FSRM 588: Financial Data Mining Homework 5

Fall 2015

Due 6:40pm on November 30

NOTES: Submit all your code with your assignment. Please read and follow the instructions carefully. I reserve the right to refuse homework that is deemed (by me) to be excessively messy.

Instructions:

- HW5 contains two parts: Part I is theoretical and Part II is computer exercises. Your answers to Part II MUST be typed up and printed out but your answers to Part I can be either typed up or written up. Submit the answers to both Part I and Part II in class.
- Use set.seed(1) every time you start a new CV. We consider K-fold CV with $K = 5$.
- Submit the R code you use to the email address fsrm588@gmail.com as an attachment before class. The subject of the email should be "hw5 + your last name + your first name". For instance, "hw5 Yang Dan". Only .r files are allowed and the title of the attached document should be "hw5 + your last name + your first name.r". For instance, "hw5YangDan.r"
- The content of the email can be empty. There is no need to say thank me and I won't reply to the emails sent to fsrm588@gmail.com. fsrm588@gmail.com is only used for the purpose of submission of homework, project, and exam. Please reach me via dyang@stat.rutgers.edu if you have questions or need help.

PART I: Theory

1. *Concavity of the dual function.* Consider the minimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0 \quad i = 1, 2, \ldots, m \\ & h_i(x) = 0 \quad i = 1, 2, \ldots, n \end{array}$$

   Show that the dual function $q(\lambda, \gamma)$ is concave on the set $\{(\lambda, \gamma) : \ q(\lambda, \gamma) > -\infty\}$.

2. Exercise 15.1 of the textbook.

PART II: Computer

3. *Local method; kernel smoothing.* Consider the zipcode data, the 2's, 3's, 8's. The zipcode data are available from Sakai. Use the following code to load the data: load("hw3.RData"). The names of the variables are self-explanatory.

   Note:

   - As we explained in class, it did not make sense to make the norms of all predictors the same, because they were already on the same scale.
   - There is no need to exclude V16.
   - To avoid confusion, let us first center the data, both the training and the test data. Note that we should use the sample mean from the training data, rather than from the test data to center both the training and the test data.

Perform *local* FDA on the centered data, retain the leading two scores, and perform *local* LDA on the leading two scores. At each query point x_0, the training data receive weights from a weighting kernel: tri-cube kernel. Use CV to choose among a series of five pre-chosen values of window size λ. Use the chosen value of λ, plot the leading two scores from local FDA. Compare them with the leading two scores from PCA and *non-local* FDA. Report the training and testing errors from local LDA with local FDA. Compare them with the results from previous hw on LDA with and PCA and FDA scores respectively.

4. *SVM.* Implement support vector classifier, SVM with the radial basis kernel, and SVM with polynomial kernel on the zipcode data in Problem 2. Use the `tune()` function to choose the regularization parameter C and the kernel parameter for both radial basis kernel and polynomial kernel. Report the misclassification rates on both the training set and the test set. Read the help file for the `tune()` function.

5. *Kernel PCA.* Implement kernel PCA with radial Basis kernel on the zipcode data in Problem 2. Choose the best tuning parameter according to some reasonable criterion that you can think of. Plot the first two scores obtained from the kernel PCA with the best tuning parameter. Compare the plot with the plain PCA and FDA.

6. *RF.* Implement random forest on the zipcode data. Perform CV on m. Plot the CV, training, and testing errors against m. What are the training and test errors with the best m?

7. *Boosting.* Implement adaboost on the zipcode data. Perform CV on the number of trees. Plot the CV, training, and testing errors against the number of trees. What are the training and test errors with the best number of trees? Compare the results from Problem 6.