

Statistical Data Mining I

Homework 1

Due: Wednesday September 18th (11:59 pm)

Directions: Submit all source codes with write up. Please see UB Learns guidelines

- 1) (10 points **MS and MPS**) Consider the cereal dataset in UBLearn. Suppose that you are getting this data in order to build a predictive model for nutritional rating. Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed in class. Pre-process this data (e.g. transformations, elimination of outliers) and justify your choices in your write up. Submit the cleaned dataset as an *.RData file.
- 2) (10 points **MS and MPS**) Perform a multiple regression on the dataset you pre-processed in question one. The response variable is rating. Use the `lm()` function in R.
 - a) Which predictors appear to have a significant relationship to the response.
 - b) What does the coefficient variable for “sugar” suggest?
 - c) Use the `*` and `:` symbols to fit models with interactions. Are there any interactions that are significant?
- 3) (10 points **MS only**) ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the *zipcode* data. In particular, consider only the 2's and 3's for this problem, and $k = 1, 3, 5, 7, 9, 11, 13, 15$. Show both the training and the test error for each choice of k . The *zipcode* data is available in the ElemStatLearn package. Note that you do not have to divide the data into test and training as it is done for you.

```
> ls("package:ElemStatLearn")
> ?zip.test
> ?zip.train
```
- 4) (10 points **MS and MPS**) ISL textbook exercise 2.10 modified: This exercise concerns the boston housing data in the MASS library (`>library(MASS)` `>data(Boston)`).
 - a) Make pairwise scatterplots of the predictors, and describe your findings.
 - b) Are any of the predictors associated with per capita crime rate?
 - c) 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.
 - d) 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.