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.

Pr(O, Q) = Pr(O|Q)Pr(Q)
= Pr(A|R)Pr(C|G)Pr(G|G)Pr(T|R)Pr(R|initial)Pr(G|R)Pr(G|G)Pr(R|G)
$= \frac{1}{4} \times \frac{3}{10} \times \frac{3}{10} \times \frac{1}{4} \times \frac{1}{2} \times \frac{1}{4} \times \frac{3}{4} \times \frac{1}{4}$
= 0.000131836

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).

i)

	A	T	T	C
N	$0.75 \times 0.25 = 0.1875$	$\text{Max}\{(0.1875 \times 0.9 \times 0.25), (0.075 \times 0.1 \times 0.25)\} = 0.042$	$\text{Max}\{(0.042 \times 0.9 \times 0.25), (0.0203 \times 0.1 \times 0.25)\} = 0.0095$	$\text{Max}\{(0.0095 \times 0.9 \times 0.25), (0.0055 \times 0.1 \times 0.25)\} = 0.0021$
D	$0.25 \times 0.3 = 0.075$	$\text{Max}\{(0.1875 \times 0.1 \times 0.3), (0.075 \times 0.9 \times 0.3)\} = 0.0203$	$\text{Max}\{(0.042 \times 0.1 \times 0.3), (0.0203 \times 0.9 \times 0.3)\} = 0.0055$	$\text{Max}\{(0.0095 \times 0.1 \times 0.2), (0.0055 \times 0.9 \times 0.2)\} = 0.0010$

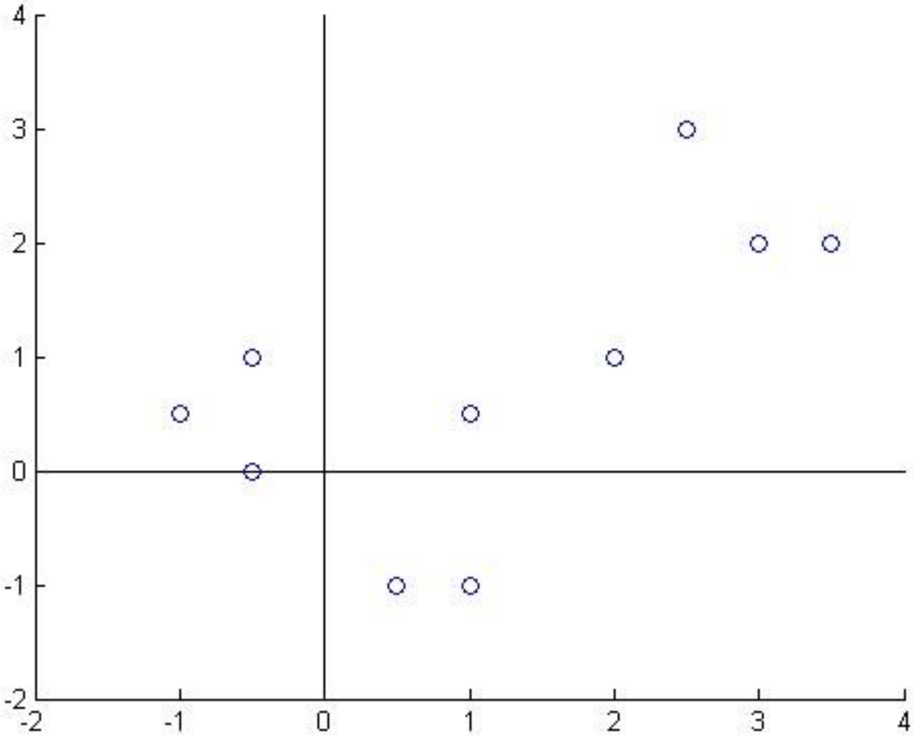
ii) The probability Pr(O, Q) for the best Q would be 0.0021.

iii) The best Q would be NNNN.

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.

K = 3
P1 = (0, 0)
P2 = (2, 3)
P3 = (1.5, 1)
Then, we form 3 clusters by assigning each point to its closet centroid.
By using Euclidean distance, we can get their distance.

	P1 = (0, 0)	P2 = (2, 3)	P3 = (1.5, 1)
(1, 0.5)	1.12	2.69	0.71
(2.5, 3)	3.91	0.5	2.24
(2, 1)	2.24	2	0.5
(3, 2)	3.61	1.41	1.80
(3.5, 2)	4.03	1.80	2.24
(-0.5, 0)	0.5	3.91	2.24
(-0.5, 1)	1.12	3.20	2
(-1, 0.5)	1.12	3.91	2.55
(1, -1)	1.41	4.12	2.06
(0.5, -1)	1.12	4.27	2.24

Cluster 1: (-0.5, 0), (-0.5, 1), (-1, 0.5), (1, -1), (0.5, -1)
Cluster 2: (2.5, 3), (3, 2), (3.5, 2)
Cluster 3: (1, 0.5), (2, 1)

Right now, we need to Re-compute the centroids (mean point) of each cluster.

P1 = ((-0.5-0.5-1+1+0.5)/5, (0+1+0.5-1-1)/5) = (-0.1, -0.1)
P2 = ((2.5+3+3.5)/3, (3+2+2)/3) = (3, 2.33)
P3 = ((1+2)/2, (0.5+1)/2) = (1.5, 0.75)

Repeat the previous step to get their distance.

	P1 = (-0.1, -0.1)	P2 = (3, 2.33)	P3 = (1.5, 0.75)
(1, 0.5)	1.25	2.71	0.56
(2.5, 3)	4.05	0.84	2.46
(2, 1)	2.37	1.66	0.56
(3, 2)	3.74	0.33	1.95
(3.5, 2)	4.17	0.60	2.36
(-0.5, 0)	0.41	4.20	2.14
(-0.5, 1)	1.17	3.74	2.02
(-1, 0.5)	1.08	4.40	2.51
(1, -1)	1.42	3.88	1.82
(0.5, -1)	1.08	4.16	2.02

Cluster 1: (-0.5, 0), (-0.5, 1), (-1, 0.5), (1, -1), (0.5, -1)
Cluster 2: (2.5, 3), (3, 2), (3.5, 2)
Cluster 3: (1, 0.5), (2, 1)

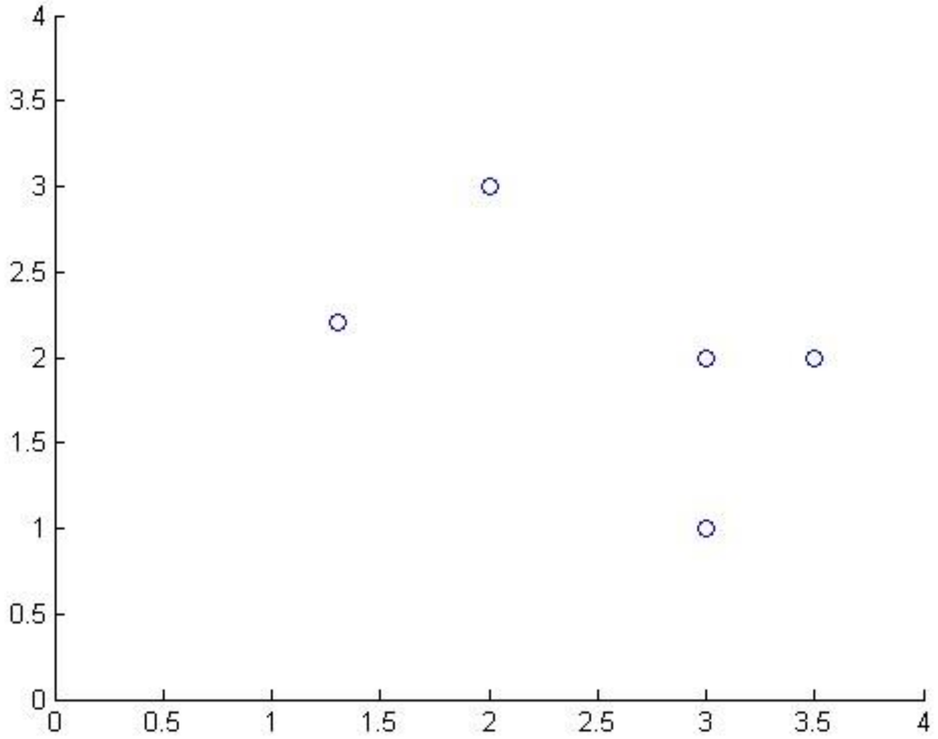
We found the result is same with previous steps. Therefore, the convergence criterion is satisfied. The loop is ended.

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



The pairwise distances are same, so we take only the lower triangular matrix.

we can compute the distance matrix using the Euclidean distance formula.

Now, we can start clustering by taking the minimum of the distances and then combining them:

	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	0				
Point 2	1.06	0			
Point 3	2.08	2.24	0		
Point 4	1.71	1.41	1	0	
Point 5	2.21	1.80	1.12	0.5	0

Minimum distance is with Point 4 and point 5, so we can combine Point 45 by taking the minimum of the distance of Point 4 and Point 5 with the other points as:

For point 1, we have min(dis(point1, point4), dis(point1, point5))
= min(1.71, 2.21)
= 1.71
For point 2, we have min(dis(point2, point4), dis(point2, point5))
= min(1.41, 1.80)
= 1.41
For point 3, we have min(dis(point3, point4), dis(point3, point5))
= min(1, 1.12)
= 1
The result matrix we get will be:

	Point 45	Point 1	Point 2	Point 3
Point 45	0			
Point 1	1.71	0		
Point 2	1.41	1.06	0	
Point 3	1	2.08	2.24	0

Now, we find that Point 3 and point 45 get the minimum distance in this matrix, so we can combine point 45 and point 3.

For point 1, we have min(dis(Point45, 1), dis(point3, 1))
= min(1.71, 2.08)
= 1.71
For point 2, we have min(dis(Point45, 2), dis(point3, 2))
= min(1.41, 2.24)
= 1.41
The result matrix we get will be:

	Point 345	Point 1	Point 2
Point 345	0		
Point 1	1.71	0	
Point 2	1.41	1.06	0

Now, we find that Point 1 and point 2 get the minimum distance in this matrix, so we can combine point 1 and point 2.

Now, we can find the distance of point 12 from Point 345:

for point 345, we have min(dis(Point345, 1), dis(point345, 2))
= min(1.71, 1.41)
= 1.41

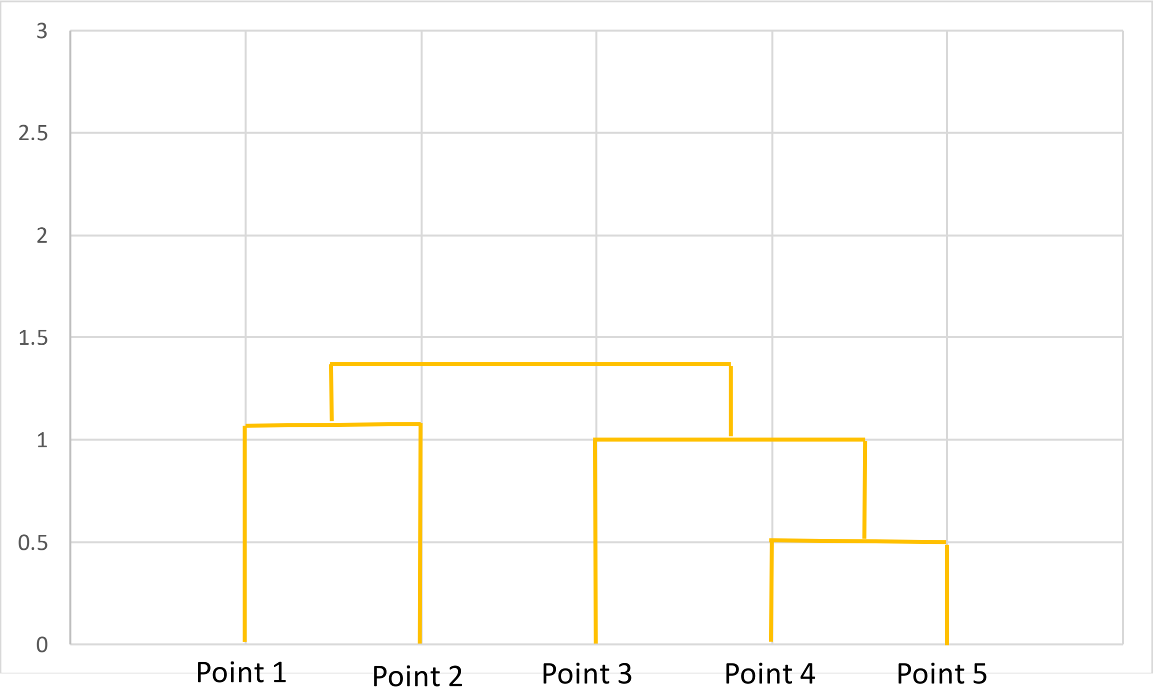
The result matrix will be:

	Point 12	Point 345
Point 12	0	
Point 345	1.41	0

Right now, the minimum distance is 1.41 from Point 12 to Point 345.
The result matrix will be:

	Point 12345
Point 12345	0

Therefore, we will have the dendrogram as follows:



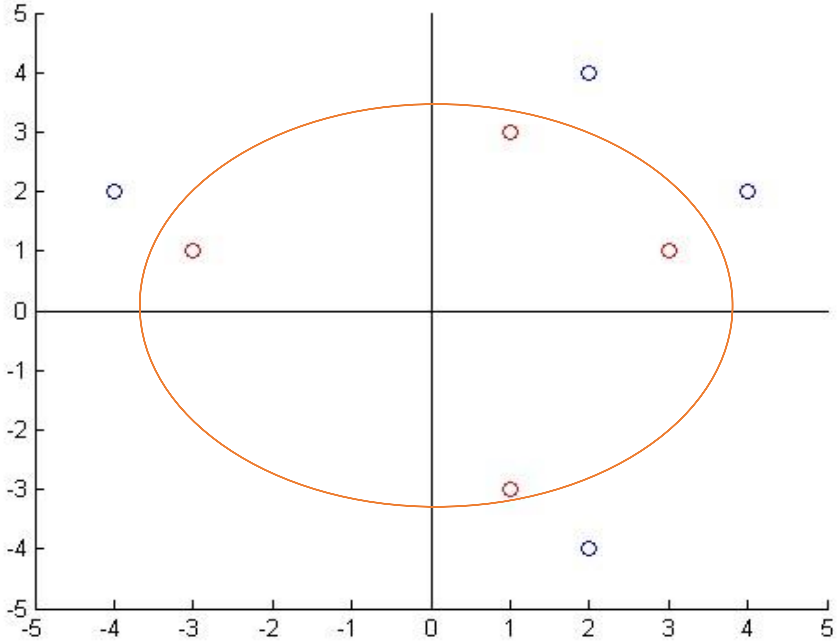
Therefore, we can find that we have cluster 1(Point 1, Point 2) and cluster 2(Point 345).

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.

We can mapping (x, y) to (x, y, x^2+y^2)