Supervised( outcome)

1. Regression x independent Y depe ( numbers )
2. Classification x inde Y depe ( classes)

UnSupervised( outcome)

1. x independent , no y variable
2. Divide them different groups /clusters
   1. When = Unknown patterns discovery
      1. Eg: C1, c2, c3 ( three clusters )

What:

1. If Only x values , we need to calculate the distance between datapoints to create clusters
2. If Only x values , divide them into groups
   1. If group number / no of clusters is already defined = K Means clustering
   2. If group number / no of clusters is not defined = Hierarchial clustering

|  |  |
| --- | --- |
| Sno | Marks |
| 1 | 45 |
| 2 | 56 |
| 3 | 56 |



How

K Means: K = 2 based on the distance



Steps to follow for K Means

1. Decide on Number of clusters , K = 2
2. Randomly assign each data point to the cluster
3. Compute the cluster centroid ( location center of my cluster )
4. Re assign each point to the closest cluster centroid
5. Re compute the cluster centroid
6. repeat the 4 and 5 steps ( iterative process )



usecases

1. Customer segmentation
2. Customer behavior Netflix

Why : If group number / no of clusters is not defined = Hierarchial clustering

What :

1. initially each datapoint as one new cluster, from many clusters to one or two clusters
2. initially one cluster and divide them slowly into many clusters ( one to many )

How :

1. Divisive HC ( one to many ) split cluster at each step
2. Agglomerative HC ( many to one /2 ) from individual datapoints to one ( merge)

Example AHC ( many to one ) ( each data point is cluster )

|  |  |
| --- | --- |
| Sno | Attendance ( in hrs ) |
| 1 | 10 |
| 2 | 7 |
| 3 | 28 |
| 4 | 20 |
| 5 | 35 |

Proximity matrix ( 5data points \*5) PM1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SNO | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 3(10-7) | 18(10-28) | 10(10-20) | 25(10-35) |
| 2 | 3(10-7) | 0 | 21(7-28) | 13(7-20) | 28(7-35) |
| 3 | 18(10-28) | 21(7-28) | 0 | 8(28-20) | 7(35-28) |
| 4 | 10(10-20) | 13(7-20) | 8(28-20) | 0 | 15(35-20) |
| 5 | 25(10-35) | 28(7-35) | 7(35-28) | 15(35-20) | 0 |

Dsitance calculation Euclidean formula : ( 10-7)2  = 3 \* 3 = 9 = srt of 9 = 3

(35-28) =7 \* 7 = 49 = sqrt 49 = 7

Formation of 1-2 cluster

|  |  |
| --- | --- |
| Sno | Attendance ( in hrs ) |
| 1&2 | 10( max of 7 and 10) |
| 3 | 28 |
| 4 | 20 |
| 5 | 35 |

Grouping both 1 and 2 into one cluster , then we will have only4 clusters

PM#2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNO | 1&2 | 3 | 4 | 5 |
| 1&2 | 0 | 18(10-28) | 10(10-20) | 25(10-35) |
| 3 | 18(10-28) | 0 | 8(28-20) | 7(35-28) |
| 4 | 10(10-20) | 8(28-20) | 0 | 15(35-20) |
| 5 | 25(10-35) | 7(35-28) | 15(35-20) | 0 |



Formation of 3 &5 cluster

|  |  |
| --- | --- |
| Sno | Attendance ( in hrs ) |
| 1&2 | 10( max of 7 and 10) |
| 3&5 | 35(max of 28, 35) |
| 4 | 20 |

Grouping both 3and 5 into one cluster , then we will have only3 clusters

PM#3

|  |  |  |  |
| --- | --- | --- | --- |
| SNO | 1&2 | 3&5 | 4 |
| 1&2 | 0 | 25(10-35) | 10(10-20) |
| 3&5 | 25(10-35) | 0 | 15(35-20) |
| 4 | 10(10-20) | 15(35-20) | 0 |

|  |  |
| --- | --- |
| Sno | Attendance ( in hrs ) |
| 1&2 and 4 | 20(Max of 10 and 20 ) |
| 3&5 | 35(max of 28, 35) |



PM4

|  |  |  |
| --- | --- | --- |
| SNO | 1&2 &4 | 3&5 |
| 1&2 &4 | 0 | 15(20-35) |
| 3&5 | 15(20-35) | 0 |

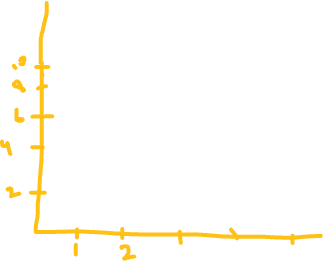
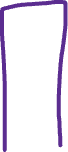
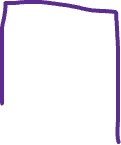
Steps to follow for AHC

1. Create proximity matrix which stores distance info between datapoints
2. Each data point is a cluster ( 5 clusters as we hve 5 datapoints )
3. Find the closet datapoint and merge to form one cluster
4. Repeat steps and 1,2,3 ( iterative process)



After repeating all above steps, we have 3 clusters

To decide on the number of clusters we use dendogram



1. Reg
   1. SSE
   2. RMSE
   3. MAE
   4. COD/ R2 = 0-1= NEAREST TO 1 => is the better
2. Class
   1. pRecision
   2. recall
   3. AUC



* 1. Accuracy



* 1. Confusion matrix



1. Clustering



* 1. SC( Sillohette quiotient) -> whether clustering technique is good or not
     1. Values = -1 to 1
        1. Value = 1 , cluster visibility is very clear ( highly dense clustering )
        2. 0 = overlapped, clusters are indifferent , distance between cluster to cluster is very small/ not that significant



* + - 1. -1= completely wrong ,assigning itself wrong
         1. Incorrect clustering
    1. Better to have 0-1, more the nearer to 1 again is the best HC
    2. Sc score = (b-a) / max ( a,b)
       1. A = in the Same cluster , mean distance between sample and all other datapoints
       2. b = in the nearest cluster , mean distance between sample and all other datapoints

if its AHC, no of clusters = 2 , SC score = 0.75

if its AHC, no of clusters = 3, SC score = 0.65

if its AHC, no of clusters = 4 , SC score = 0.45

if its AHC, no of clusters = 5 , SC score = 0.4

if its AHC, no of clusters = 6 , SC score = 0.35

if its AHC, no of clusters = 2 , SC score = 0.65

if its AHC, no of clusters = 3, SC score = 0.75

if its AHC, no of clusters = 4 , SC score = 0.35

if its AHC, no of clusters = 5 , SC score = 0.4

if its AHC, no of clusters = 6 , SC score = 0.55

Hamming distance out 6 characters in both 2 strings , only one common char ( A ) in thr 5th place

Kishan raajah



Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated



Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated



Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated



Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated