Homework 1
You are required to type the solution for the homework
Submit your solution through D2L Dropbox labeled "Homework 1".

1. Given the following training data where each data is described by two features and a class label:
   D1 (8, 6, negative)
   D2 (11, 8, positive)
   D3 (1, 9, positive)
   D4 (3, 2, positive)
   D5 (8, 9, positive)
   D6 (8, 5, negative)
   D7 (12, 10, positive)
   D8 (7, 2, negative)
   D9 (6, 4, negative)
   D10 (5, 1, negative)
   D11 (1, 3, positive)
   D12 (2, 6, positive)

   (1) Build a KD-Tree based on this data
   (2) Using KNN (K=3), for Query data 1 Q1(1, 5), what should be the class label for this data? Show how you use the KD tree built above to find the K nearest neighbors and derive the class label.
   (3) Repeat question (2) for Query data 2: Q2(10, 5)

2. Compute the distance between the two customers shown below:

	Age	Income	Purchased Before	Category	State of resident	Education level	Marital status
Customer1	29	54000	Yes	Entry level	TN	College	Not married
Customer2	40	112000	No	Silver level	KY	Graduate	married

   From the overall data, it is known that the mean and standard deviation of age is m=36, s=5; the mean and standard deviation of income is m=68000, s=12000, customers are defined by these 5 categories: Entry, level 1, level 2, silver, gold; Education levels: high school, college, and graduate;

3. Apply Levenshtein's edit distance two compute the edit distance between the two strings:
   S1: ATRLTL
   S2: ARATEK
   Show the table of distance values computed using dynamic programming.

4. We have collected a data set of 14 data objects representing 14 different mushrooms. Each mushroom is labeled by domain experts whether it is edible or poisonous. We would like to learn a decision tree that will help us determine for any mushroom we may find in the future whether it is edible or poisonous. The three attributes chosen for describing the mushrooms, together with the possible values for each attribute are shown below:

   - Cap Shape: bell, flat, or convex
   - Cap Color: brown, grey
   - Odor: almond, spicy, foul

Here is the data set:
Data:
Object Cap Shape Cap color Odor class

D1	bell	brown	almond	edible
D2	flat	grey	almond	edible
D3	convex	grey	spicy	poisonous
D4	bell	brown	almond	edible
D5	flat	grey	almond	edible
D6	flat	grey	spicy	edible
D7	convex	grey	almond	edible
D8	bell	brown	almond	edible
D9	convex	brown	foul	poisonous
D10	bell	brown	spicy	edible
D11	bell	grey	almond	edible
D12	convex	grey	spicy	poisonous
D13	flat	brown	almond	edible
D14	flat	grey	foul	poisonous

(1) Use *information gain* as the attribute selection criterion to build a decision tree for the data. Show the computation involved in the attribute selection that results in a partial decision tree with 2 levels (root is considered as level 1). Draw the partial decision tree and label each node in the 2nd level of tree with a class label.

(2) Classify the following three mushrooms using the decision tree learned from the previous step:

Data:

Object	**Cap Shape**	**Cap color**	**Odor**	**class**
q1	bell	grey	almond	??
q2	convex	grey	spicy	??
q2	flat	brown	foul	??

(3) Is the best attribute selected above considered statistically significant using the Chi Squared test for predicting the class labels? (assuming the confidence level chosen is 5%) Show the computation to support your answer.