CS 76140 - Spring 2017 Olga Vitek

Homework 4

Due on Blackboard before 5pm on Tuesday April 18, 2017.

Note: Use any tool of your choice (including Word, latex, Markdown, or pencil+paper) to prepare the solutions. The answers should be easy to find and grade. Unreadable hand-written solutions will be given 0 points, at the grader's discretion, regardless of the correctness of the answer. For each problem, use the appropriate notation for random variables, probabilities etc. State the full formula in addition to the numerical conclusions. For data analysis problems, use reproducible research tools such as R Markdown whenever possible.

Note: This homework will use the 'South African Heart Disease" dataset, which was also used in Homework 3. Use the same partition into the training and validation sets for all the methods in both homeworks, to obtain comparable results.

1. JWHT Chapter 7, Problem 1
2. JWHT Chapter 7, Problem 3
3. In this problem we will investigate the use of Naïve Bayes classifier.
   (a) Implement Naive Bayes classifier described in HTF Chapter 6.6.3, using Gaussian kernel for density estimation. (I.e., do not use the existing implementation, but write your own code. Include the code as an appendix to the homework).
   (b) Consider the dataset "South African Heart Disease", used in Homework 3. As in Homework 3, randomly partition the dataset on the training and validation set. *[Note: use the same partition of the observations as in Homework 3.]* Use the training set, and the continuous predictors only, to classify the variable `chd`. Evaluate the bandwidth parameter λ of the kernel on a grid, and select the best parameter using cross-validation. Report the predictive performance on the validation set.
   (c) Compare the performance of the Naïve Bayes classifier on the validation set to that of LDA on the same validation set. Visualize the data and/or the results of the model fit, to provide reasons for same (or different) performance.
4. In this problem, we will investigate the use of tree-based methods.
   (a) Answer the questions in JWHT Chapter 8, Problem 9, but using the "South African Heart Disease" dataset, and the same training and validation sets as in the problem above.
   (b) Perform bagging on the training set, with a range of tree numbers B.
      i. Produce a plot with different values of B the x-axis and the corresponding training set % of correct predictions on the y-axis.
      ii. Produce a plot with different values of B on the x-axis and the corresponding test set % of correct predictions on the y-axis. Compare the results.

    iii. Compare the predictive performance on the training and the validation set to the performance of the single pruned tree. How does the value of B influence the comparison?

  (c) Perform random forest on the training set, with a range of tree numbers B, and with two different numbers of selected predictors m of your choice.

    i. Produce a plot with different values of B and the x-axis and the corresponding training set % of correct predictions on the y-axis, for one m. Overlay the plot for the second m.

    ii. Produce a plot with different values of B and the x-axis and the corresponding test set % of correct predictions on the y-axis, for one m. Overlay the plot for the second m.

    iii. Compare the predictive performance on the training and the validation set to those of the bagging, and of the single pruned tree. How do the values of B and m influence the comparison?

  (d) Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ.

    i. Produce a plot with different shrinkage values on the x-axis and the corresponding training set % of correct predictions on the y-axis.

    ii. Produce a plot with different shrinkage values on the x-axis and the corresponding test set % of correct predictions on the y-axis. Compare the results.

    iii. Which variables appear to be the most important predictors in the boosted model? Are these the same variables chosen in the best single-tree approach?

5. JWHT Chapter 9, Problem 3.

6. In this problem, we will investigate the use of support vector machines and neural networks.

  (a) Train support vector machine the “South African Heart Disease” dataset, and evaluate its performance on the validation set.

  (b) Train neural networks on the training set of the “South African Heart Disease” dataset, and evaluate its performance on the validation set.

7. Summarize the classification results obtained for the “South African Heart Disease” in homework 3 and 4. Which method performed better, which performed worse? Discuss the possible reasons.