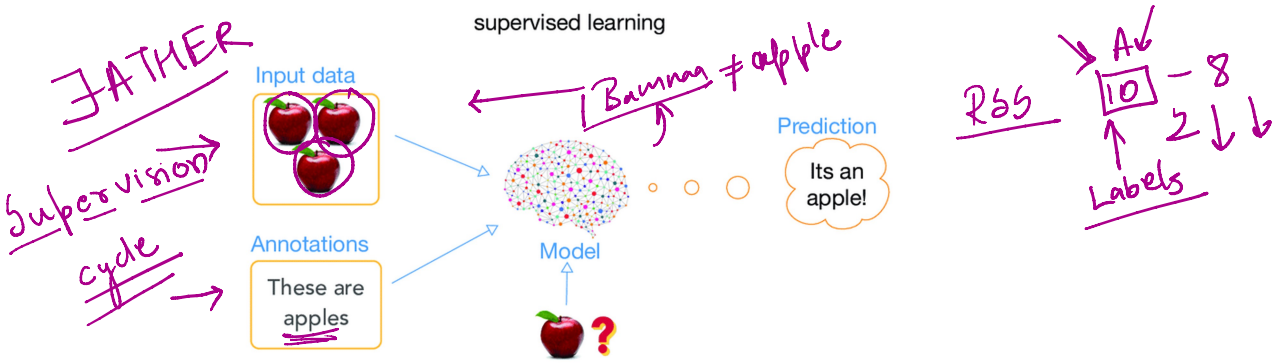
Unsupervised Learning

Sunday, January 8, 2023 8:29 PM

Difference → Supervised and Unsupervised

Labels

Without Labels



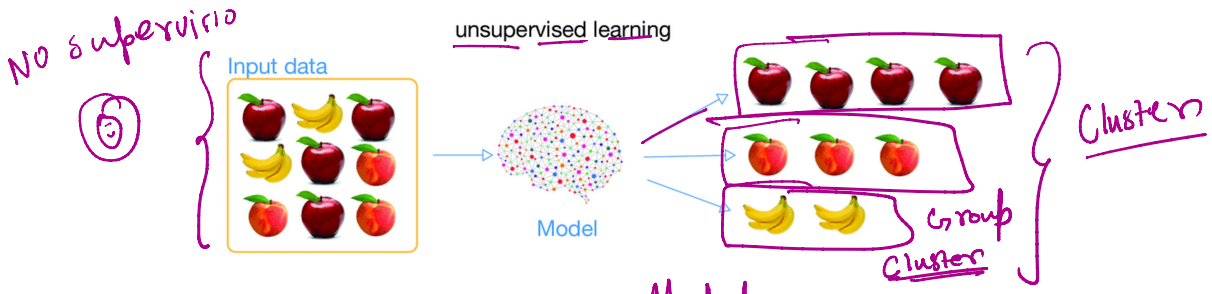
FATHER

Supervision

cycle

RSS

Labels



No supervision

Model is going to learn

Amazon : Labels

100M → Unsupervised : Labels ~~Labels~~ : 100M

1		10,000	Tshirt Branded
2		299	Tshirt Non Branded

High
Low
Medium

20 { Team Data Labelling

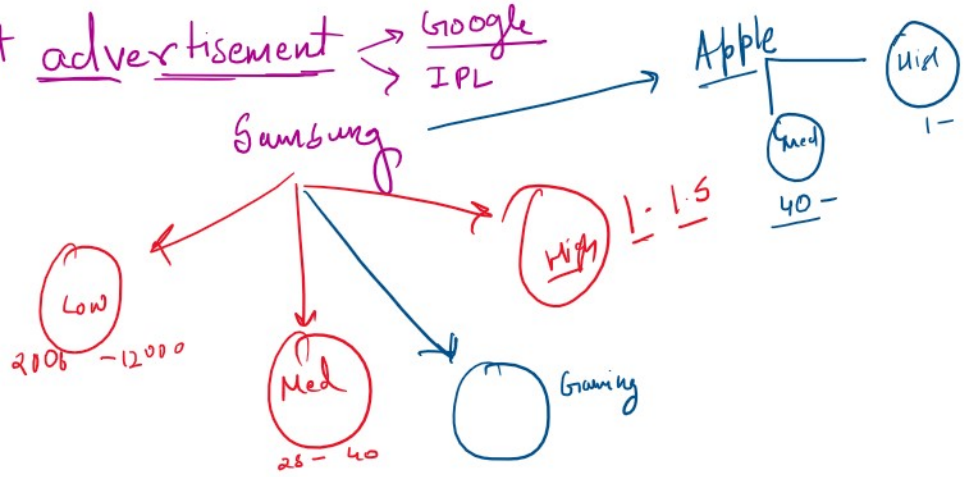
1 Month → T, M, Delayed

Branded { 10,000, 20,000, 15,000

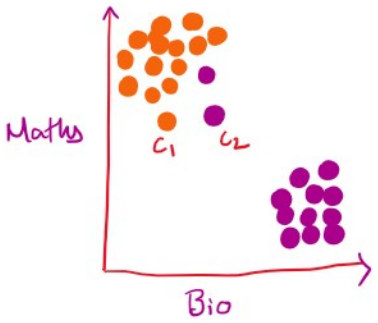
200000 } 1 copy

2000 5000 } Non D

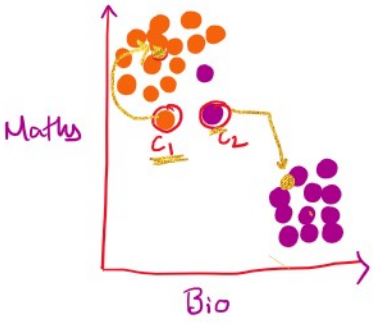
• Target advertisement → Google
→ IPL



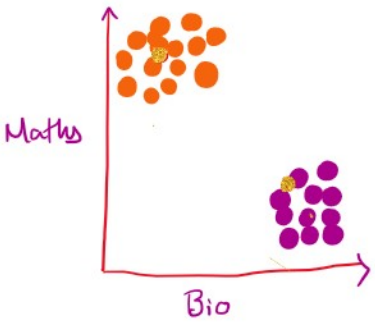
K Means



2 Step
- → 1 step
  : Randomly assign Clusters Centroid
- → 2 step
  Assignment Step



2 Step
- → 1 step Centroid for cluster
  : ~~Randomly assign Clusters Centroid~~
- → 2 step
  Assignment Step

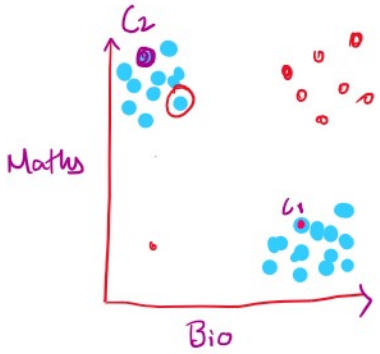


2 Step
→ 1 step centroid for cluster
: ~~Randomly assign Clusters Centroid~~
→ 2 step Assignment Step

- Pick the first centroid point (C_1) randomly.
- Compute distance of all points in the dataset from the selected centroid. The distance of x_i point from the farthest centroid can be computed by

From <https://towardsdatascience.com/understanding-k-means-k-means-and-k-medoids-clustering-algorithms-ad9c9fbf47ca>

K Means ++ } Smart Cluster centroid

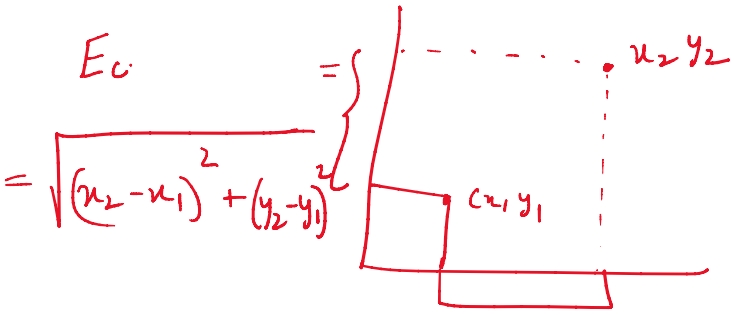


1. Randomly select 1 Point from the given Point

(2) dist is higher, highest would be your next centroid

(3) Calc the dist to its nearest centroid

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



K Means

Centroid

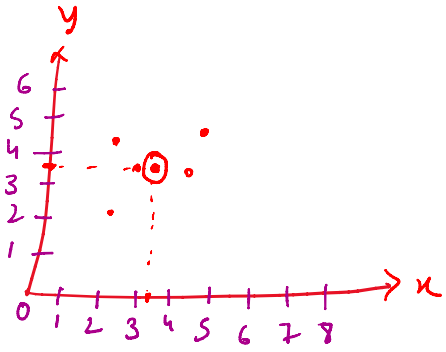
↓

No of cluster

Centroid

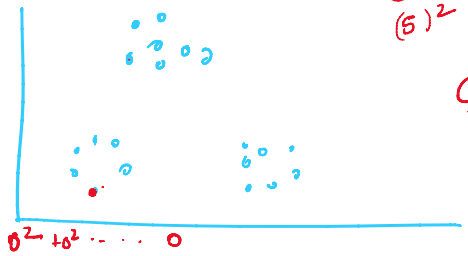
$$x = \frac{2+2+3+4.5+5}{5} = 3.3$$

$$y = \frac{2+3.5+4.5+3.5+4.7}{5} = 3.6$$



$SSD_{C_1} + SSD_2 \rightarrow V_2$

$d_1^2 + d_2^2$



$(10)^2$

$(5)^2$

Elbow → Sum of square distance

Cluster K=1 → SSD = V_1

K=2 SSD = V_2

K=3

⋮

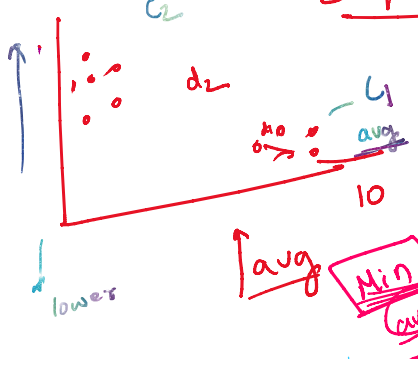
K=n SSD = 0

Sum $d_1^2 + d_2^2 \cdots d_n^2$ n

{Silhouttee, Hierarchical} → {PCA, LDA} TS

Association Rules

Silhouette Score



$$\frac{b-a}{Max(a,b)}$$

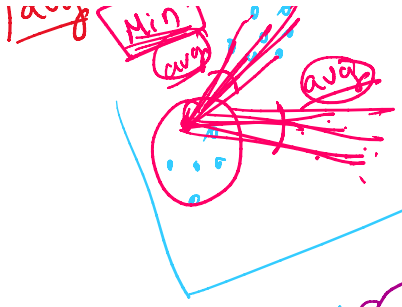
avg → Min (avg, avg)

a = intra → cohesion

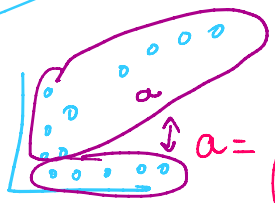
a = avg diff within a cluster → Lower

b = inter → Sepration

lower



b = inter Sepration
↳ average distance between cluster

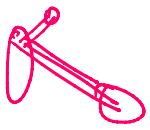


a = intra

a → lower
b → higher

$\frac{b-a}{Max(a,b)}$

b = inter



$\frac{1-10}{Max(1,10)}$ → $\frac{10-1}{10}$

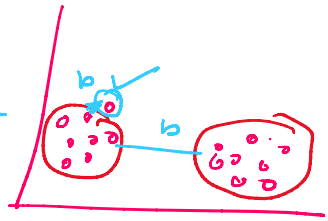
$\frac{-9}{10}$

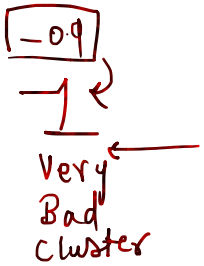
$\frac{9}{10}$

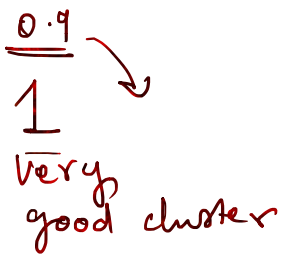
−0.9
Bad

0.9
good

b = 100

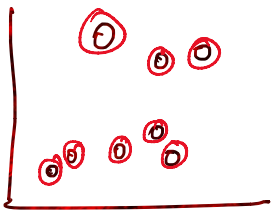




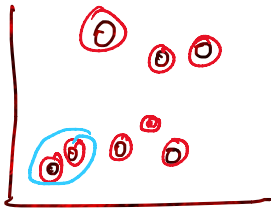


X

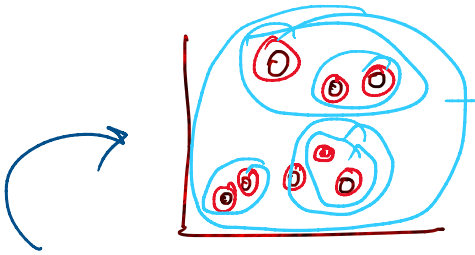
Hierarchical Clustering



Every Point is consider as own cluster



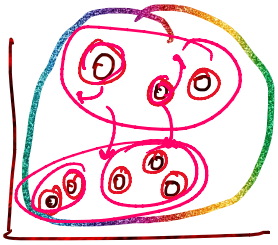
You Merge the cluster which are closer



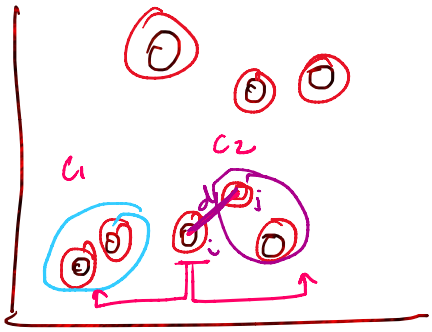
Agglomerative

Divisive

All Points belong to one cluster



Diving the cluster



distance between a Point and a cluster } Linkage

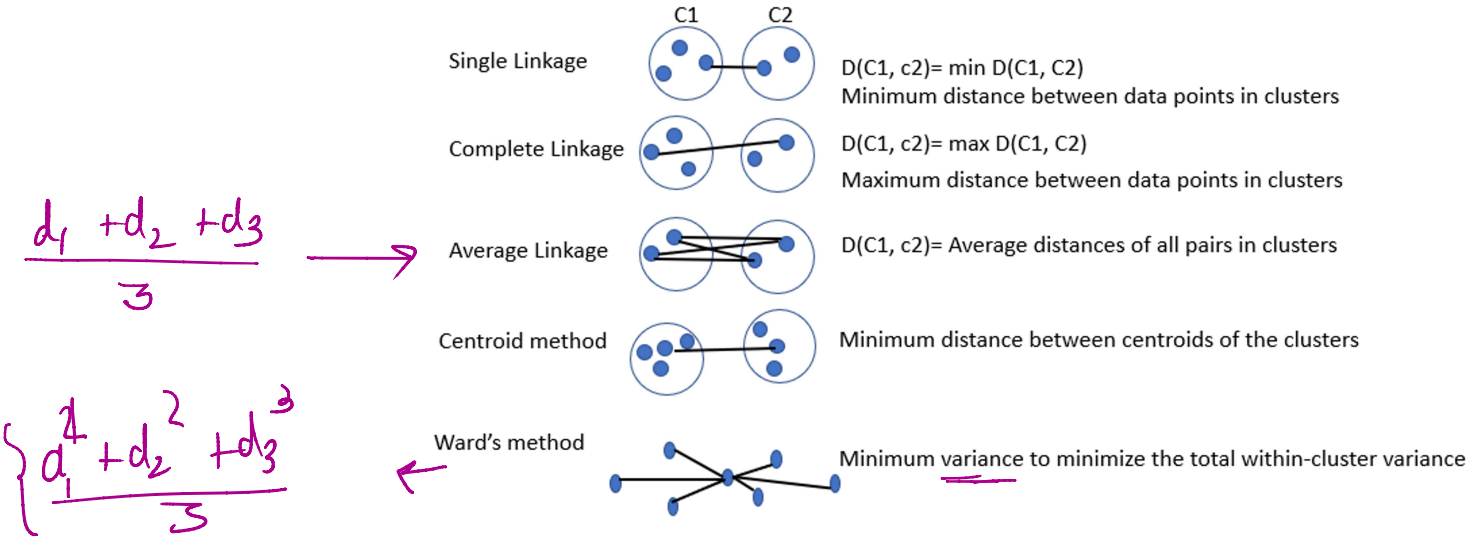
Single linkage (Min (d_1, d_2)

Average linkage (avg (d_1, d_2)

complete linkage (Max (d_1, d_2)

complete linkage ($\max(d_1, d_2)$)

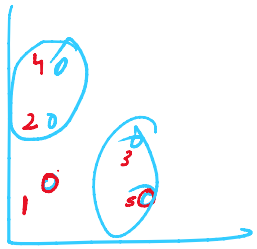
Ward $\left(\frac{\sum Sim(P_i, P_j)}{|C_1|\,|C_2|} \right)^2$



$\frac{d_1 + d_2 + d_3}{3}$ →

$\left\{ \frac{d_1^4 + d_2^2 + d_3^3}{3} \right.$ ←

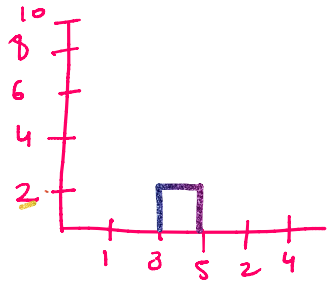
	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0



1 step

find the closest point

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0



	35	1	2	4
35	0			

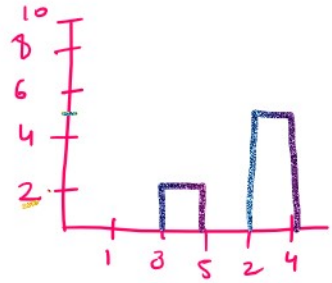
2nd dist of Point and cluster.

Complete Linkage

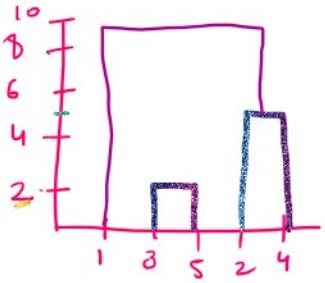
	35	1	2	4
35	0			
1	11	0		
2	10	9	0	
4	9	6	5	0

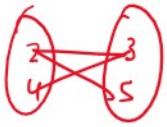
3 step

	35	1	2	4
35	0			
1	11	0		
2	10	9	0	
4	9	6	5	0



	35	24	1
35	0		
24	10	0	
1	11	9	0





2 3
5
4 3
4 5

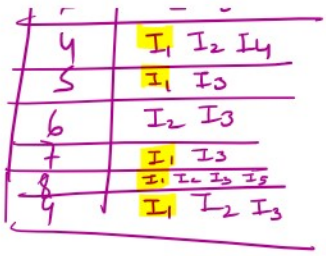
2 3 - 7
2 5 - 10
4 3 - 9
4 5 - 8

Association Rules — Apriori

Id	Name of Product
1	I_1 I_2 I_5
2	I_2 I_4
3	I_2 I_3
4	I_1 I_2 I_4
5	I_1 I_3

→

1 item set	Frequency
I_1	- 6
I_2	- 7

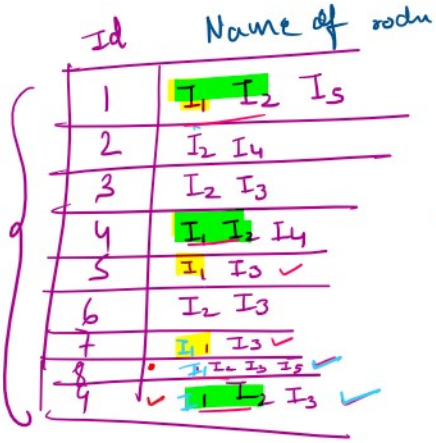


4	I_1 I_2 I_4
5	I_1 I_3
6	I_2 I_3
7	I_1 I_3
8	I_1 I_2 I_3 I_5
9	I_1 I_2 I_3

→

I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Threshold = 2



Id	Name of items
1	I_1 I_2 I_5
2	I_2 I_4
3	I_2 I_3
4	I_1 I_2 I_4
5	I_1 I_3
6	I_2 I_3
7	I_1 I_3
8	I_1 I_2 I_3 I_5
9	I_1 I_2 I_3

→

2 Item Freq Set

I_1 I_2	4
I_1 I_3	4
I_1 I_4	1
I_1 I_5	2
I_2 I_3	4
I_2 I_4	2
I_2 I_5	2
I_3 I_4	0
I_3 I_5	1
I_4 I_5	0

3 Freq.

I_1 I_2 I_3	2
I_1 I_2 I_4	1
I_1 I_2 I_5	2
I_1 I_3 I_4	0
I_1 I_3 I_5	1
I_1 I_4 I_5	0
I_2 I_3 I_4	0
I_2 I_3 I_5	1
I_2 I_4 I_5	0
I_3 I_4 I_5	0

$I_1\ I_2\ I_3$
$I_1\ I_2\ I_5$

Freq Item 3 → {1, 2, 3} {1, 2, 5}

{1} {2} {3} ✓
{1, 2} {1, 3} {2, 3}

✓ Support → 2/9
✓ confidence → 50%

Rule 1 → {1} → {2, 3}

CD
1 → Buy Mobile

Confidence = $\frac{\text{Suppor}(1,2,3)}{\text{Support}(1)} = \frac{2/9}{6/9} = 33\%$

1 ⌒ 2, 3

~~33 < 50~~

Rule 2 {2} → {1, 3}

$\text{con} = \frac{\text{Suppor}\{1,2,3\}}{\text{Support}(2)} = \frac{2/9}{7/9} = 28\%$

~~28 < 50~~

Rule 3 → {3} → {1,2} → = 33.3

Rule 4 →

$$\{1,2\} \rightarrow \{3\} = \frac{\text{support}(1,2,3)}{\text{support}(1,2)}$$

$$= \frac{2/9}{4/9} = 50\%$$

50% ≥ 50%

Valid Rule

{1,2} → {3}

* PCA
* Demo Apriori → TG
* LDA