CSC570AG Machine Learning
Homework 4

Linear Regression

(due by the end of the day on Sunday, April 16th)

In this homework, you need to implement a linear regression model.

1. Load Boston.cvs data set. It records 14 variables for 506 neighborhoods around Boston:

`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 centers.
`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.

2. Do any of the suburbs of Boston appear to have particularly high median home values? Low crime rates? Low pupil-teacher ratios? Comment on the range of each of these variables. Which suburbs have high median home values and low pupil-teacher ratios and low crime rates?

3. How many suburbs in this data set bound the Charles river?

4. Fit a multiple regression model to predict the response variable `medv` using all other variables. For which variables the corresponding regression coefficients are likely to be significant?

5. On the basis of your response to the previous question, fit a smaller linear regression model that only uses predictors for which there is evidence of association with `medv`.

6. How well the models in 4 and 5 fit the data?

7. Try to reduce the set of predictors by making correlation and scatterplot matrices.

8. Fit even a smaller model using the predictors of your choice.

9. Investigate possible interactions between the variables in the last model. Try to find a model with R^2>77.

Write a short report providing the answers to the questions above. Your submission must consist of two text files:

- a text file, report.txt, no longer than a page,

- a text file, script.txt, with the clean history (without any drafts, errors, or debugging) of your session. Provide short comments for your commands.