Statistical Data Mining II
Homework 1

Due: Friday February 16th (11:59 pm)
30 points

Directions: Submit all source codes with write up. You must provide thorough explanations with output. See "homework guidelines" on UB learns for detailed information.

(1) (10 points) (Adopted from http://infolab.stanford.edu/~ullman/mmds/ch9.pdf exercise 9.3.1) Consider the following "utility matrix":

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

(a) Treat the utility matrix as Boolean and compute the Jaccard distance, and the cosine distance between users.

(b) Use a different discretization: treat ratings 3,4,5 as 1, and ratings 1, 2, and blank as 0. Compute the Jaccard distance and cosine distance and compare to that of part A.

(c) Normalize the matrix by subtracting from each nonblank entry the average value for its user. Using this matrix, compute the cosine distance between each pair of users.

(2) (10 points) Consider the Boston Housing Data. This data can be accessed in the ElemStatLearn package (available through CRAN).

```
> library(ElemStatLearn)
> data(boston)
> head(boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

The variables are as follows:

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 nitric 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 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
B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT % lower status of the population
MEDV Median value of owner-occupied homes in $1000's

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 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 (3 points): 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?

(3) (10 points) (Modified Exercise 14.4) Cluster the demographic data of Table 14.1 using a classification tree. Specifically, generate a reference sample the same size as the training set, by randomly permuting the values within each feature. 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.

*Note: you may use a variety or R libraries for tree construction e.g., rpart or tree Computational lab for tree building is available upon request.