Day 8

1. K-means ← Unsupervised ✓
2. KNN ← Supervised. ✓
3. Overfitting
4. Underfitting

} concepts

1. K-means

K ← # of clusters.

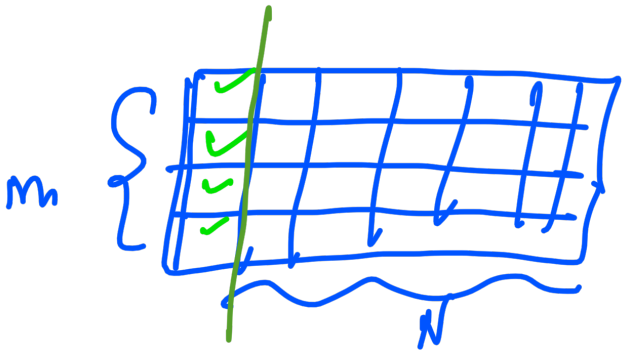
K = 2

Dataset → # samples

m x N

↓

attributes

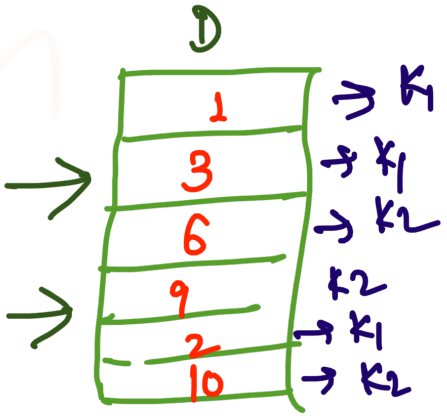


2D K=3

Example 1D m x 1

m = # samples

1 ← feature



6 K1

1. K = 2 ←
2. $K_1 = 3$ $K_2 = 9$
3.

		K_1	K_2
D_1	1	$(1-3)^2 = 4$	$(1-9)^2 = 64$
D_2	3	✗	✓
D_3	6	$(6-3)^2 = 9$	$(6-9)^2 = 9$
D_4	9	✗	✓
D_5	2	$(2-3)^2$	$(2-9)^2$
D_6	10	$(10-3)^2$	$(10-9)^2$

→ $K_1 = \{1, 3, 2\}$ ←
→ $K_2 = \{6, 9, 10\}$ ←

1st iteration

1. Centroid } ←
2. Mean }

K_1 Centroid = 2
mean = 2

K_2

$K_1 = \{1, 3, 2\}$

$K_2 = \{6, 9, 10\}$

1st $K_1 = 3$ $K_2 = 9$

2nd $K_1 = 2$ $K_2 = 9$

$K_1 = \{1, 2, 3\}$
$K_2 = \{6, 9, 10\}$

K_1	K_2
2,	9

$D_1 = [1-2]^2 \quad [1-9]^2$

$D_2 \quad [3-2]^2 \quad [3-9]^2$

:

Iteration: Stop → { Centroid, K, is not changing; Cluster assignment is not changed }

Steps (K-means):

1. Select the number of clusters (K)
2. Select the no. of data points = K
3. Assign each datapoint to one of the 'K'
   - 3.1. Distance
     - ✓ — L_2 norm $[a-b]^2$
     - — L_1 norm $a-b$
     - — $|a-b|$

     } Research area
   - 3.2 Find the min^m and assign it to that 'K'
4. Find the cluster centroid.
5. Repeat step 3 untill convergence.

Convergence

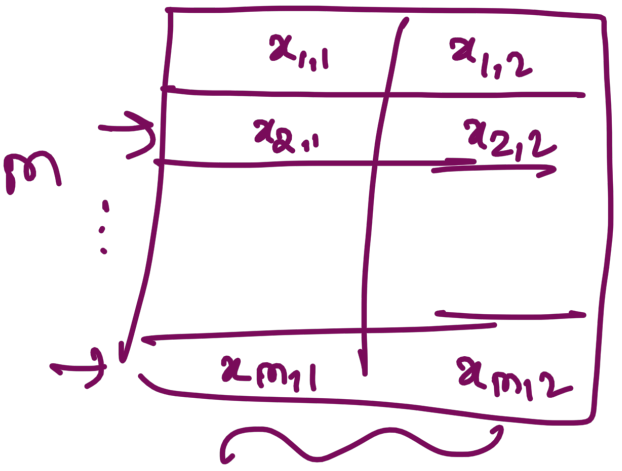
- 'k' clusters centroid is not changed
- assignment of data points are not changed.

Dataset $m \times 2$ 2 features

Clusters assignment

$K_1 = (x_{N,1}, x_{N,2})$

$K_2 = (x_{N',1}, x_{N',2})$



Euclidean distance - (ED) 2

1st point: $\sqrt{(x_{1,1} - x_{N,1})^2 + (x_{1,2} - x_{N,2})^2}$

1st point, (1, 2)

K_1 (3, 4)

$(ED) = \sqrt{(1-3)^2 + (2-4)^2} = x$

$N < m$

$N' < m$

$N \neq N'$

$N, N' \subseteq m$

K-Means

Applications

S = Symptoms

D = Disease

→ Popular

→ Social media Analysis.

→ Healthcare $S_1 S_2 S_3 \rightarrow C_1 \rightarrow D_1$

→ Segmentation

→ Market analysis

Major Research Problem (K) ?

Unsupervised → It will not have class label.

How to decide the value 'K' ?

Supervised

Classification

class labels

↓

discrete

Regression

class labels

↓

continuous.

Unsupervised

→ Clustering
→ Kmeans.
(pattern)

→ Association.
→ grouping.
→ Hierarchical

Advantage of using an association:
→ the value of k need not to be given.

Hierarchical Clustering -

→ Agglomerative (A)
→ Divisive. (top down approach)

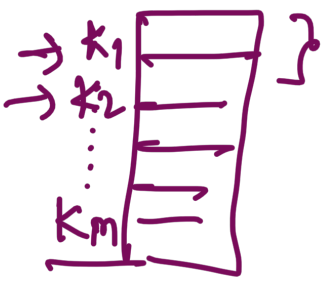
(A) - Consider the entire dataset
#datapoint = # K

bottom-up

end up with 1 cluster

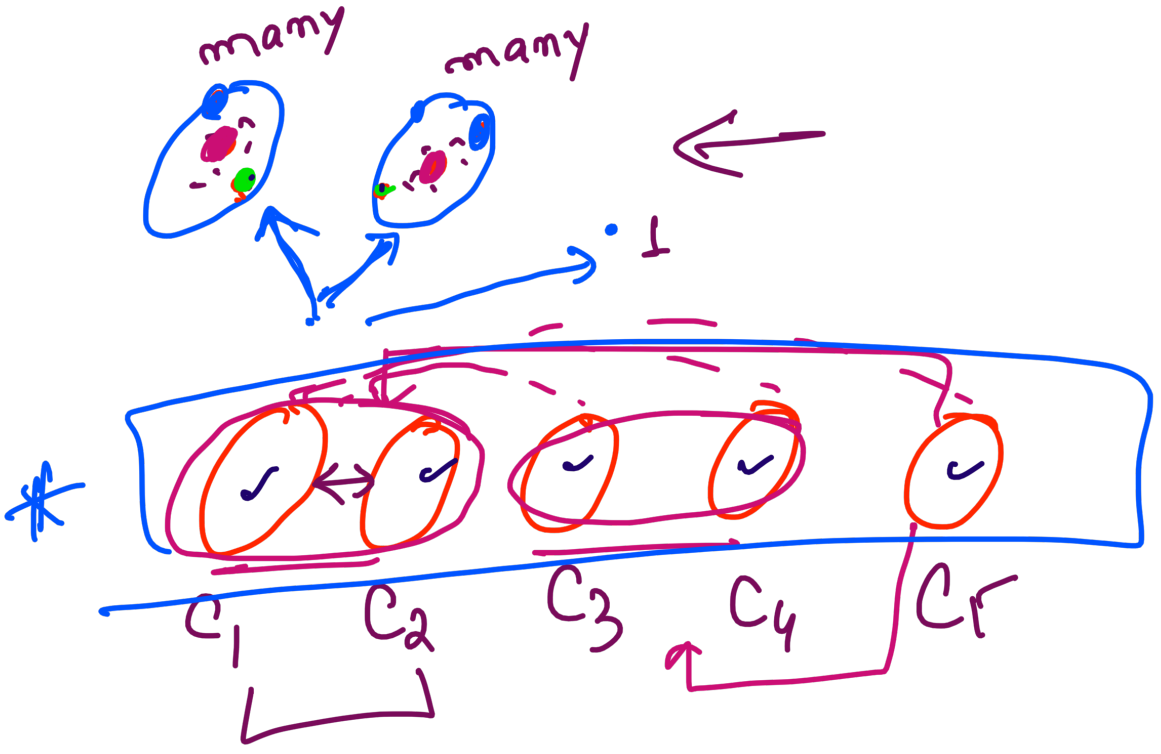
intial = K centroids

Step 1 : Assign each data point as cluster centroid



Step 2: (k_1, k_5) : $(k-1)$ → centroid

Step 3: (k_2, k_6) : $(k-2)$ → remaining.



Step 4: Find the distance between cluster

→ Single Linkage ✓
- shortest distance between two clusters

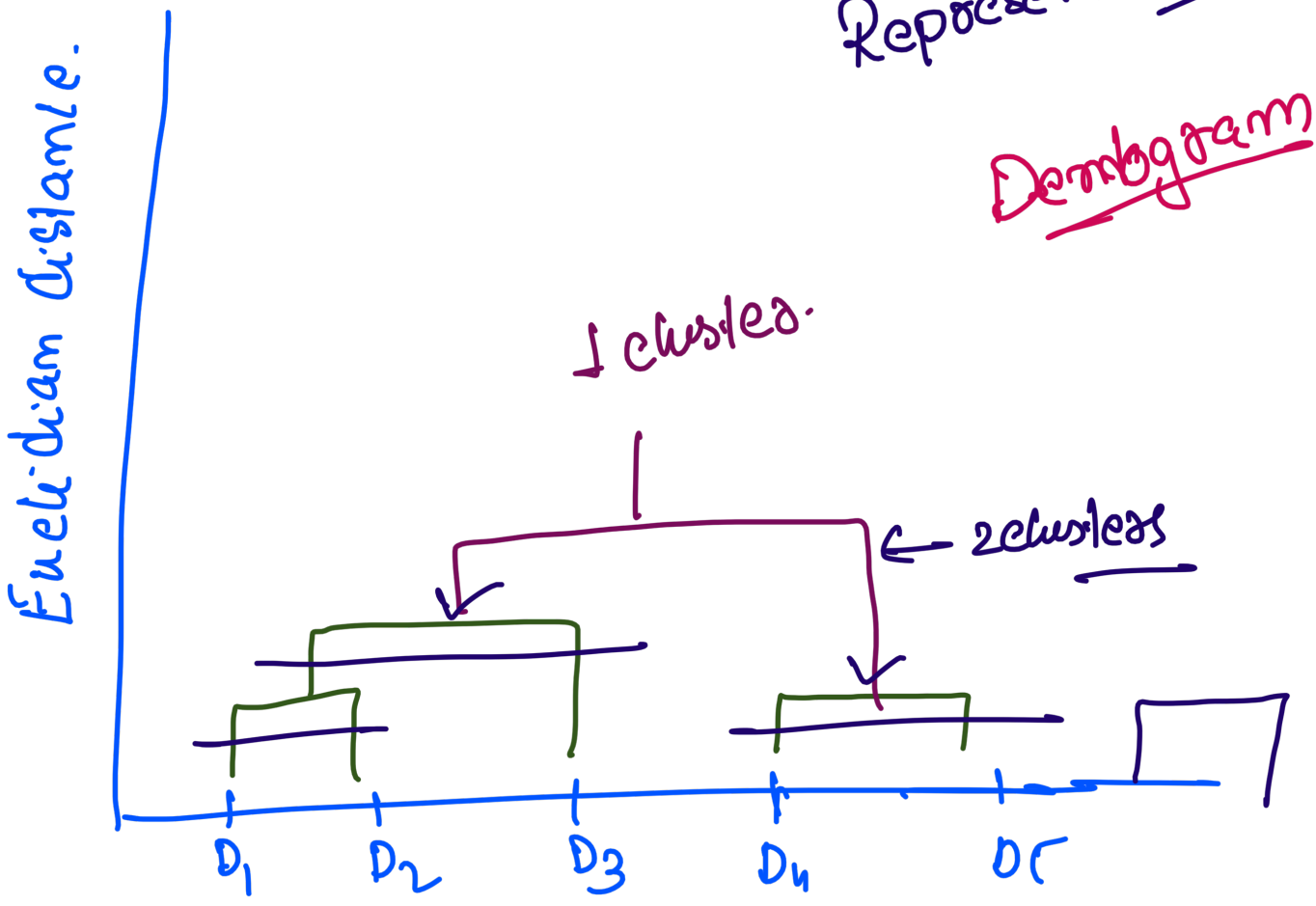
→ Complete linkage: ✓
- Farthest distance

→ Average / Centroid linkage ✓

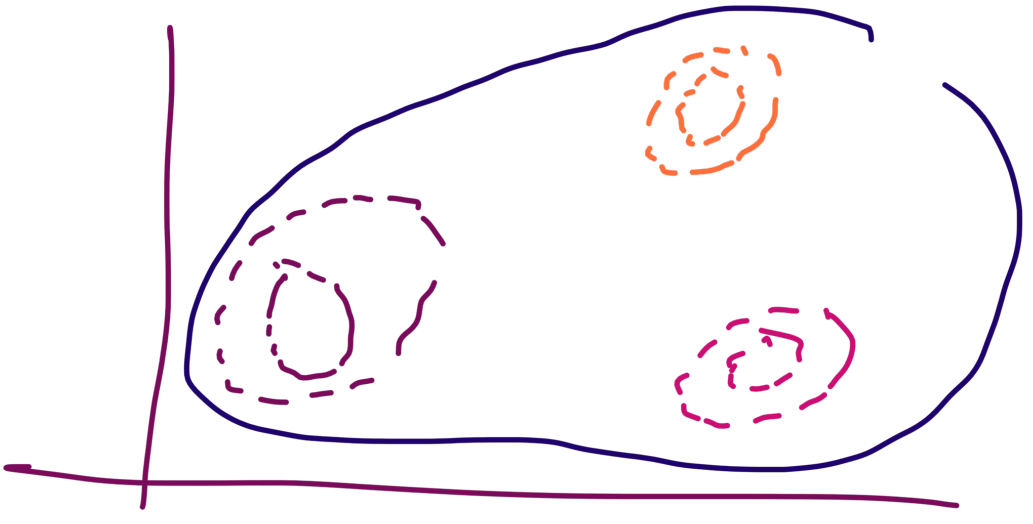
*

Representation

Dendogram

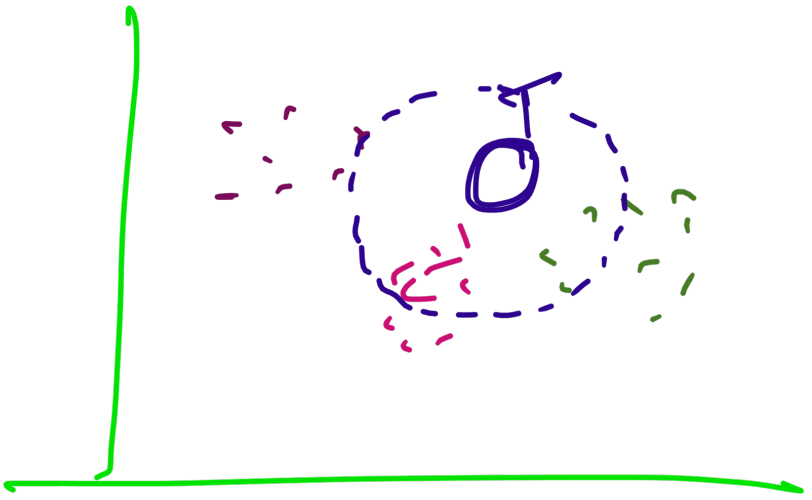


→ Cluster Similarity.



K-NN Nearest Neighbour

(Supervised)



K ← datapoints

K = 5 ←

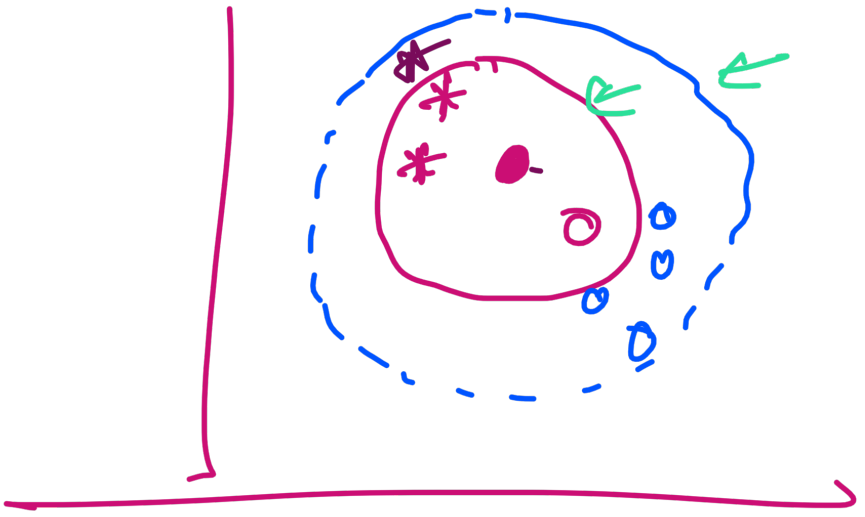
T it will take L_2 distance with 5 datapoint which falls nearer to it

$(T - D_1)^2 = a$

T
- $D_1 \rightarrow$ (a)
- $D_2 \rightarrow$ (b)

} → C_1

- $D_3 \rightarrow c$
- $D_4 \rightarrow d$
- $D_5 \rightarrow e$



K=2

* ← *

K=8

O ← O

Lazy algorithm

---

1. Overfitting
2. Underfitting.

Training (Good)

Dataset → Algorithm → model.

Learning

Testing

Inference — it not getting good results

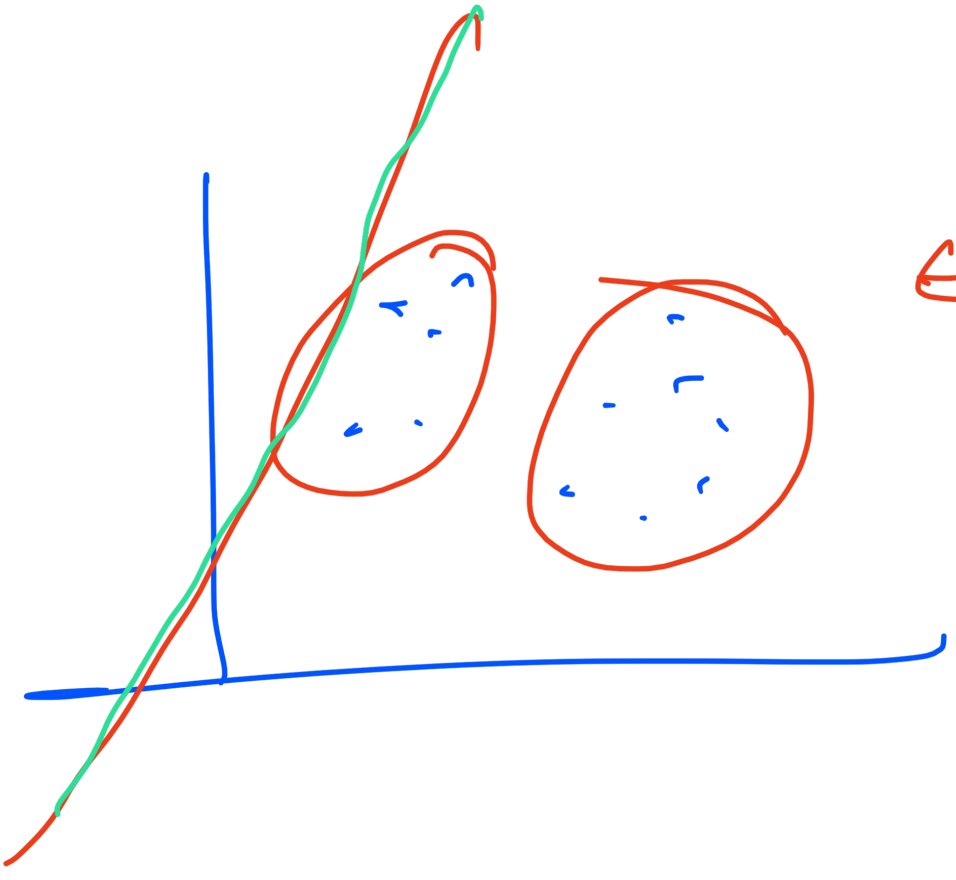
Overfitting

2. Underfitting

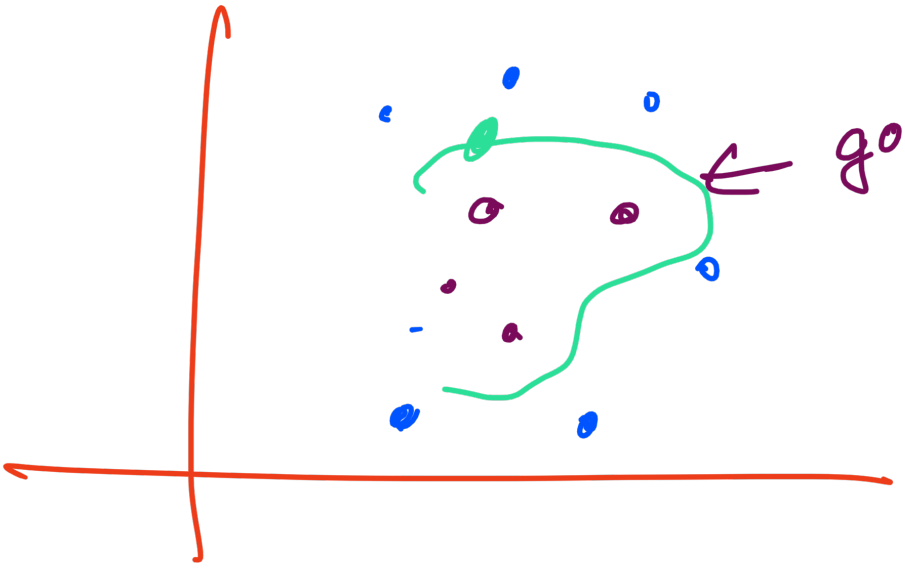
Training (bad)

Testing (bad)

model is not learn any point



← Underfitting



good for training

testing good result X

overfitting