MSBA7027 Machine Learning

Homework 1

Due 11:59 pm Dec. 20, 2023

Notes:

- You are required to submit 1) original R Markdown file and 2) a knitted HTML or PDF file via Moodle. Please provide comments for R code wherever you see appropriate. In general, be as concise as possible while giving a fully complete answer. Nice formatting of the assignment will receive extra points.
- Remember that the Class Policy strictly applies to homework. You are encouraged to work in groups and discuss with fellow students. However, each student has to know how to answer the questions on her/his own.
- Degree of freedom in this homework refers to degree of freedom in R.
- Please allow some buffer time and do not submit homework at the last moment. You will have points deducted if you submit the above two items late (even by one minute).
- Please note that to be fair to all students, the instructor and the TAs can only answer clarification questions about the assignment.

Question 1. This question uses the variables horsepower and mpg from the Auto data as part of the ISLR package. We will treat horsepower as the predictor and mpg as the response.

(a) Use the poly() function to fit a cubic polynomial regression to predict mpg using horsepower. Report the regression output, and plot the resulting data and polynomial fits.

(b) Use the bs() function to fit a cubic spline to predict mpg using horsepower. Report the output for the fit using six degrees of freedom. How did you choose the knots? Plot the resulting fit.

(c) Now fit a cubic spline for degrees of freedom ranging from 4 to 12, and plot the resulting fits as well as the resulting RSS. Describe the results obtained.

(d) Perform cross-validation to select the best degrees of freedom for a cubic spline on this data. Describe your results.

(e) Use the ns() function to fit a natural cubic spline to predict mpg using horsepower. Report the output for the fit using degrees of freedom s.t. number of knots is the same as (b). Plot the resulting fit.

(f) Now fit a natural cubic spline for a range of degrees of freedom, and plot the resulting fits as well as the resulting RSS. Describe the results obtained.

(g) Perform cross-validation to select the best degrees of freedom for a natural cubic spline on

this data. Describe your results.

(h) Compare your results of (d) and (g), choose the best model for cubic spline and natural cubic spline respectively. Which one performs the best? With how many knots?

Question 2. This problem involves the Boston data set which is from the MASS package. We will fit classification models in order to predict whether a given suburb has a crime rate above or below the median.

Preliminary step: Create an extra column crimAbvMed which is 1 if crime rate is above median and 0 otherwise. Then delete the crime rate column crim.

(a) Create a training set containing a random sample of 80% of the observations, and a test set containing the remaining observations.

(b) Fit a linear support vector classifier to the training data. Use the tune() function to select an optimal cost. Consider values in the range of 0.001 to 100.

(c) Compute the training and test error rates using this new value of cost.

(d) Repeat parts (b) through (c) using a support vector machine with a radial kernel. Tune both gamma and cost.

(e) Repeat parts (b) through (c) using a support vector machine with a polynomial kernel. Tune both degree and cost.

(f) Overall, which approach seems to give the best results on this data set?