Department of Mechanical Engineering

ME 781: Engineering Data Mining and Applications

Assignment-2

With reference to the given data set **a2-data-set.csv** perform the following:

1. Read in the data set into a data frame
2. From the data frame create a random sample **s1** consisting of 40 observations
3. Using the **trainControl** function of the **caret** package set up a ***train control*** object with the following settings:
    - 10-fold cross validation, retain all re-sampling results
4. Perform the following regressions on this smaller data set:
    a) Run Ordinary Least Squares Regression using **s1** and capture the results in **lm1**
    b) Run the **train** function on s1 using the train control object defined above, and capture the results as indicated below:
        i. Use the OLS regression method capture the results in tr.lm
        ii. Use the svm (Radial) regression method capture the results in tr.svmRadial
        iii. Use the svm (Linear) regression method and capture the results in tr.svmLinear
        iv. Use the ridge regression method and capture the results in tr.ridge
        v. Use the lasso regression method and capture the results in tr.lasso
        vi. Use the elasticnet regression method and capture the results in tr.enet
5. Create the following Table to compare and contrast the results of all the above regression methods

Method	RMSE	MAE	Regression Coefficients (Where available)
lm1			
tr.lm			
tr.svmRadial			
tr.svmLinear			
tr.ridge			
tr.lasso			
tr.enet			

6. In all the above cases where the model has been trained, create a plot of the outcome (eg. tr.svmRadial) and state your observations / learnings from the plot.
7. Based on the above, write a short note on each of the following:
    - What does it mean to 'train' the model, and the role played by the train control object?
    - What is cross validation and re-sampling?
8. Additionally, in all the above cases, generate the following plots:
    - **y-predicted v/s y-given** and comment on whether or not the two values are correlated. What is the correlation coefficient?
    - **residuals v/s y-predicted** and comment on whether the residuals are random, or show a pattern? What are your conclusions in each case?
9. Repeat steps 2 through 5 for another sample, s2, of 400 observations. Do you see any change?
10. Repeat steps 2 through 5 for the entire data set of 1000 observations and compare the results with those obtained from s1 and s2.

Note:

- Submit your answers to the assignment submission point in Moodle.
- This is one of a set of 5 Problems that you will have to solve. Submission of all assignments will fetch you max 5 marks.
- The Test scheduled on Nov-8-2017 will assume you have done this assignment.