Zoom Workplace Meeting View Edit Window Help

Machine learning

Supervised

— Labelled data

— Target is available

y - continuous

Regression

y categorical

Classfn

Unsupervised

→ unlabelled dat

→ No target

No prediction

Metrics

✓ MSE

✓ RMSE

✓ R^2

— MPG → HP, SP, WT, VOL

Cont

✓ Linear Reg

Dtree Reg

Random Reg

XGBoost Reg

→ Checked → Area, income, gender, time, usage

Cat

Logistic Regression ✓

Dtree clf

KNN clf

XGBoost clf

→

Metrics

Accuracy →

Precision →

Recall →

F1_Score →

AUC ROC →

Regressn

model

	x_1	x_2	x_3	y	$\hat{y}$	$(y-\hat{y})^2$
train				35	29	+6
				43	47	−4
				76	79	−3

actual

MSE

65	61	4
73	70	3
57	54	3

→ unseen data

MSE

$y = mx + c$

$y = \beta_0 + \beta_1 x$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \quad + \beta_m x_m$

β's → Model Parameters

Classfn

	x_1	x_2	x_3	y	$\hat{y}$
train				0	0
				1	0
				1	1
test				0	0
				1	0

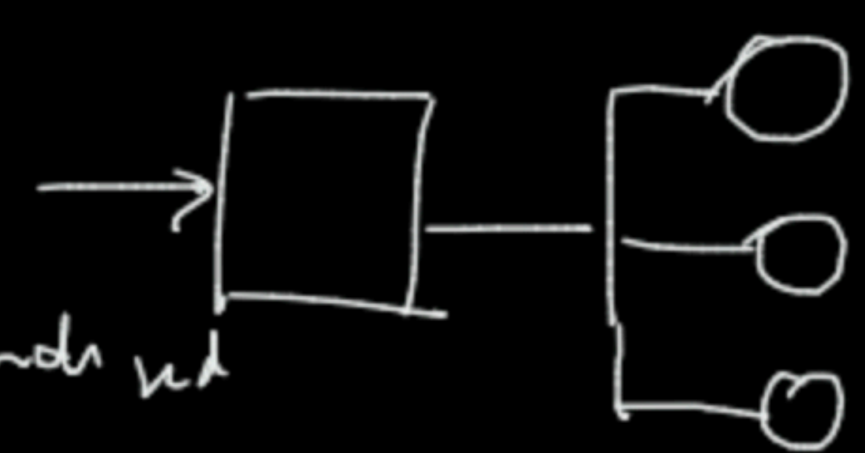
→ unseen

→ quality

Acc,

Acc $\left(\frac{3}{5}\right)$ ←

TP	FP
FN	TN



Similarity

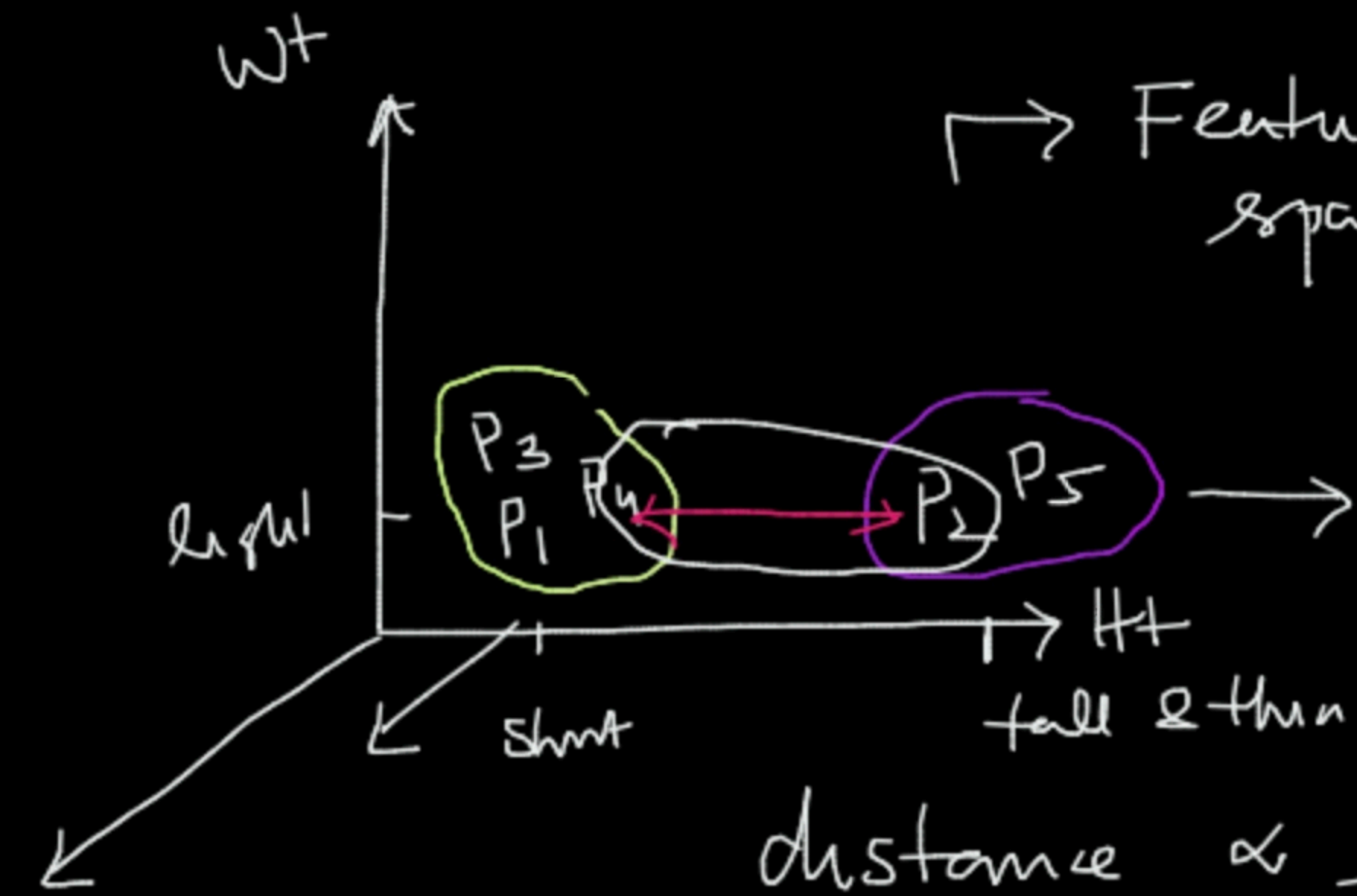
unstandardized

Clustering → WCSS ✓ — 5

Silhouette score ✓ (−1 to +1)

Dunn Index ✗ — Standardized

	Ht	Wt	Age	[illegible]	[illegible]
✓ P_1	—	—			
✓ P_2	—	—			
✓ P_3	—	—			
✓ P_4	—	—			
✓ P_5	—	—			



→ Feature space

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

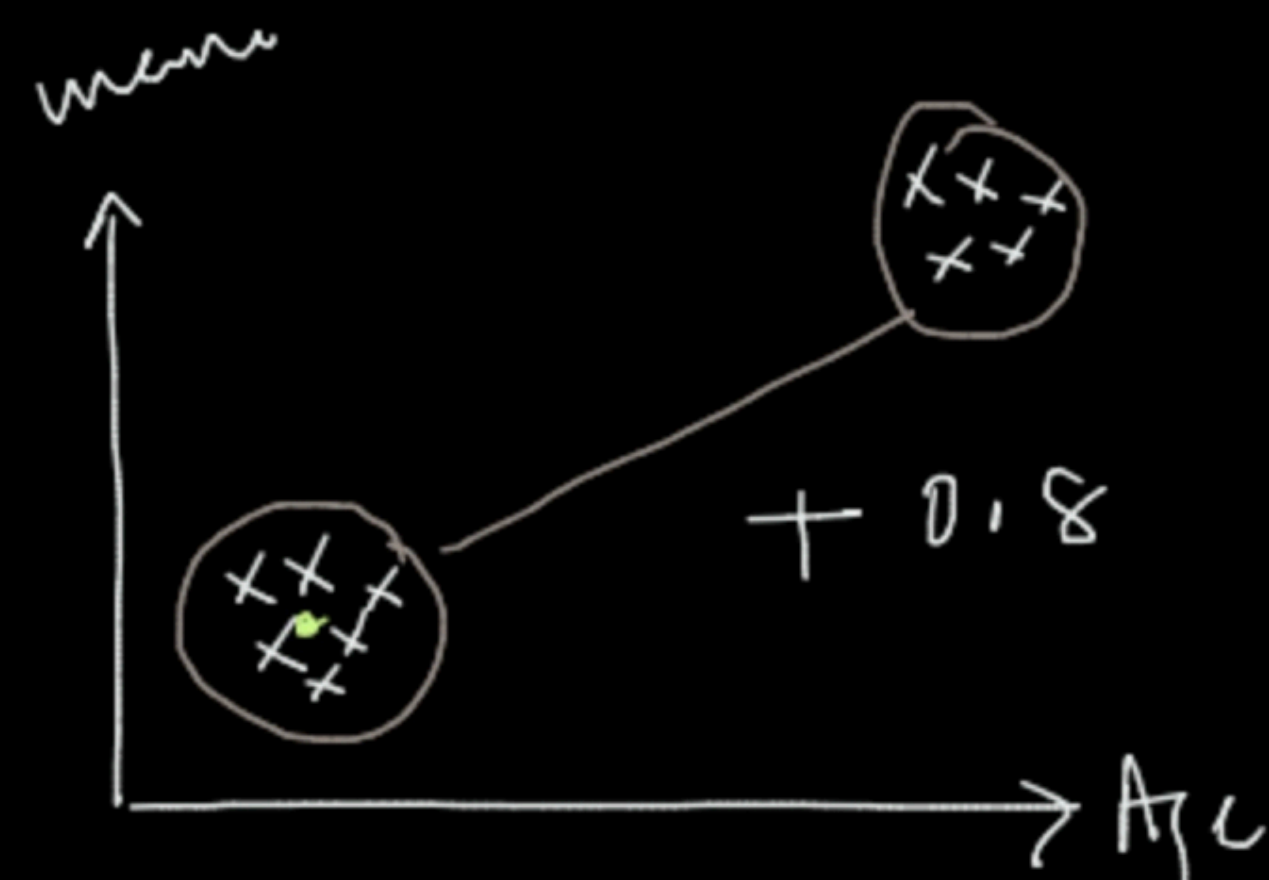


Fig 1

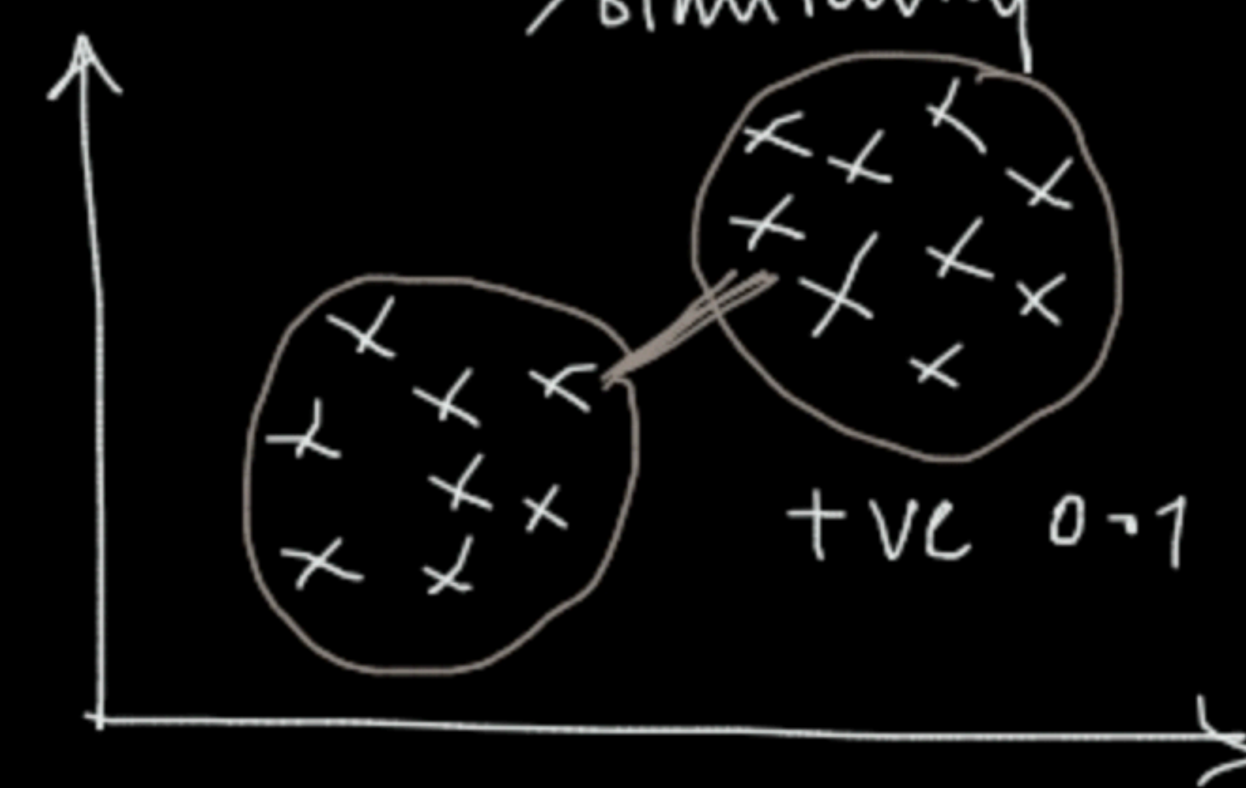


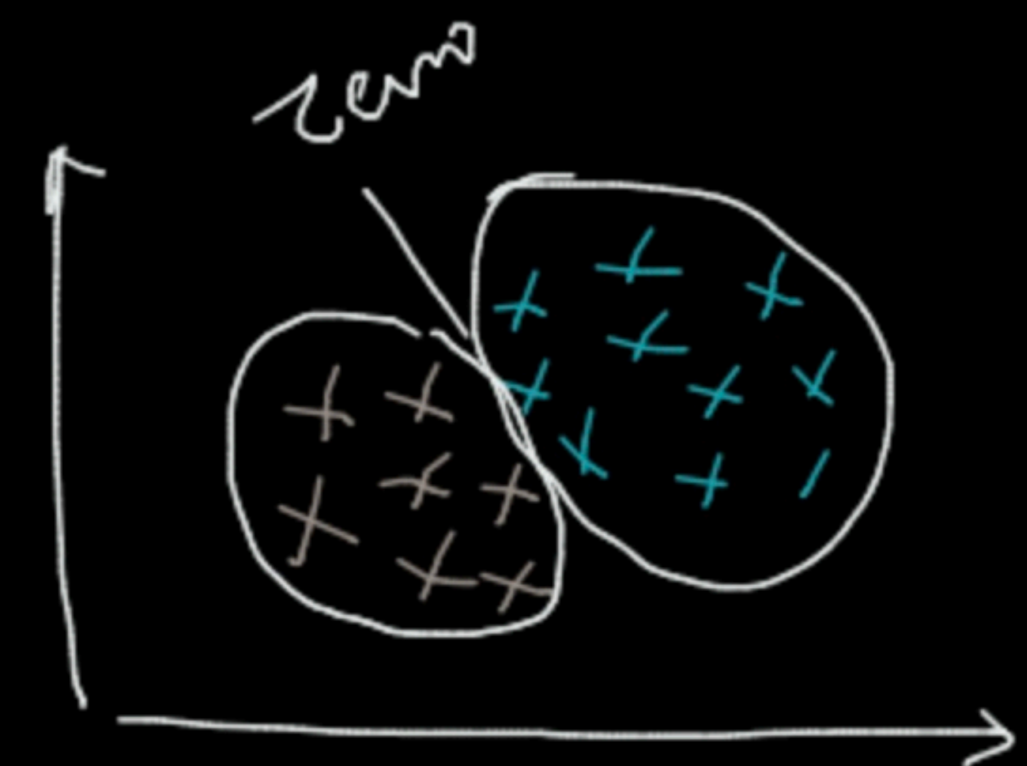
Fig 2

(intra cluster distance)

1 WCSS → As small as possible

2 Intercluster distance — As large as possible

→ Silhouette score → −1 to +1



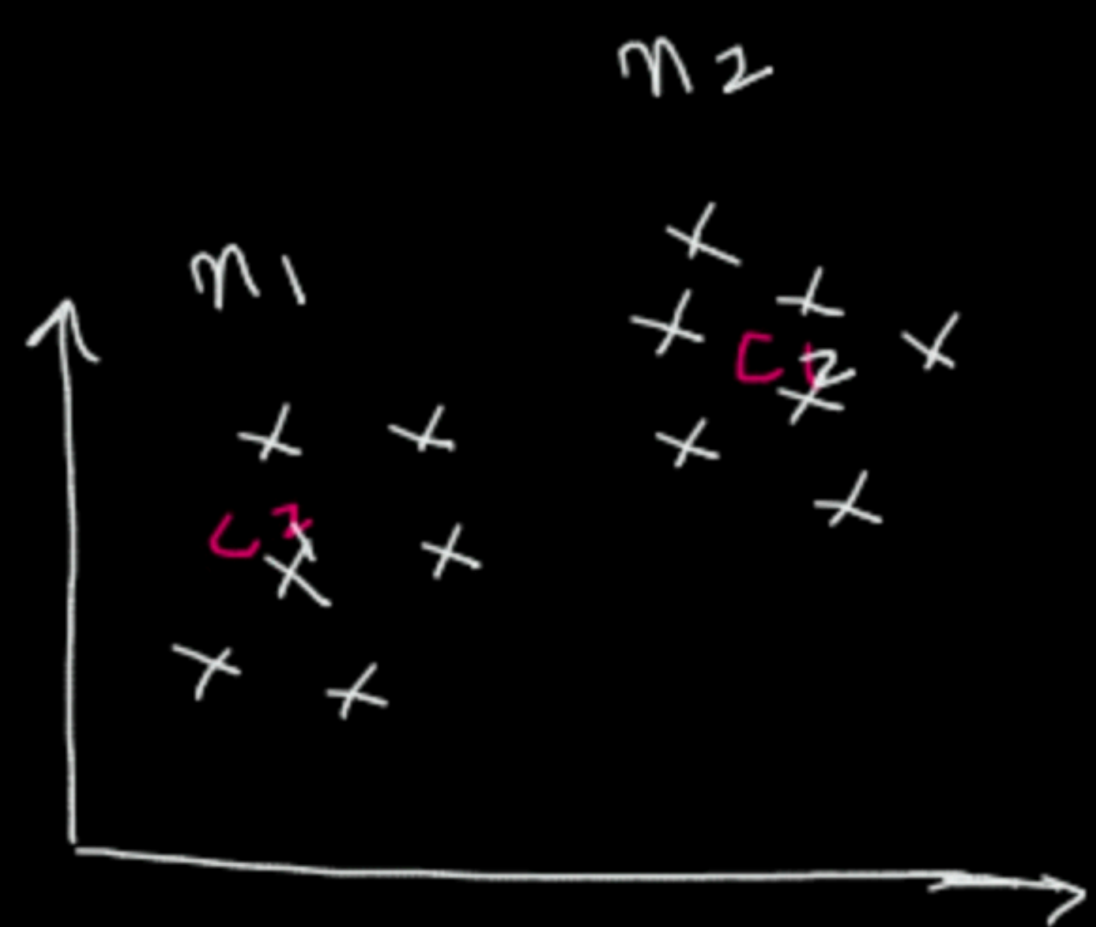
−ve silhouette score

→ KMeans clustering

→ DBSCAN

K Means Clustering

↳ No of clusters → 2



WCSS — Within Cluster Sum of squared distances

Task:

Find $\underline{k}$ $\underline{\text{centroids}}$ s.t the $\boxed{\text{WCSS}}$ is minimum

$$\sum_{i=1}^{n_1}(x_i-c_1)^2 + \sum_{i=1}^{n_2}(x_i-c_2)^2 + \qquad + \ k \text{ Clusters}$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_i-c_j)^2 \rightarrow \text{WCSS}$$

NP Hard problems

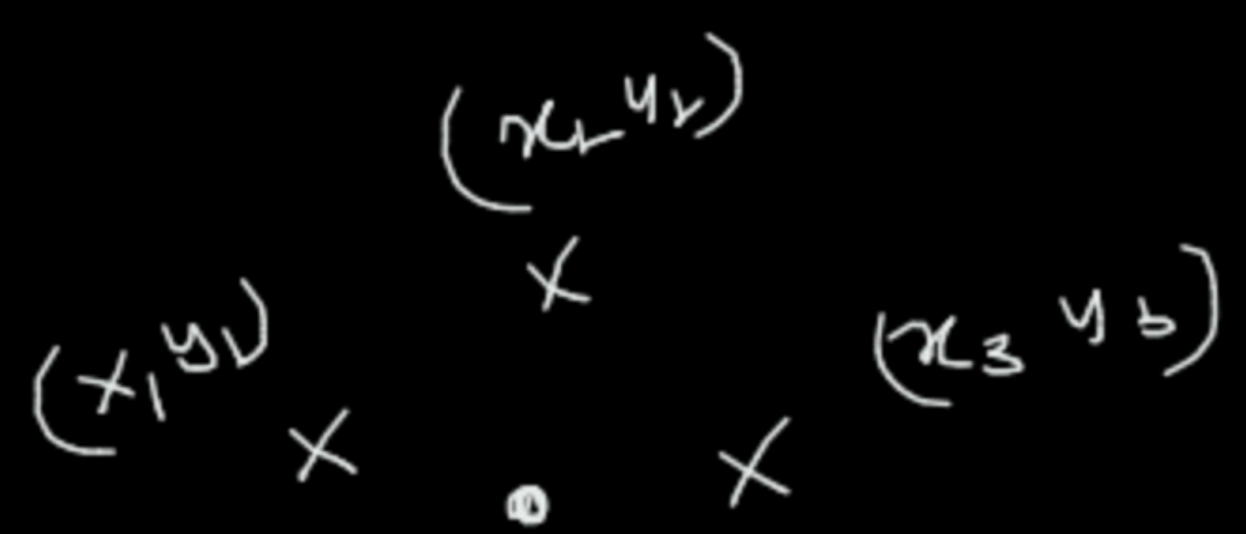
— Lloyd's approximation

Step 1 Randomly choose k-points from the dataset as centroids

Step 2 Find the distance betn every dp and the k centroids

Step 3 Assign points to the closest centroids

Step 4 Recalculate the centroids



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

KMeans ++

Problems

1 Initialisation sensitivity

— Final cluster depend on the choice of initial centroids

2 Outliers are not handled properly

Hyper parameters

— We supply to the model

Model parameters

— estimated by the model from the data

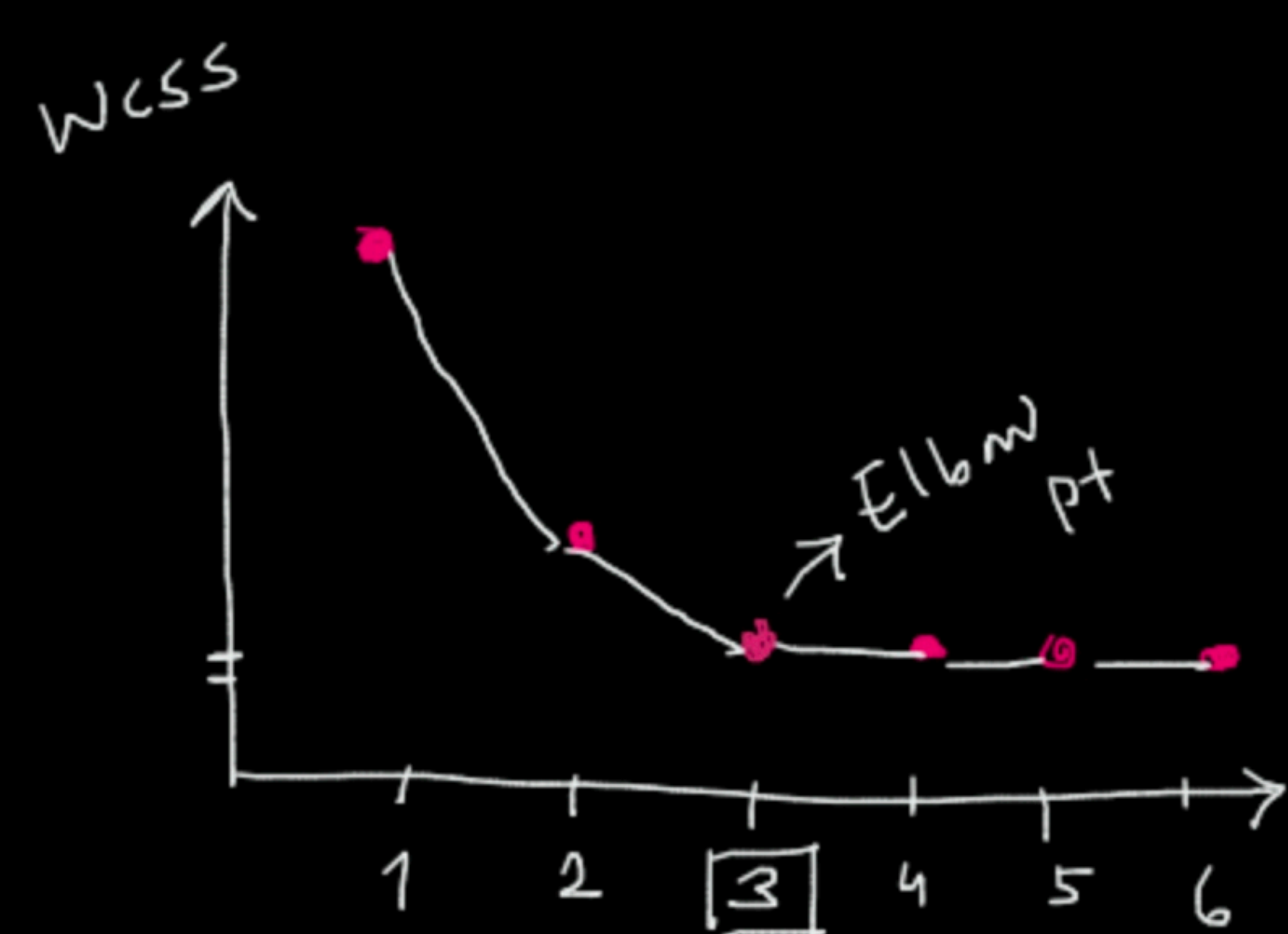
Hyperparameter $\longrightarrow$ $\boxed{}$ $\longrightarrow$ Model parame

K $\rightarrow$ Hyperparameter

$\rightarrow$ No. of clusters

$\rightarrow$ Elbow Method

Athul Ansu Assign Not sharp



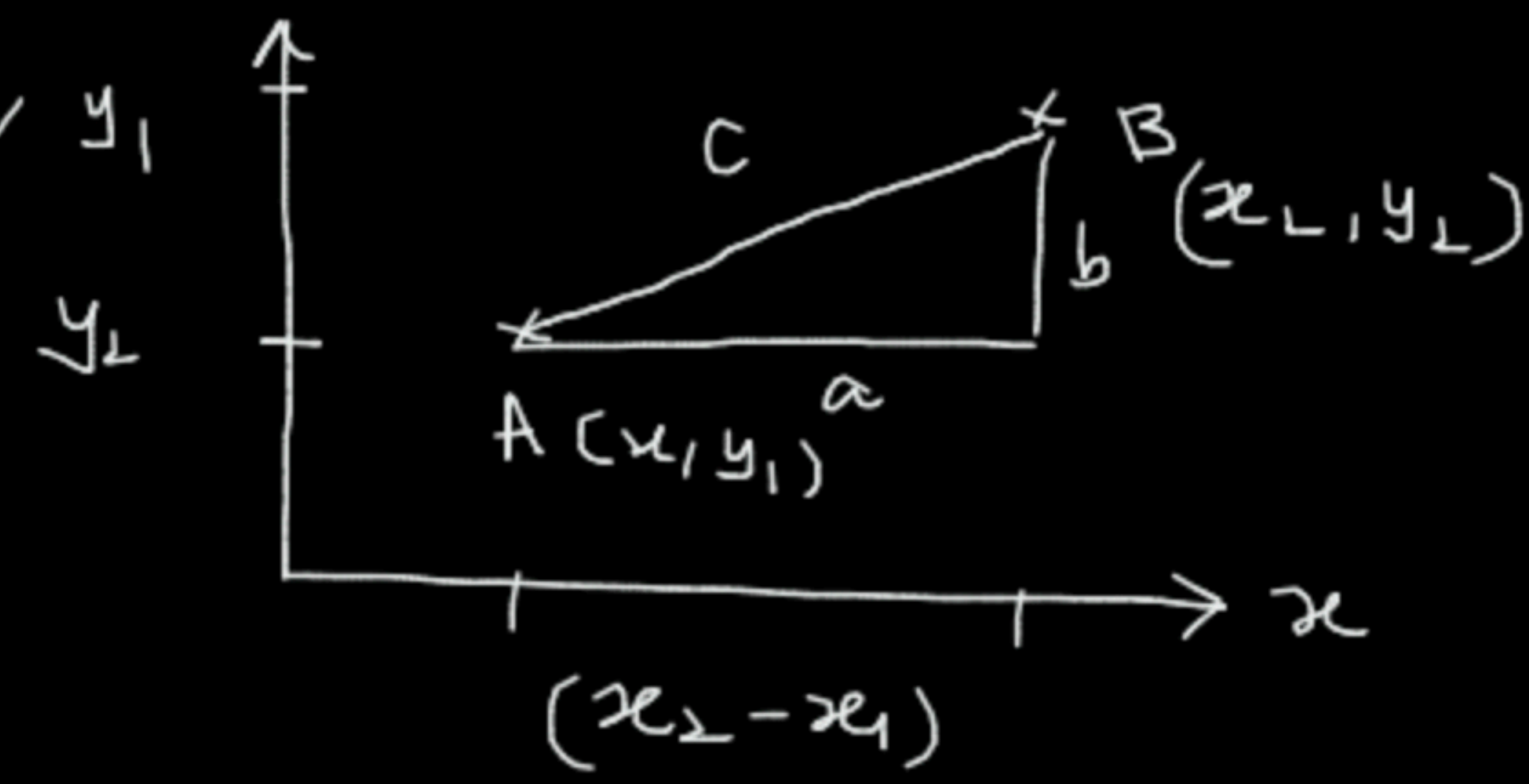
Euclidean distance

Ht Wt Gender y_1

1

2

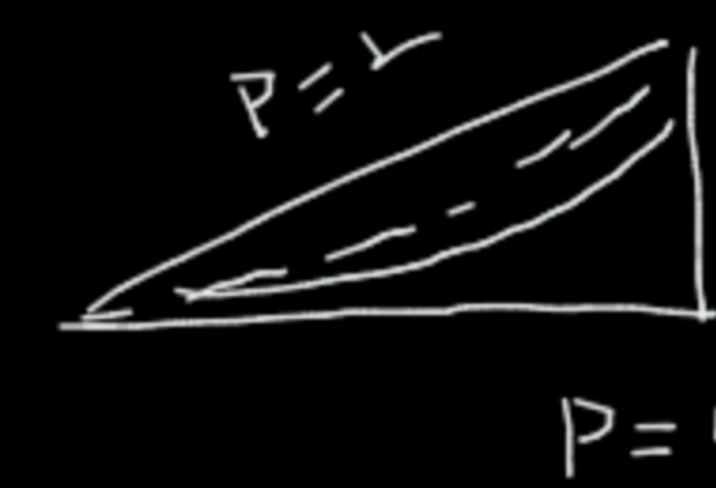
3



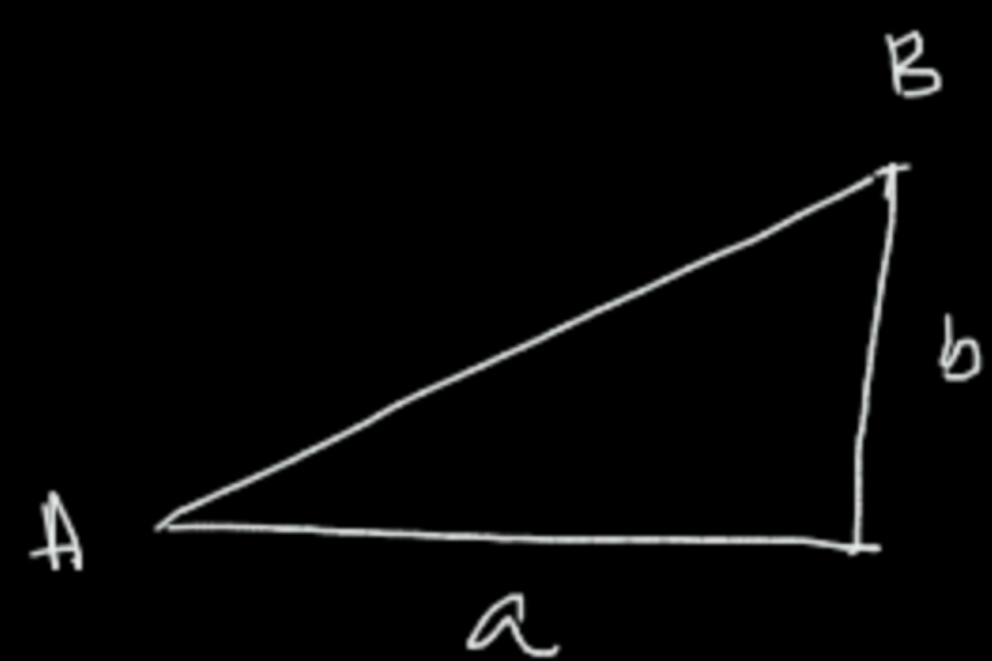
$$C = \sqrt{a^2 + b^2}$$

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

'As the crow flies'



Manhattan Distance — taxi distance



$$C = a + b$$

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

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

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

$p = 1$ $(x_2 - x_1) + (y_2 - y_1)$

PowerPoint File Edit View Insert Format Arrange Tools Slide Show Window Help

All Features are Nr

— Euclidean

— Manhattan

— Minkowski

— Mahalanobis

Ht Wt age

P_1

P_2

P_3

Categorical

Binary Euclidean

Simple Matching Co-efficient

Jacquard's dist

Gender Job Married Smoker

P_1

P_2

Mix of Nr & Cat

Gower's dissimilarity index

Gend Age income Married

P_1

P_2

P_3

9 / 9

Opera File Edit View History Bookmarks Developer Window Help

	Age ✓	Income ✓
P_1	28 ×	15L
P_2	32 ×	25L

$$\sqrt{(28-32)^2 + (15L - 25L)^2}$$

$$= \underline{\underline{16}} + 10,00,000\ 000000$$

Scaling

- Standardize $\rightarrow$ $z_x = \dfrac{x-\mu}{\sigma} \rightarrow -3 \ \& \ +3$
- Normalizing $\rightarrow$ $N_x = \dfrac{x - x_{min}}{x_{max} - x_{min}} \rightarrow 0 \ \& \ 1$

Age	NAge
27	
32	
45	
20 →	0
54	1

$$\frac{27-20}{54-20} = 7/34 \rightarrow 0$$

$$\frac{54-20}{54-20} = 1$$

$$\frac{20-20}{34} \rightarrow 0$$

10 / 10