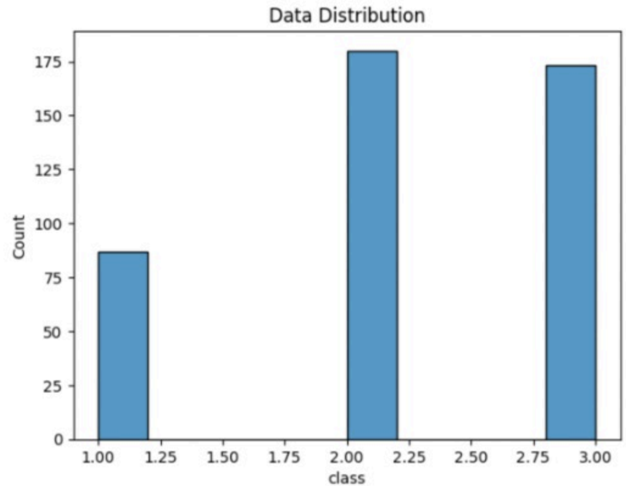
Supervised Algorithms

KNN

Quiz time!

Time Left: 12s



What can be said about the data ?

40 users have participated

A	Multi-class balanced data	10%
B	Multi-class imbalanced data	83%
C	Binary-class imbalanced data	7%
D	Binary-class balanced data	0%

End Quiz Now

Quiz time!

Quiz Ended!

polynomial

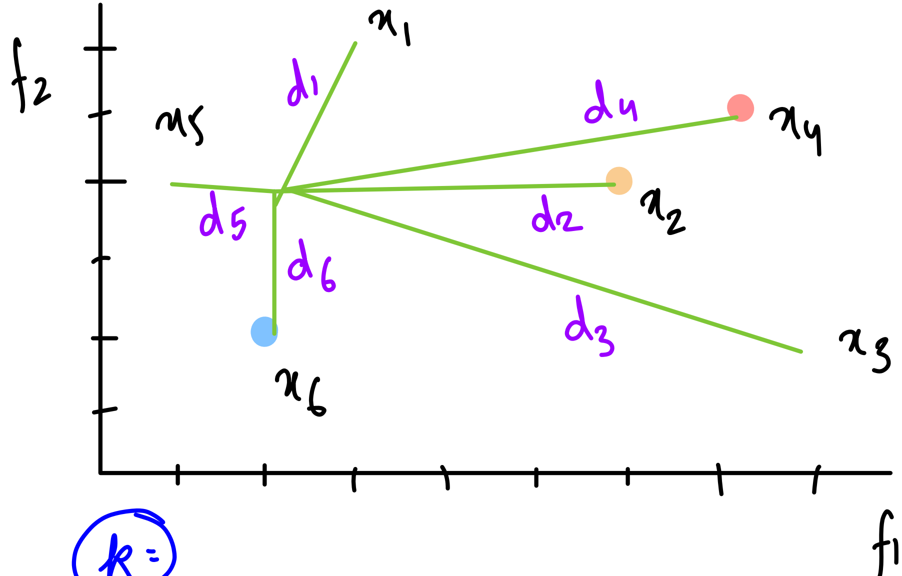
OvR

How will Logistic Regression handle non-linear, multi-class data ?

43 users have participated

✓	A	Polynomial, OneVsRest	37%
X	B	Linear, OneVsRest	23%
	C	OneVsRest, Polynomial	26%
	D	OneVsRest, Linear	14%

	f_1	f_2	y
x_1	3	6	1
x_2	6	4	1
x_3	8	2	3
x_4	7	5	3
x_5	1	4	2
x_6	2	2	2

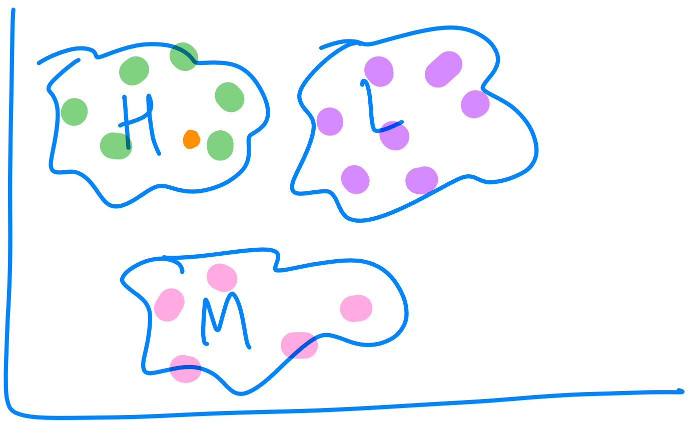


$k =$

$$\begin{array}{ccccccccccc} d5 & < & d6 & < & d1 & < & d2 & < & d4 & < & d3 \\ 2 & & 2 & & 1 & & 1 & & 3 & & 3 \end{array}$$

* We try not to choose even values of k

* Always odd values



$k = 1$ $x_q \Rightarrow 2$

$k = 2$ $x_q \Rightarrow 2$

$k = 3$ $x_q \Rightarrow 2$

$k = 4$ $x_q \Rightarrow 2/1$

$k = 5$ $x_q \Rightarrow 2/1$

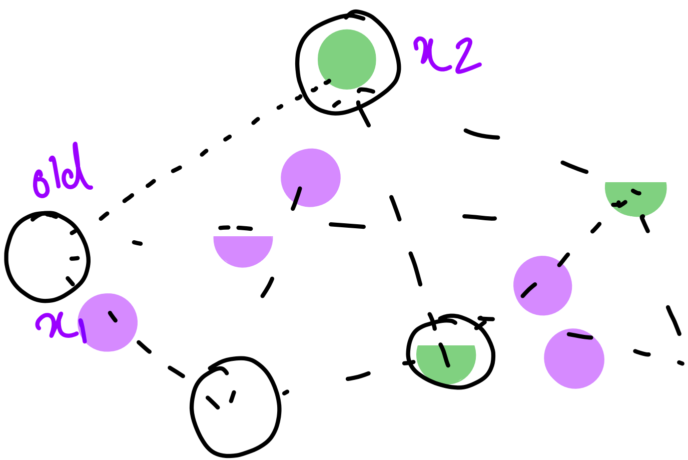
$k = 6$ $x_q \Rightarrow 2/1/3$

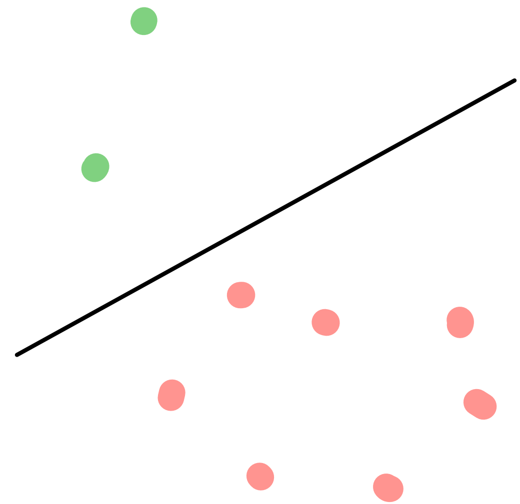
Blinkit

MAP

Route Distance

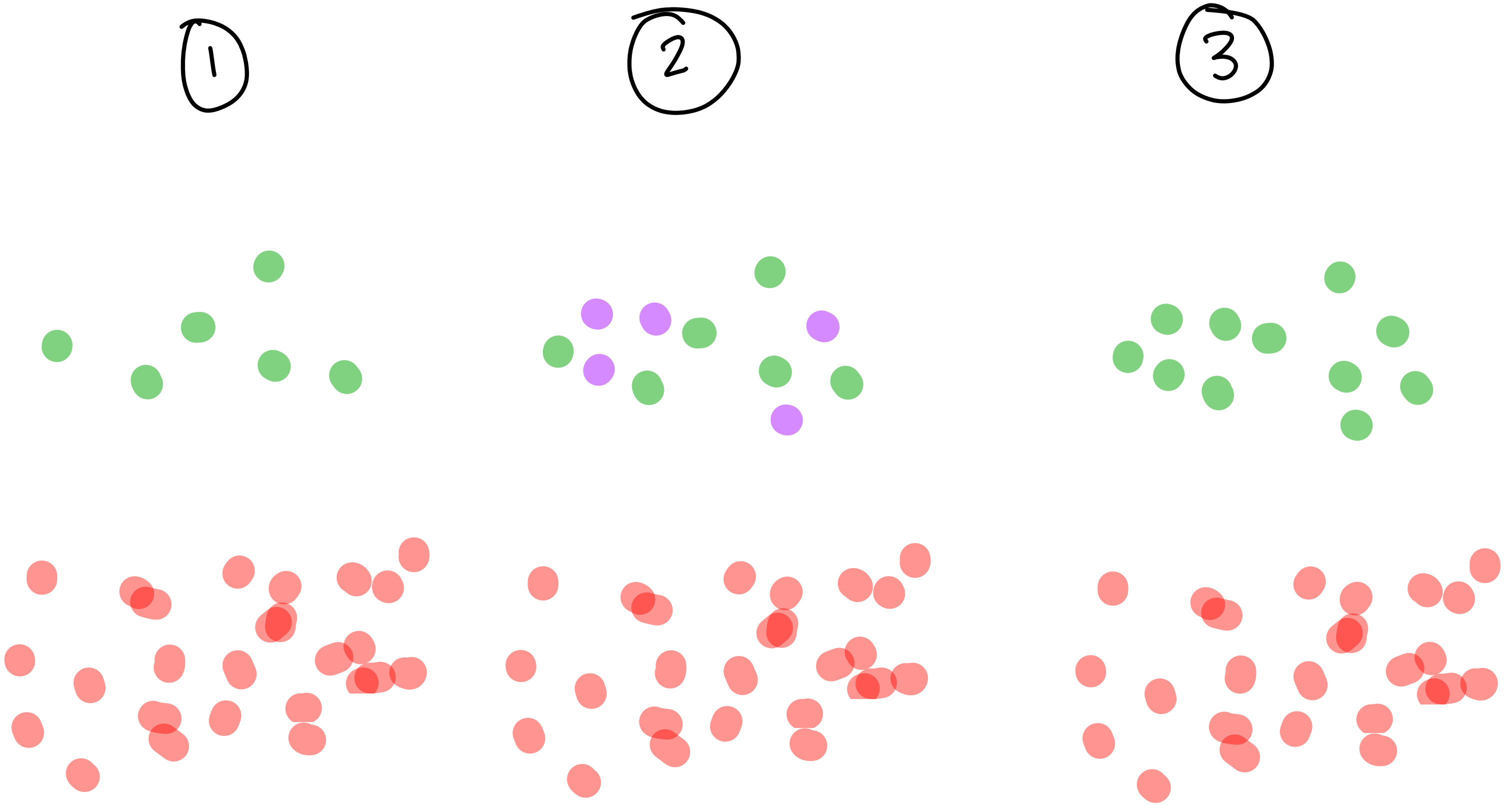
SMOTE [Synthetic Minority Oversampling Technique.]





1. Pick any value of $k = 2$
2. Pick of any value of α, $\alpha \in [0,1]$
3. Newpoint

$$x_{new} = x_{old} + \alpha \cdot \text{distance b/w old \& } x_2$$


1
2
3

KNN

↳ Are we learning anything? → parameters. w_1, w_2, w_3, w_4 ... etc.

NO

NON PARAMETRIC

Lin Reg w_1

Log Reg w

At inference

↳ Calculate distances for all unknown points everytime at inference

Extremely slow

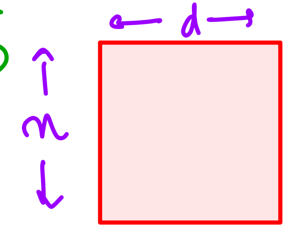
fast → training

slow → inference.

Algorithm

100 M datapoint

$\sqrt{O+O+O+O}$



1. For each point
   → Calc distance ⇒ $O(nd)$
   → Append ⇒ X
2. Sorting distance in Ascending order ⇒ $O(n \log n)$
3. Pick top k values along with their classes ⇒ $O(k)$ ↓↓
4. Majority voting ⇒

$$\left.\begin{array}{ll} +ve & 3 \\ -ve & 2 \end{array}\right] \Rightarrow +ve.$$

At ~~inference~~ = testing

$O(nd + n\log n)$

KNN → slow at inference

Advantages

- Simple + intuitive
- Powerful
- Non Linear Boundaries
- Multi Class Classification.

Quiz time!

Quiz Ended!

Arrange the statements in correct order based on kNN alg

s1- find majority vote 3

s2- perform euclidean distance 1

s3- sort and select k datapoints 2

s4- give class to x_q datapoint 4

$S_2 \to S_3 \to S_1 \to S_4$

32 users have participated

A	s2,s1,s3,s4	9%
B	s2,s3,s4,s3	0%
C	s2,s3,s1,s4	69%
D	s2,s4,s3,s1	22%

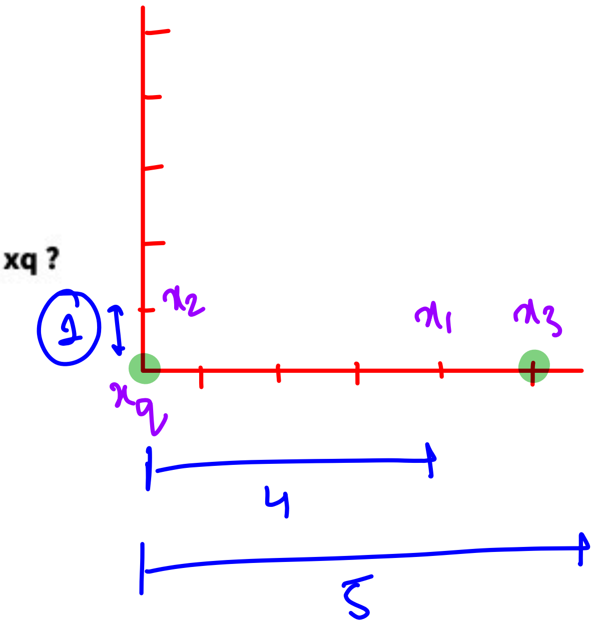
Quiz time!

Quiz Ended!

If x1 at (4,0) , x2 at (0,1) and x3 (5,0) and xq at (0,0). Which is nearest point to xq ?

29 users have participated

A	x1	3%
B	x2	93%
C	x3	3%



Assumption →

Data in my neighbourhood is homogenous.

Quiz time!

Quiz Ended!

how kNN is better than logistic regression. Select the correct option

33 users have participated

A	kNN has less time complexity	12%
B	kNN classifies data better	12%
C	kNN handles most noise/outlier	0%
D	kNN handles multi-class problem	76%

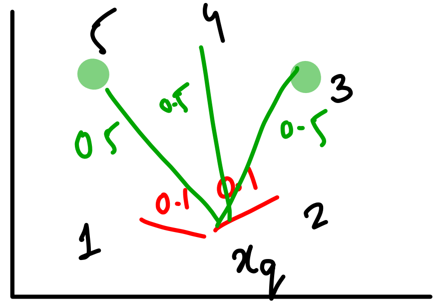
log

OVR

weighted KNN

$k=5$

$w_i = \frac{1}{d_i}$



$\rightarrow x_q$

$\rightarrow 1$ (+ve)

$\rightarrow 0$ (-ve)

$$w_1 = \frac{1}{0.1} = 10$$

$$w_2 = \frac{1}{0.1} = 10$$

$$w_3 = \frac{1}{0.5} = 2$$

$$w_4 = \frac{1}{0.5} = 2$$

$$w_5 = \frac{1}{0.5} = 2$$

Red = 20

Green = 6

Majority = Red

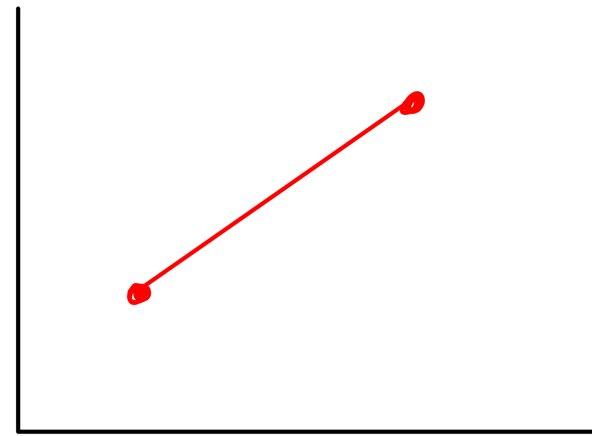
Distance

Euclidean

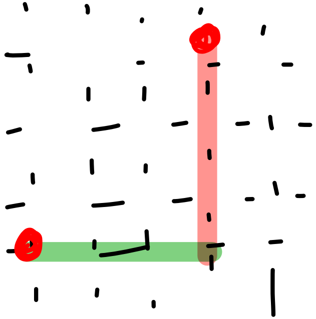
$$\sqrt{\sum_{i=1}^{n}(x_{1i} - x_{ij})^2}$$

Euclidean Distance

$$\left[\sum_{j=1}^{d}(x_{1j} - x_{2j})^2\right]^{1/2}$$



↳ Doesnot work when d is very high.



$$d_M = \left[\sum_{j=1}^{d} |x_{1j} - x_{2j}|\right]^{1/1}$$

General $\Rightarrow d(x_1, x_2, p) = \left[\sum_{j=1}^{d} (x_{1j} - x_{2j})^p\right]^{1/p}$

Minkowski distance

$p=1 \Rightarrow \left[\sum_{j=1}^{d} |(x_{1j} - x_{2j})^1|\right]^{1/1} \Rightarrow$ Manhattan

$p=2 \Rightarrow \left[\sum_{j=1}^{d} (x_{1j} - x_{2j})^2\right]^{1/2} \Rightarrow$ Euclidian.

L2 RIDGE

$$\sum_{j=1}^{d} |w_j|^2$$

L1 LASSO

$$\sum_{j=1}^{d} |w_j|$$

$$\sum_{j=1}^{d} |w_j|^p \Rightarrow \begin{array}{l} p=1 \rightarrow \\ p=2 \rightarrow \end{array}$$