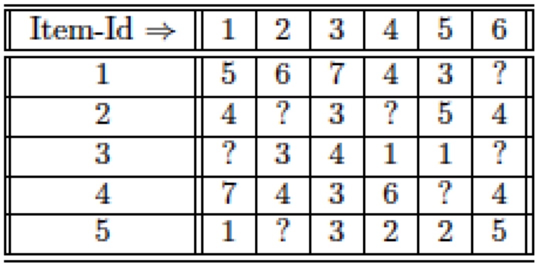
Data Mining II
Homework 1

Due: Wednesday February 27th (11:59 pm)
30 points

(1) (10 points) (Adopted from Recommender Systems, Aggarwal)
Consider the following ratings table between five users and six items.

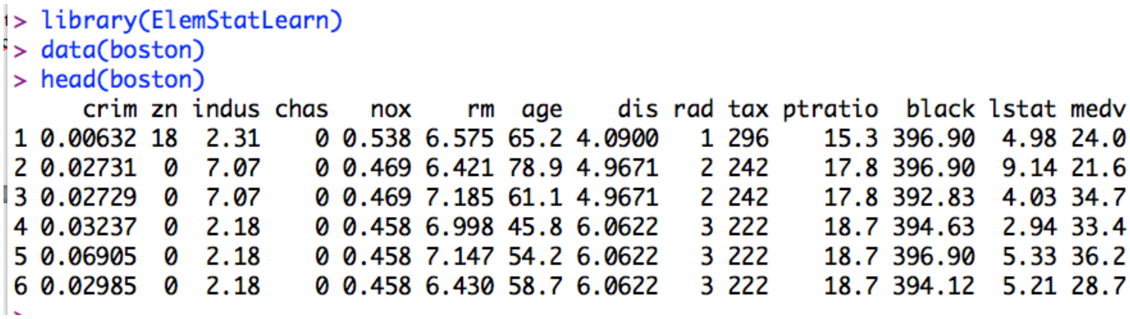


Item-Id ⇒	1	2	3	4	5	6
1	5	6	7	4	3	?
2	4	?	3	?	5	4
3	?	3	4	1	1	?
4	7	4	3	6	?	4
5	1	?	3	2	2	5

(a) Predict the values of unspecified ratings of user 2 using user-based collaborative filtering. Use the Pearson correlation with mean-centering.

(b) Predict the values of unspecified ratings of user 2 using item-based collaborative filtering algorithms. Use the adjusted cosine similarity.

(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?
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) Cluster the demographic data of Table 14.1 using a classification tree. Specifically, generate a reference sample the same size as the training set. 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.

~~~~~~

EXTRA CREDIT

Consider the MovieLense data in the “recommenderlab” package.

```
>library(recommenderlab)
>data(MovieLense)
```

Design and evaluate your own recommendation system based on the following principles:

- For each user “i” and each movie “j” they did not see, find top “k” most similar users to “i” who have seen “j” and then use them to infer the user “i” ’s rating on movie. Handle all exceptions in a reasonable way and report your strategy if you did so; e.g., if you cannot find “k” users for some movie “j”, then take all users who have seen it.

- Test the performance of your system using cross-validation. For each data set, the MovieLens database already provides a split of the initial data set into N = 5 folds. This means you will run your algorithm N times; in each step, use the training partition to make predictions for each user on all terms rated in the test partition (by that user). When you complete all N iterations, you will have a large number of user-movie pairs from the 5 test partitions on which you can evaluate the performance of your system. Measure the performance of your recommendation system.
~~~~~~