K-Means Clustering

Dataset :

- Point 1 : (2,10)
- Point 2 : (2,5)
- Point 3 : (8,4)
- Point 4 : (5,8)
- Point 5 : (7,5)
- Point 6 : (6,4)

Initial Centroid :

- Centroid 1 : (2,10)
- Centroid 2 : (5,8)

Iteration 1

Point	Coordinates	Distance to centroid 1	Distance to centroid 2	Assigned Cluster
1	(2,10)	$\sqrt{(2-2)^2+(10-10)^2}$ = 0	$\sqrt{(2-5)^2+(10-8)^2}$ = 3.61	1
2	(2,5)	$\sqrt{(2-2)^2+(5-10)^2}$ = 5	$\sqrt{(2-5)^2+(5-8)^2}$ = 4.24	2
3	(8,4)	$\sqrt{(8-2)^2+(4-10)^2}$ = 8.49	$\sqrt{(8-5)^2+(4-8)^2}$ = 5	2
4	(5,8)	$\sqrt{(5-2)^2+(8-10)^2}$ = 3.61	$\sqrt{(5-5)^2+(8-8)^2}$ = 0	2
5	(7,5)	$\sqrt{(7-2)^2+(5-10)^2}$ = 7.07	$\sqrt{(7-5)^2+(5-8)^2}$ = 3.61	2
6	(6,4)	$\sqrt{(6-2)^2+(4-10)^2}$ = 7.21	$\sqrt{(6-5)^2+(4-8)^2}$ = 4.12	2

Date :

Step 2:

Update centroids

Cluster 1 (Point 1)

∴ New centroid 1 = (2, 10) [remain the same]

Cluster 2 (Point 2, 3, 4, 5, 6, 8).

∴ New centroid 2 = $\left(\frac{2+8+5+7+6}{5}, \frac{5+4+8+5+4}{5}\right)$

= (5.6, 5.2)

Iteration 2

Point	Coordinates	Distance to centroid 1	Distance to centroid 2	Assigned Cluster
1	(2, 10)	$\sqrt{(2-2)^2+(10-10)^2}$ = 0	$\sqrt{(2-5.6)^2+(10-5.2)^2}$ = 6	1
2	(2, 5)	$\sqrt{(2-2)^2+(5-10)^2}$ = 5	$\sqrt{(2-5.6)^2+(5-5.2)^2}$ = 3.61	2
3	(8, 4)	$\sqrt{(8-2)^2+(4-10)^2}$ = 8.49	$\sqrt{(8-5.6)^2+(4-5.2)^2}$ = 2.68	2
4	(5, 8)	$\sqrt{(5-2)^2+(8-10)^2}$ = 3.61	$\sqrt{(5-5.6)^2+(8-5.2)^2}$ = 2.86	2
5	(7, 5)	$\sqrt{(7-2)^2+(8-10)^2}$ = 7.07	$\sqrt{(7-5.6)^2+(5-5.2)^2}$ = 1.414	2
6	(6, 4)	$\sqrt{(6-2)^2+(8-10)^2}$ = 7.21	$\sqrt{(6-5.6)^2+(4-5.2)^2}$ = 1.26	2

Step 2: Update centroid:

Cluster 1 (Point 1)

∴ New centroid 1 = (2, 10) [Unchanged]

Cluster 2 (2, 3, 4, 5, 6)

∴ New centroid 2 = $\left(\frac{2+8+5+7+6}{5}, \frac{5+4+8+5+4}{5}\right)$

= (5.6, 5.2) ⓐ [Unchanged]

Date : / /

After 2 iterations : Cluster 1 : Point 1 with centroid (2,10)

and cluster 2 : Point (2,3,4,5,6) with centroid (7, 4.33)

The algorithm has converged as the centroid remain unchanged after the 2nd iteration.

CT-1

Manhattan Distance →

$$L_1 = |x_2 - x_1| + |y_2 - y_1|$$

Euclidean Distance →

$$L_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Here, suppose the value of K is 3.

Suppose the points A(2,1), B(1,3), C(7,4), D(7,3) are given. A and B from cluster I, where C and D from cluster II.

Apply the KNN algorithm using K=3 and Euclidean distance to determine which cluster (I or II) the query point E(8,2) belongs to. Also Illustrate a graph to visualize the clusters.

Date :

Points	x_1	Y_1	class
A	2	1	cluster I
B	1	3	cluster I
C	7	4	cluster II
D	7	3	cluster II

We know, Euclidean distance

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

Given, E (8,2)

For A, $\sqrt{(8-2)^2 + (2-1)^2} = 6.08$

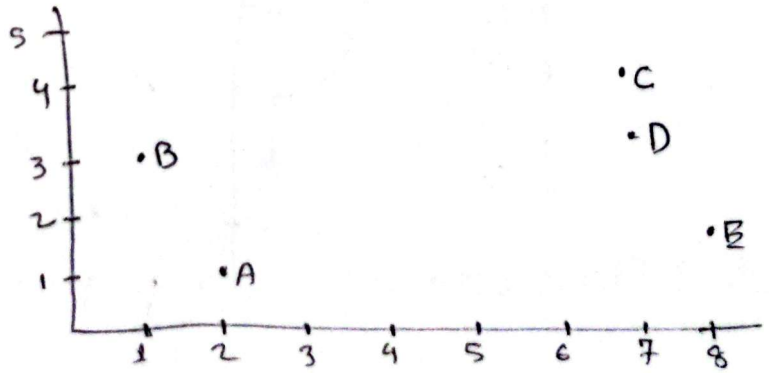
B, $\sqrt{(8-1)^2 + (2-3)^2} = 7.07$

C, $\sqrt{(8-7)^2 + (2-4)^2} = 2.24$

D, $\sqrt{(8-7)^2 + (2-3)^2} = 1.41$

Sorting the distance, D (1.41)
C (2.24)
A (6.08)
B (7.07)

∴ three nearest neighbors are D, C, A and 2 of them belongs to the cluster B, we can classify E(8,2) into cluster II.



Date : / /

For Manhattan

	x_1	y_1	class
A	2	1	clusterI
B	1	3	clusterI
C	7	4	clusterII
D	7	3	clusterII

∴ We know Manhattan distance

$$= |x_2 - x_1| + |y_2 - y_1|$$

E = (3,2)

∴ We cakulate the distance from E(3,2) to all given points.

For A = $|3-2| + |2-1| = 2$

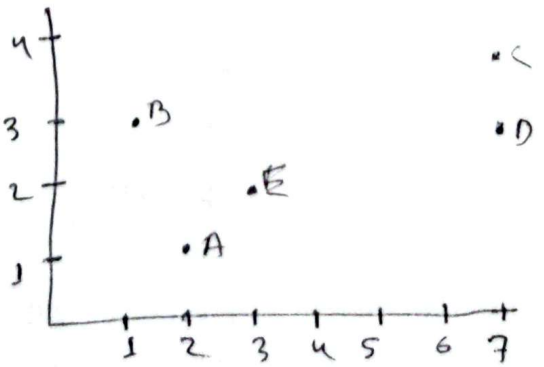
B = $|3-1| + |2-3| = 3$

C = $|3-7| + |4-2| = 6$

D = $|3-7| + |2-3| = 5$

∴ Sorting the distance.

A (2)
B (3)
D (5)
C (6)



∴ E(3,2) = clusterI.

Date :

CT-2

* Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

* Pencision = $\frac{TP}{TP+FP}$

* Sensitivity/ Recall = $\frac{TP}{TP+FN}$

* Specivity = $\frac{TN}{TN+FP}$

∴ F-measure/ F-score = $\frac{2 \times (\text{pencision} \times \text{recall})}{(\text{Pencision} + \text{recall})}$

Instance	1	2	3	4	5	6	7	8	9	10	11	12
Actual classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Positive ← Predicted classification
Negative

Date :

CT-2

Instance	x	Actual Label (x+3)%2	Predicted Label (x+3)%2	Result
1	13	1	16%2 = 0	FN
2	13+1=14	0	1	FP
3	13+7=20	0	1	FP
4	13+3=16	0	1	FP
5	13+4=17	1	0	FN
6	13+2=15	1	0	FN
7	13+1=14	0	1	FP
8	13+7=20	0	1	FP

$$\therefore \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{0+0}{0+0+5+3} = 0$$

$$\therefore \text{Precision} = \frac{TP}{TP+FP} = \frac{0}{0+5} = 0$$

$$\therefore \text{Sensitivity/Recall} = \frac{TP}{TP+FN} = \frac{0}{0+3} = 0$$

$$\therefore \text{Specifity} = \frac{TN}{TN+FP} = \frac{0}{0+5} = 0$$

Here,

TN = 0

FP = 5

FN = 3

TP = 0

Date :

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

confusion KNN = $\begin{bmatrix} 35 & 2 \\ 8 & 25 \end{bmatrix}$,

TP FN
FP TN

TP = 35
FP = 8

$\therefore$ accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

$= \frac{35+25}{35+25+8+2} = 0.85$

$\therefore$ Precision = $\frac{TP}{TP+FP} = \frac{35}{35+8} = 0.813$

$\therefore$ Recall = $\frac{TP}{TP+FN} = \frac{35}{35+2} = 0.945$

$\therefore$ Specifity = $\frac{TN}{TN+FP} = \frac{2}{2+8} = 0.2$

Date : / /

Explain the concept of Dataset and label and feature.

Dataset: A dataset in ML is a collection of data used to train and test a model. It contains multiple examples. (also called instances)

[A collection of samples]

Features: Features are the input variable that describe each example in the dataset. The model learns pattern from these features.

[The input data]

Label: The Label (also called the target variable) is the value or category that the model is trying to predict.

[The expected Output]

Date :

Explain the weight and bias update process in linear regression.

Ans: In linear regression, we predict output(y) using:

$$y = wx + b$$

w = weight (how much influence x has)
b = Bias (shift the line up/down)

How do we update weight and bias?

We use Gradient Descent to adjust w and b to minimize error. The rules are.

$$w = w - \alpha \frac{\partial L}{\partial w}$$

$$b = b - \alpha \frac{\partial L}{\partial b}$$

Steps:

1. Compute the error (difference between predicted and actual values
2. Find Gradients to loss w.r.t. w & b.
3. Update w and b using the above formula.
4. Repeat until the error is minimized.

Date : / /

Explain the role of derivatives in the gradient descent algorithm.

1. Gradient = Derivative → shows how fast a function changes.

2. Direction of Change →

   * If the derivative is positive, the function is increasing - move left
   * If the derivative is negative, the function is decreasing → move right.
   * If the derivative is zero, we may have reached a minimum.

3. Update Rule: Adjust parameters using:

$$\theta := \theta - \alpha \nabla J(\theta)$$

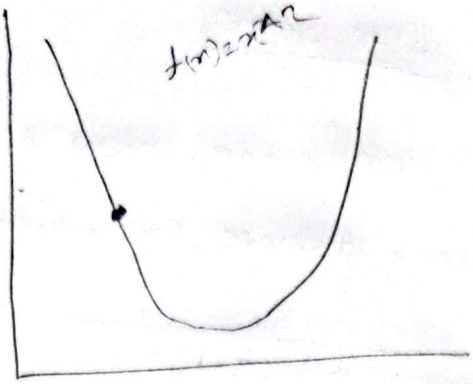
where α is the learning rate

4. Repeat Until Convergence → keep updating until loss is minimized.

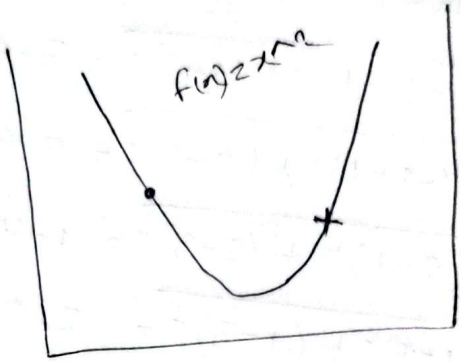
Derivatives tell us how to adjust model parameters to reduce error.

Date :

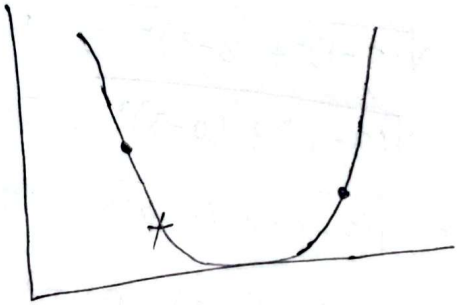
1



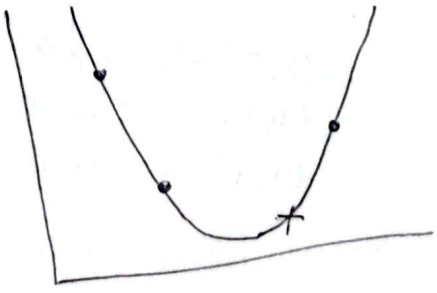
2



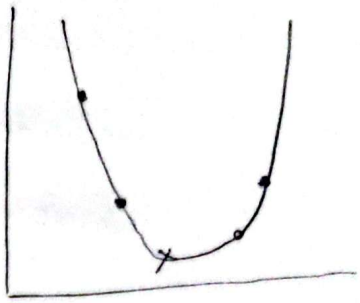
3



4



5



6

- previous

Date : / /

Mid Solve

(2)

We need to classify the new data point (1,3) using Euclidean distance, (3-nearest Neighbors).

~~For B, $\sqrt{(\quad)^2 + (3-1)^2}$~~

For (-2,-1), For A, $\sqrt{(-2-1)^2 + (1-3)^2} = 3.61$

For (-1,0), For A, $\sqrt{(-1-1)^2 + (0-3)^2} = 3.61$

For (0,2), $\sqrt{(0-1)^2 + (2-3)^2} = 1.41$

For (1,1), $\sqrt{(1-1)^2 + (1-3)^2} = 2.00$

For, (2,3), $\sqrt{(2-1)^2 + (3-3)^2} = 1.00$

For (3,0), $\sqrt{(3-1)^2 + (0-3)^2} = 3.61$

For (3,2), $\sqrt{(3-1)^2 + (2-3)^2} = 2.23$

K=3. Here, 3 closet points are,

(2,3) → class B
(0,2) → class B
(1,1) → class A

∴ the new (1,3) is classified as B

Impact of Increasing K :

1. If we increase (K=5 or 7), the model will consider more neighbors.
2. This may lead to smoothen decision boundaries but can also increase bias if the dataset is small.
3. If K is too large compared to the dataset size, the model may favor the majority class. ignoring nearby points.

Date :

Explain the difference between linear regression and multivariate linear regression for predicting relationships between independent and dependent variables. Use two different dataset examples.

* Difference between Linear Regression and Multivariate Linear Regression:-

1. Linear Regression

- Involve one Independent variable (X) and one dependent variable (Y).
- The relationship : $y = mX + b$

Example Dataset for Linear Regression:

(X)	Y
1	50
2	55
3	65
4	70

2. Multivariate Linear Regression:

- Involves multiple independent variables, predicting one dependent variable (Y).

$$Y = W_0 + W_1X_1 + W_2X_2 + \cdots + W_pX_p + b$$

Example Dataset for Multivariate Linear Regression:

Square Feet (X_1)	Bedrooms (X_2)	House Price
1200	2	2000000
1500	3	3000000
1800	4	5000000

Date : / /

#Given the following scenarios, identify whether the issue is likely due to high variance, high bias, on both;

(i) A model performs ~~exp~~ exceptionally well on the the training set but poorly on the test set.

(ii) A model performs poorly on both the training and test sets.

(iii) A model performs decently on the training set but its performance fluctuates significantly across different test sets.

Scenario	Issue	Explanation
I	High Variance	Overfits training data, performs poorly on test data.
II	High Bias	Underfits, performs poorly on both training and test data
III	High Variance	Performance ~~fluctuat~~ fluctuates across, test sets.