DS 5220 / CS6140-02 - Fall 2017 Olga Vitek

Homework 4

Each part of the problems 5 points
Due on Blackboard before midnight on Tuesday November 14, 2017.

1. JWHT Chapter 7, Problem 1
2. JWHT Chapter 7, Problem 4
3. JWHT Chapter 7, Problem 5
4. JWHT Chapter 9, Problem 3.
5. In this problem we 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.
   (a) Implement Naïve 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 from scratch. Include the code as an appendix to the homework).
   (b) Use the training set to classify the variable `chd` with all the predictors, and with your implementation of Naïve Bayes above. 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.
   (d) Train support vector machine (using an existing implementation) on the "South African Heart Disease" dataset, and evaluate its performance on the validation set.
   (e) Summarize the classification results obtained for the "South African Heart Disease" with Naïve Bayes, LDA, and SVM. Which method performed better, which performed worse? Discuss the possible reasons.