KNN

* K-Nearest Neighbour
* Can do both Regression as well as Classification
* No assumptions.
* Multi-class classification
* Reg → KNN Regressor
  Classification → KNN Classifier

General

K → hyperparamete

K = 3

Regression

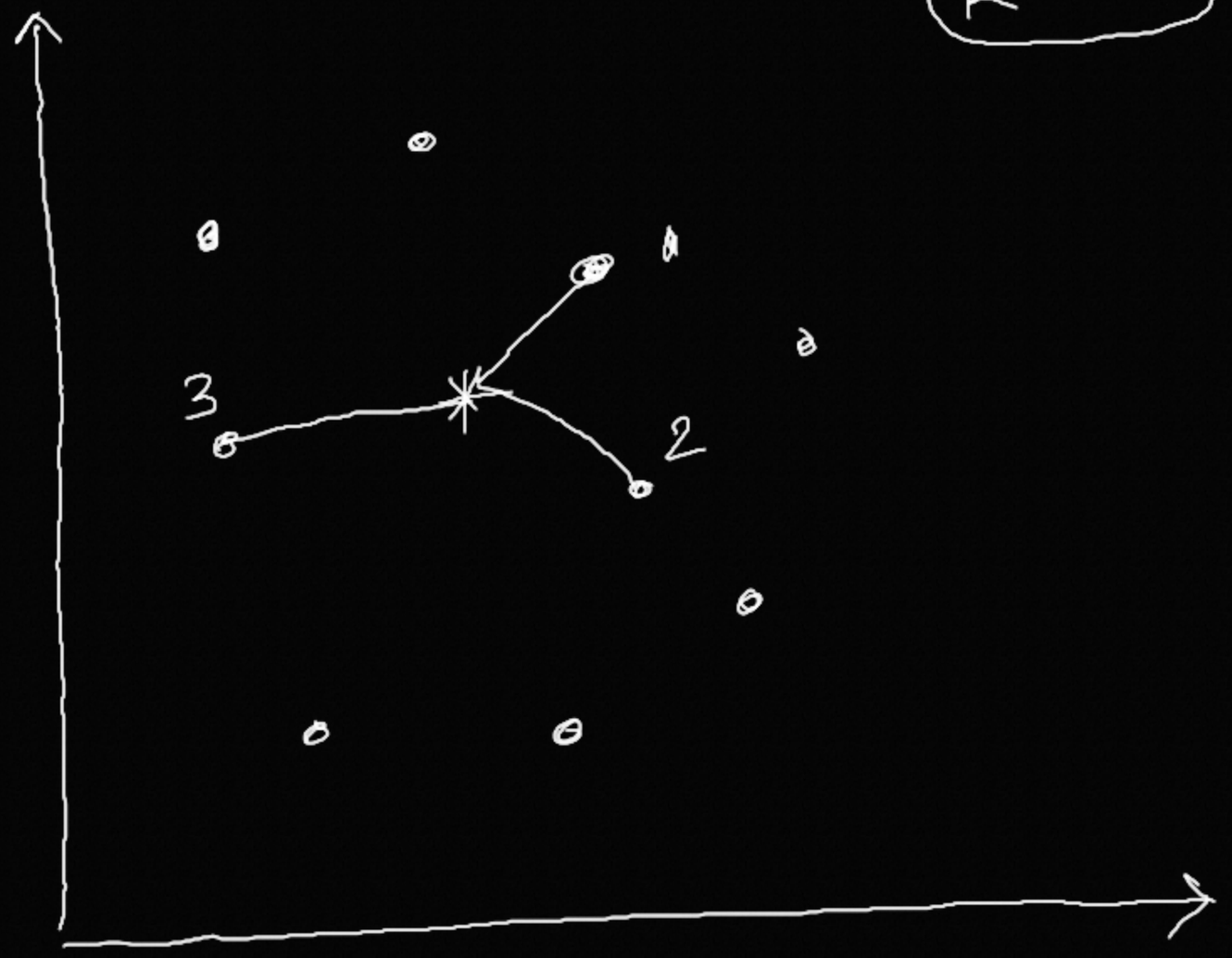
↳ Average / Mean

Classification

↳ Voting

1 → A
2 → A
3 → B

} → A



Mathematical (classification)

Data

X_1	X_2	Y
Weight (g)	Size (cm)	fruit
150	7	Apple
170	8	Apple
180	8.5	Pear
120	6	Apple
200	9	Pear

* New data point ⇒ $X_1 = 160g$

$X_2 = 7.5\ cm$

New point ⇒ $(160, 7.5)$

Current ⇒ $(170, 8)$

① Calculate Euclidean distance

formula :- $\sqrt{(X_2-X_1)^2 + (Y_2-Y_1)^2}$

$d_1 = \sqrt{(160-150)^2 + (7.5-7)^2} = 10.01$

$d_2 = \sqrt{(160-170)^2 + (7.5-8)^2} = 10.01$

$\boxed{d_3 = 20.02}$

$\boxed{d_4 = 40.01}$

$\boxed{d_5 = 40.03}$

② Select a K value

We set K = 3 (3 nearest points)

d_1 Apple

d_2 Apple

d_3 Pear

③ Do Voting

$\boxed{\text{Prediction} \Rightarrow \text{Apple}}$

Mathematical (Regression)

Data

x_1	x_2	y
Size (m)	Bedroom	Price
50	1	200
60	2	220
70	2	240
80	3	260
90	3	280

* New point ⇒ $X_1 = 65m$
$X_2 = 2$

New point = (65, 2)
Current - (50, 1)

① Calculate euclidean distance

$$d_1 = \sqrt{(65-50)^2 + (2-1)^2} = 15.03$$

$$d_2 = 5$$

$$d_3 = 5$$

$$d_4 = 15.03$$

$$d_5 = 25.02$$

② Select K

We set $K = 3$

d_2 220

d_3 240

d_1 200

③ Do Average

$$\text{Prediction} = \frac{220 + 240 + 200}{3}$$

$$\boxed{\text{Prediction} = 220}$$

Hyperparameters

* metric ⇒ "minkowski"

formula ⇒ $$d = \left[\sum |x_2 - x_1|^p \right]^{1/p}$$

* p ⇒ power parameter ⇒ default = 2 → euclidean

p = 1 → manhattan

Manhattan ($p=1$)

$$d = \sum |x_2 - x_1|$$

* Better with high-D data
* I/p colums are categorical or discrete
* Perform well with outliers.

Euclidean ($p=2$)

$$d = \left[\sum (x_2 - x_1)^2 \right]^{1/2}$$

$$= \sqrt{\sum (x_2 - x_1)^2}$$

* Poor with high-D data
* I/p colum are continou
* Very sensitive to outlier.

When to use

* Simple & easy
* Small or medium sized datasets
* Do both reg & classification tasks

When Not to Use

* Large datasets
* High-D dataset (curse of Dimensionality)
* Data has outliers (Sensitive to outliers)
* Slow → computationally intensive (Store overall data in your memory)

* Imbalanced dataset
* If you have not done proper feature scaling it will not perform