Introduction to Data Science - Spring 2019 Takehome Assignment

Due Date - Friday 27th September 2019 (Week 10) Midnight

Instructions:

Each part of the question requires code, output and a logical clear explanation.

Write the resulting model equation to the relevant questions.

You can write the answers in any word processing system (eg. Word or R-markdown) and should be submitted online via the link in vUWS. (Include a cover sheet from the Learning Guide)

Question 01

Consider "***CPU_Performance.csv***" dataset to answer the following questions.

1. Explore the given dataset and identify the attributes of CPU that have linear association with CPU performance.
2. Select the most suitable attribute of CPU that can be used to predict accurately the performance of CPU using Simple Linear Regression. Justify your choice of the attribute.
3. Model the performance of CPU using the attributes in the dataset, obtain the optimal model and Interpret your findings.
4. Carry out model diagnostics and comment your findings.
5. Suggest an appropriate transformation to overcome the issues in the previous method.

Question 02

Consider "***CPU_Performance.csv***" dataset to answer the following questions.

1. Select the most suitable attribute of CPU that can be used to predict accurately the performance of CPU using Polynomial Regression. Justify your choice of the attribute.
2. Use 10-fold cross-validation to select the optimal polynomial regression model.
3. Comment on the accuracy of the model.
4. Carry out model diagnostics and comment your findings.

Question 03

Consider "***Wine_Quality.csv***" dataset to answer the following questions.

1. Divide the dataset into Training set with 4000 observations and assign rest of the observations into Test set. [Use set.seed as 10 to generate same randomness.]
2. Build a decision tree model to predict the Quality of Wine. Hence, identify the attributes that contribute in creating a Quality Wine.
3. Comment on the performance of the model obtained.
4. Manufacturer classifies the wine quality as high if WineQuality > 6 and low otherwise. Create a new variable to categorise it as high or low and name it "Wine_Cat". Build a decision tree model to classify the Quality of Wine.

12 Sep SVM readings

5. Comment on the performance of the model obtained.

Question 04

15 Sep start

Consider "***CPU_Performance.csv***" dataset to answer the following questions.

Suppose the researcher is interested only in whether the CPU performance was high or low and not on the numerical value. He categorised Performance > 500 as high and low otherwise.

1. Design a svm model to classify the CPU Performance. Consider the different types of SVM and select the optimal one.
2. Discuss the performance of the optimal SVM model you designed.

NOTE: Unsupervised Learning Questions will be covered in the Final Exam in addition to rest of the contents covered in this course.

17 Sep Draft

20 Sep Done