Problem 1

1. True or False: Transition probability and emission probability are the only two parameters in an HMM model.
2. Of the following two probability expression, which is joint probability, and which is marginal probability: P (X, Y) and P(X)
3. True or False: The running time of a Viterbi algorithm is O(MN) where M is the length of the sequence and N is the number of states.
4. What's difference between classification and regression in terms of their label?
5. True or False: For point a = (x1, y1) and b= (x2, y2), the Euclidian distance is defined as d (a, b) = sqrt((x1-x2)^2+(y1-y2)^2)
6. True or False: In agglomerative clustering, we always merge the two closest clusters.
7. True or False: Labeled data in a classification task is always linear separable.
8. True or False: P(A, B) = P(A|B)P(B)
9. True or False: Viterbi Algorithm is used to estimate the hidden state give a sequence and an HMM.
10. True or False: In K-means clustering, using different K values, it's possible for us to get the same prediction for an unlabeled data point.

HMM
Problem 2

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:

$$a_{RG} = a_{GR} = \frac{1}{5}$$

$$a_{RR} = a_{GG} = \frac{4}{5}$$

The **emission** probabilities of this HMM are:

$$e_R(A) = e_R(C) = e_R(G) = e_R(T) = \frac{1}{4}$$

$$e_G(A) = e_G(T) = \frac{1}{8}$$

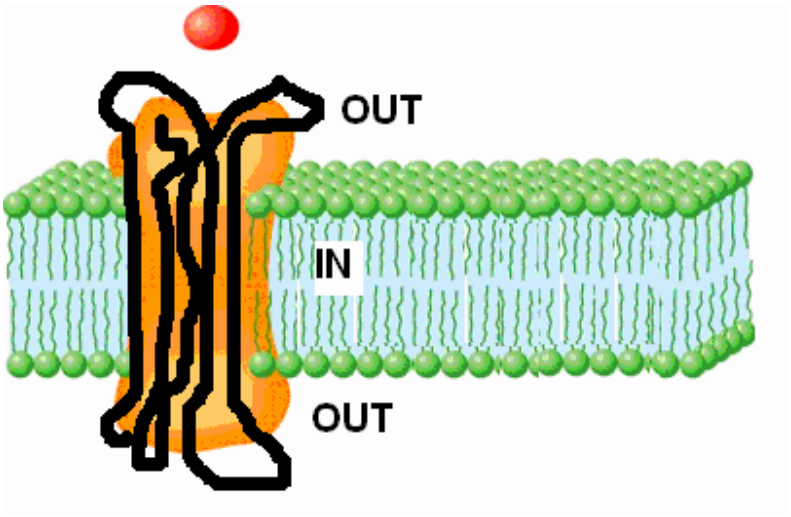
$$e_G(C) = e_G(G) = \frac{3}{8}$$

Assume that the **initial state** of the HMM is “R” or “G” with equal probabilities.

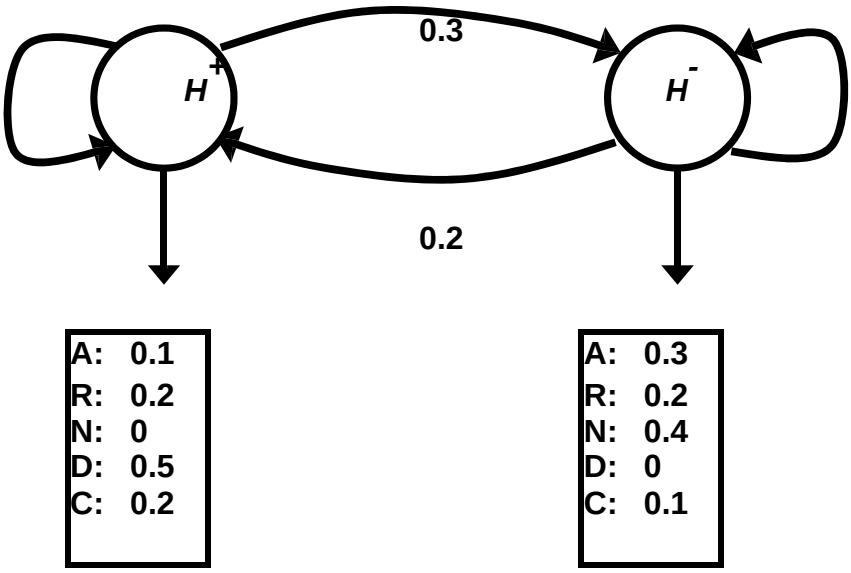
Given a sequence $S = AGT$ and an HMM path $\pi = RGR$, calculate the probability $Pr(S, \pi)$ of the sequence and the path, and the marginal probability $\Pr(S)$ of the observation sequence.

Problem 3

Consider an odorant receptor, i.e. a protein molecule that sticks in the cell membrane, extending in both the interior and exterior of the cell. This is visualized in the picture below, where the fat curved line represents the folded protein.



The floating of this molecule in the cell membrane is caused by parts of the molecule that 'likes' to be in the membrane (hydrophylic) and parts that definitely not like this (hydrophobic). Therefore, we consider that a part of the molecule can be in two states: H^+ = hydrophylic, and H^- = hydrophobic.
We employ a Hidden Markov model to estimate the hydrophobic and hydrophylic parts of the molecule, represented as:



Suppose that we have the following protein sequence: **NARNRDCCRN** Determine the most likely hidden sequence of hydrophobicity-states over the molecule using the *Viterbi algorithm*.

Your answer should include
(i) the dynamic programming calculations as a table.
(ii) the probability you computed for the most likely sequence of hidden states.
(iii) the most likely hidden sequence.

Classification
Problem 4.

Consider the following points on a 2-D plane:

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

1) Now using k-nearest neighbor, what would you classify the following three data points if k = 1:

x	y	Label
-3	0	?
3	3	?
0	2	?

You don't need to justify your answer, but you might want to plot all the data points to help you derive the answer.

2) Again, using the k-nearest neighbor, would you get a different if k=3? If so, what will be the new label?

x	y	Label
-3	0	?
3	3	?
0	2	?

Again, you don't need to justify your answer, but you might want to plot all the data points to help you derive the answer.

Problem 5.

Consider the following points on a 2-D plane:

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

Four of these points are labeled positive, and four are labeled negative. Find a linear classifier for these data points. (It can be *any* linear classifier.)

Your answer should be a *linear* function of x and y that evaluates to a positive number for all points of one label and to a negative number for all points of the other label.

Problem 6.

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	-

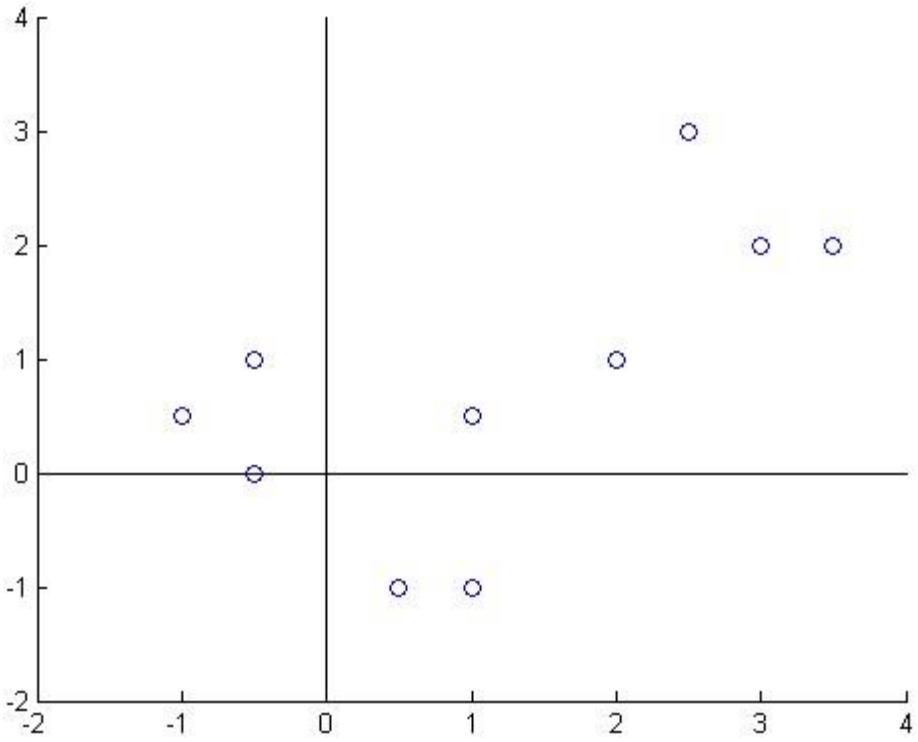
Four of these points are labeled positive, and four are labeled negative. We want to learn a linear classifier for this data set. Note that in the 2D "input space", there is no straight line separating the positive and negative points, which means that these data points are not linearly separable.

Your goal is to map the 2D input space to a 3D feature space such that the positives and negatives are separable by a plane. Find two different mappings $(x, y) \rightarrow (x, y, z)$ 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.

Clustering

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



Problem 7.

Show the steps (and final result) of the K-means clustering for this data set, with K=2 and with initial cluster centers set to (0,0), (2,2). 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 8.

A. Show the steps and final result of the Hierarchical Clustering algorithm applied to this data set. Define the distance between two clusters as the maximum 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.

B. Show the steps and final result of the Hierarchical Clustering algorithm 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.

Regression
Problem 9.

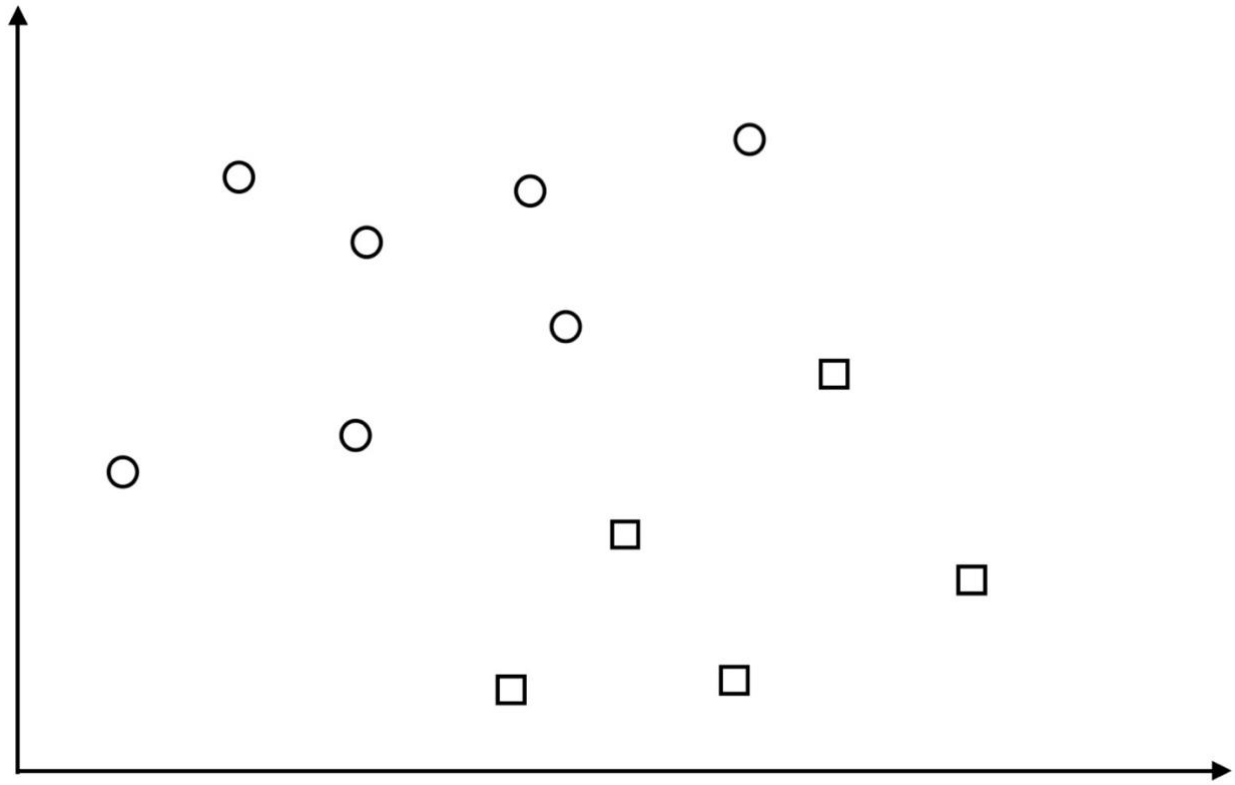
Consider linear regression on a set of N data points with 2D input vectors ($x_i = [x_{i,1}, x_{i,2}]$) and continuous labels (y_i). We hope to find a linear function $f(x_i) = w_1*x_{i,1} + w_2*x_{i,2}$ to fit label y_i.

A. Write down the sum of squared predicted error of this linear regression problem where your true label is y_i, and the prediction is defined by $f(x_i)$.
B. Now formulate this linear regression problem as an optimization problem. Hint: minimize the sum of squared predicted error defined above.
C. Derive the solution (w_1 and w_2). You can also show your answer using matrices and vectors.

SVM
Problem 10

In SVM, we want to find a line to separate the label with the largest margin.

1) For the following data points where circle and square are two classes, please draw the ***two*** support vectors for the circle class and square class.



2) Now based on your support vector, draw the line that you will use to separate the two classes with the largest margin.

3) Once you have the three lines, show us what the “margin” is in this SVM.

Assembly
Problem 11

Suppose you are given a string: ATTTGGGCATTTGGGC

a) Enumerate all 4-mers from this string.

b) Create a De Bruijn graph of these 4-mers that may be used to reconstruct the original string.
c) Reconstruct a string that covers the above 4-mers by constructing a Eulerian path.

d) Is the reconstructed string identical to the original one? If not, can you explain why?

Problem 12

Suppose you are given a string: ATTTGGGCATTTGGGC

a) Enumerate all 4-mers from this string.

b) Create an overlap graph of these 4-mers with overlap size >=1.
c) Reconstruct a string using the graph constructed in b).

d) Create an overlap graph of these 4-mers with overlap size >=2.
e) Reconstruct a string using the graph constructed in d).

f) Create an overlap graph of these 4-mers with overlap size >=3.
g) Reconstruct a string using the graph constructed in f).

h) Are the reconstructed strings in c), e), and g) identical to the original string? Can you explain why?