

# INFO 7390–Machine Learning and Data Sciences

Northeastern University, Fall 2017

PROBLEM SET 1, DUE: OCTOBER 07, 2017

## Problem Set Rules:

1. Each student should hand in an individual problem set at the beginning of class.
2. Discussing problem sets with other students is permitted. Copying from another person or solution set is *not* permitted.
3. Late assignments will *not* be accepted. No exceptions.

1. This exercise involves the Auto data set. Make sure that the missing values have been removed from the data.
  - (a) Which of the predictors are quantitative, and which are qualitative?
  - (b) What is the range of each quantitative predictor? You can answer this using the `range()` function.
  - (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.
2. This exercise involves the Boston housing data set.
  - (a) How many rows are in this data set? How many columns? What do the rows and columns represent?
  - (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.
  - (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.
  - (e) How many of the suburbs in this data set bound the Charles river?
  - (f) What is the median pupil-teacher ratio among the towns in this data set?
  - (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.
3. This question should be answered using the Carseats data set.
- Fit a multiple regression model to predict Sales using Price, Urban, and US.
  - Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!
  - Write out the model in equation form, being careful to handle the qualitative variables properly.
  - For which of the predictors can you reject the null hypothesis  $H_0 : \beta_j = 0$ ?
  - On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.
  - How well do the models in (a) and (e) fit the data?
  - Using the model from (e), obtain 95% confidence intervals for the coefficient(s).
  - Is there evidence of outliers or high leverage observations in the model from (e)?
4. This problem involves the Boston data set. We want to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.
- For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.
  - Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis  $H_0 : \beta_j = 0$ ?
  - How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the  $x$ -axis, and the multiple regression coefficients from (b) on the  $y$ -axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the  $x$ -axis, and its coefficient estimate in the multiple linear regression model is shown on the  $y$ -axis.
  - Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor  $x$ , fit a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$