CS 445 Homework 1

Due date:

March 16th, 2018 by 17:00.

Deliverables:

The following problems should be completed and submitted by the due date and time specified above. You will lose 10% for each day they are late and assignment submissions will not be accepted after two days. Solutions (along with your source code) must be typed (**no hand written solutions**) and should include clear, detailed explanations. You should print your solutions and submit them **by hand to your instructor**.

Problems:

Given below the descriptions of two problems, pick your favorite method and solve **only one** of them to get 100 points. If you solve both, you will receive extra 20 points.

a) This problem uses the Boston Housing dataset from the CMU StatLib Library **partially**. It concerns the prices of housing in Boston suburbs. Original dataset contains thirteen features, however this problem focuses only on **three features** to predict the value of housing. The data description is given below and the data are in the file **TrainingDat1.txt**.

i) First column of the given data contains the average number of rooms per dwelling,
ii) second column contains the weighted distances to five Boston employment centres,
iii) third column is for the percentages of lower status of the population, and
iv) the last column lists the median value of owner-occupied homes in $1000's.

For instance, the first training example (first row of data) has the following values:

6.5630 2.8470 5.68 32.50

In this example, there are 6.5630 rooms (including kitchen, bathroom(s), bedroom(s), etc.) in average per dwelling in a building, group of buildings or community; the weighted distance from this location to five Boston employment centres is 2.8470 miles; 5.68 percent of the population living here place in lower status of the overall population and finally the corresponding median value of these owner-occupied homes is $32500.

Given the training set, **predict the value of a house** with 6 rooms, 5 miles as its weighted distance to five Boston employment centres and about 8 percent of the people in its neighbourhood placed in lower status of the population.

b) In this problem, the original data-set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

We will use the data **partially** to come up with a way **predicting the risk factor** given some particular characteristics. The data description is given below and data are given in the file **TrainingDat2.txt**.

i) First entry (of a row) is the risk factor,
ii) the second entry is the make of the car,
iii) third one is for the body-style of the car, and
iv) the last one is for the price of the car in dollars.

Given this data-set, make a **prediction for the risk factor** of a sedan peugot brand car with a price of $9000.