301044 Data Science: Final Examination

School of Computing Engineering and Mathematics, Western Sydney University

Spring 2018: Duration 1 hour 45 minutes including reading time

Students please note that you are expected to answer the questions clearly and give the R commands used within the answer. Once completed submit the answer scripts as a pdf file via TurnItin link within vUWS site.

Question 1 (3+2+2+2+3+3+3+4 = 20)

This Question uses the data set Q1.The data relates to life in America's small cities.

The crime data (X1, X2, X3, X4, X5, X6) are for each city.

X1 = reported violent crime rate per 100,000 residents

X2 = annual police funding in $/resident
X3 = % of people 25 years+ with 4 yrs. of high school
X4 = % of 16 to 19 year-olds not in high school and not high school graduates. X5 = % of 18 to 24 year-olds in college
X6 = % of people 25 years+ with at least 4 years of college

a. Construct the matrix plot and correlation matrix and comment

b. Derive a multiple linear regression model to describe "reported violent crime rate per 100,000 residents" in terms of other numeric variables and give the estimated model.

c. Add the interactive term X2*X3 to the model in part b and derive the estimated model.

d. Add the polynomial term X2*X2 of order 2 to the model in part c and derive the best model to predict "reported violent crime rate per 100,000 residents".

e. Test the significance of each slope of the model and discuss the results.

f. Give the resultant best model and describe its accuracy.

g. Further discuss overall findings from parts a through to f.

h. List the model assumptions and test for two of these assumptions.

Question 2 (2+3+2+2+ 2 + 2+ 2 = 15)

This Question uses the data set Q1. The data relates life in America's small cities. The

crime data (X1, X2, X3, X4, X5, X6,) are for each city.

X1 = reported violent crime rate per 100,000 residents X2
= annual police funding in $/resident
X3 = % of people 25 years+ with 4 yrs. of high school
X4 = % of 16 to 19 year-olds not in highschool and not highschool graduates. X5
= % of 18 to 24 year-olds in college
X6 = % of people 25 years+ with at least 4 years of college

a. Construct and plot a Regression Tree model to classify "reported violent crime rate per 100,000 residents" in terms of other associated variables given in the data set.

b. Use the training and testing data sets to select the best tree (model) after pruning and give the corresponding measure of model accuracy. Discuss the model accuracy.

c. Give classification rules from the above decision tree and summarize the results.

d. Transform the target variable "reported violent crime rate per 100,000 residents" as a factor variable to indicate when X1>675 as High Crime Rate = Yes and when X1 <= 675 as High Crime Rate = No.

e. Construct and plot a Classification Tree to classify according to High Crime Rate in terms of other associated variables given in the data set.

f. Construct the misclassification table and give the misclassification rate and discuss the model accuracy.

g. Give classification rules from the best tree and summarize the results.

Question 3 (3+3+3+4+2 = 15)

This Question uses the data set "College" within ISLR library.

The data includes777 observations on 18 variables:

a. Use K means Clustering method and identify clusters with K=3.

b. Calculate the cluster size and their means, within cluster sum of squares by cluster, ratio "Between Sum of Squares" / "Total Sum of Squares" and describe the results.

c. Plot the principal components and color according to the k-means class.

d. Repeat b and c after scaling. Comment on the plots and compare the results with and without scaling.

e. Describe the difference between the three methods, single, average and complete in Hierarchical Clustering.

END