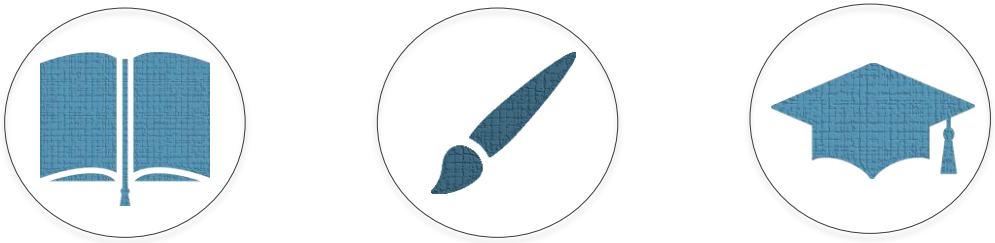


Data Science Concepts

Lesson05–Clustering and Segmentation

Objective

After completing this lesson you will be able to:

- Explain Clustering and its applications
- Describe hierarchical clustering and K means clustering.

Cluster Analysis

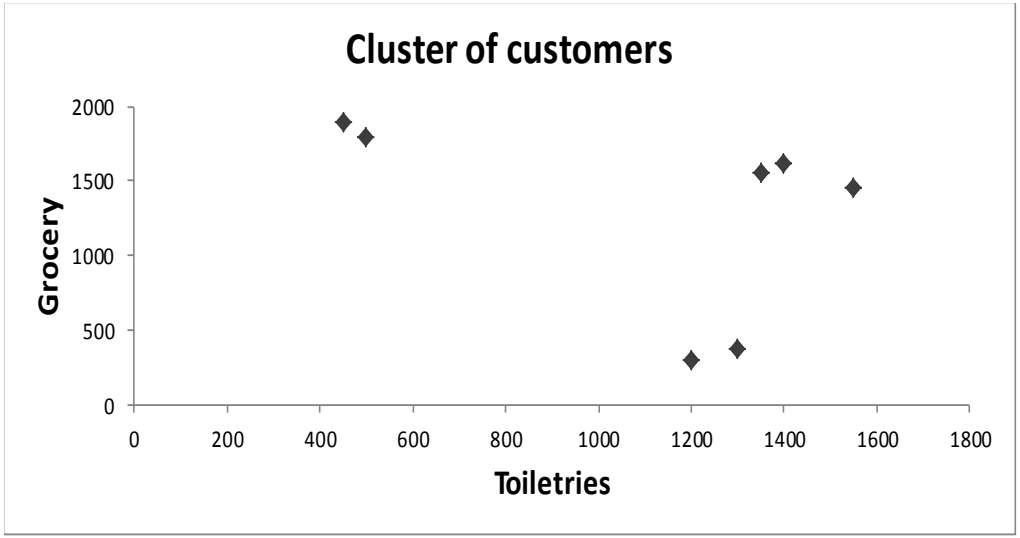
- Used in marketing for creating product segmentation and customer segmentation.
  - Is helpful to understand the product spread and understand which products are cannibalizing. Either internal or of the competitors.
  - Helps in creating customer profiles for targeted marketing.
  - The marketing expense can be optimized and utilized effectively.

- Clustering:
  - Putting similar things into one single group.
  - Clustering is performed by looking into different characteristics which may be helpful in bringing out a pattern.

Types of clustering being discussed:
- Hierarchical clustering
- K means clustering

Hierarchical Clustering–Concept Development

Customer	Groceries	Toiletries
1	1200	300
2	1300	380
3	500	1800
4	450	1900
5	1350	1560
6	1400	1620
7	1550	1450



Three distinctively different categories:

- Low on toiletries and high on grocery
- High on toiletries and low on grocery
- High on toiletries and high on grocery

How do you do this for 10000 customers and 20 products?

Hierarchical Clustering–Concept Development

- Closer the points were, more similarity within the customers. Father the points, more dissimilarity within the customers.
  - Distance between the points is a measure of similarity or dissimilarity

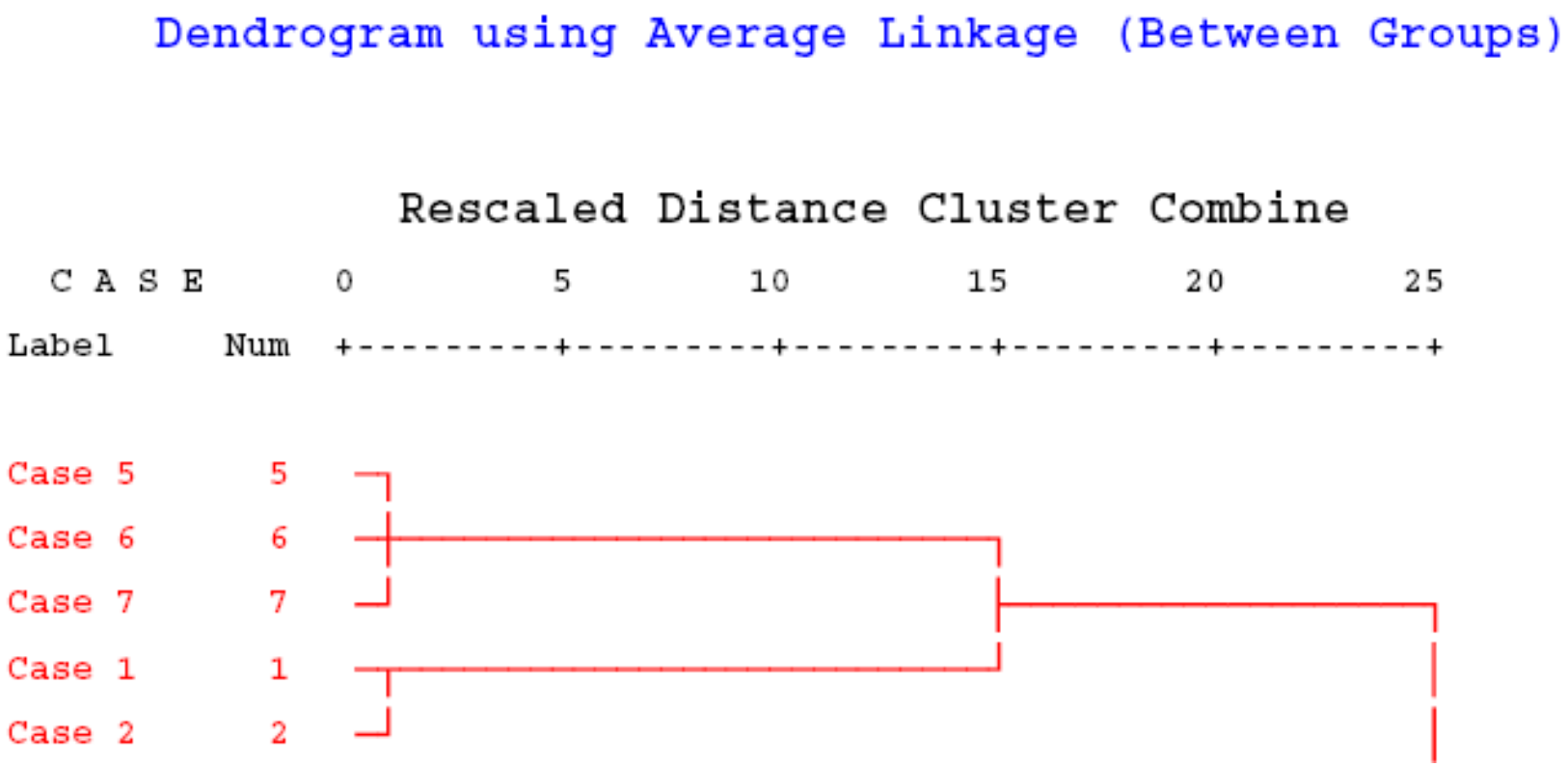
- Calculate the linear distance between all the customers:
  - Distance between Customer 1 to Customer 2, 3, … , 7.
  - Distance between Customer 2 to Customer 1, 3, … , 7.
  - Distance between Customer 3 to Customer 1, 2, … , 7.
  - .
  - .
  - Distance between Customer 7 to Customer 1, 2, … , 6.

- 7*7 matrix is formed. Pick the smallest number. This forms the first cluster.

- Now one cluster and 5 customers. Total six entities for which above steps are repeated. Cluster formation happens in hierarchy and thus the name

Hierarchical Clustering–Concept Development

- When to stop the clustering:
  - The variation within the cluster is low and variation across cluster is very high.
  - Dendrogram gives this output in graphical form.
  - Farther distance travelled on dendrogram, more dissimilar entities are being clustered.



Hierarchical Clustering–Beer Data Case

- 20 brands of beer with Calorie content, sodium content, Alcohol content and Cost.

Perform Hierarchical Clustering

ID	BEER	CAL	SOD	ALC	COST
1	Budweiser	144	15	4.7	0.43
2	Schlitz	151	19	4.9	0.43
3	Lowenbrau	157	15	4.9	0.48
4	Kronenbourg	170	7	5.2	0.73
5	Heineken	152	11	5	0.77
6	Old Mil	145	23	4.6	0.28
7	Augsburger	175	24	5.5	0.4
8	Strohs	149	27	4.7	0.42
9	Miller lite	99	10	4.3	0.43
10	Bud light	113	8	3.7	0.44
11	Coors	140	18	4.6	0.44
12	Coors lite	102	15	4.1	0.45
13	Michelob light	135	11	4.2	0.5
14	Becks	150	19	4.7	0.76
15	Kirin	149	6	5	0.79
16	Pabst	68	15	2.3	0.38
17	Hamms	136	19	4.4	0.43
18	Heilemans	144	24	4.9	0.43
19	Olympia	72	6	2.9	0.46
20	Schilitz lite	97	7	4.2	0.47

Hierarchical Clustering–Beer Data Case

- Linear distance from Budweiser
- These distances are to be calculated for each beer brand

ID	BEER	CAL	SOD	ALC	COST	Total
1	Budweiser	0	0	0	0	0
2	Schlitz	49	16	0.04	0	65.04
3	Lowenbrau	169	0	0.04	0.0025	169.0425
4	Kronenbourg	676	64	0.25	0.09	740.34
5	Heineken	64	16	0.09	0.1156	80.2056
6	Old Mil	1	64	0.01	0.0225	65.0325
7	Augsburger	961	81	0.64	0.0009	1042.641
8	Strohs	25	144	0	0.0001	169.0001
9	Miller lite	2025	25	0.16	0	2050.16
10	Bud light	961	49	1	0.0001	1011
11	Coors	16	9	0.01	0.0001	25.0101
12	Coors lite	1764	0	0.36	0.0004	1764.36
13	Michelob light	81	16	0.25	0.0049	97.2549
14	Becks	36	16	0	0.1089	52.1089
15	Kirin	25	81	0.09	0.1296	106.2196
16	Pabst	5776	0	5.76	0.0025	5781.763
17	Hamms	64	16	0.09	0	80.09
18	Heilemans	0	81	0.04	0	81.04
19	Olympia	5184	81	3.24	0.0009	5268.241
20	Schilitz lite	2209	64	0.25	0.0016	2273.252

Hierarchical Clustering–Beer Data Case

- Square of distance from budewiser.

ID	BEER	CAL	SOD	ALC	COST	Total
1	Budweiser	0	0	0	0	0
2	Schlitz	49	16	0.04	0	65.04
3	Lowenbrau	169	0	0.04	0.0025	169.0425
4	Kronenbourg	676	64	0.25	0.09	740.34
5	Heineken	64	16	0.09	0.1156	80.2056
6	Old Mil	1	64	0.01	0.0225	65.0325
7	Augsburger	961	81	0.64	0.0009	1042.641
8	Strohs	25	144	0	0.0001	169.0001
9	Miller lite	2025	25	0.16	0	2050.16
10	Bud light	961	49	1	0.0001	1011
11	Coors	16	9	0.01	0.0001	25.0101
12	Coors lite	1764	0	0.36	0.0004	1764.36
13	Michelob light	81	16	0.25	0.0049	97.2549
14	Becks	36	16	0	0.1089	52.1089
15	Kirin	25	81	0.09	0.1296	106.2196
16	Pabst	5776	0	5.76	0.0025	5781.763
17	Hamms	64	16	0.09	0	80.09
18	Heilemans	0	81	0.04	0	81.04
19	Olympia	5184	81	3.24	0.0009	5268.241
20	Schilitz lite	2209	64	0.25	0.0016	2273.252

The total column is called the Euclidian distance

$$D(X_1, X_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \ldots + (x_{1n} - x_{2n})^2}$$

Hierarchical Clustering–Beer Data Case

Euclidian distance matrix for all the brands

BEER	Budweise	Schlitz	Lowenbra	Kronenbo	Heineken	Old Mil	Augsburge	Strohs	Miller lite	Bud light	Coors
Budweise	0	65.04	169.04	740.34	80.21	65.03	1042.64	169	2050.16	1011	25.01
Schlitz	65.04	0	52	505.18	65.13	52.11	601.36	68.04	2785.36	1566.44	122.09
Lowenbra	169.04	52	0	233.15	41.09	208.13	405.37	208.04	3389.36	1986.44	298.09
Kronenbo	740.34	505.18	233.15	0	340.04	881.56	314.2	841.35	5050.9	3252.33	1021.44
Heineken	80.21	65.13	41.09	340.04	0	193.4	698.39	265.21	2810.61	1531.8	193.27
Old Mil	65.03	52.11	208.13	881.56	193.4	0	901.82	32.03	2285.11	1249.84	50.03
Augsburge	1042.64	601.36	405.37	314.2	698.39	901.82	0	685.64	5973.44	4103.24	1261.81
Strohs	169	68.04	208.04	841.35	265.21	32.03	685.64	0	2789.16	1658	162.01
Miller lite	2050.16	2785.36	3389.36	5050.9	2810.61	2285.11	5973.44	2789.16	0	200.36	1745.09
Bud light	1011	1566.44	1986.44	3252.33	1531.8	1249.84	4103.24	1658	200.36	0	829.81
Coors	25.01	122.09	298.09	1021.44	193.27	50.03	1261.81	162.01	1745.09	829.81	0
Coors lite	1764.36	2417.64	3025.64	4689.29	2516.91	1913.28	5411.96	2353.36	34.04	170.16	1453.25
Michelob	97.25	320.49	500.49	1242.05	289.71	244.21	1770.7	452.26	1297.01	493.25	74.16
Becks	52.11	1.15	65.12	544.25	68.09	41.24	650.77	65.12	2682.27	1491.1	101.11
Kirin	106.22	173.14	145.11	442.04	34	305.42	1000.4	441.23	2516.62	1301.81	225.28
Pabst	5781.76	6911.76	7927.77	10476.53	7079.44	5998.3	11540.24	6710.76	990	2075.96	5198.29
Hamms	80.09	225.25	457.25	1300.73	320.48	97.06	1547.21	233.09	1450.01	650.49	17.04
Heileman	81.04	74	250	965.18	233.13	2.11	961.36	34.04	2221.36	1218.44	52.09
Olympia	5268.24	6414	7310	9610.36	6429.51	5620.92	10939.76	6373.24	746.96	1685.64	4770.89
Schilitz lit	2273.25	64	0.25	0	2337.5	2560.2	6374.69	3104.25	13.01	257.25	1970.16

But there may be a problem if clustering is done without standardizing the data. Why?

Hierarchical Clustering–Beer Data Case

Amalgamation or Linkage Rules: Once several objects have been linked together, how do we determine the distances between those new clusters?

Single linkage (nearest neighbor):

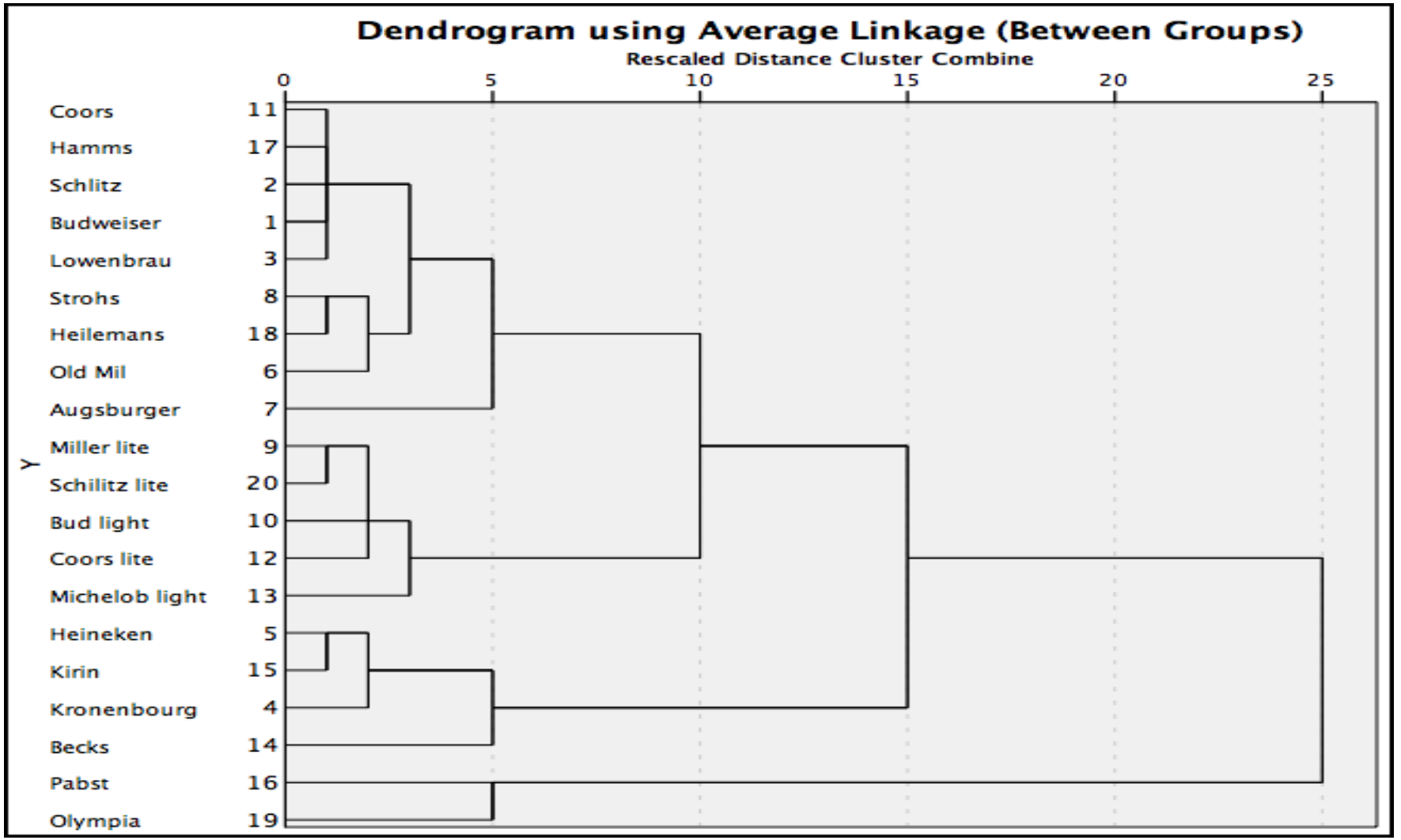
- The distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters.
- This rule will, in a sense, string objects together to form clusters, and the resulting clusters tend to represent long "chains.“

Complete linkage (furthest neighbor):

- The distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors").

Hierarchical Clustering Using Beer Data

Hierarchical Clustering–Beer Data Case

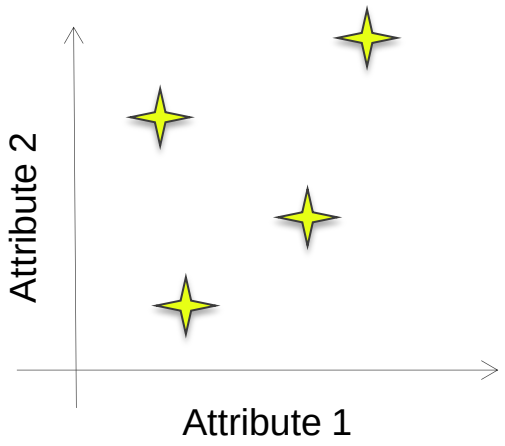


K Means Clustering–Concept Development

- Assume 70, 000 customer data having two attributes which needs to be segmented in 2 clusters.
- Here 2 depicts the K value for the cluster to be formed. Steps in K – means clustering
  1. Map 70K data points to 2 random clusters centroids. Called cluster assignment step.
  2. Once 70K numbers are mapped, calculate the new centroid. There will be two new centroids.
  3. Remove the data points and keep the new centroids.

- Repeat step 2 to 3 till the centroid movement stops.

Output is 2 clusters mapping 70, 000 customer data.



K Means Algorithm

Randomly initialize K Cluster Centroids $(\mu_1, \mu_2, \mu_3 \ldots . \mu_k)$

Training set is $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots . x^{(m)}\}$

Cluster assignment step

$Repeat \{ i = 1 \; to \; m$

$c^{(i)}$ = index from 1 to K of cluster centroid closet to $x^{(i)}$

}

Move Centroid

$Repeat \{ k = 1 \; to \; K$

μ_k = average of points assigned to cluster k

}

K Means Clustering Using Car Data

Clustering with Categorical Variables

- Hierarchical and K means clustering cannot handle categorical variables.

Why?

Partitioning around Mediods (PAM) using ‘Gowers’ as the distance measure rather than ‘euclidian’ as the distance measure.

Two step clustering technique (SPSS) can be applied to handle data with a mix of continuous and categorical variable.

Other Distance Measure

Assume that the data has n attributes, then the distance between two n-dimensional observations X_1 (x_{11}, x_{12}, …, x_{1n}) and X_2 (x_{21}, x_{22}, …, x_{2n}) can be calculated:

Manhattan distance

$$DM(X_1, X_2) = \sum_{i=1}^{n} |X_{1i} - X_{2i}|$$

- Useful while measuring distance between different locations (for example distance between two shops in a city).

Other Distance Measure

Minsowski distance

$$\mathrm{MinkowskiD}(X_1, X_2) = \left(\sum_{i=1}^{n} \left| X_{1i} - X_{2i} \right|^{p} \right)^{1/p}$$

- It is a generalized distance measure between two cases in the dataset. When p = 1, Minkowski distance is same as the Manhattan distance and for p = 2, Minkowski distance is same as the Euclidian distance.

Jaccard Similarity Coefficient (Jaccard Index)

$$J(X_1, X_2) = \frac{n(X_1 \cap X_2)}{n(X_1 \cup X_2)}$$

- Where n(X1 ∩ X2) is the number of observations in (X1 ∩ X2), n(X1 ∪ X2) is the number of observations in (X1 ∪ X2).
- Data has to be pre-processed to convert the data into binary or Boolean representation.

Other Distance Measure

Cosine Similarity

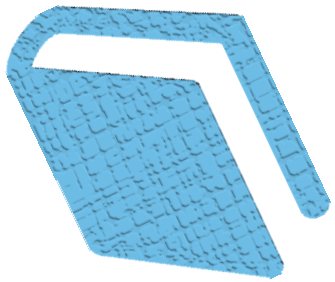
$$Cosine\_Similarity(X_1, X_2) = \frac{X_1 \bullet X_2}{\|X_1\| \bullet \|X_2\|} = \frac{\sum_{i=1}^{n} X_{1i} \times X_{2i}}{\sqrt{\sum_{i=1}^{n} X_{1i}^2} \times \sqrt{\sum_{i=1}^{n} X_{2i}^2}}$$

- It may be useful to compare customers likening for products based on the ratings given by customers on a 5-point scale.
- A lower cosine similarity indicates low similarity between two observations. Completely similar observations will have a cosine similarity value of 1.

$distance(...)function\ from\ 'philentropy'\ package\ in\ R$ implements 46 different distance/similarity measures to quantify the distance.

Summary

Summary of the topics covered in this lesson:



- Clustering is one of the most used unsupervised learning algorithm.
- Hierarchical clustering is useful when comparing various brands, products on certain parameters.
- K means clustering is useful when the number of observations runs in thousands say customer footfall into supermarket, bank etc.
- Both Hierarchical and K means clustering with Euclidian distance measure, cannot be used for grouping data with categorical variable.

QUIZ TIME

© Copyright 2015 All rights reserved.

Quiz Question 1

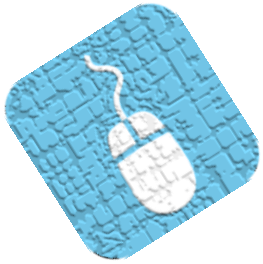
Quiz 1	What is the distance measure for measuring dissimilarity when the data has both numeric and categorical variable?

a. Gower distance.

b. Euclidian distance.

c. Both Gower and Euclidian distance can be used.

d. Distance between categorical variable cannot be measured.



Quiz Question 1

Quiz 1	What is the distance measure for measuring dissimilarity when the data has both numeric and categorical variable?

a. Gower distance.

b. Euclidian distance.

c. Both Gower and Euclidian distance can be used.

d. Distance between categorical variable cannot be measured.

Correct answer is: *a*

Gower distance can be used as a distance measure in such cases.

End of Lesson05–Clustering and Segmentation