CS-482 MACHINE LEARNING

HOMEWORK QUESTIONS

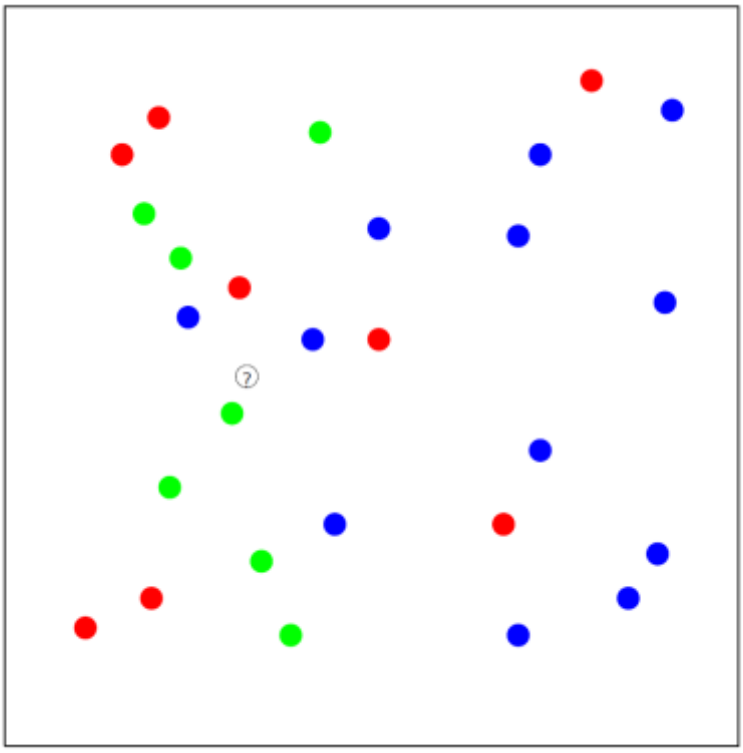
CHAPTER 2

1. Compute Minkowski distance (p=2) between data at row 5 and row 6 for the following data set using only those predictors whose values are numbers. (not categorical data). The row index begins at 1.

Table 2-1. Example data about customers

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

2. In this problem we have circles of three different colors and we want to classify a new circle (indicated in black) as one of these colors. The red, blue and green circles are illustrated below along with an unknown circle which we want to classify.



a. Classify the unknown circle as red, green or blue using a single nearest neighbor. Justify your answer by drawing circle(s) about the unknown point. Explain.
b. If possible, classify uniquely the unknown circle as red, green or blue using KNN with k = 2. Justify your answer by drawing circle(s) about the unknown point. If you are unable to uniquely classify it, explain why.
c. If possible, classify uniquely the unknown circle as red, green or blue using KNN with k = 3. Justify answer by drawing circle(s) about the unknown point. If you are unable to uniquely classify it, explain why.
d. If possible, classify uniquely the unknown circle as red, green or blue using KNN with k = 3. Justify answer by drawing circle(s) about the unknown point. If you are unable to uniquely classify it, explain why.

3. In this problem we are given the location of each object and the type of object (triangle, circle or square) along with an identifying name (e.g., T1, C1, S1). The data is tabulated below. The goal is to classify the object centered at the point (2, -1) as a triangle, circle or rectangle using k nearest neighbors (KNN).

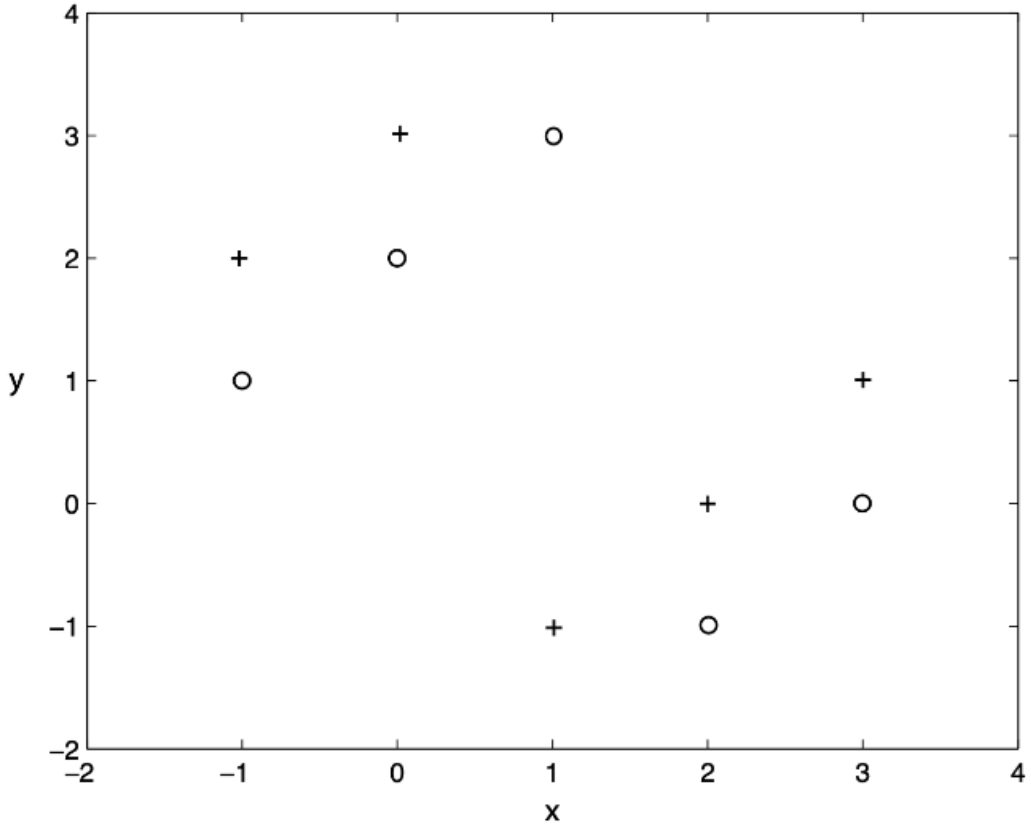
   Shape ID & Location
   Triangles: T1: (-3,-2) T2: (3,3) T3: (0,3)
   Circles: C1: (3,2) C2: (-2,-1)
   Squares: S1: (7,4) S2: (4,5) S3: (3,0)

(a) Fill in the table below which gives the Euclidean distance from the point (2,-1) to each of the points in the data set. Leave your answer in terms of the square root. Recall that to find the Euclidean distance between two points $(x_1; y_1)$ and $(x_2; y_2)$ we use the formula below.

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}\,.$$

distance	T1: (-3,-2) $\sqrt{(-3-2)^2+(-2+1)^2} = \sqrt{25+1} = \sqrt{26}$	T2: (3,3)	T3: (0,3)
distance	C1: (3,2)	C2: (-2,-1)	
distance	S1: (7,4)	S2: (4,5)	S3: (3,0)

4. Consider K-NN using Euclidean distance on the following data set (each point belongs to one of two classes: + and◦).



a) What is the leave one out cross validation error when using 1-NN?
b) Which of the following values of k leads to the minimum leave-one-out cross validation error: 3, 5 or 9? What is the error for that k? (If there is a tie, please elaborate)

5. Consider k-fold cross-validation. Let's consider the tradeoffs of larger or smaller k (the number of folds). Please select one of the multiple choice options. With a higher number of folds, the estimated error will be, on average

a) Higher b) Lower c) Same d) Can't tell
Explain the reason for your choice

6. In a KNN classification problem, assume that the distance measure is not explicitly specified to you. Instead, you are given a "black box" where you input a set of instances P_1, P_2, .. P_n and a new example Q, and the black box

outputs the nearest neighbor of Q, say P_i and its corresponding class label C_i. Is it possible to construct a k-NN classification algorithm (w.r.t the unknown distance metrics) based on this black box alone? If so, how and if not, why not?