University of Michigan, Dept of Statistics

Stat 503, Instructor: Long Nguyen

Homework Assignment 3
(issued Feb 12, due Feb 22, 2017)
in class or Canvas dropbox by 1pm

1. **Optimal Bayes classifier**

   (a) Suppose there are two classes in a population and the prior probability for the first class is $\pi_1 = 2/3$. There is a single predictor X, which given class label $Y = k$ ($k = 1, 2$) follows the exponential distribution with parameter λ_k. Derive the Bayes optimal classification rule, and derive a formulae for the (optimal) Bayes risk. In addition, compute the numeric values for the boundaries and the Bayes risk when $\lambda_1 = 1$ and $\lambda_2 = 2$.

   (b) Instead of having one single predictor variable we now have two predictors, X_1 and X_2. Moreover, X_1 and X_2 are conditionally independent given the class label Y. Given $Y = k$, the conditional distribution of X_1 is exponentially distributed with parameter λ_k, while X_2 is normally distributed with unit variance and mean μ_k. Derive the Bayes optimal classification rule, and compute the (optimal) Bayes risk. In addition, compute the numeric values for the boundaries and the Bayes risk when $\lambda_1 = 1, \lambda_2 = 2$, and $\mu_1 = 1$, $\mu_2 = -1$.

   Reminder: the probability density function of an exponential random variable X with parameter $\lambda > 0$ is given by $p(x) = \lambda e^{-\lambda x}$ for $x > 0$.

2. **Classification** Recall the data set contained in `auto-mpg.dat` on Canvas. In this exercise, we will apply a number of classification methods to obtain a classification rule, which allows us to classify the cars into three different classes of cylinders based on information provided by other continuous input features (predictors). Specifically, define class label $Y = 1$ if the number of cylinders is 5 or less, $Y = 2$ if the number of cylinders is 6, and $Y = 3$ otherwise.

   We will also preprocess the data set in a number of ways: (i) standardize all continuous valued columns so they all have zero mean and unit variance; (ii) perform principal component analysis on these the continuous features and represent the data based on the first two or more principal components (note that you already did this in Homework 1).

   (a) Randomly select from each class 75% of the data which shall be used for training/validation purposes. The remaining 25% will be left aside as the test set. Provide a few summary statistics/ scatter plots to make sure that the training set and the test set are visually "compatible".

   (b) Apply the LDA and QDA to the original data, standardized data, and PCA-preprocessed data. What are the error rates? Make a plot of the data projected onto the first two discriminant directions and comment on any interesting features. Use different symbols for different classes.

   (c) For each version of the data, fit a logistic regression model. Report the mean error rate on both the training and test sets.

   (d) Apply the nearest neighbor classifier to each of the three versions of the data. What are your choices of distance metric in determining the nearest neighbors? Use cross-validation on the training data set to choose the value of k. Do not use the test data to choose k, but do report the test error for comparison (include a plot of training, cross-validated, and test errors as a function of k).

   (e) Comment compare the performance among the three classification methods. Please report also the role of data processing on each method.

Instructions The solution to problem 1 can be either typed up or written by hand. The solution to data analysis questions (Problem 2) may be written as a data analysis report. The report needs to be clear, concise, and to the point. There should be no graphs or tables that are not commented on in the text. Please include your R code (at least the main parts) in a *separate* appendix at the end of the report.