Problem 1 (20pts)

Consider an HMM with two possible states, “R” and “G” (for “regulatory” and “gene” sequences respectively). Each state emits one character, chosen from the alphabet
{A,C,G,T}.

The transition probabilities of this HMM are:

aRG = aGR = 1/4

aRR = aGG = 3/4

The emission probabilities are:

eR (A)= eR (C)= eR (G) = eR (T)=1/4

eG (A)= eG (T)= 2/10 and eG(C)=eG(G)=3/10

Assume that the initial state of the HMM is “R” or “G” with equal probabilities. Given a sequence O = ACGT and an HMM path Q = RGGR, calculate the probability Pr(O, Q) of the sequence and the path.

Problem 2 (30pts)

Consider an HMM with two possible states, “N” and “D” (for “noncoding” and “coding” sequences respectively). Each state emits one character, chosen from the alphabet {A,C,G,T}.
The transition probabilities of this HMM are:

aND = aDN = 0.1
aNN = aDD = 0.9

The emission probabilities are:

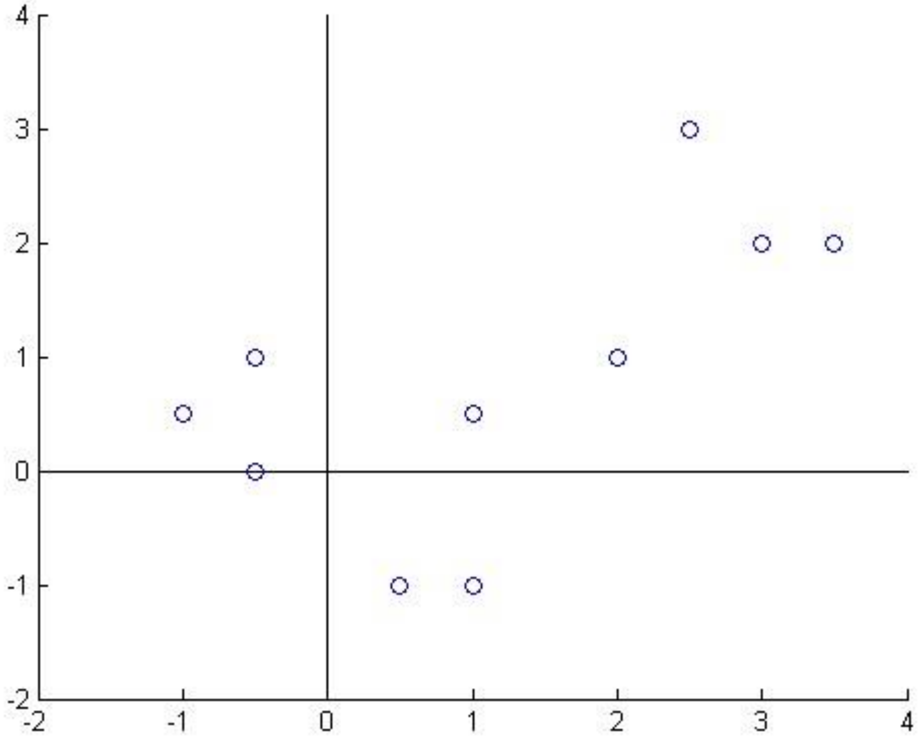
eN (A)= eN (C)= eN (G)= eN (T)=1/4
eD (A)= eD (T)= 3/10 and eD (C)= eD (G)=2/10

Assume that the initial state of the HMM is “N” with probability 0.75 and “D” with probability 0.25 respectively.
Now, you are given that the sequence emitted by this HMM is O = ATTC. Show the calculations of the Viterbi algorithm to derive the most likely sequence of states, i.e., Q =q1 q2 q3 q4 , where each qi is either “N” or “D”, that maximizes Pr(O, Q). Your answer should include
(i) The dynamic programming calculations as a table. (15 points).
(ii) The probability Pr(O, Q) you computed for the best Q (5 points).
(iii) The best Q (10 points).

Problem 3 (20pts)

Consider the ten data points (in 2D) listed below. A plot of the ten points is also shown below, for your convenience.

x	y
1	0.5
2.5	3
2	1
3	2
3.5	2
-0.5	0
-0.5	1
-1	0.5
1	-1
0.5	-1



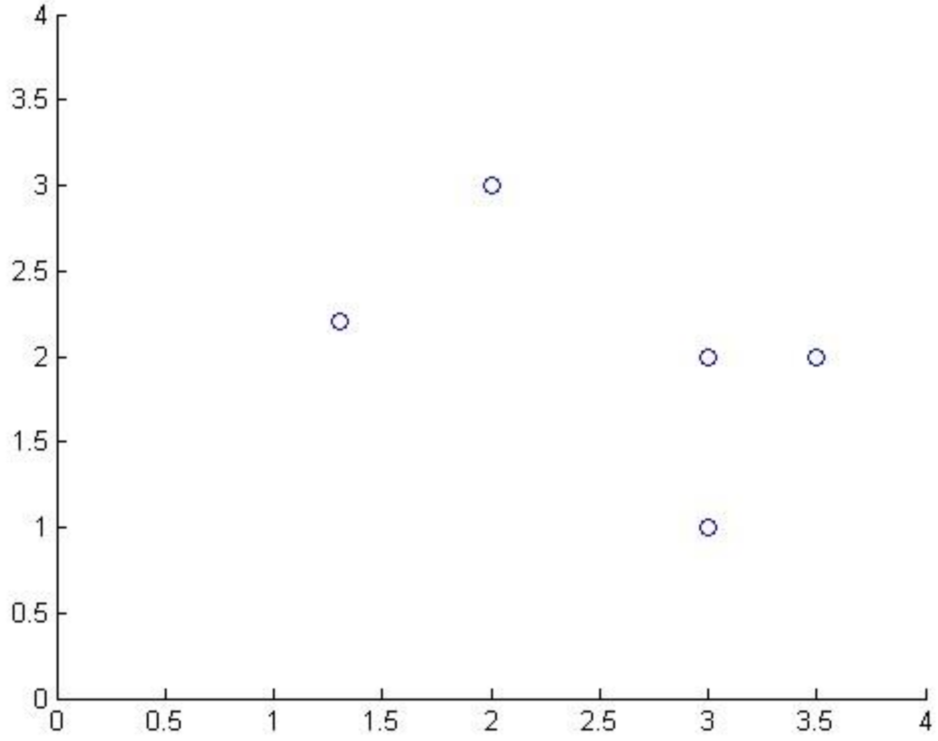
Show the steps (and final result) of K-means clustering for this data set, with K=3 and with initial cluster centers set to (0,0), (2,3) and (1.5,1). For each step, show the current cluster centers, cluster assignment of each point, and the distance calculations you used in making this cluster assignment.

Problem 4 (20pts)

Consider the five points listed in the table below. A plot of these five points is shown below, for your convenience. Show the steps and final result of the Hierarchical Clustering algorithm, as discussed in class, applied to this data set. Define the distance between two clusters as the minimum distance between a pair of points, one in each cluster. In each step, show the pairwise distance matrix between the current set of clusters, as well as the pair of clusters chosen for merging.

Table:

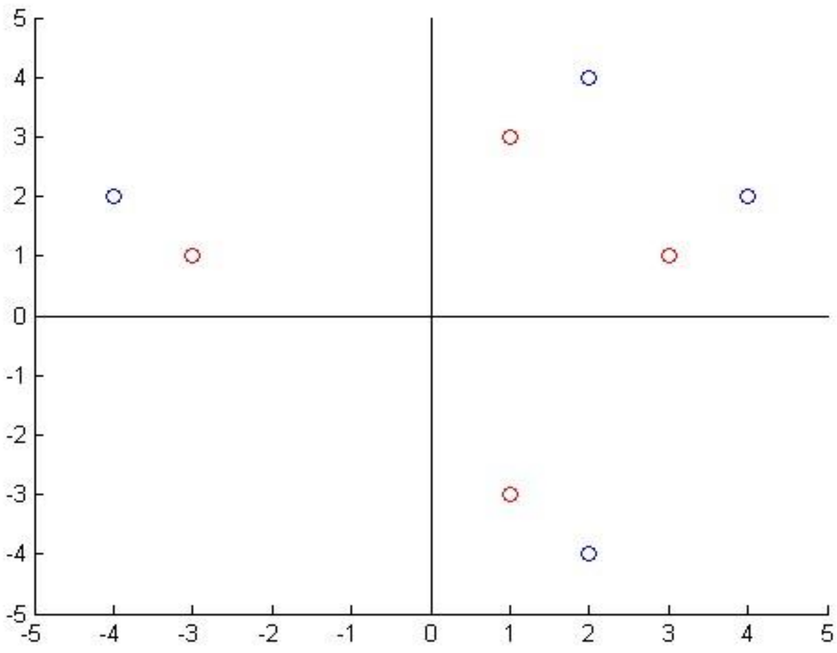
	X	Y
Point 1	1.3	2.2
Point 2	2	3
Point 3	3	1
Point 4	3	2
Point 5	3.5	2



Problem 5 (20pts)
Consider the following eight points on a 2D plane:

x	y	Label
3	1	+
1	3	+
-3	1	+
1	-3	+
2	4	-
4	2	-
-4	2	-
2	-4	-

Four of these points are labeled positive, and four are labeled negative. We want to learn a linear classifier for this data set. Note that in the 2D "input space", there is no straight line separating the positive and negative points:



Your goal is to map the 2-D input space to a 3D feature space such that the positives and negatives are separable by a plane. Find a mapping (x, y) → (x, y, z) where the first two coordinates remain unchanged and the third coordinate z is a function of x and y, such that the positives and negatives are separable by a plane.