Statistical and Machine Learning (Spring 2018)

Mini Project 6

Instructions:

- Due date: April 26, 2018.
- Total points = 40.
- Submit a typed report.
- It is OK to discuss the project with your old partner and other students in the class, but each person must write their own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- **Explain all assertions, interpret all results, and justify all conclusions**.
- You must use the following template for your report:

  Mini Project #
  Name
  Names of group members
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. Consider the `OJ` dataset from the `ISLR` package that was used in the previous project. It consists of `Purchase` as a binary response variable and a number of other variables as predictors.

   (a) Divide the data into a training set consisting of the first 870 observations and a test set consisting of the last 200 observations. This split is just like the one in the previous project.

   (b) Fit a support vector classifier to the training data with cost parameter chosen optimally using 10-fold cross-validation. Evaluate its performance on the test data. Summarize your results.

   (c) Repeat (b) using a support vector machine with a polynomial kernel of degree two with cost parameter chosen optimally.

   (d) Repeat (b) using a support vector machine with a radial kernel with both γ and cost parameters chosen optimally.

   (e) Compare the results from the above three methods and also with the results from the method you recommended for these data in the previous project. Which method seems to perform the best on these data?