Clustering

K-Means Clustering (Partitioning algorithms)

Steps:

Initialization: Select k initial cluster centroids (randomly).

Assignment: Assign each point to nearest cluster - distance (e.g., Euclidean distance).

Update: Recalculate the centroid of each cluster by mean points.

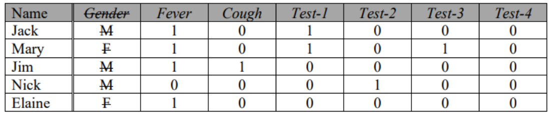
Repeat: Continue assignment and update steps until no significant change.

Jaccard Coefficient

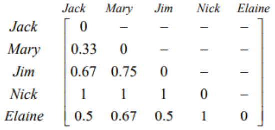
Equation: $JC = \frac{P(01)+P(10)}{P(11)+P(01)+P(10)}$

Simple Matching Coefficient

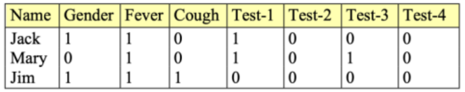
Equation: SMC= $\frac{P(01)+P(10)}{P(11)+P(01)+P(10)+P(00)}$



Name	~~Gender~~	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	~~M~~	1	0	1	0	0	0
Mary	~~F~~	1	0	1	0	1	0
Jim	~~M~~	1	1	0	0	0	0
Nick	~~M~~	0	0	0	1	0	0
Elaine	~~F~~	1	0	0	0	0	0



$$\begin{array}{c|ccccc} & Jack & Mary & Jim & Nick & Elaine \\ \hline Jack & 0 & - & - & - & - \\ Mary & 0.33 & 0 & - & - & - \\ Jim & 0.67 & 0.75 & 0 & - & - \\ Nick & 1 & 1 & 1 & 0 & - \\ Elaine & 0.5 & 0.67 & 0.5 & 1 & 0 \end{array}$$



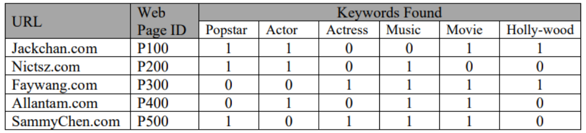
Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	1	0	1	0	0	0
Mary	0	1	0	1	0	1	0
Jim	1	1	1	0	0	0	0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

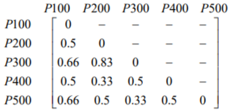
$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Only asymmetric variables are considered!!!



URL	Web Page ID	Keywords Found					
		Popstar	Actor	Actress	Music	Movie	Holly-wood
Jackchan.com	P100	1	1	0	0	1	1
Nictsz.com	P200	1	1	0	1	0	0
Faywang.com	P300	0	0	1	1	1	1
Allantam.com	P400	0	1	0	1	1	0
SammyChen.com	P500	1	0	1	1	1	0



$$\begin{array}{c|ccccc} & P100 & P200 & P300 & P400 & P500 \\ \hline P100 & 0 & - & - & - & - \\ P200 & 0.5 & 0 & - & - & - \\ P300 & 0.66 & 0.83 & 0 & - & - \\ P400 & 0.5 & 0.33 & 0.5 & 0 & - \\ P500 & 0.66 & 0.5 & 0.33 & 0.5 & 0 \end{array}$$

Categorical:

$d(i,j) = \frac{p-m}{p}$ /One-hot encoding

Transactional

Basic ideas:

- Let $T_1 = \{A,B,C\}$, $T_2=\{C,D,E\}$ where A-E denote items
- Similarity function defined as:

$$Sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

where ∩ & ∪ denote the intersection and union of two transaction records respectively.

- For our example, we have

$$Sim(T_1, T_2) = \frac{|\{C\}|}{|\{A,B,C,D,E\}|} = \frac{1}{5}$$

dissim(T1, T2)
= 1- Sim(T1, T2) = 1-1/5 = 4/5

Variants: K-modes (replace with modes)

K-Medoids (representative objects)

CLARA (large dataset)

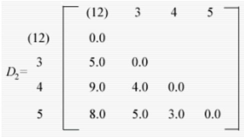
Hierarchical Clustering Methods

Single-link agglomerative (ANGES) - fusion

Divisive (DIANA) – division

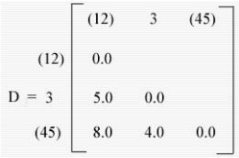
results – dendrogram

- $d(12)3 = \min[d13, d23] = d23 = 5.0$
- $d(12)4 = \min[d14, d24] = d24 = 9.0$
- $d(12)5 = \min[d15, d25] = d25 = 8.0$

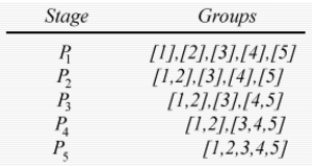


$$D_2 = \begin{array}{c|cccc} & (12) & 3 & 4 & 5 \\ \hline (12) & 0.0 & & & \\ 3 & 5.0 & 0.0 & & \\ 4 & 9.0 & 4.0 & 0.0 & \\ 5 & 8.0 & 5.0 & 3.0 & 0.0 \end{array}$$

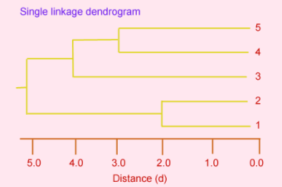
- $d(12)3 = 5.0$ as before
- $d(12)(45) = \min[d14, d15, d24, d25] = 8.0$
- $d(45)3 = \min[d34, d35] = d34 = 4.0$



$$D = \begin{array}{c|ccc} & (12) & 3 & (45) \\ \hline (12) & 0.0 & & \\ 3 & 5.0 & 0.0 & \\ (45) & 8.0 & 4.0 & 0.0 \end{array}$$

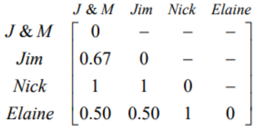


Stage	Groups
P_1	[1],[2],[3],[4],[5]
P_2	[1,2],[3],[4],[5]
P_3	[1,2],[3],[4,5]
P_4	[1,2],[3,4,5]
P_5	[1,2,3,4,5]



b) Cluster the data records using the single-link agglomerative clustering algorithm and the Jaccard coefficient matrix computed in part (a). Make your own assumption(s) if necessary.

Merging Jack and Mary (d=0.33), we have



$$\begin{array}{c|cccc} & J\,\&\,M & Jim & Nick & Elaine \\ \hline J\,\&\,M & 0 & - & - & - \\ Jim & 0.67 & 0 & - & - \\ Nick & 1 & 1 & 0 & - \\ Elaine & 0.50 & 0.50 & 1 & 0 \end{array}$$

If merging of more than 2 records is allowed, J&M, Jim and Elaine should be merged next. Thus, the last record being grouped is Nick.

b) Based on the coefficient matrix completed in part (a), cluster the data records using the single-link agglomerative hierarchical clustering algorithm.

Answer:

1st round: Merging P200 & P400 (distance=0.33)

2nd round: Merging P300 & P500 (distance=0.33)

3rd round: Merging C1(P200,P400) to C2(P300, P500) (distance=0.5) or
Merging C1(P200,P400) to P100 (distance=0.5)

4th round: Merging the remaining two clusters

Detail steps are omitted here.

Density Based Clustering

Identify clusters of arbitrary shapes.

Define clusters as dense region separated by low-density area

ε-Neighborhood:

- All points within the ε radius of a point p.

- A point's density is "high" if ε-neighborhood contains at least MinPts.

Point Types: Core Points, Border Points, Outliers

Density Reachability:

- Directly Density-Reachable: point q is within the ε-neighborhood of a core point p.

- Indirectly Density-Reachable: point q reachable through chain of (d) reachable points.

- Density Connectivity: 2 points are (d) connected if commonly (d) reachable from third point.

DBSCAN

2 Parameters

ε (Epsilon): Radius defining the neighborhood of a point.

MinPts: Minimum points required within ε-neighborhood to qualify as dense.

Steps:

1. Arbitrarily pick a point p.
2. Retrieve all points density-reachable from p (within ε and MinPts).
3. Mark p as a core point if it meets density criteria.
4. If p is a border point, it does not expand a cluster.
5. Mark isolated points as noise.

Output: Clusters formed by density-connected points.