V-S

SML

classmate
Date
Page

1 K-means clustering

- Clustering is the process of grouping observations in such a way that members of same cluster are grouped together and members of different cluster are grouped together.
- Clustering is commonly used to explore the datasets & identify the patterns
- K-means clustering algorithm is an iterative process of moving the centers to means positions until there is no significant changes.
- The cost function of k-means is determined by Euclidean distance between the observation

- If there is one cluster ($K=1$), then the distance b/w all obs. are compared with its single mean

- If the clusterer is increased to 2 ($K=2$), then, two means are calculated and few of them are assigned to cluster 1 and remaining to cluster 2.

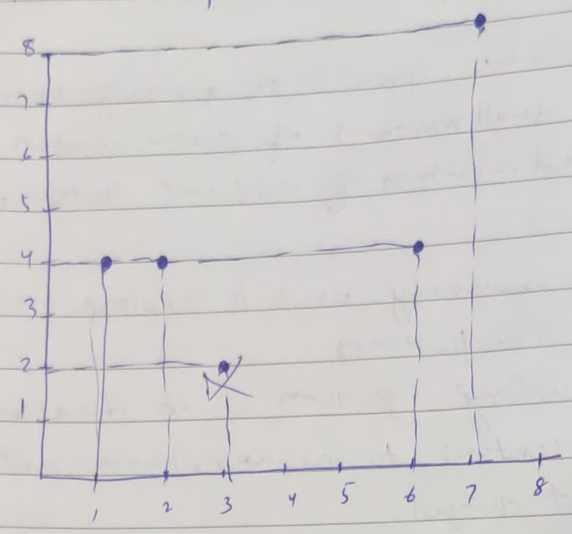
$$J = \sum_{K=1}^{K} \sum_{i \in C_K} ||x_i - \mu_K||^2$$

Example

Lets take 4 instances, with their x & y values

Instances	x	y
1	7	8
2	2	4
3	6	4
4	1	4

→ Plot the data points on 2D chart



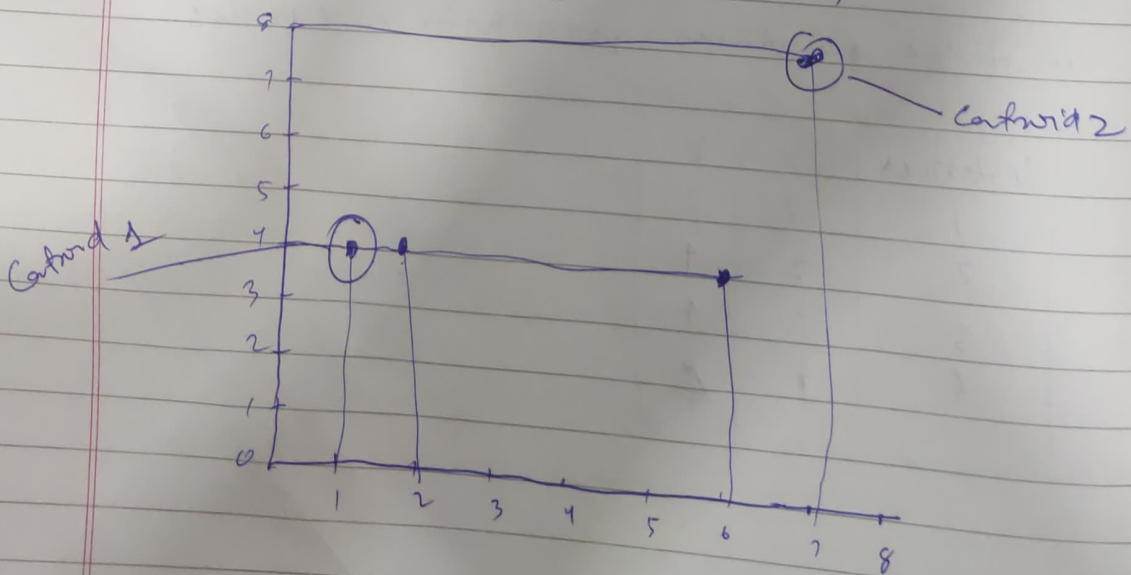
→ Iteration 1: Let's choose instance 1 (X=7, Y=8) → centroid 2
instance 4 (X=1, Y=4) → centroid 1

Calculate euclidean distance

$4+6$
$16+36$ $\quad 4+5$

Instance	X	Y	Centroid 1 dist	Centroid 2 dist	Assign cluster
1	7	8	7.21	0	C2
2	2	4	1	6.40	C1
3	6	4	5.00	4.12	C2
4	1	4	0	7.21	C1
Centroid 1	1	4			
Centroid 2	7	8			

euclidean distance $= \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$



→ Iteration 2: Calculate the new centroids

Instance	X	Y	Assigned cluster
1	7	8	C2
2	2	4	C1
3	6	4	C2
4	1	4	C1
Centroid 1	1.5	4.0	
Centroid 2	6.5	6	

Centroid 1 coordinates = Avg coordinates (instances 2, 4)

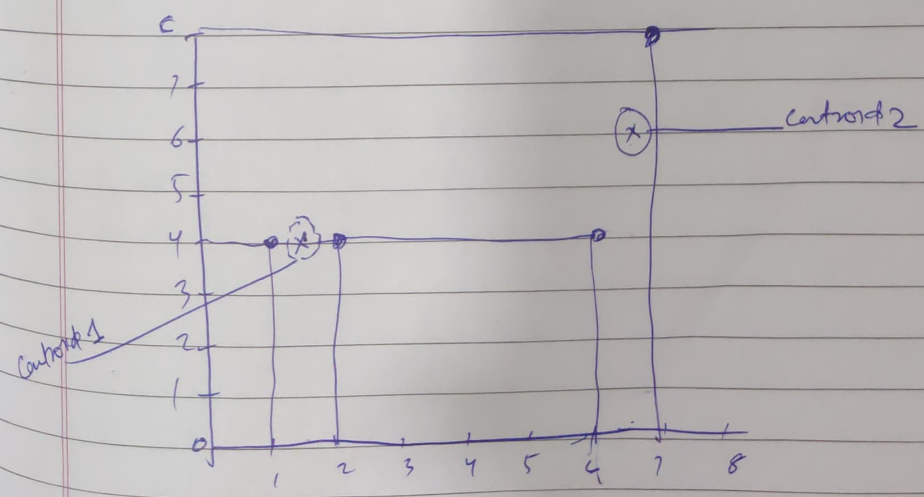
Centroid 2 coordinates = Avg coordinates (instances 1, 3)

$$\text{Centroid 1 } x = \frac{2+1}{2} = \frac{3}{2} = 1.5$$

$$\text{centroid 1 } y = \frac{4+4}{2} = \frac{8}{2} = 4.0$$

$$\text{centroid 2 } x = \frac{7+6}{2} = \frac{13}{2} = 6.5$$

$$\text{centroid 2 } y = \frac{8+4}{2} = \frac{12}{2} = 6$$



30.25 ✓

Iteration 3: Check for the new clusters is new or same and calc. the centroids distance

C1(1.5, 4.0) C2(6.5, 6)

Instance	X	Y	C1 dist	C2 dist	old clusters	New Clusters	Changed?
1	7	8	6.80	2.06	C2	C2	No
2	2	4	0.5	4.92	C1	C1	No
3	6	4	4.5	1.5	C2	C2	No
4	1	4	0.5	5.85	C1	C1	No

Since there is no change
Hence soln is converged.

(2) PCA (Principal component Analysis)

↳ It is dimensionality reduction method
↳ used to reduce the dimensions of large dataset by transforming the large dataset to smaller dataset
↳ While reductions it may lead in reducing accuracy but using PCA it will not effect
↳ It maximizes the variance of the data
↳ It minimizes the mean squared distance

Need:

→ Removes inconsistent data
→ Removes redundant data
→ Makes data processing faster
→ No loss of information

PLA Steps:

① Standardization of the data
② Computing covariance matrix
③ calculating eigenvectors and eigenvalues
④ Computing principal components
⑤ Reducing the dimensions of data set

Step 1: Standardization of the data

↳ It is about scaling the data
↳ Get some data & plot it.
↳ Take the mean(x) & Mean(Y)

Mean(x) = 3
Mean(Y) = 6.5

X	Y
1.5	2.4
2.5	3.6
3.5	4.2
4.5	2.8

Step2: computing covariance matrix

$$\begin{bmatrix} cov(a,a) & cov(a,b) \\ cov(b,a) & cov(b,b) \end{bmatrix}$$

→ The covarince values depnds:
(1) -ve covariance - indirectly proportional
(2) +ve covarince - directly proportional

→ It defines the correlation b/w the different variable in dataset.
→ This is called DATA ADJUSTMENT

Step3: calculating eigenvectors and eigenvalues

eigenvectors: These are the vectors when performed on them their direction does not changes
eigenvalues: denotes the scaler of eigenvectors.

They are the mathmatical concepts that can be calculated using covariant matrix.

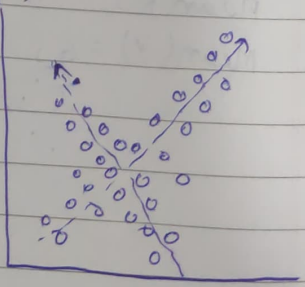
calculate covarint matrix

$$cov(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

	x	y
x	cov(x,x)	cov(x,y)
y	cov(y,x)	cov(y,y)

Step4: Computing principal compnts.

→ PC1 is most significant and stores max possible info's
→ PC2 is second most significant and stores the remaining max info.



→ After calc. eigen vectors and values, we need to arrange them in descending order

Steps: Reorg the dimns:

- Rearrange the orgnl data
- The orgnl data will be reduced
- The eigen vectors with highest eigen value will be the PCA

PCA from first principles

(1) Take instances with (X and Y)

Instnce	X	Y
1	0.1	0.6
2	0.2	0.7
3	0.3	0.8
4	0.4	0.9
5	0.5	0.1
Mean	0.3	0.62

3.1
5x10
6.2

(2) Remove the scale factors.

X	Y
0.1-0.3 = -0.2	0.6-0.62 = -0.02
0.2-0.3 = -0.1	0.7-0.62 = 0.08
0.3-0.3 = 0	0.8-0.62 = 0.18
0.4-0.3 = 0.1	0.9-0.62 = 0.28
0.5-0.3 = 0.2	0.1-0.62 = -0.52

③ $Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{n-1}$

$cov(x,x) = [-0.2)^2 + (-0.1)^2 + (0)^2 + (0.1)^2 + (0.2)^2 / 5-1 =$

$cov(x,Y) = (-0.2 * -0.02)^2 + (-0.1 * 0.05)^2 + \cdots\cdots / 5-1 =$

$cov(Y,x) = (\quad) + (\quad) + \cdots\cdots / 5-1 =$

$cov(Y,Y) = (\quad) + (\quad) + (\quad) \cdots / 5-1 =$

covarue Matrix, $C = \begin{bmatrix} cov(x,x) & cov(x,Y) \\ cov(Y,x) & cov(Y,Y) \end{bmatrix}$

Calculating the eigen vectors and vaues

Eigen vaues = [ ]

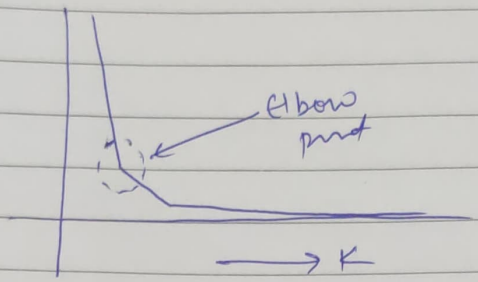
Eigen vectors = [ ]

The eigen vectors with hight eigen value will be the PCA.

4m

① The elbow method

- It is the method used to find the optimal number of clusters
- It plots the values of cost function
- For value K, if K ↑, the avg. distortion will ↓
- The value of K where it is improved is called Elbow
- At this point we should stop dividing the data.



② Evaluation of clusters.

- It can be done with ~~silhouette~~ silhouette coefficient.
- It is the measure of compactness and seperation of clusters
- Higher value sep. better quality of clustn
- values lies from -1 to +1
- Higher value will be better

$$s = \frac{ba}{\max(a,b)}$$

a is mean dist in clustn
b is mean dist of next clustn

③ PCA

Pros: : Dimentionality Reduction
: Data visualization
: Improve data quality
: easy to implmt

Cons: : Less accuracy
: difficult to get result
: slow