Lebanese University
Faculty of information
Section 1

Final Exam: Machine Learning

Time: 120 minutes
Dr. Abbas RAMMAL

Problem 1: K-means & Hierarchical clustering

Use the K-means algorithm and Euclidean distance to cluster the 8 data points given into K = 3 clusters. The distance matrix based on the Euclidean distance is given in Table 1.

The coordinates of the data points are:

$x_1 = (2,8)$ $x_2 = (2,5)$ $x_3 = (1,2)$ $x_4 = (5,8)$

$x_5 = (7,3)$ $x_6 = (6,4)$ $x_7 = (8,4)$ $x_8 = (4,7)$

Table 1: Distance matrix for training dataset

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0	3.000	6.082	3.000	7.071	5.656	7.211	2.236
x_2		0	3.162	4.242	5.385	4.123	6.082	2.828
x_3			0	7.211	6.082	5.385	7.280	5.831
x_4				0	5.385	4.123	5.000	1.414
x_5					0	1.414	1.414	5.000
x_6						0	2.000	3.605
x_7							0	5.000
x_8								0

1. Suppose you are initializing K-means method, that is, you initialize the cluster centers to K randomly chosen data points. Let's assume that points x_3, x_4 and x_6 were chosen. Perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids.

   At the end of this epoch show:

a) The new clusters (i.e. the examples belonging to each cluster)
b) The centers of the new clusters
c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
d) How many more iterations are needed to converge? Draw the result for each epoch.

2. Use single and complete link agglomerative clustering to group the data described by the previous distance matrix. Show the dendrograms.

Problem 2: Naive Bayes Classifier

1. In a medical study, 100 patients all fell into one of three classes: Pneumonia, Flu, or Healthy. The following database indicates how many patients in each class had fever and headache. Consider a patient with a fever but no headache.

a. What values would a Bayes' optimal classifier assign to the three diagnoses? (A Bayes' optimal classifier doesn't make any independence assumptions about the evidence variables.) Again, your answers for this question need not sum to 1.

b. What values would a naive Bayes classifier assign to the three possible diagnoses? Show your work. (For this question, the three values need not sum to 1. Recall that the naive Baye's classifier drops the denominator because it is the same for all three classes.)

c. What probability would a Bayes optimal classifier assign to the proposition that a patient has Pneumonia. Show your work. (For this question, the three values should sum to 1.)

d. What probability would a naive Bayes classifier assign to the proposition that a patient has Pneumonia. Show your work. (For this question, the three values should sum to 1.)

Pneumonia		
Fever	Headache	count
T	T	5
T	F	9
F	T	4
F	F	1
total:		19

Flu		
Fever	Headache	count
T	T	8
T	F	6
F	T	2
F	F	2
total:		20

Healthy		
Fever	Headache	count
T	T	2
T	F	3
F	T	7
F	F	50
total:		71

Problem 2: Simple Linear Regression

The yield y of a chemical process is a random variable whose value is considered to be a linear function of the temperature x. The following data of corresponding values of x and y is found:

Temperature	0	25	50	75	100
Chemical process	14	38	54	76	95

The average and standard deviation of temperature and yield are

$$\bar{x} = 50, \qquad S_{xx} = 39.528, \qquad S_{xy} = 31.662, \qquad \bar{y} = 55.4, \qquad S_{yy} = 31.667$$

a) Draw a scatter diagram of the data. Does a simple linear regression model seem appropriate here?
b) Fit the simple linear regression model using the method of least squares. Find an estimate of $\hat{\sigma}$
c) Estimate the standard errors of β_0 and β_1.
d) Test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ using the analysis of variance procedure with $\alpha = 0.05$ and $t_{3,0.025} = 3.182$.
e) Find a 95% confidence interval of β_0 and β_1.