**CS 251/340 - Machine Learning**

Spring 2019, AUA

**Homework No. 01**

Program files submission - Due date: 23:55, February 14, 2019

1. **(score =4) The sample size is extremely large, and the number of predictors is small. In general, which statistical learning method will perform better?**

**flexible**

**inflexible**

1. **(score =4) The number of predictors is extremely large and the number of observations is extremely small. In general, which statistical learning method will perform better?**

**flexible**

**inflexible**

1. **(score =4) A training data has extremely high variance with lots of outliers. In general, which statistical learning method will perform better?**

**flexible**

**inflexible**

1. **(score =4) The relationship between the predictors and response is highly non-linear. In general, which statistical learning method will perform better?**

**flexible**

**inflexible**

1. **(score =4) The variance of the error term is extremely high. In general, which statistical learning method will perform better?**

**flexible**

**inflexible**

1. **Assume set of data with observations containing a single predictor and a quantitative response. Assume linear regression model to the data, as well as a separate cubic regression, i.e. . If the true relationship between X and Y is linear which answer is correct?**
2. **(score =6)**

**Training RSS for the linear regression is lower than training RSS for the cubic regression**

**Training RSS for the cubic regression is lower than training RSS for the linear regression**

**Training RSS for the linear regression is the same as training RSS for the cubic regression**

**There is no enough information to answer**

1. **(score =6)**

**The test RSS for the linear regression is lower than the test RSS for the cubic regression**

**The test RSS for the cubic regression is lower than the test RSS for the linear regression**

**The test RSS for the linear regression is the same as the test RSS for the cubic regression**

**There is no enough information to answer**

1. **Explain whether the scenarios below are a classification or regression problem, and indicate whether we are most interested in inference or prediction. Provide response, predictors and number of observations:**
2. **(score =6) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.**

**regression  inference n = 500, p = 3**

**classification  prediction**

1. **(score =6) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, and competition price.**

**regression  inference n = 20, n = 3**

**classification  prediction**

1. **(score =6) We are interested in predicting the change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2012. For each week we record the change in the USD/Euro, the change in the US market, the change in the British market, and the change in the German market.**

**regression  inference n = 52, p = 3**

**classification  prediction**

***The next problems are for coding. Write down the solutions in R script file and upload together with this word file with the corresponding answers.***

1. **Exercise involves the Auto data set.**

**Remove the missing values (score =3).**

1. **(score =3) Which of the predictors are quantitative, and which are qualitative?**

**All variables except name and origin are quantitative. The mentioned two are qualitative.**

1. **(score =3) What is the *range* of each quantitative predictor?**

**Mpg - 9.0 to 46.6**

**Cylinders – 4 to 8**

**Displacement – 68 to 455**

**Horsepower – 46 to 230**

**Weight – 1613 to 5140**

**Acceleration – 8.0 to 24.8**

**Years – 70 to 82**

1. **(score =3) What is the mean and standard deviation of each quantitative predictor?**

**Mpg - 23.45 and 7.805007**

**Cylinders – 5.472 and 1.705783**

**Displacement – 194.4 and 104.644**

**Horsepower – 104.5 and 38.49116**

**Weight – 2978 and** **849.4026**

**Acceleration – 15.54 and 2.758864**

**Years – 75.98 and 3.683737**

1. **(score =14) Investigate the predictors graphically, using scatterplots or other tools of your choice from ggplot library. Create some plots highlighting the relationships among the predictors. Comment on your findings. Plot the same graphs using the basic plot functionality, for comparison.**

**Weight, displacement and horsepower are negatively correlated with mpg. Horsepower and weight are pos. correlated. Japanese cars have higher mpg than US or European cars.**

1. **(score =4) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.**

**Yes. As Mentioned there is correlation with weight, displacement and horsepower. Those can be useful. Plus the origin is also somehow useful.**

1. **Exercise involves the Boston housing data set (MASS library).**
2. **(score =4) How many rows are in this data set? How many columns? What do the rows and columns represent?**

**506 rows, i.e n = 506 and 14 columns, i.e. p = 14**

1. **(score =4) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.**

**When household is further from employment centers, there is higher crime rate. Higher the building age higher the crime rate. Other two plots show almost nothing.**

1. **(score =4) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.**

**We found two in the prev question -> Age and distance. There is also a relationship between lstat and crim -> lower the status (higher the % in data) more the crimes.**

1. **(score =4) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.**

**Only 18 rows with high crime rate. 5 rows with very high tax rate (690+) and 137 rows with high tax rate (650+). 201 rows with high ptratio (20+). Ranges are as follows:**

**Crim range -> 0.00632 88.97620**

**Tax range -> 187 711**

**Ptratio -> 12.6 22.0**

1. **(score =4) How many of the suburbs in this data set bound the Charles river?**

**35 suburbs**