STAT 6340 (Statistical and Machine Learning, Spring 2020)

Mini Project (Bonus)

Instructions:

- Due date: May 4, 2020.
- Total points = 15.
- Submit a typed report.
- It is OK to discuss the project with others and to search on the internet, but each student must write their own code and answers. If your 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.
- You must use the following template for your report:

  Mini Project #
  Name
  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.
- Section 1 of the report must be limited to **3** pages. Also, only those output should be provided in this section that are referred to in the report.

---

1. Consider the business school admission data from Mini Project 2. Split the data into training and test sets just as in that project.

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

   (b) Repeat (a) using a support vector machine with a polynomial kernel of degree two.

   (c) Repeat (a) using a support vector machine with a radial kernel with both γ and cost parameter chosen optimally.

   (d) Compare results from the above three methods and also the method you recommended for these data in Mini Project 2. Which method would you recommend now? For this method, display the data with decision boundary superimposed.

2. Consider the wine quality dataset from Mini Projects 3 and 4. As in those projects, take all data as training data.

   (a) Fit a support vector classifier to the data with cost parameter chosen optimally using 10-fold cross-validation. Summarize key features of the fit.

   (b) Repeat (a) using a support vector machine with a polynomial kernel of degree two.

   (c) Repeat (a) using a support vector machine with a radial kernel with both γ and cost parameter chosen optimally.

   (d) Compare results from the above three methods and also from the method you recommended for these data in Mini Projects 3 and 4. Which method would you recommend now?