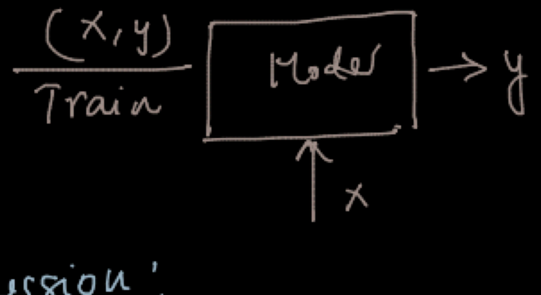
Machine Learning -

Supervised Models

— Model is trained using both X & y.



Regression :

— 'y' is Continuous.

(Linear Regression)

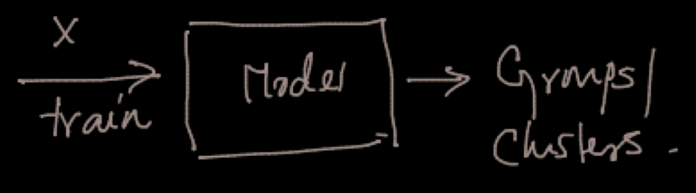
Classification

— 'y' is categorical

(Logistic Regression)

Unsupervised Model

— Model is trained using only X. Values.

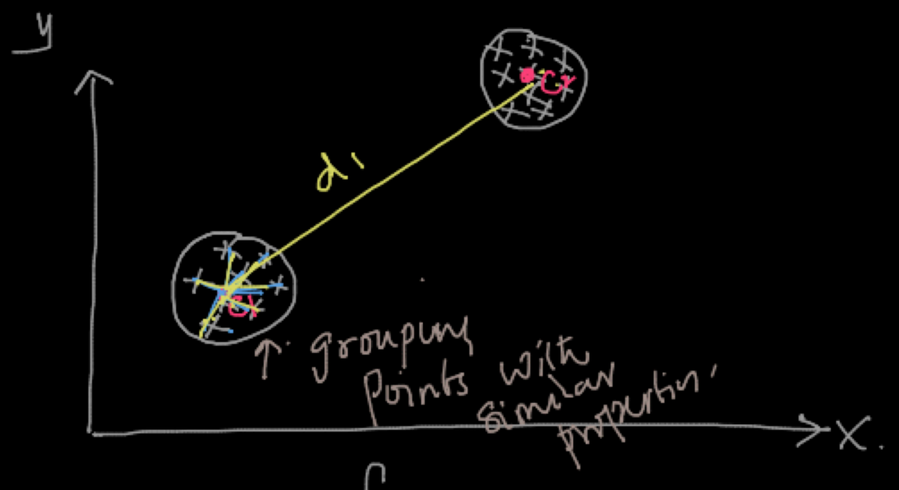


— K-Means

— Hierarchical

— DBSCAN

Clusters.



↑ grouping Points with similar properties.

fig 1.

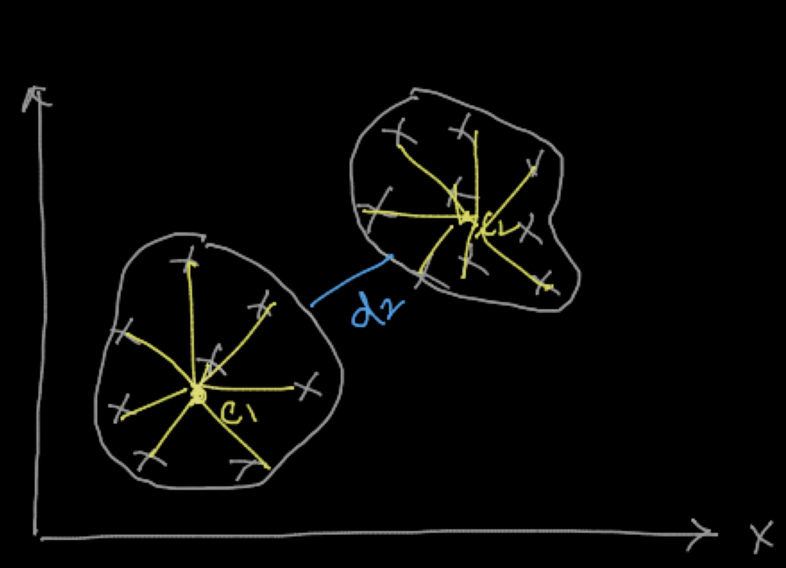


fig 2

Good clusters :

1. Intra cluster distance — Minimum.
2. Inter cluster distance → Maximum.

	x_1 Height	x_2 Weight	x_3 wc	..... x_d
P1	190	95	63	. . .
P2	130	55	95	. . .
P3	189	100		

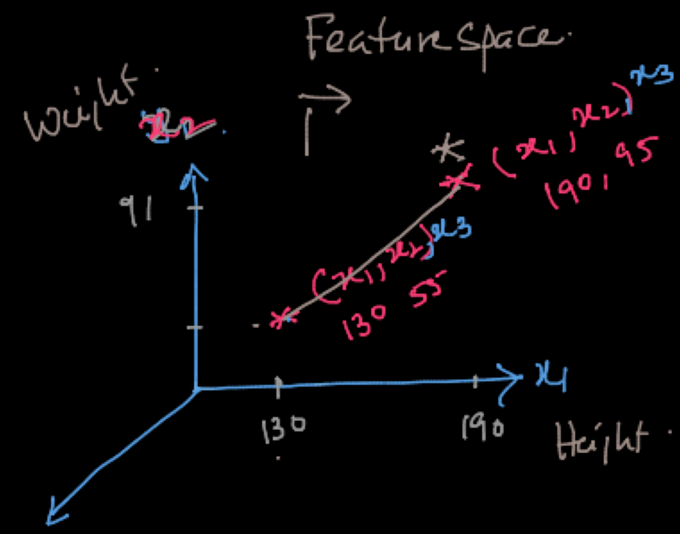
↳ High dimension

Similarity $\propto \frac{1}{\text{distance}}$

Binary.

Voted	Cand
Yes	0
No.	1.

x_1	x_2	x_3
23	M	Numeric-



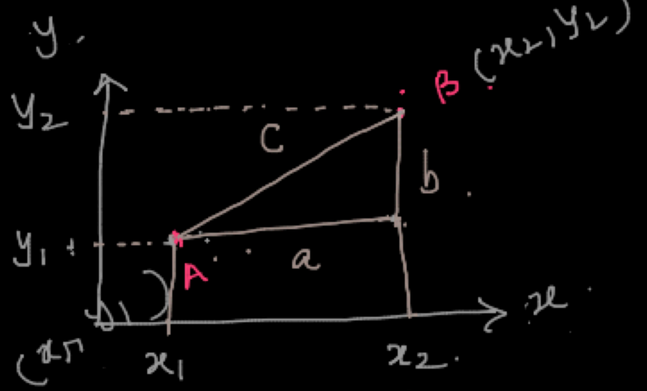
10.20 am

Break.

Distance Measures.

- Numeric.
  - Euclidean
  - Manhattan
  - Minkowski
  - Mahalanobis
  - Cosine-
- Text-
  - Hamming distance
  - Cosine
- Binary.
  - Simple matching Co-eff-
  - Binary Euclidean
  - Jaccard's Co-eff.
- Mixed.
  - Gowers Dissimilarity index-

Euclidean distance

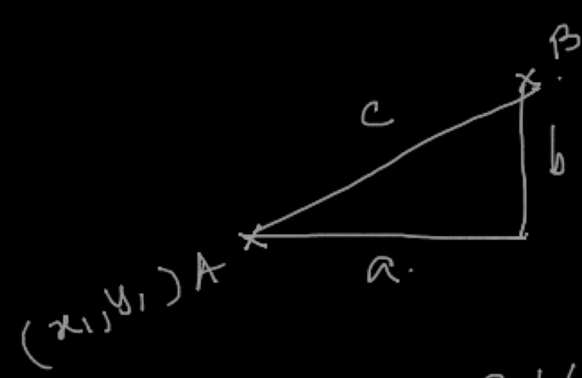


$$c = \sqrt{a^2 + b^2}$$

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

$$c = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + \cdots}$$

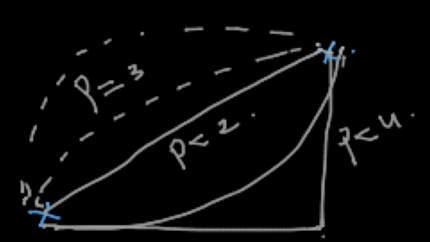
Manhattan Distance



$$c = a + b$$

$$c = (x_2 - x_1) + (y_2 - y_1)$$

Minkowski



$$c = \left[(x_2 - x_1)^p + (y_2 - y_1)^p \right]^{1/p}$$

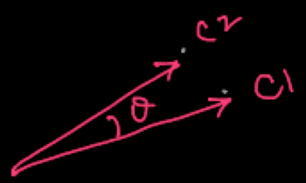
$p = 1, 2, 3, \ldots$

$$c = (x_2 - x_1) + (y_2 - y_1)$$

$p = 2$ — Euclidean

$$c = \left[(x_2 - x_1)^2 + (y_2 - y_1)^2 \right]^{1/2}$$

Cosine Distance



$\theta = 0 \quad 90$

$\cos(\theta) \Rightarrow 1 \quad 0$

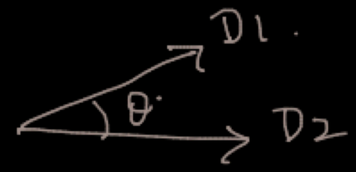
Text Values.

Corpus.
- D1: The car is driven on the road → document
- D2: The truck is driven on highway → document.

Vocab: The car truck is driven on road highway.

	The	car	truck	is	driven	on	road	highway
D1:	1	1	0	1	1	1	1	0
D2:	1	0	1	1	1	1	0	1

Hamming distance = 2



Cos θ ⇒ distance both D1 & D2

Scaling → Bring all Columns to the same range.

	SAT	Top 10	AR.	SFR	Expensen	Gradrate.
	1310	29	22	6	450000	94
	1200	83	34.	25	10000	73

1. Standardize.
2. Normalize.

x y.

SFR x_1	Expenses y_1
6 x_2	45000 y_2
25	10000
13	20000

Euclidean

$$d_1 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$= (25-6)^2 + (35000)^2$$

$$\Rightarrow [(19)^2 + (35000)^2]^{1/2}$$

$$(13-6)^2 + (25000)^2$$

$$49 + \longrightarrow$$

$$[361 + 1225000000]$$

SFR	Expenses
6 →	45K.
13	20 K
25 →	10 K.
10	33 k

$\bar{X}_{SFR} =$ ——

σ_{SFR}

Standardize → StandardScaler

$$Z_{SFR} \quad \frac{6 - \bar{X}_{SFR}}{\sigma_{SFR}} = -3$$

$$\frac{13 - \bar{X}_{SFR}}{\sigma_{SFR}} = +3$$

$$Z_{Expenses} \quad \frac{45K - \bar{X}_{Exp}}{\sigma_{Exp}} = -3$$

$$= +3$$

SFR	Expen.
−2.3	2.56
1.7	0.15
0.85	−1.32

Normalize

MinMax Scaler

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

$$\frac{6-6}{25-6} = 0 \rightarrow \quad 0$$

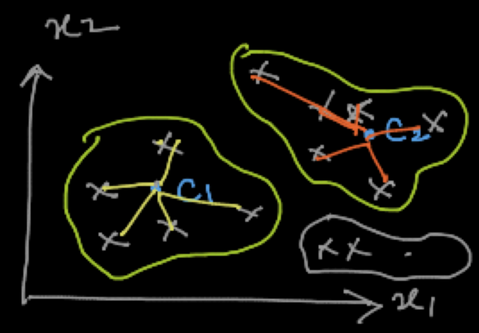
0.25

0.37

$$\frac{25-6}{25-6} = 1 \rightarrow \quad 1$$

K-Means Clustering.

No. of groups. ← (Hyperparam) $K = 2$



WCSS. (Within Cluster Sum of Squares)

$$= \sum_{i=1}^{n_1} (x_i - C_1)^2 + \sum_{i=1}^{n_2} (x_i - C_2)^2 + \dots + \sum_{i=1}^{n_K} (x_i - C_K)^2$$

$$\text{WCSS} = \sum_{j=1}^{K} \sum_{i=1}^{n_K} (x_i - C_j)^2$$

→ Lloyd's Algorithm

Task: Find those centroids which will minimize WCSS

- NP Hard problem
- Approximations (Lloyd's algorithm)

$K = 2$

$A(x_1, y_1)$, $B(x_2, y_2)$, $C(x_3, y_3)$

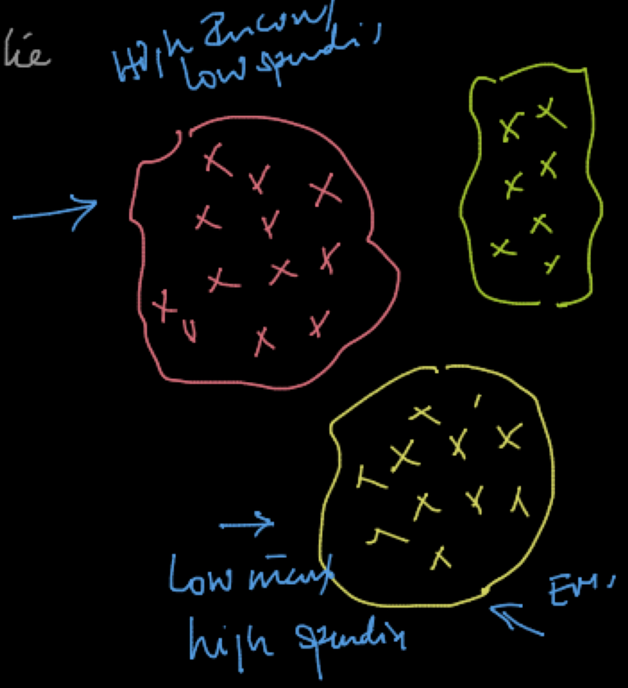
$$C = \left(\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3} \right)$$

$$C = (x, y)$$

Initialization Sensitivity

— Based on the choice of initial centroids the final clusters would be difference

x_1 x_2 x_3 x_4



K-Means++

→ Guidelines to choose the initial points

Lloyd's Algorithm

Step 1 : Randomly choose k-points as centroids.

Step 2 : Find the distance between every dp and the k-centroids.

Step 3 : Assign the dp to the closest centroid.

Step 4 : Recalculate the centroids based on the cluster created.

Repeat until Convergence

Convergence : The centroids don't change
— Stable clusters.

Step 1 : Define the objective function
— Usually the loss function ✓
— Some form of residual (Error - $(y - \hat{y})$)

Step 2 : Find model parameters which minimize / maximize the value of objective fn.

Step 1 : Define the objective function
— usually the loss function ✓
— Some form of residual / Error - $(y - \hat{y})$

Step 2 : Find model parameters which minimize / maximize the value of objective fn.

1. Which obj. fn. — Squared loss
2. What are we finding ? → β's which minimize the squared loss
3. What algo ? → OLS.
4. Metrics → R^2, MSE, RMSE, MAE.