KNN Regressor

100 → data

iq | cgpa | package

predict

$\boxed{\text{Gpa} | iq}$ → Palkage
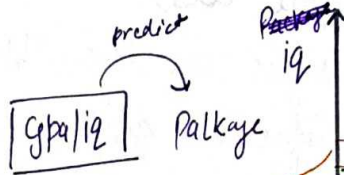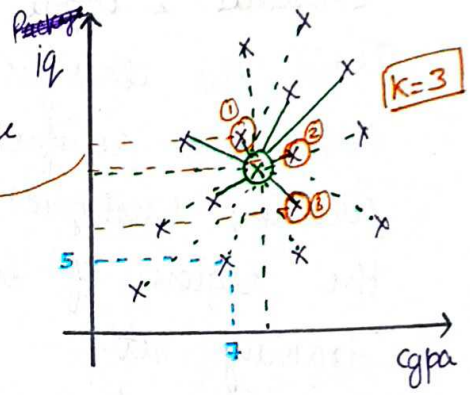
(Nearest point) Palkage

3
5
4

Average → $\overline{12/3} = \boxed{4 \text{ LPA}}$

Median → if outlier

Steps: find distance x_q
and all the x-train

Step2: Sort

$k \downarrow\downarrow$ = high variance and overfitting
$k \uparrow\uparrow$ = high biased and underfitting

Hyperparameter
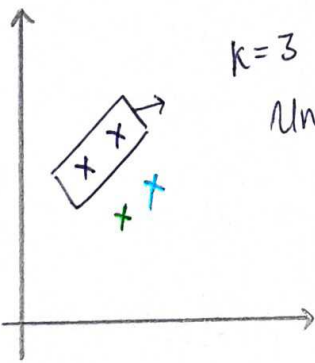
$k = 3$
Uniform Weight $\Big\}$ Simple KNN

Weighted KNN (option in sklearn)

$k=5$

$$W = \frac{1}{distance}$$

blue ⤷ option → distance

green

2

①

⑦3

⑦4

⑦5
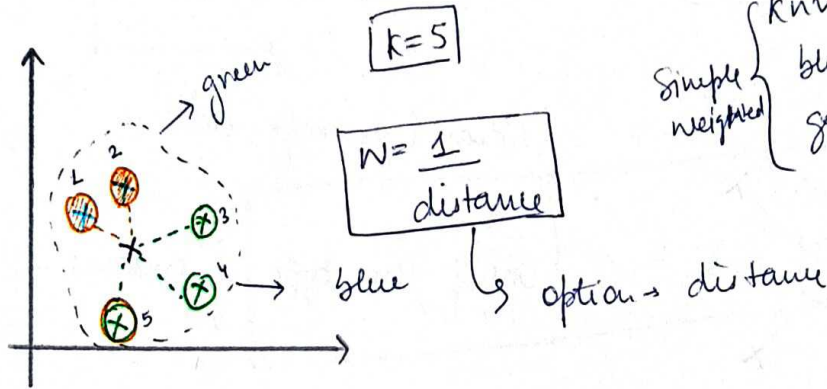
blue

Simple {knn → Uniform
weighted { blue = 2
green = 3 → label

		dist	W (Weighted)
1	Blue →	0.2 →	5
2	Blue →	0.5 →	2

Blue = 5+2 = 7

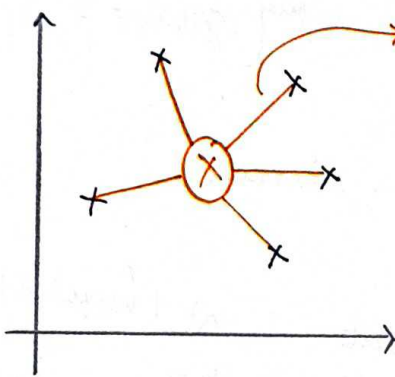| 3 | green → | 1 → | 1 |
| 4 | green → | 2 → | 0.5 |
| 5 | green → | 3 → | 0.33 |

green
1+0.5+0.33
= 1.83

Blue = 7   green = 1.83

* Blue choose for new point in Weighted (distance)
* green choose for new point in Uniform KNN (simple)

Hyperparameter tunning

1) Simple KNN (k=1)

2) Distance (weighted)

Types of Distance

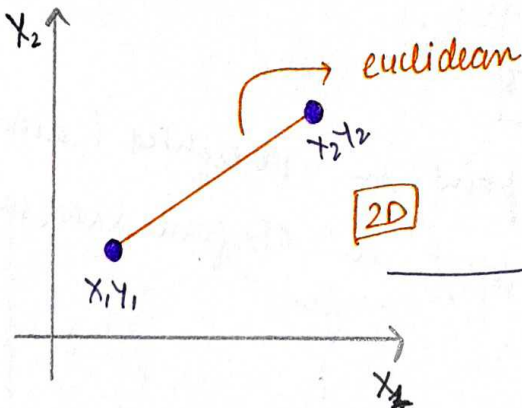

euclidean (generally use)

└ one of the type of Distance

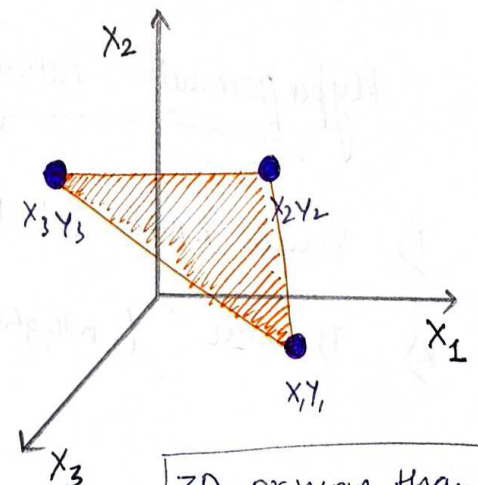**Sklearn** → metric = " minkowski "

↓

$P = 2$ → euclidean (default)

Distance

euclidean

2D

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

$X_1 \ X_2 \ X_3 - - - - X_d$

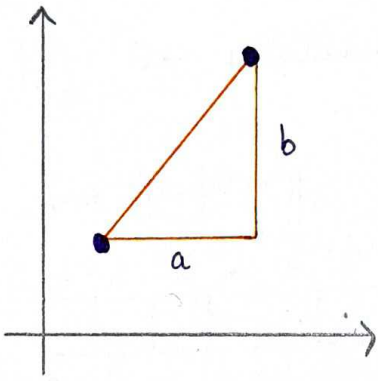$$dist = \sqrt{\sum_{i=1}^{d} (X_{2i} - X_{1i})^2}$$
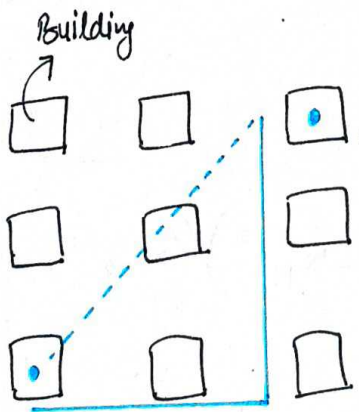
→ for multi dimension

3D or more than

x_2

$x_{21} \, x_{22} \, x_{23}$ $x_{11} \, x_{12} \, x_{13}$ $\Rightarrow$ $\sqrt{(x_{21}-x_{11})^2 + (x_{22}-x_{12})^2 + (x_{23}-x_{13})^2}$

x_1

3D

x_3
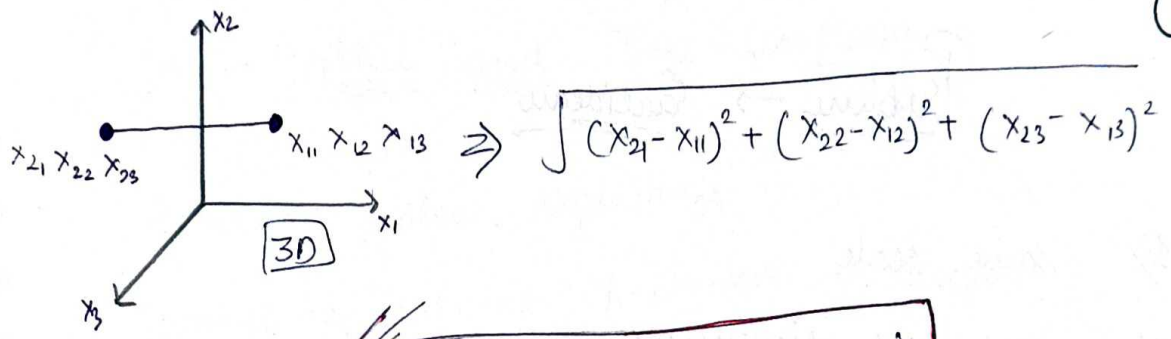
$$\text{dist} = \left(\sum_{i=1}^{d} (x_{2i} - x_{1i})^2 \right)^{\frac{1}{2}} \rightarrow \begin{array}{l} n \text{ dim} \\ \text{points} \end{array}$$

L2 norm

Manhattan Distance

b

a

$a+b$ $\rightarrow$ manhattan Distance

$\downarrow$

taxi cab Distance
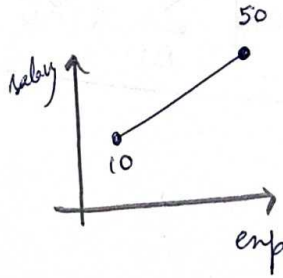
Building

$$m = \sum_{i=1}^{d} |x_{2i} - x_{1i}|$$

$$|x_{2b} - x_1| + |y_2 - y_1|$$

Problem → Euclidean

1.) same scale

2.) curse of dimension

$(0 - cr)$ salary
(experience)

$$\sqrt{(10)^2 - (50)^2} \underset{\uparrow}{} \text{salary}$$

different scale
create prob if
salary is larter
square become complex

* higher dimension create problem
  so use, __Manhattan Distance__.
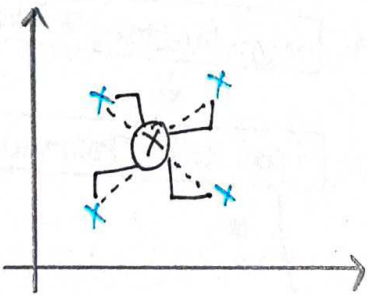
Minkowski

euclidean → $\left(\sum_{i=1}^{d} (x_{2i} - x_{1i})^2 \right)^{1/2}$ → $P = 2$

Manhattan → $\left(\sum_{i=1}^{d} |x_{2i} - x_{1i}|^1 \right)^{1/1}$ → $P = 1$

General → $\left(\sum_{i=1}^{d} (|x_{2i} - x_{1i}|^p) \right)^{1/p}$ → $P = $ any value

$$\boxed{P \geq 0}$$

Space and Time complexity

KNN → slow algorithm

Distance → Sort → K-Nearest → major

Time complexity of KNN → $O(nd)$

n → no. of Row in training data

d → no. of feature

Space complexity of KNN → $O(nd)$

KD-Tree (Introduction)

time complexity → $O(d \log n)$

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

⇒ data store in array

X
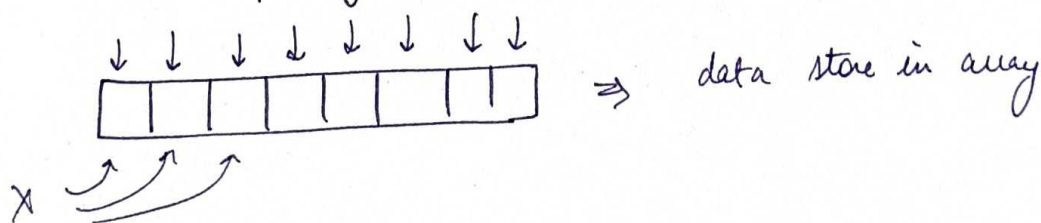
KD-Tree
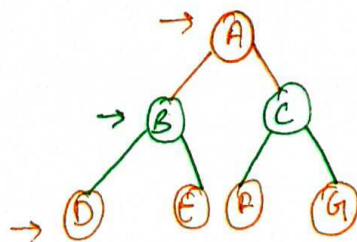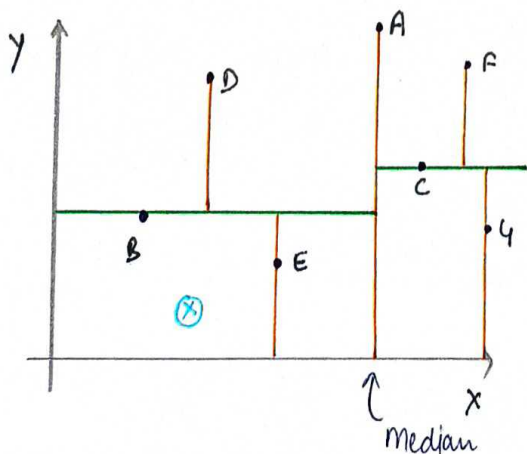
Steps:- find median of all the point and draw line at median according to X (A).

Step:- find median of Y at the right sid of A and left side of A (B)(C)

Step: Again find median according to x and draw line
(D, E, F, G).