21, 27, 35, 43) and CV = T. Use unweighted Euclidean distance as a dissimilarity measure. Choose the priors to reflect the class frequencies in the training data. What is the optimal choice of k and the corresponding training error rate?

(d) (10 points) Calculate the test set error rate.