Data Mining: 36-462/36-662
Homework 6

Due Thursday April 25 2013
(at the beginning of lecture)

Append your R code to the end of your homework. In your solutions, you should just present your R output (e.g. numbers, table, figures) or snippets of R code as you deem it appropriate. Make sure to present your results (i.e., your R ouput) in a clear and readable fashion. Careless or confusing presentations will be penalized.

Problem 1

Here you'll learn how to draw more flexible boundaries than simple linear ones, using logistic regression. Download the file "hw6prob1.Rdata" from the course website and load it into your R session. Now you should have the predictor matrix `x`, which contains $n = 800$ points in $p = 2$ dimensions. Each point falls into either class 0 or 1. There are two scenarios for the class labels, given by `y1` and `y2`.

(a) Plot the data in `x` with the class labels given by `y1`. (Use the option `col` or `pch` or both to distinguish between the classes.) Run logistic regression, using the `glm` function with `family="binomial"`, to build a prediction rule. What is the training misclassification rate of this rule? (Hint: use the `predict` function; to read the relevant documentation, type `?predict.glm`.)

(b) Draw the decision boundary in $\mathbb{R}^2$ of the logistic regression rule from (a), on top of your plot from (a). What shape is it? Does this boundary look like it adequately separates the classes?

(c) Run logistic regression on the predictors in `x`, as well as the predictor `x[,1]^2`. This is analogous to adding a quadratic term to a linear regression. To do this, define a new predictor matrix `x.quad = cbind(x,x[,1]^2)`, and run a logistic regression of `y1` on `x.quad`. What is the training misclassification rate of this rule? Why is this better than the rule from (a)?

(d) In $\mathbb{R}^2$, i.e., in the space `x[,1]` versus `x[,2]`, what is the shape of the decision boundary of the logistic regression rule from (c)? Draw this decision boundary on top of a plot of the (appropriately color-coded or pch-coded) data `x`. What shape is it?

(e) Plot the data in `x` with the labels given by `y2`. Try a running logistic regression of `y2` on `x`, and also on `x.quad = cbind(x,x[,1]^2)`. What are the training misclassification rates of these rules? Draw the decision boundaries of each rule on top of a plot of the data.

(f) Why are neither of the decision boundaries from (e) adequate? What additional predictors can you pass to logistic regression in order for it to do a better job of separating the classes? (Hint: draw a curve between the classes by eye... what shape does this have?) Run a logistic regression with these additional predictors, report the training misclassification rate, and draw the new decision boundary. What shape is it?

(e) If adding polynomial terms seems to improve the training misclassification rate of the logistic regression rule, why don't we generally just keep including polynomial terms of higher and higher order? How could we choose how many polynomial terms to include in a principled manner?

Problem 2

In class, we learned of three measures for impurity used by classification trees. We are given training data (x_i, y_i), $i = 1, \dots n$, where $x_i \in \mathbb{R}^p$ are the feature vectors and $y_i \in \{1, \dots K\}$ are the class labels. For a region R, remember that the proportion of points in R that are of class k is simply

$$\hat{p}_k = \frac{1}{n_R} \sum_{x_i \in R} 1\{y_i = k\},$$

where n_j is the total number of points in region R. Let

$$k^* = \underset{k=1,\dots K}{\text{argmax}} \ \hat{p}_k,$$

the most common class among points in R. The three measures of impurity for R are:

$$\begin{aligned} \text{Misclassification error:} \quad & 1 - \hat{p}_{k^*} \\ \text{Gini index:} \quad & \sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k) \\ \text{Cross-entropy:} \quad & -\sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k. \end{aligned}$$

(a) Suppose that there are only two classes, $K = 2$. Let p denote the proportion of points in R that are in the first class, i.e., $p = \hat{p}_1$. Write out each of the above impurity measures as a function of p (and only p).

(b) Plot the expressions for misclassification error, Gini index, and cross-entropy that you found in (a) as functions of p. When you plot cross-entropy, you can scale it by a constant so that it is equal to $1/2$ when $p = 1/2$. Are the curves similar?

Problem 3

Suppose that we are looking at the set of the first n positive integers $\{1, \dots n\}$. We decide to randomly sample from this set B times with replacement. In other words, for a total of

B times, we randomly select a number between 1 and n, placing equal probability on each number. (The "with replacement" terminology refers to the fact that the same number—e.g., the number 2—can be chosen multiple times in this process.) In statistics, we call the resulting sample a *bootstrap* sample. Let's write this bootstrap sample as $\{i_1, \ldots i_B\}$.

(a) What is the expected number of numbers in $\{1, \ldots n\}$ that do not appear in the random sample $\{i_1, \ldots i_B\}$? (Hint: first determine the probability that a given number, say i, does not appear in $\{i_1, \ldots i_B\}$.) Hence, what is the expected proportion of numbers in $\{1, \ldots n\}$ that do not appear in the random sample $\{i_1, \ldots i_B\}$?

(b) Suppose that $B = n$, i.e., we draw n numbers from $\{1, \ldots n\}$ with replacement. What is now your formula from (a) for the expected proportion of numbers that do not appear in the random sample? What does this approach as $n \to \infty$? (Hint: go dig up some of your old notes from calculus and find a formula that expresses e^x as a limit.)

These last two problems are from Section 8.4 in the Introduction to Statistical Learning book. They are helping us ramp up our coding skills up for the final project. You will have to download the ISLR package to use the data sets in each problem (remember, this is found at `http://www-bcf.usc.edu/~gareth/tempISL`; this page is password protected, use the login name: `StatLearn` and password: `book`). It will also be helpful to read through the labs in Section 8.3.

Problem 4

ISL Section 8.4, exercise 9.

Problem 5

ISL Section 8.4, exercise 10.