Data Mining II
Homework 1

(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 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?
e) 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 in ESL) Cluster the demographic data (>data(marketing in ESL package)) 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.