Data Algorithms II

Chelsea Nowlin

Assignment 1

Exercise 2, 5, 6, 8-10

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a. 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 = profit, # of employees, industry

b. 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, competition price, and ten other variables.

- Classification
- Prediction
- N = 20
- P = price, marketing budget, competition price

c. 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 = % change US market, % change British market, % change German market

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

- Generally, the more flexible the approach, the harder it is to estimate what predictors are doing
- Restrictive models are more interpretable
- Flexible approach can be used to find a nonlinear effect

- Inflexible approach can be used to interpret a regression model

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

- Parametric tests assume there are underlying distribution in the data and relies on conditions of validity to be met such as the sample following a normal distribution and the sample variances being homogenous. Non-parametric tests do not rely on a distribution and applied even if conditions of validity or not met.
- Parametric approach makes it easier to estimate f
- Parametric approach is fairly accurate in representing the relationship
- Parametric tests tend to have more statistical power and tends to have a lower p-value than the nonparametric counterpart
- Nonparametric tests tend to be more robust than parametric tests and can be considered valid in a broader range of situations of f due to lower conditions of validity.
- Disadvantages of parametric tests: not always a clear reflection of the true f, potentially can overfit the model

8. This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US.

9. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

a. Which of the predictors are quantitative, and which are qualitative?

- Mpg, cylinders, displacement, horsepower, weight, acceleration, origin, year = quantitative
- Name = qualitative

b. What is the range of each quantitative predictor? You can answer this using the range() function. range()

c. What is the mean and standard deviation of each quantitative predictor?

d. Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

e. Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

f. 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

- Weight is a good indicator of the relationship between the weight of the vehicle and the mileage it will have

10. This exercise involves the Boston housing data set.

a. To begin, load in the Boston data set. The Boston data set is part of the MASS library in R. > library(MASS) Now the data set is contained in the object Boston. > Boston Read about the data set: > ?Boston How many rows are in this data set? How many columns? What do the rows and columns represent?

- Rows = 506
- Columns = 14
- `Crim` – per capita crime rate by town.
- `Zn` – proportion of residential land zoned for lots over 25,000 sq.ft.
- `Indus` – proportion of non-retail business acres per town.
- `Chas` – Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- `Nox` – nitrogen oxides concentration (parts per 10 million).
- `Rm` – average number of rooms per dwelling.
- `Age` – proportion of owner-occupied units built prior to 1940.
- `Dis` – weighted mean of distances to five Boston employment centres.
- `Rad` – index of accessibility to radial highways.
- `Tax` – full-value property-tax rate per \$10,000.
- `Ptratio` – pupil-teacher ratio by town.
- `Black` – *1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town.
- `Lstat` – lower status of the population (percent).
- `Medv` – median value of owner-occupied homes in \$1000s.

b. Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

c. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

- Lsat and crim

d. 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.

- Yes

e. How many of the suburbs in this data set bound the Charles river? 2.4 Exercises 57

f. What is the median pupil-teacher ratio among the towns in this data set?

- Around 20

g. Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

h. In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

- More than 8 rooms, percentage is very small, suburbs with housing with 8 or more rooms tend to be newer in terms of age of the home