Statistical Data Mining 2
Homework 2

1) (20 points Modified Exercise 14.4 in ESL)
Cluster the marketing data of Table 14.1 (ESL) using a classification tree. This data is in the ISLR package, and also available on UB learns.
Specifically, generate a reference sample of the same size of the training set. This can be done in a couple of ways, e.g., (i) sample uniformly for each variable, or (ii) by randomly permuting the values within each variable independently. Build a classification tree to the training sample (class 1) and the reference sample (class 0) and describe the terminal nodes having highest estimated class 1 probability. Compare the results to the results near Table 14.1 (ESL), which were derived using PRIM.

(2) (20 points) Consider the Boston Housing Data. This data can be accessed in the MASS package (available through CRAN).

```
> library(MASS)
> data(Boston)
```

a) Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.
b) Visualize the data using the itemFrequencyPlot in the “arules” package. Apply the apriori algorithm (Do not forget to specify parameters in your write up).
c) A student is interested is a low crime area, but wants to be as close to the city as possible (as measured by “dis”). What can you advise on this matter through the mining of association rules?
d) A family is moving to the area, and has made schooling a priority. They want schools with low pupil-teacher ratios. What can you advise on this matter through the mining of association rules?

Extra Credit: Use a regression model to solve part d. Are you results comparable? Which provides an easier interpretation? When would regression be preferred, and when would association models be preferred?