

MACHINE LEARNING (ML-10)

Dr. NEERAJ GUPTA, Department of CEA, GLA University, Mathura

AGENDA

- K Nearest Neighbor

K-NEAREST-NEIGHBORS ALGORITHM

K nearest neighbors (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (distance function)

KNN has been used in statistical estimation and pattern recognition since 1970's.

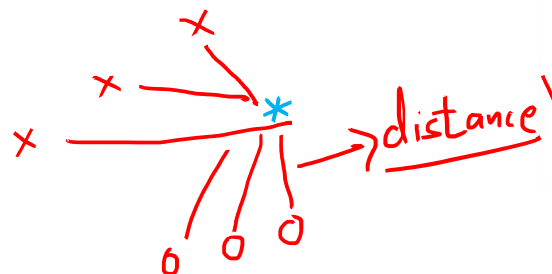
K-NEAREST-NEIGHBORS ALGORITHM

A case is classified by a majority voting of its neighbors, with the case being assigned to the class most common among its K nearest neighbors measured by a distance function.

If $K=1$, then the case is simply assigned to the class of its nearest neighbor

DISTANCE FUNCTION MEASUREMENTS

Distance functions



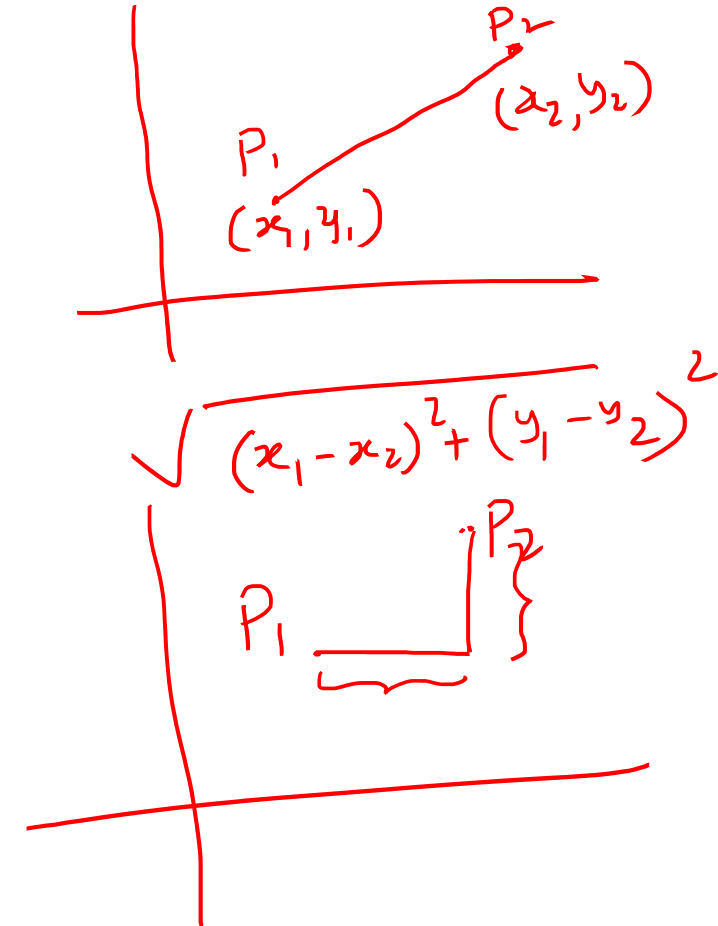
✓ **Euclidean** $\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$

✓ **Manhattan**

 $\sum_{i=1}^k |x_i - y_i|$

✓ **Minkowski**

 $\left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$



HAMMING DISTANCE

For category variables, Hamming distance can be used.

Hamming Distance

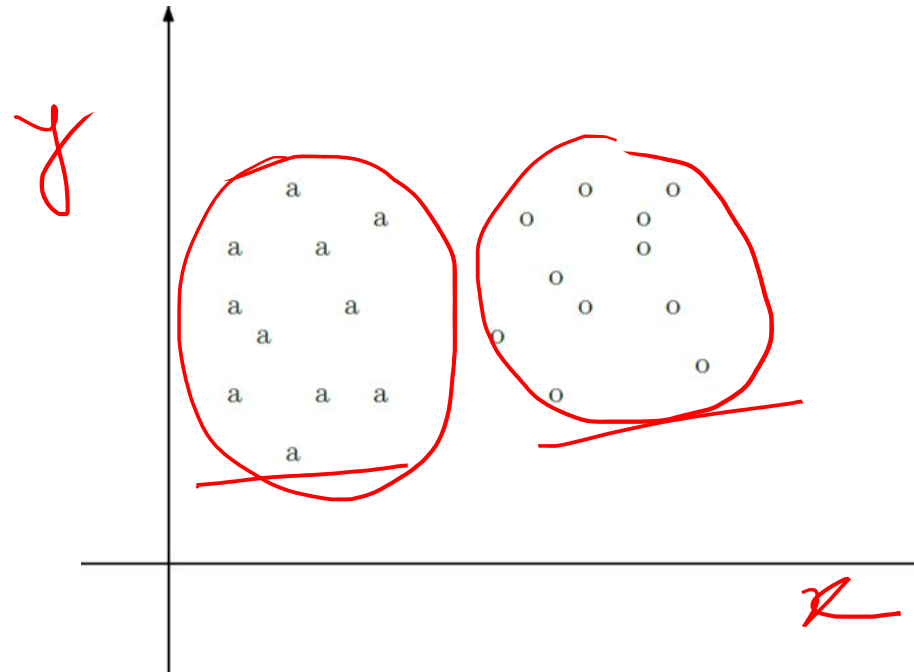
$$D_H = \sum_{i=1}^k |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

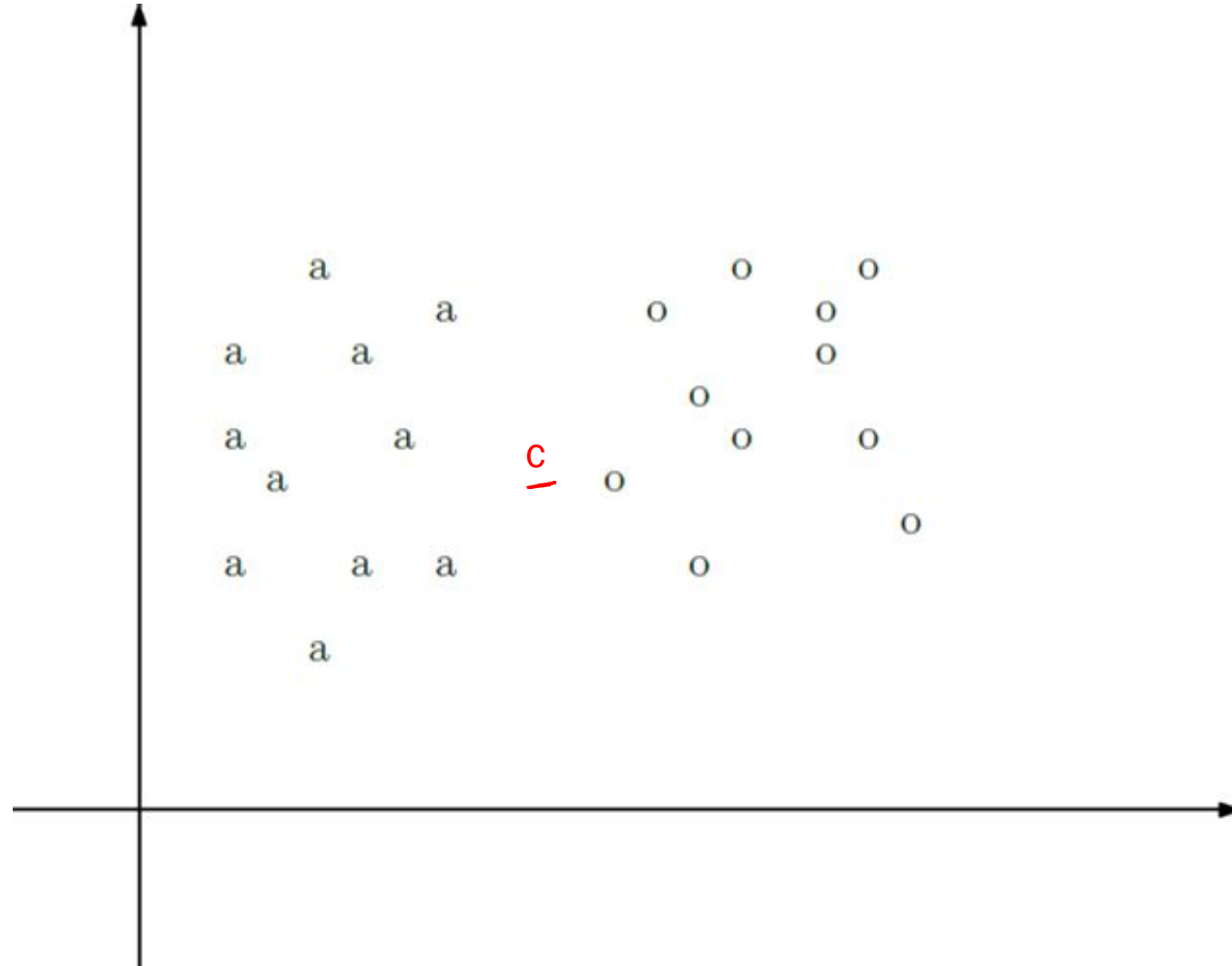
$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

K-NEAREST-NEIGHBORS



WHAT IS THE MOST POSSIBLE LABEL FOR C?



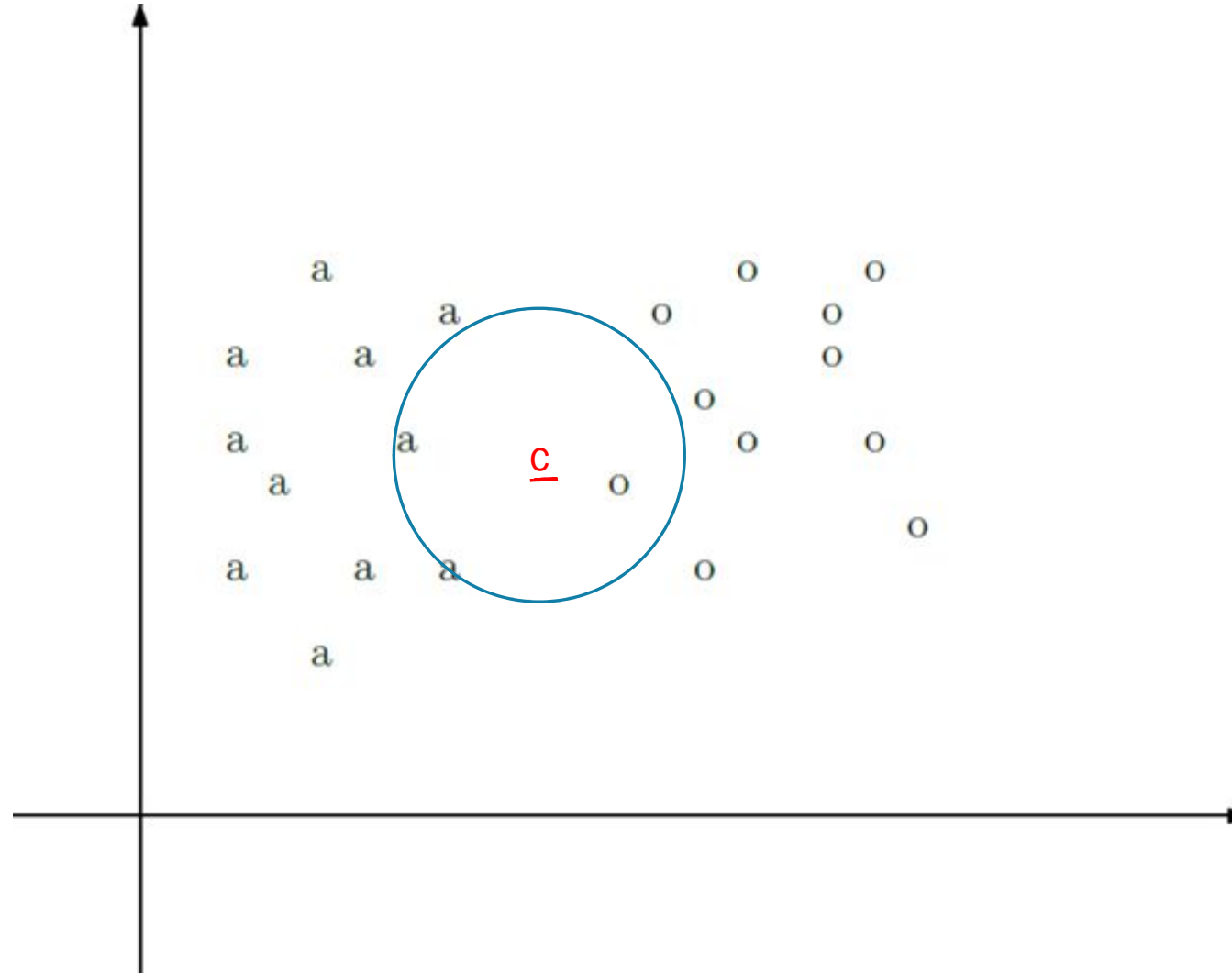
WHAT IS THE MOST POSSIBLE LABEL FOR C?

Solution: Looking for the nearest K neighbors of c.

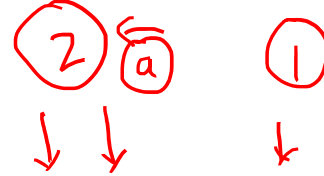
Take the majority label as c's label

Let's suppose $k = 3$:

WHAT IS THE MOST POSSIBLE LABEL FOR C?



WHAT IS THE MOST POSSIBLE LABEL FOR C?



The 3 nearest points to c are: a, a and o.

Therefore, the most possible label for c is a.

PSEUDO CODE OF KNN

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
 1. Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method.
 2. Sort the calculated distances in ascending order based on distance values
 3. Get top k rows from the sorted array
 4. Get the most frequent class of these rows
 5. Return the predicted class

REMARKS

- chose an odd k value for a 2 class problem
- k must not be a multiple of the number of classes
- the main drawback of k NN is the complexity in searching the nearest neighbors for each sample

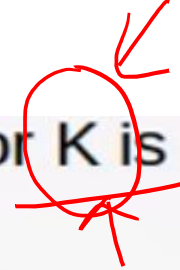
Binary



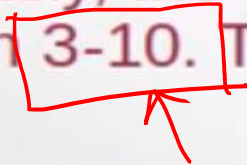
CHOOSING THE MOST SUITABLE K

- Choosing the optimal value for K is best done by first inspecting the data
- In general, a large K value is more precise as it reduces the overall noise but there is no guarantee
- Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value
- Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN

train | Test
20
Bo
K-fold

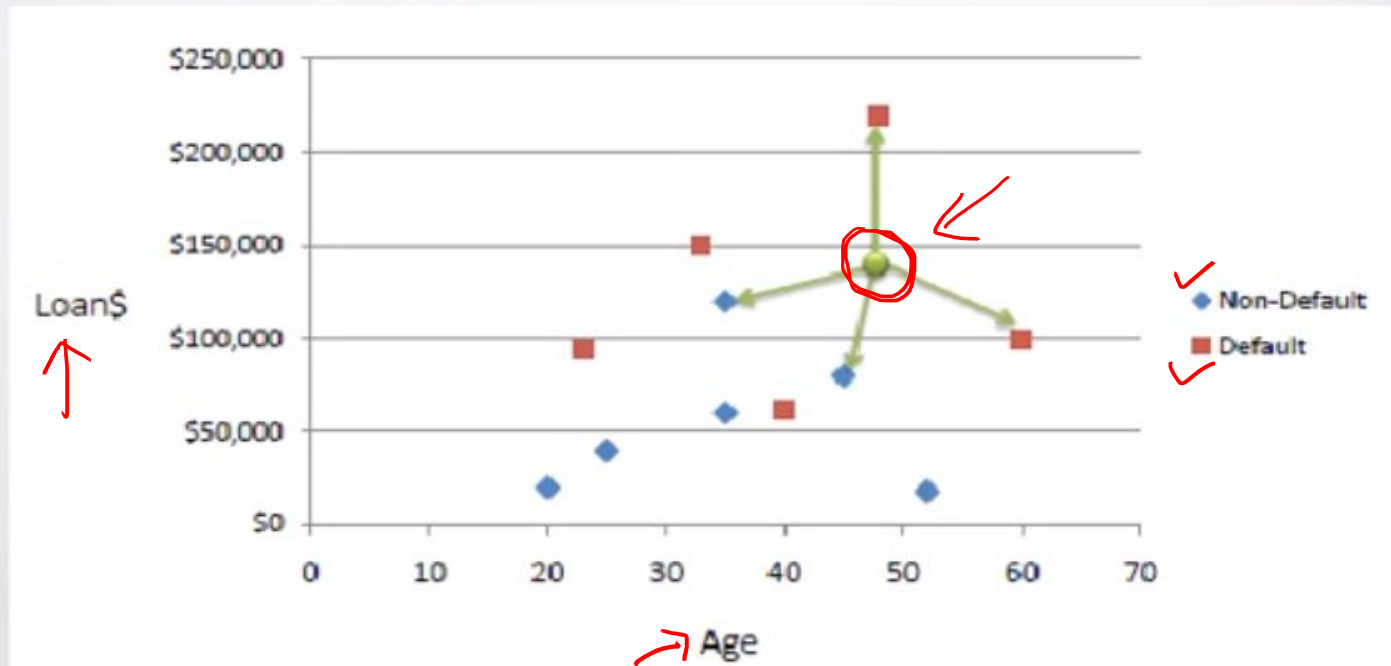


} ← K?



NORMALIZATION

- Consider the following data concerning credit default. Age and Loan are two numerical variables (predictors) and Default is the target



NORMALIZATION

- We can now use the training set to classify an unknown case (Age=48 and Loan=\$142,000) using Euclidean distance.
- If K=1 then the nearest neighbor is the last case in the training set with Default=Y.
- $D = \text{Sqrt}[(48-33)^2 + (142000-150000)^2] = 8000.01 \gg \text{Default}=Y$

Age	Loan	Default	Distance
25 ✓	\$40,000 ✓	N ✓	✓ 102000
35 ✓	\$80,000 ✓	N ✓	✓ 82000
45 ✓	\$80,000 ✓	N ✓	✓ 62000
20 ✓	\$20,000 ✓	N ✓	✓ 122000
35 ✓	\$120,000 ✓	N ✓	✓ 22000
52 ✓	\$18,000 ✓	N ✓	✓ 124000
23 ✓	\$95,000 ✓	Y ✓	✓ 47000
40 ✓	\$62,000 ✓	Y ✓	✓ 80000
60 ✓	\$100,000 ✓	Y ✓	✓ 42000
48 ✓	\$220,000 ✓	Y ✓	✓ 78000
33 ✓	\$150,000 ✓	Y ✓	✓ 8000
48	\$142,000	? Y	

Euclidean Distance

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

- With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y

Age = 48
Loan = 1,42,000

Distance from Euclidean Distance

$$D = \sqrt{(48-33)^2 + (150000-142000)^2}$$

$$= 8000$$

K=3

KNN Model

NORMALIZATION

- One major drawback in calculating distance measures directly from the training set is in the case where variables have different measurement scales or there is a mixture of numerical and categorical variables.
- For example, if one variable is based on annual income in dollars, and the other is based on age in years then income will have a much higher influence on the distance calculated.
- One solution is to standardize the training set

NORMALIZATION

Using the standardized distance on the same training set, the unknown case returned a different neighbor which is not a good sign of robustness.

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	0.61	?	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

K-NEAREST NEIGHBOR CLASSIFICATION (KNN)

Unlike all the previous learning methods, **kNN does not build model from the training data.**

To classify a test instance d , define k -neighborhood P as k nearest neighbors of d

Count number n of training instances in P that belong to class c_i

Estimate $\Pr(c_i | d)$ as n/k

No training is needed. Classification time is linear in training set size for each test case.

DISCUSSIONS

kNN can deal with complex and arbitrary decision boundaries.

Despite its simplicity, researchers have shown that the classification accuracy of kNN can be quite strong and in many cases as accurate as those elaborated methods.

kNN is slow at the classification time

kNN does not produce an understandable model

EXERCISE

Consider a dataset having two variables: height (cm) & weight (kg) and each point is classified as Normal or Underweight

Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

On the basis of the given data we have to classify the below set as Normal or Underweight using KNN

57 kg	170 cm	?
-------	--------	---

EXERCISE

Hence, we have calculated the Euclidean distance of unknown data point from all the points as shown:

Where $(x_1, y_1) = (57, 170)$ whose class we have to classify

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

EXERCISE

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.

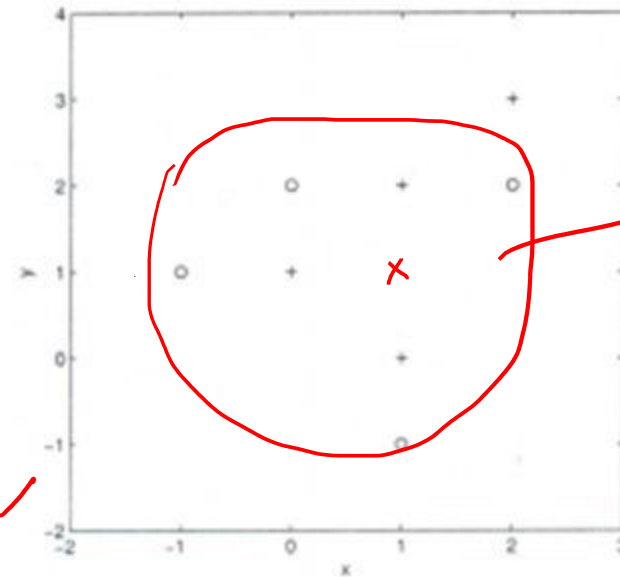
	x	y	Class
1.	-1	1	-
2.	0	1	+
3.	0	2	-
4.	1	-1	-
5.	1	0	+
6.	1	2	+
7.	2	2	-
8.	2	3	+

$$\sqrt{(-1-1)^2 + (1-1)^2} = 2$$

3

K=3

+ 3 ✓
- 0



-ve

Q1. Suppose, you want to predict the class of new data point x=1 and y=1 using euclidian distance in 3-NN. In which class this data point belong to?

EXERCISE

K=7

✓ Q2. In the previous question, you are now want use 7-NN instead of 3-KNN which of the following $x=1$ and $y=1$ will belong to?

Q2. In the previous question, you are now want use 5-NN instead of 3-KNN which of the following $x=1$ and $y=1$ will belong to?

THANKS



*Keep Learning
Keep Growing*



Dr. Neeraj Gupta
Assistant Professor, Dept. of CEA
neeraj.gupta@gla.ac.in