CIS 635 Data Mining

Homework 1

Description

In this first assignment you are to consider some scenarios where data mining is to be done and describe the process of preparing the data to be mined. You are to identify the important details of the data and describe them. Additionally, you will install R studio and do perform some fundamental commands. Write or copy/paste your answers to the handin sheet and submit on BB.

Instructions

Part 1 – general data mining

1. Assume that you work for a bank and are given the task of predicting which customers who are applying for a loan are likely to pay the loan back. You can gather information about new loan applications and previous loans (including those who paid back the loan and those who did not). Name four attributes that you would be interested in for each customer.
2. This question asks you to think about clustering. Considering an average kitchen, identify 4 clusters of objects. Name them according to their utility.

   For example, if you are asked to identify 4 clusters in a drug store, you might choose pain relievers, make-up (cosmetics), life enhancers (vitamins) and oral care (toothpaste, etc)

part 2 – getting to know R

1. setup and install
   a) create a folder on the computer you will be using for this course (cis635 is a good name).
   b) create subfolders in the main folder for homeworks and projects.
   c) install R from https://www.r-project.org/
   d) change the working directory to the folder where your homework1 files are
2. vectors
   a) create a vector named iq of 100 elements of N(15,2) (normal with a mean of 15 and std dev of 2) data.
   b) add 5 to every element of iq
   c) calculate the mean and std dev of iq

d) display the first 10 elements of iq using an index (e.g. iq[1:10])

e) display the elements 11 through 20 of iq using an index

3. tables

a) create a matrix called tmp with the following values:

-9	0	1
5	3	-2

b) read the file hw01data.txt in R and assign it to the variable x

c) calculate the mean of column 3

d) sum up the values in column 2, rows 5 through 15

e) calculate the average rate of all states where the average income is > 90000, example below

- idx=x[,3]>90000
- mean(x[idx,2])

f) calculate the average avgInc for all states where the fico is < 700

- using the method above (using idx)
- using a for loop and if statement

4. functions

a) write a function that takes a table and returns an vector that contains the maximum value of column 2