KNN Classification

[K-nearest neighbors]
Predicting classes

Eg: Predicting movie Genre

IMDb Rating	Duration (min)	Genre
8.0 (MI)	160	Action
6.2 (42)	170	Action
7.2 (R2)	168	Comedy
8.2 (0.2)	155	Comedy

Predict Genre of "Barbie" movie
→ Rating: 7.4
→ duration: 114

Sol → Euclidean Distance

Step 1 calculate distances b/w new movie (Barbie) and each movie in dataset

Distance to (8.0, 160) = $\sqrt{(7.4-8)^2+(114-160)^2}$
$\cong 46.00$

Distance to (6.2, 160) = $\sqrt{(7.4-6.2)^2+(114-160)^2}$
$\cong 56.01$

Distance to (7.2, 168) = $\sqrt{(7.4-7.2)^2+(114-168)^2}$
$\cong 54.00$

Distance to (8.2, 155) = $\sqrt{(7.4-8.2)^2+(114-155)^2}$
$\cong 41.00$

Step 2 Select K-nearest neighbors

↓
least

→ Generally K=5, but here $\boxed{K=3}$

46.00 (Action)

54.00 (comedy)

41.00 (comedy)

Step3 Assign class according to majority voting:

Action → 1 votes

Comedy → 2 votes

* *

WINNER! :[Comedy] *

*

Q Suppose you have the following dataset with two features (X & Y) and corresponding table. {CLASSIFY}

Data Point	X	Y	Label
1	2	3	A
2	3	4	A
3	5	6	B
4	7	8	B
5	10	10	A

consider a new datapoint with $x_1=6$ & $y_1=7$. using KNN with K=3, predict the labels for this new data point.

Sol^n

Step1 , calculating the distances...

~~dap~~ data point1 : $\sqrt{(6-2)^2+(7-3)^2} = 5.65$

data point2 : $\sqrt{(6-3)^2+(7-4)^2} = 4.24$

data point3 : $\sqrt{(6-5)^2+(7-6)^2} = 1.41$

data point4 : $\sqrt{(6-7)^2+(7-8)^2} = 1.41$

data point 5 : $\sqrt{(6-10)^2+(7-10)^2} = 5$

Step2 selecting 3 nearest neighbors

↓

data points:	2	3	4
	↓	↓	↓
class label :	A	B	B

Step3 majority votes are for Label "B".

Hence, data point(new) will have Label $\boxed{B}$

Q suppose you have the following dataset with two features (X & Y) and corresponding labels.

Datapoint	X	Y	~~Label~~
1	2	3	~~A~~
2	3	4	~~A~~
3	5	6	~~B~~
4	7	8	~~B~~
5	10	10	

Regression

consider new data point with $X_1 = 7.0$, using KNN, with $K = 2$ predict the value of Y.

Soln

Step 1 calculate nearest Euclidean distance for single variable X. } $X_1 = 7$

Data point	X	Y	distance
1	2	5	5
2	4	8	3
3	6	12	1
4	8	15	1
5	10	20	3

Step 2 K=2, nearest neighbors

↓

Data points:	3	4
X :	6	8
Y :	12	15

Step 3 Predicted y = mean of Y of nearest neighbors

$$Y\text{-pred} = \frac{12+15}{2} = \boxed{13.5}$$

Choosing 'K'

$$\boxed{K < \sqrt{n}}$$

{n → no. of data points}