STA 141A

Fall 2016

Homework 5

Due: December 7 (Wednesday) at 5:00 PM

Submit the assignment both electronically (through smartsite) and by submitting the printed copy. Electronic submission must be in the form of a zip folder (with extension .zip, .7z, etc.) containing two files: (i) descriptions of your analysis (as appropriate); (ii) codes used.

Honor Code: *"The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment:" (ADD: names of persons or web resources, if any, excluding the instructor, TAs, and materials posted on course website)*

1. This is a basic illustration of using Bootstrap for inference.

   (i) Generate a random sample of size $n = 100$ following the univariate regression model

   $$Y_i = -5 + 2X_i + \varepsilon_i$$

   where X_i's are independent Chi-square random variables with degrees of freedom 6, and ε_i's are i.i.d. $N(0, \sigma^2)$ with $\sigma = 1$. Fix a random seed to ensure that the results are reproducible.

   (ii) Fit the least squares regression line to the data and obtain the estimate of $(\beta_0, \beta_1, \sigma^2)$.

   (iii) Obtain resampling-based 95% confidence intervals for β_0 and β_1 by using a parametric (i.e., residual-base) bootstrap procedure with 400 bootstrap replicates.

   (iv) How do the confidence intervals in (iii) compare with the theoretical confidence intervals for β_0 and β_1 ? [To compare the accuracy of the confidence intervals, repeat the procedure in steps (i)–(iii) 10 times (using different random seed for each simulation run) and report the average lengths of the bootstrap confidence intervals and that of corresponding theoretical confidence intervals.]

2. In this example, compare k-NN classification method, linear discriminant analysis and logistic regression in a two-class classification problem. For this consider the **iris** data available in R.

   (i) Extract the data corresponding to flower types *setosa* and *versicolor*, numbering a total of 100 flowers. Set aside the last 10 measurements for each flower type as test data and use the remaining data consisting of 80 measurements as training data.

(ii) Fit a logistic regression model to the training data, using the variable Sepal.Length as predictor. Obtain the estimates of the model parameters. Compute the *confusion matrix* for the test data set.

(iii) Compute the decision boundary for linear discriminant analysis, using Sepal.Length as the predictor variable. Compute the *confusion matrix* for the test data set.

(iv) Use k-nearest neighbors classification method with $k = 3, 4, 5$, again using Sepal.Length as the predictor variable. In each case, *confusion matrix* for the test data set.

(v) Write a very brief summary of the comparative performance of different classification procedures.

Reference

1. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. [Chapters 3,4 & 5]