Assignment 1 Part 1: Support Vector Classifiers

BUAD 5082 Section 1 – Spring 2018

---

1. Objectives

The purpose of this assignment is to exercise recently developed skills in Support Vector Machines.

2. What You Will Need

- Access to a Windows computer with a recent version of R installed.

3. What You Will Hand In

When all Assignment 1 parts are complete, submit your script as Assignment1.R via Blackboard – Assignment1.

4. Due Date

TBA.

5. Note on Collaboration

This is a Category C individual assignment.

6. Preliminaries:

To get set up for the lab, follow these steps:

1. As the first statement in your script file, enter rm(list=ls())
2. Your response to each lettered question in the assignment should begin with the following three comment lines, where *n* is the Section number and m is the Part number:

   ```
   ######################
   #### Part m
   ######################
   ```

3. I should be able to run your script on my computer without errors or interruptions. For this to happen, you must:
   a. Avoid entering file path information…my files will be located in a different location that yours, and so your code will fail on my machine. Instead, always refer only to files in your working directory.
   b. Do not use functions like file.choose(), fix(),edit(), or q()
   c. Do not include install.packages() functions (or comment them out)
4. Do not create console output other than what is asked of you explicitly. For example, in your final script, remove any statements that you used to verify the contents or structure of data that were not requested.

7. Tasks:

Section 1: Support Vector Machines

This Section involves the OJ data set which is part of the ISLR package.

a) Use the following code to create a set of indices containing a random sample of 800 integers representing the training subset of set OJ, and a set of test indices representing the remaining observations:

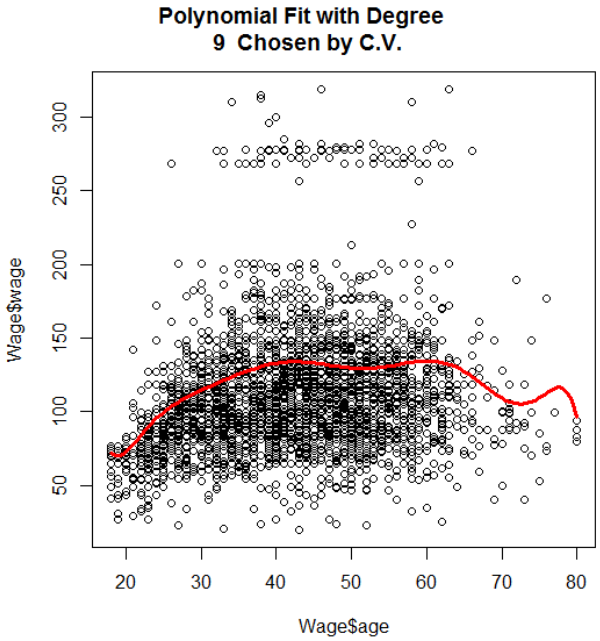
```
set.seed(5082)
n = dim(OJ)[1]
train_inds = sample(1:n,800)
test_inds = (1:n)[-train_inds]
```

b) Fit a support vector classifier to the training data using cost=0.01, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics, and describe the results obtained.

c) Compute and display the training and test error rates?
d) Use the tune() function to select an optimal cost. Consider values in the range 0.01 to 10.
e) Compute and display the training and test error rates using this new value for cost.
f) Repeat parts (b) through (e) using a support vector machine with a radial kernel. Use the default value for gamma.
g) Repeat parts (b) through (e) using a support vector machine with a polynomial kernel. Set degree=2.
h) Overall, which approach seems to give the best results on this data?

Section 2: Polynomial and Step Function Regression

a) Work through the lab in ISLR Section 7.8.1 (beginning on pp. 288) – **do not include this code in your submission**.
b) Using the Wage dataset from the ISLR package, perform polynomial regression to predict wage using age with polynomial degrees from 1 to 10.
c) Set the seed to 5082, then use 10-fold cross-validation to select the optimal degree *d* for the polynomial (recall cv.glm() function (bottom of page 192). What degree was chosen?
d) Plot the resulting polynomial fit and the original data.
    i. To plot the fit, break the range of age into 100 partitions, use the model to predict these points and plot the points against the predictions. My plot looks like this:



e) Fit a step function to predict wage using age (recall the cut() function used in the last step of the lab on Splines)
    i. Set the seed to 5082, then investigate step functions using steps from 1 to 12. Use 10-fold cross validation to choose the optimal number of steps.

        Note that the cut() function wants the number of intervals to cut the data into, not the number of steps inside the range. So cut(age,1) produces an error and cut(age,2) has one step.

ii. Plot the fit obtained with this optimal number of cuts. I get 8 as the optimal number and my plot looks like this: