DDA2001: Assignment 5

1. The assignment is due at Monday **11:59 pm, May 9, 2022**.
2. Please submit your solution in **PDF** form. **Any other forms of solution will not be accepted and will be graded as 0.** Please leave enough time to make sure you have uploaded your solution as requirement before due.
3. If you submit the assignment late, you will get 0 for this assignment. **No excuses will be accepted for any late submission.**
4. **Please make sure that your file could be downloaded successfully from BB after uploading your solution file.**

---

1 Multiple Choice [20 points]

1. **[5 points]** What is correct about unsupervised learning?
   A. Number of clusters is always known.
   B. Label of data points is explicitly stated.
   C. Neither label nor number of clusters is known.
   D. None of the above.

   Solution: C. □

2. **[5 points]** What is the advantage of K-means clustering?
   A. Being independent on the initial values.
   B. Being able to deal with outliers.
   C. K can be chosen randomly.
   D. Guarantees convergence.

   Solution: D. □

3. **[5 points]** Which of the following are true?
   A. 5-NN is always more accurate than 1-NN.
   B. 5-NN is more robust to outliers than 1-NN.
   C. K-NN will always give a linear decision boundary.
   D. Logistic Regression will always give a linear decision boundary.

   Solution: B/D (B, D or BD all taken as a correct answer.) □

4. **[5 points]** Consider a labeled data set with 3 classes: $1, 2, 3$ and the numbers of data points for three classes are $N_1 = 16$, $N_2 = 32$ and $N_3 = 64$, respectively. Now we apply the K-Nearest neighbor (KNN) classifier and set $K = N_1 + N_2 + N_3$. Then what is the label of a new data point?
   A. 1

B. 2
C. 3
D. No enough information.

Solution: C. Since $K = N_A + N_B + N_C$, the new point belongs to the class with the most number of data points. □

2 Computation Questions [80 points]

1. **[9 points]** We are given the following labeled data set with points of three different classes:

Points	x_1	x_2	class
A	11	3	1
B	10	0	1
C	4	-5	2
D	5	-4	2
E	-3	5	3
F	-1	7	3

We perform a KNN classification.
(1) Classify the new point $(4, 3)$ with $K = 3$ using the L_1-norm as the distance measure.
(2) Classify the new point $(4, 3)$ with $K = 3$ using the L_2-norm as the distance measure.
(3) Classify the new point $(4, 3)$ with $K = 3$ using the L_∞-norm as the distance measure.
Hint: If there is a tie, then we can randomly choose from the tied classes. The L_1, L_2, and L_∞-norms between two points $\boldsymbol{x} = \begin{bmatrix} x_1, & x_2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} y_1, & y_2 \end{bmatrix}$ are defined as

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = |x_1 - y_1| + |x_2 - y_2|$$
$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$
$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max\{|x_1 - y_1|, |x_2 - y_2|\}.$$

Solution: The distances between the new point and other points, its 3-nearest neighbor and its classification are:

	A	B	C	D	E	F	3-nn	class
L_1-norm	7	9	8	8	9	9	A, C, D	2
L_2-norm	7	6.7082	8	7.0711	7.2801	6.4031	F, B, A	1
L_∞-norm	7	6	8	7	7	5	F, B, $A/D/E$,	1/2/3

□

2. **[12 points]**Consider the following labeled training set in the 2-dimensional space with two labels $-$ and $+$:

x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$

Figure 1 shows a visualization of the data.

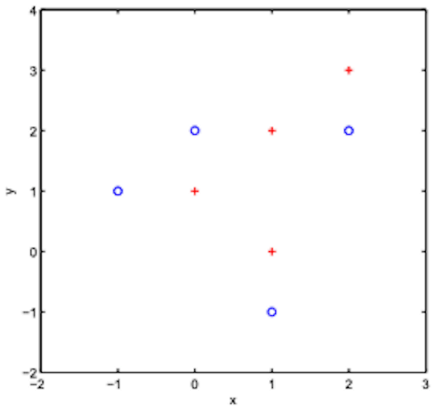

Figure 1: Dataset for Problem 2.

(a). What is the prediction of the 3-NN classifier at a new point $(1, 1)$?

(b). What is the prediction of the 5-NN classifier at a new point $(1, 1)$?

(c). What is the prediction of the 7-NN classifier at a new point $(1, 1)$?

(d). Which of the following statements are true for K-NN classifiers (select all answers that are correct).

A. The classification accuracy is better with larger values of K.

B. The decision boundary is smoother with smaller values of K.

C. K-NN is a type of instance-based (sample-based) learning.

D. K-NN does not require an explicit training step.

E. The decision boundary is linear.

Solution:

(a). $+$;

(b). $+$;

(c). $-$;

(d). CD.

□

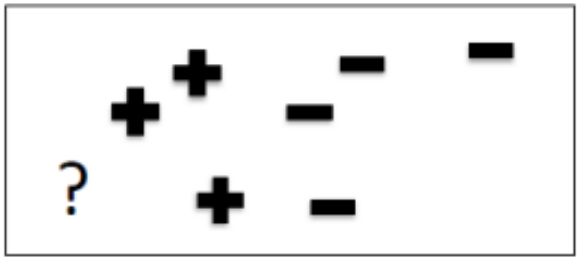

Figure 2: Labeled dataset

3. **[7 points]**Given the following labeled dataset (Figure 2)

   Suppose we use KNN to classify the point "?". For what (minimal) value of k will be the query point "?" be labeled negative ('-')? Ties are broken at random, and you need to avoid them. Give 0 as your answer if this is impossible.

   Solution:

   The answer is 7. Analysis can be outlined as follows.

   - If $k = 1$, then the cloest neigbors would be: $\{+\}$, and the majority rule would classify this query point as $+$.
   - If $k = 2$, then the cloest neigbors would be: $\{+, +\}$, and the majority rule would classify this query point as $+$.
   - If $k = 3$, then the cloest neigbors would be: $\{+, +, +\}$, and the majority rule would classify this query point as $+$.
   - If $k = 4$, then the cloest neigbors would be: $\{+, +, +, -\}$, and the majority rule would classify this query point as $+$.
   - If $k = 5$, then the cloest neigbors would be: $\{+, +, +, -, -\}$, and the majority rule would classify this query point as $+$.
   - If $k = 6$, then the cloest neigbors would be: $\{+, +, +, -, -, -\}$, and this query point would be arbitrarily classified.
   - If $k = 7$, then the cloest neigbors would be: $\{+, +, +, -, -, -, -\}$, and the majority rule would classify this query point as $-$.

   □

4. **[12 points]**For the logistic regression model, we know that the the likelihood function of y_i is

$$f\left(y_i \mid x_i, \theta, b\right) = \left(\frac{1}{1+\exp\left(-\theta x_i - b\right)}\right)^{y_i} \left(\frac{\exp\left(-\theta x_i - b\right)}{1+\exp\left(-\theta x_i - b\right)}\right)^{1-y_i},$$

   where $y_i \in \{0, 1\}$. Then given samples $\{(x_i, y_i)\}_{i=1}^m$, we fit the logistic regression model by maximizing the log-likelihood function

$$l(\theta, b) = \log \Pi_{i=1}^m f\left(y_i \mid x_i, \theta, b\right),$$

   with respect to (θ, b). Show that $l(\theta, b)$ is concave for $(\theta, b) \in \mathbb{R}^2$.

Solution:

$$l(\theta, b) = \sum_{i=1}^{m} [-y_i \log(1 + \exp(-\theta x_i - b)) + (1 - y_i)(-\theta x_i - b) - (1 - y_i) \log(1 + \exp(-\theta x_i - b))]$$
$$= \sum_{i=1}^{m} [(1 - y_i)(-\theta x_i - b) - \log(1 + \exp(-\theta x_i - b))] .$$

The first term $(1 - y_i)(-\theta x_i - b)$ is obviously concave in θ and b.
We let $(\theta, b) = (\theta_0, b_0) + t\boldsymbol{e}$ where $\boldsymbol{e} = (\theta_1, b_1)$ is any unit vector and (θ_0, b_0) is any point. Then we define

$$\begin{aligned} h(t) &= \log(1 + \exp(-\theta x_i - b)) \\ &= \log(1 + \exp(-\theta_0 x_i - b_0) + t(-\theta_1 x_i - b_1)) \\ &= \log(1 + \exp(C_1 + tC_2)). \end{aligned}$$

If follows that

$$h''(t) = \frac{C_2^2}{(1 + \exp(C_1 + tC_2))^2} \exp(C_1 + tC_2) \geq 0.$$

Therefore the second term is $-h(t)$ and is also concave. □

5. **[10 points]**We consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x_1, x_2) = (x_1 - 1)^2 + x_2^2.$$

Suppose we use gradient descent to minimize the function f. Starting at $\boldsymbol{x}^{(0)} = [2, 1]$, conduct one iteration of gradient descent with step size $\alpha = 1$ and find $\boldsymbol{x}^{(1)}$.

Solution: The partial derivative of $f(x_1, x_2)$ is

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2(x_1 - 1)$$

and

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = 2x_2.$$

So $x_1^{(1)} = x_1^{(0)} - \alpha 2(x_1^{(0)} - 1) = 0$ and $x_2^{(1)} = x_2^{(0)} - \alpha 2 x_2^{(0)} = -1$. Thus $\boldsymbol{x}^{(0)} = [0, -1]$. □

6. **[12 points]**Suppose we would like to use the K-means algorithm and L2-norm distance (Euclidian distance) to cluster the 8 data points given in Figure 3 below into K = 3 clusters. The L2-norm distance between points $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ is $d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. The coordinates of the data points are:

$$x^1 = (2, 8) \quad x^2 = (2, 5) \quad x^3 = (1, 2) \quad x^4 = (5, 8)$$
$$x^5 = (7, 3) \quad x^6 = (6, 4) \quad x^7 = (8, 4) \quad x^8 = (4, 7)$$

(a) Let's assume that points x^3, x^4 and x^6 were chosen as the initial cluster center. Perform one iteration of the K-means algorithm. Report the coordinates of the updated cluster centers and assign each data point according to the new cluster centers.

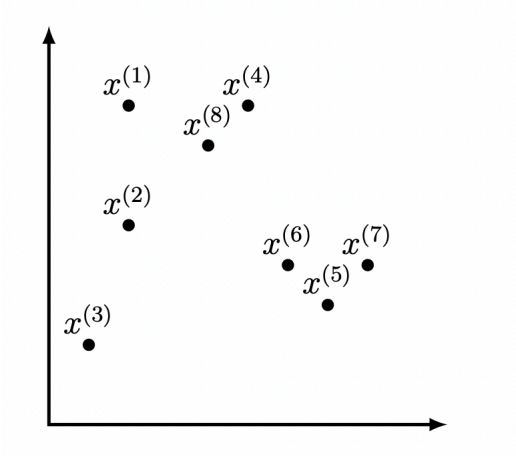

Figure 3: Data points

(b) Denote c_1, c_2, c_3 as the center for 3 clusters, and for each data point x^i, $s^i \in \{1, 2, 3\}$ is the cluster to which x^i belongs (i.e., the cluster center c_{s^i} is the nearest to x^i). Calculate and compare the loss function $l(c_1, c_2, c_3) = \frac{1}{8}\sum_{i=1}^{8}(d(x^i, c_{s^i}))^2$ before and after the first iteration.

Solution:

(a) i. Initialization: $c^1 = x^3 = (2, 8)$, $c^2 = x^4 = (5, 8)$, $c^3 = x^6 = (6, 4)$

ii. Cluster assignment: For each point x_i, assign it to cluster s^i by:

$$s^i = \arg\min_{j=1,2,3} d(x^i, c_j)$$

Then we can obtain that

$$s^2 = s^3 = 1$$
$$s^1 = s^4 = s^8 = 2$$
$$s^5 = s^6 = s^7 = 3$$

iii. Center adjustment: Then we adjust each center c_j by:

$$c_j = \frac{1}{|\{i : s^i = j\}|}\sum_{i:s^i=j} x^i$$

Therefore, we have the new centers:

$$c^1 = \frac{1}{2}(x^2 + x^3) = \begin{pmatrix} 1.5 \\ 3.5 \end{pmatrix}$$
$$c^2 = \frac{1}{3}(x^1 + x^4 + x^8) = \begin{pmatrix} 3.67 \\ 7.67 \end{pmatrix}$$
$$c^3 = \frac{1}{3}(x^5 + x^6 + x^7) = \begin{pmatrix} 7 \\ 3.67 \end{pmatrix}$$

Therefore the points are assigned to 3 clusters: $\{x_2, x_3\}$, $\{x_1, x_4, x_8\}$, $\{x_5, x_6, x_7\}$.

(b) Denote the loss function before as L_0 and after as L_1

Before: $L_0 = \frac{1}{8}(9 + 10 + 2 + 4 + 2) = 3.375$

After: $L_1 = \frac{1}{8}(2.9 + 2.5 + 2.5 + 1.9 + 0.44 + 1.11 + 1.11 + 0.56) = 1.625$

□

7. **[18 points]**Suppose you are given the following (x, y) pairs. You will simulate the K-means algorithm to identify TWO clusters in the data.

Table 1: Data pairs (x, y)

Data #	1	2	3	4	5	6	7	8	9	10
x	1.90	1.76	2.32	2.31	1.14	5.02	5.74	2.25	4.71	3.17
y	0.97	0.84	1.63	2.09	2.11	3.02	3.84	3.47	3.60	4.96

Suppose you are given initial cluster center as $\{cluster1 : \#1\}$, $\{cluster2 : \#10\}$ – the first data point is used as the first cluster center and the 10-th as the second cluster center. Please simulate the K-means (K=2) algorithm for ONE iteration (useing Euclidean distance). What are the cluster assignments after ONE iteration? Repeat the procedure until it stops and report what are the cluster assignments until convergence? (Fill in the table below)

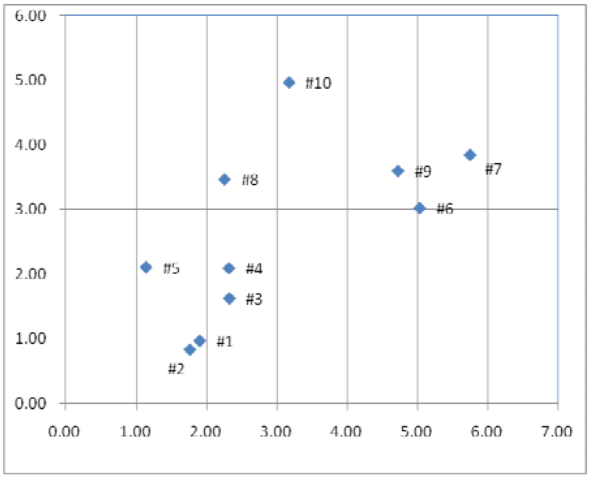

Figure 4: Point pairs (x, y)

Table 2: K-Means Iteration

Data #	1	2	3	4	5	6	7	8	9	10
Cluster Assignment after One Iteration										
Cluster Assignment after convergence										

Solution:

Table 3: KMeans Iteration

Data #	1	2	3	4	5	6	7	8	9	10
Cluster Assignment after One Iteration	1	1	1	1	1	2	2	2	2	2
Cluster Assignment after convergence	1	1	1	1	1	2	2	2	2	2

□