Clustering & Application trends in Data Mining.

Page No.:
Date:

- Distance - used for determining clusters

Classification
↓
Decision tree
↙ ↘
Entropy Information gain

Association Rule
↓
Apriori Algorithm
↙ ↓
Support confidence

Clustering
1) Distance

Lift = XUY / X * Y

Page No.: Date: YOUVA

4,4 2,2

~~slope~~ $y = mx+c$

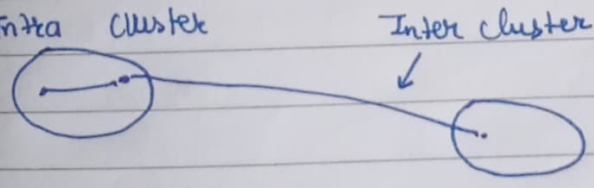
$d = \frac{(y_2-y_1)^2}{(x_2-x_1)^2}$

(centroid point (x) = $\frac{(x_1+x_2 \ldots x_n)}{n}$

(centroid point (y) = $\frac{(y_1, y_2, \ldots y_n)}{n}$

- Clustering is unsupervised learning as we don't know class labels here in advance.

Q: What are different methods of clustering (2M) R. Level [6 types]



- Ideal cluster: same elements
  - distance within cluster must be minimum
  - distance within different cluster elements must be maximum

Extrinsic measures, Intrinsic measures.

** Agglomerative / divisive clustering (theory) 4M

*** DBSCAN ~~(theory)~~ (theory 4M)

** Partioning

* Hierarchical.

** K-means algorithm problems / how it works / steps.

**

Page No.:
Date:
YOUVA

Example:

Data:

$x_1 = (0,2)$

$x_2 = (0,0)$

$x_3 = (1.5,0)$

$x_4 = (5,0)$

$x_5 = (5,2)$

Random distribution of samples:

$C_1 = \{x_1, x_2, x_4\}$ and $C_2 = \{x_3, x_5\}$

Centroids

$M_1 = \{(0+0+5)/3, (2+0+0)/3\} = \{1.66, 0.66\}$

$M_2 =$

$$M_k = \frac{1}{u_k} \sum_{i=1}^{u_k} x_{ik}$$

$$e_k^2 = \sum_{i=1}^{u_k} (x_{ik} - M_k)^2$$

$$E_k^2 = \sum_{k=1}^{K} e_k^2$$

→ Centroid: $x_1\ x_2\ x_4$ $y_1\ y_2\ y_4$

$M_1 = \{(0+0+5)/3, (2+0+0)/3\}$

$= \{1.66, 0.66\}$

$M_2 = \{(1.5+5)/2, (0+2)/2\}$ ($x_3\ x_5$; $y_3\ y_5$)

$= \{3.25, 1\}$

$M_{k(1)} = (1.67, 0.67)$ // Centroid of cluster 1.

$M_{k(2)} = (3.25, 1)$

Within-cluster variations :

$e_1^2 = [(0-1.66)^2 + (2-0.66)^2] + [(0-1.66)^2 + (0-0.66)^2] + [(5-1.66)^2 + (0-0.66)^2] = 19.33$

$e_2^2 = [(1.5-3.25)^2 + (0-1)^2] + [(5-3.25)^2 + (2-1)^2] = 8.125$

Total square error $= 19.33 + 8.13 = 27.46$

Reassign all samples :

$d(M1, x1) = \sqrt{(0-1.66)^2 + (2-0.66)^2} = \sqrt{4.55} = 2.13$

$d(M_2, x_1) = \sqrt{(0-3.25)^2 + (2-1)^2} = \sqrt{(-3.25)^2 + (1)^2} = \sqrt{11.56} = 3.40.$

Similarly :

$d(M_1, X_2) = 1.79$ & $d(M_2, X_2) = 3.40$

$d(M_1, X_3) = 0.83$ & $d(M_2, X_3) = 2.01$

$d(M_1, X_4) = 3.41$ & $d(M_2, X_4) = 2.01$

$d(M_1, X_4) = 3.60$ & $d(M_2, X_5) = 2.01$

New clusters =

$C_1 = \{x_1, x_2, x_3\}$

$C_2 = \{x_4, x_5\}$

New Centroids:

$M_1 = \{(0+0+1.5)/3, (0+0+2)/3\}$

$M_1 = (0.5, 0.67)$

$M_2 = \{(5+5)/2, (0+2)/2\}$

$M_2 = \{5, 1\}$

Errors:

$e_1^2 = 4.17$

$e_2^2 = 2.00$

C1 C2

D = { (5,3), (10,15), (15,12), (24,10), (30,45), (85,70), (71,80), (60,78), (55,52), (80,91) }

Iteration 1

Sno	Data Points	Euclidean distance from cluster centroid C1 = (5,3)	Euclidean distance from cluster centroid C2 = (10,15)	Assigned cluster
1	(5,3)	0	13	C1
2	(10,15)	13	0	C2
3	(15,12)	13.45	5.83	C2
4	(24,10)	20.24	14.86	C2
5	(30,45)	48.88	36.05	C2
6	(85,70)	104.35	93	C2
7	(71,80)	101.41	89.14	C2
8	(60,78)	93	80.43	C2
9	(55,52)	70.01	58.25	C2
10	(80,91)	115.62	103.32	C2

$C_2(x) = (10+15+24+30+85+71+60+55+80)/9$

$= 47.78$

$C_2(y) = (15+12+10+45+70+80+78+52+91)/9$

$= 50.33$

// Table again but updated C2 value.