**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**

**Answer:** The cubic regression model is more flexible than the linear regression model. Therefore, we would expect the cubic model to fit much better the data (maybe even overfit), and thus to have lower training RSS.

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**

**Answer:** If the true relationship between X and Y is linear, a cubic regression model is too flexible for it, and we would expect the method to fit test data poorly. It is the opposite case of the train RSS, where it got the noise as well. Therefore, we would expect the cubic model to have a higher test RSS.

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:**

**Number of observations – n.**

**Predictors – p.**

1. **(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**

**classification  prediction**

**Answer:** Since salary is a continuous variable (has non-discrete values), it is a regression problem. Moreover, as we are not predicting anything but trying to understand a relationship in data it is an inference task. n = 500 (500 firms), p = 3 (profit, number of employees and industry).

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**

**classification  prediction**

**Answer:** We have two categories for products (success and failure) and we want to classify our new product, so this is obviously a classification problem. Here we are making a prediction about a future product based on the data we have, so this is a prediction task. n = 20 (20 similar products), p = 13 (price charged for the product, marketing budget, competition price, and ten other variables).

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**

**classification  prediction**

**Answer:** We are working with a continuous quantitative variable (% change) and, hence, this is a regression problem. Here we are making a prediction about future changes of exchange rate based on previous changes, so this is a prediction task. n = 52 (52 weeks in 2012), p = 3 (% change in US, % change in Britain, % change in Germany).

***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?**
2. **(score =3) What is the *range* of each quantitative predictor?**
3. **(score =3) What is the mean and standard deviation of each quantitative predictor?**
4. **(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.**
5. **(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.**
6. **Exercise involves the Boston housing data set (MASS library).**
7. **(score =4) How many rows are in this data set? How many columns? What do the rows and columns represent?**
8. **(score =4) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.**
9. **(score =4) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.**
10. **(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.**
11. **(score =4) How many of the suburbs in this data set bound the Charles river?**