Unsupervised Learning

Clustering

Machine Learning (Clustering)

1. Clustering is primarily an exploratory technique to discover hidden structures of the data, possibly as a prelude to more focused analysis or decision processes
   a. A way to decompose a data set into subsets with each subset representing a group with similar characteristics
   b. When we cluster observations we seek to partition them into distinct groups such that objects in the same group are more similar to each other in some sense than to objects of different groups
   c. The groups are known as clusters and each cluster gets distinct label called cluster id, the centroid of the cluster, and inertia

2. Clustering is often used as a lead-in to classification. Once the clusters are identified, labels can be applied to each cluster to classify each group based on its characteristics

3. Can also be used to decide whether there is a need for separate models representing each cluster (will be so if clusters are very different on the attributes), or a single model can be reliable. If separate models is required, will the attributes have equal importance

4. Clustering helps simplifying the data representation through the centroid methods.

Applications of clustering

Some specific applications of k-means are image processing, medical, and customer segmentation

a. **Image processing** : used to cluster of pixels representing objects in each frame. The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame. Successive frames are examined to identify any changes to the clusters. These newly identified clusters may indicate unauthorized access to a facility.

b. **Medical** : Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters under various health conditions

c. **Customer segmentation** : Cluster customers on basis of frequency of purchase, recency of purchase, value of purchase and look for common attributes among high value customers. Target all potential customers who have similar attributes

Clustering types

1. Two broad categories of clustering include hierarchical (agglomerative, divisive) and non hierarchical

2. Hierarchical clustering
   a) Agglomerative clustering algorithm uses a bottom-up approach and merges smaller clusters into larger ones
   b) Divisive clustering uses top-bottom approach to break a large cluster into smaller clusters

3. Non-hierarchical / partitional clusters are formed on assumption that the clusters are disjoint and there is no hierarchical relation between them. K Means is an example

Machine Learning (Clustering – Distance calculations)

1. Irrespective of the clustering algorithm, we need a way of defining similarity/ dissimilarity

2. Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between objects being clustered

3. Clustering method attempts to group the objects based on the definition of similarity supplied to it. The definition uses distance calculation functions for the same

4. We also need a way to define and calculate distance between clusters of objects

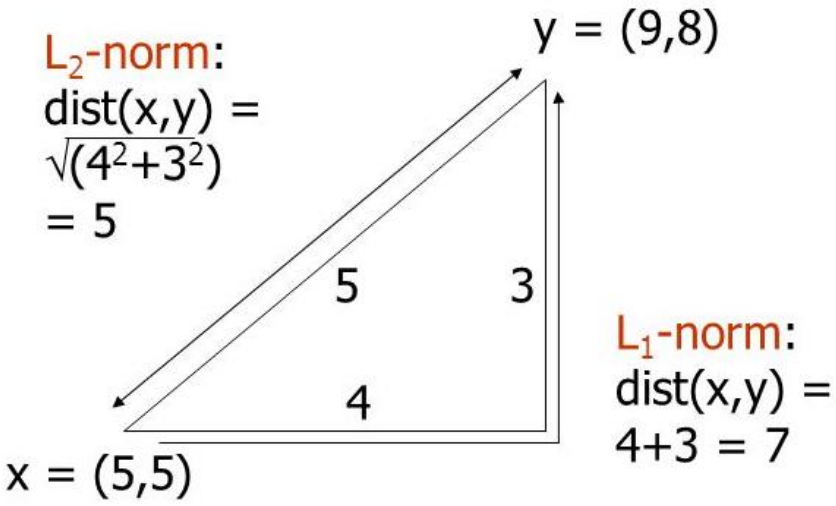
5. The lesser the distance, more similar the objects are and more suited to form a larger cluster

6. There are many ways of calculating distance between two points i.e. if d = f(x,y) then there are many ways in which f can be implemented

Machine Learning (Clustering – Distance calculations)

1. Euclidean Distance
   a. L2 norm : d(x,y) = square root of the sum of the squares of the differences between x and y in each dimension. The most common notion of “distance”. If there are two dimensions x and y, the distance between two point A and B is –

2. Manhattan / Taxi Distance
   a. L1 norm : sum of the differences in each dimension. Manhattan distance = distance if you had to travel along coordinates only



Machine Learning (Clustering – Distance calculations)

1. Other non Euclidean Distance
    a. Jaccard distance for sets = 1 minus ratio of sizes of intersection and union.
    b. Cosine distance = angle between vectors from the origin to the points in question.
    c. Edit distance = number of inserts and deletes to change one string into another
    d. Mahalanobis distance – takes into account the covariance between attributes (Ref: http://mccormickml.com/2014/07/22/mahalanobis-distance/)

2. Euclidian Distance … Some points
    a. The measures computed in Euclidian methods are highly influenced by the scale of each variable
    b. Variables with larger scale have much greater influence over the total distance. This may or may not be good for clustering

Machine Learning (Clustering Loss Function)

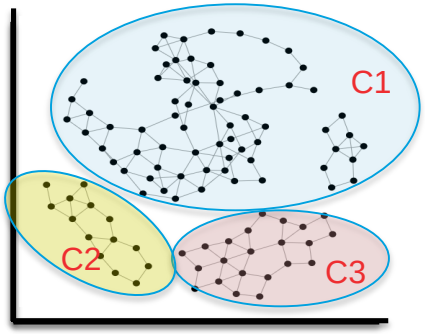
1. Kmeans clustering partitions data into K disjoint sets or clusters where K is a pre-specified number. It can range from 1 to n where n is number of data points

2. Let the clusters be C1, C2, … , Ck

3. C1 U C2 U … U Ck = { 1,2,….,n} data set

4. Ci n Cj = 0 for all i not equal to j

For e.g.
K = 3



5. A good clustering is the one where the within cluster variation is minimal

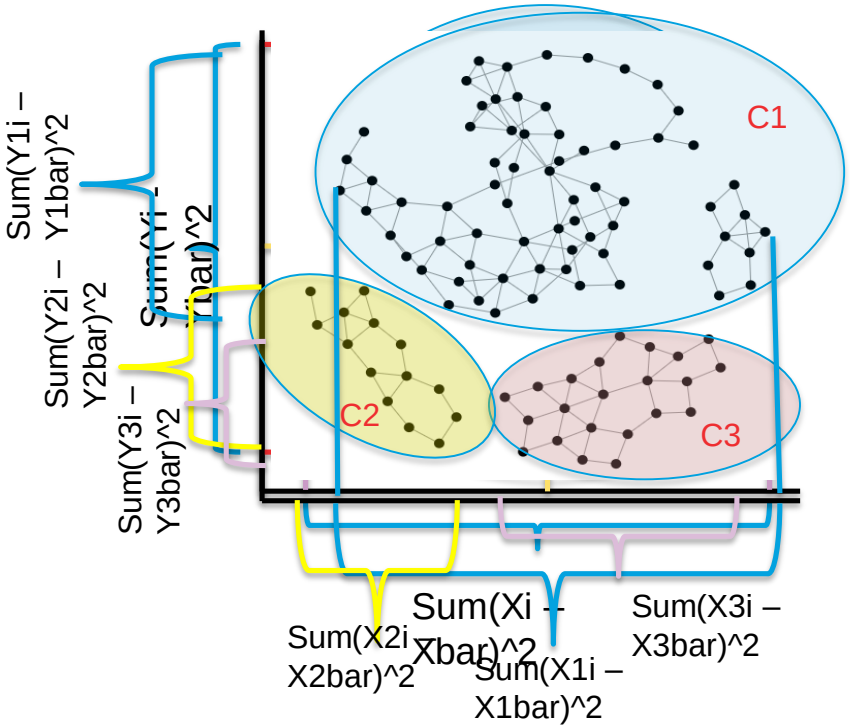
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

6. In words, minimize the total sum of all within cluster variations

7. The within cluster variation is defined as $W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$

8. Thus the optimization problem becomes $\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$

Unsupervised Machine Learning



When super cluster is broken into smaller clusters then the sum of within cluster variance i.e. $\left\{\sum_{k=1}^{K} W(C_k)\right\}$ is reduced and becomes less than $\left\{\sum_{k=1}^{1} W(C_k)\right\}$

The difference in sum of within cluster variance between super cluster and the three cluster represents inter cluster variance

The sum of within cluster variance will approach zero as number of clusters approach number of data points i.e. each data point becomes one cluster

Such clusters will be of no use as interesting properties are found in groups

Note: $\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$

Machine Learning (Clustering Loss Function)

9. Given that the total variance T across the entire data set is a constant, T = BC + WC (T = between cluster variations (BC) and within cluster variance (WC)

10. Thus WC = T – BC

11. Minimizing WC = Maximizing BC. Thus the objective of clustering is to get tightest and farthest clusters. Achieving one will automatically get the other i.e. tightest will make the clusters farthest. In the process ensure the clusters are natural, meaningful

12. Unfortunately, there is no well defined algorithm to achieve this. The problem of finding the tightest and farthest clusters in a data set belongs to a family of problems called the NP Hard problems (Non deterministic, Polynomial-time Hard problems). Ref: http://jeffe.cs.illinois.edu/teaching/algorithms/notes/30-nphard.pdf

13. For this reason, the clustering algorithms usually converge to sub-optimal solutions or converge to local optima. However, they are still powerful techniques in highlighting hidden structures

Dissimilarities and attributes

1. Let Xij represent measurements for Xi data point where i = 1,2,3,…n is index of the data points and j= 1,2,3,…p is the index of the variables or attributes

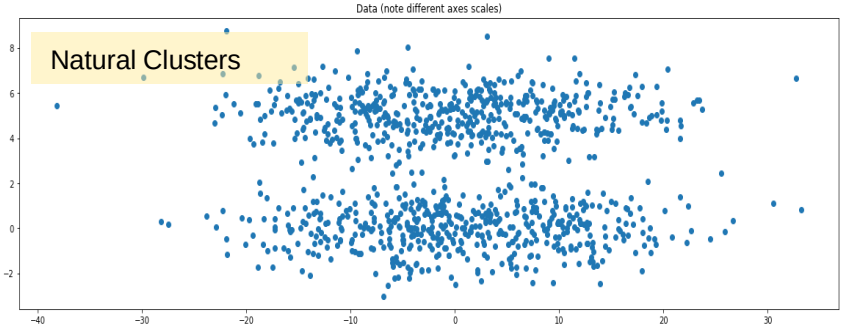
2. Thus distance D between two data points Xi , Xi' is $D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$
   a. dj is distance between the points on j attribute
   b. d is a function for calculating dissimilarity
   c. There are multiple mathematical functions that can do the job of d
   d. D (in capitals) is the sum of dissimilarity on all the concerned attributes (dj)

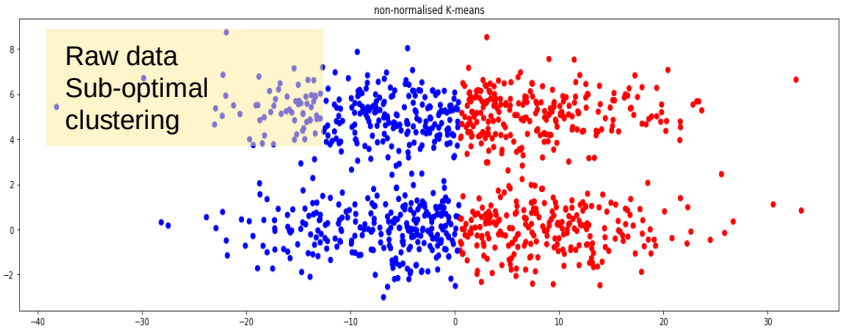
3. Scaling data and it's impact
   a. Scaling is the process of representing all the attributes on the same standard scale. For e.g. Zscoring. Usually scaling helps making "d" effective in representing dissimilarities
   b. Scaling ensures all attributes have same units and more importantly, have same degree of influence on the overall distance measure between points
   c. However, some attributes may exhibit more grouping tendency than others. Thus, they should get more weightage than others in clustering
   d. Thus, when we scale, the natural tendency of an attribute to cluster may be diluted which may distort the clusters

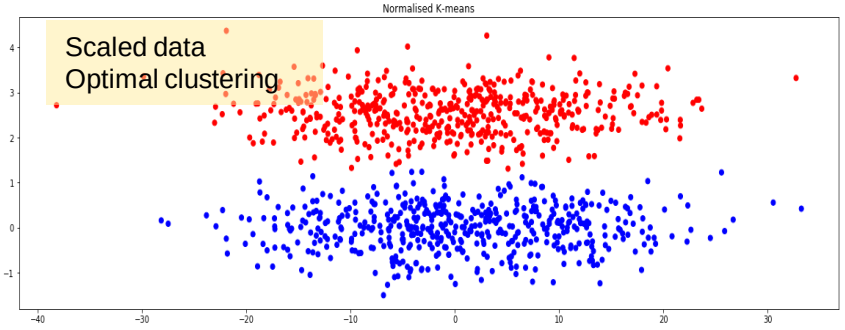
Unsupervised Machine Learning

greatlearning
Learning for Life

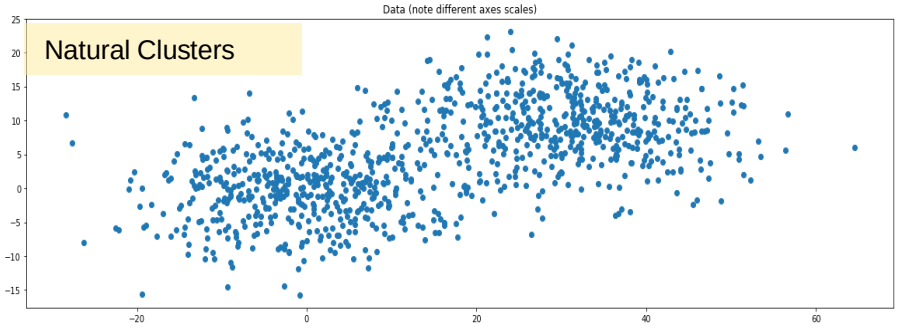
Normalization is good

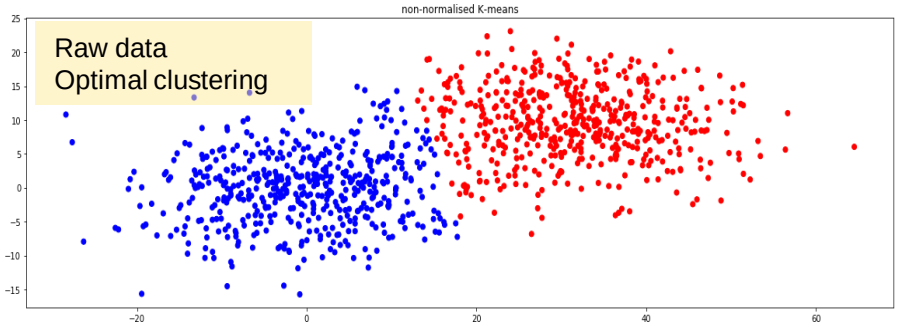


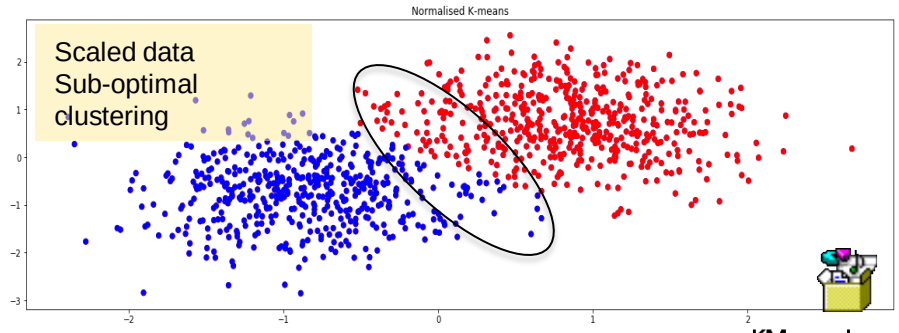




Normalization is not good







KMeans_Impact_o

Machine Learning (Clustering – Distance calculations)

1. Distance measures and some key points:
    a. Choice of distance measures play a key role in cluster analysis

    b. Knowledge of the distribution of data (gaussian or otherwise) will help

    c. Are the various attributes independent or influence each other

    d. Are their outliers in the data on the various dimensions

    e. Though Euclidian distance is the most commonly used distance metric, it has three main features that should be kept in view
        a. It is highly scale dependent. Changing the units of one variable can have a huge influence on the results. Hence standardizing the dimensions is a good practice
        b. It completely ignores the relationship between measurements (Refer to Mahalanobis distance diagram)
        c. It is sensitive to outliers. If the data has outliers that cannot be handled or removed, use of Manhattan distance is preferred

    f. <u>KMeans algorithm implements only Euclidian distance</u>

Machine Learning (K Means Clustering – Some considerations)

1. K-Means (a.k.a Lloyd's algorithm) clusters data by separating data points into groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squared errors

2. It requires the number of clusters to be specified, hence the term "K" in its name

3. It divides the samples into K disjoint clusters Ci, each described by the mean of the samples in the cluster. The means are commonly called "centroids" (they are not the points from the data)

4. The K-Means algorithm chooses centroids that minimizes the within cluster inertia (variations) across all the clusters

Machine Learning (K Means Clustering – Some considerations)

5. From a computational perspective, the k-means algorithm is indifferent to the units of measure for a given attribute (for example, meters or centimeters for a patient's height). However, the algorithm will identify different clusters depending on the choice of the units of measure.

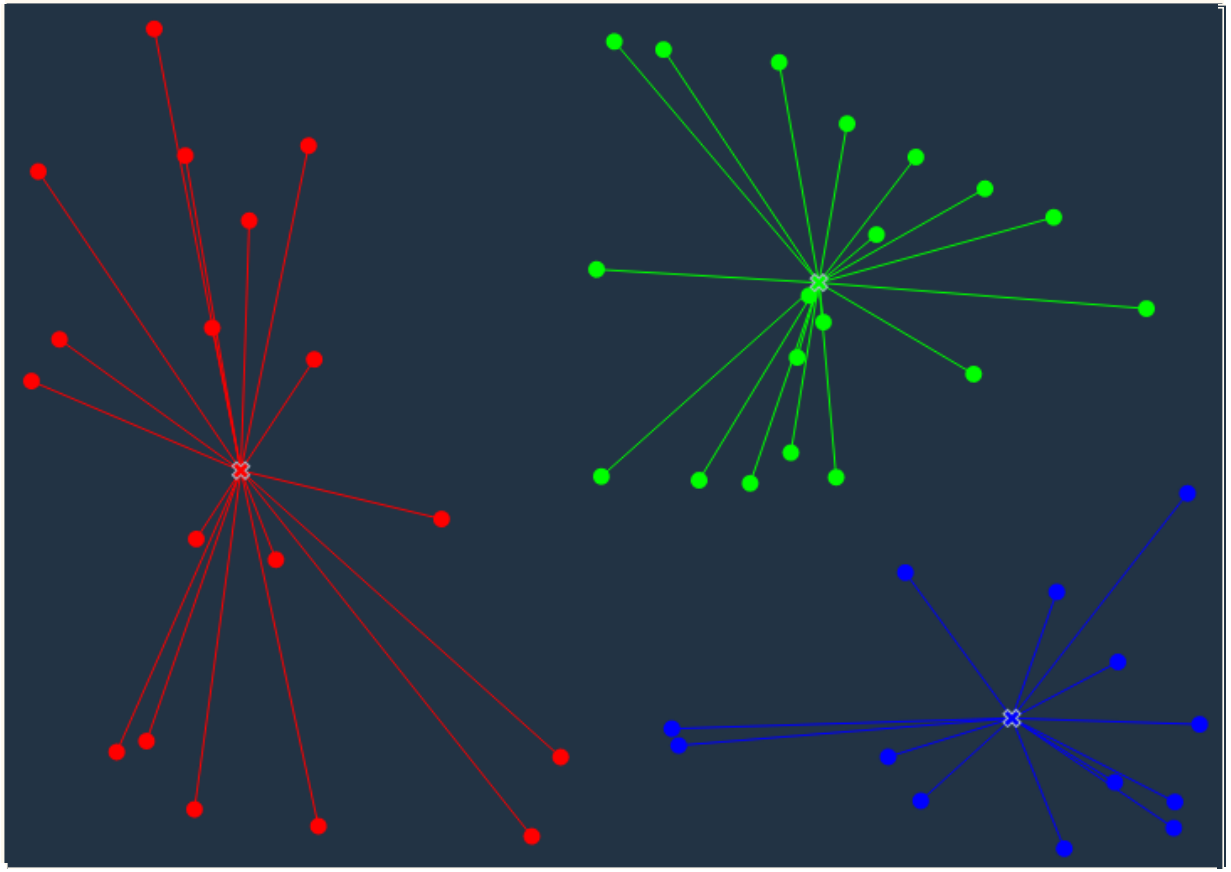
6. Choosing different starting points can result in different clusters. The algorithm is sensitive to the initial starting condition

7. Given enough time, K-means will always converge, however this may be a local minimum. This is highly dependent on the initialization of the centroids

8. Scikit-learn has implemented K-mean++ initialization scheme, which initializes centroids to be distant to one another which provably leads to better results

Machine Learning (K-means Clustering)

K-means treats feature values as coordinates in the multi-dimensional feature space



In the update phase the average position of each cluster is calculated (based on position of the points in the cluster)

Note : based on the distance of some boundary points from the new centroid, they get reallocated to different cluster

Ref: http://tech.nitoyon.com/en/blog/2013/11/07/k-means/

Machine Learning (K-means Clustering)

K Means Algorithm

1. Randomly pick K points in the feature space as initial centroids. Or randomly assign values 1 – K to each data point. This is the first version of K clusters. Obviously the worst clusters

2. Iterate until centroids stop moving or cluster assignment stops changing –
    a. For each of the K clusters, compute the centroid i.e. avg on each attribute. The point in the mathematical space where these averages meet, is the new centroid. The Kth cluster centroid is thus a vector of P feature means for the observations in the Kth cluster

    b. Re assign each data point to the centroid which is closest to them. Closeness is defined using Euclidian distance

Machine Learning (K-means Clustering)

K Means Algorithm

1. The Algorithm is guaranteed to decrease the objective function

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

2. We can re-write the objective function as $\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$

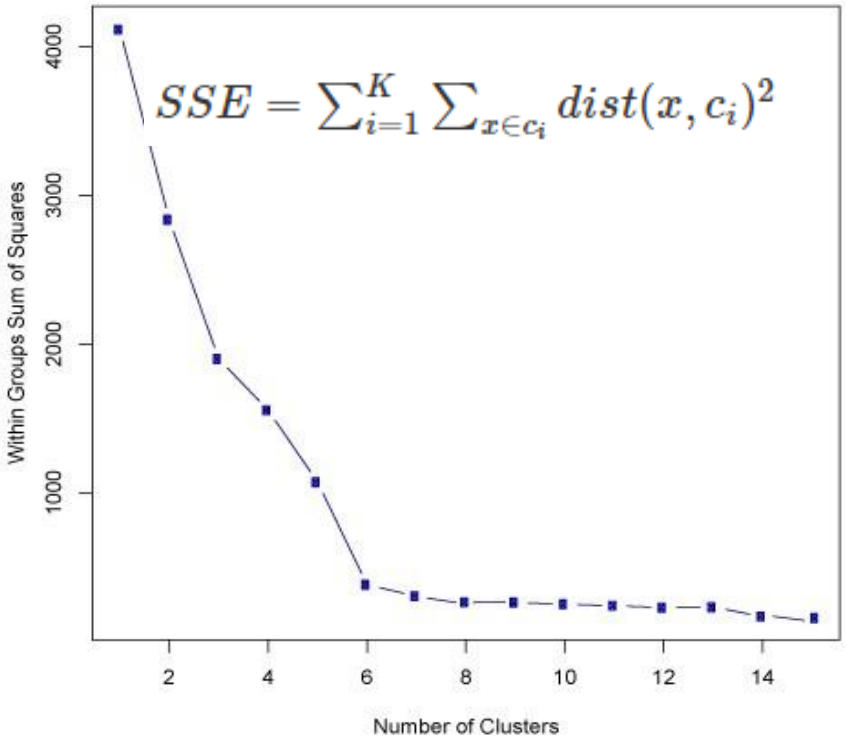
3. The objective function is now represented using the attribute average of a cluster

4. Since the algorithm recomputes and re-assigns the data points to the cluster closest, the term $\sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$ minimizes in every iteration

5. However, the final clusters so found, depend on the initial starting points and hence, the algorithm may get into **local minima**

Machine Learning (K-means Clustering)

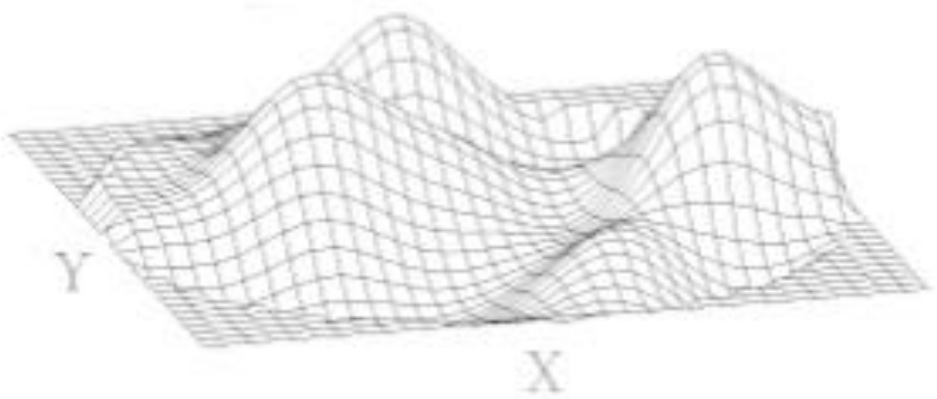
Without apriori knowledge, one can use elbow method that measures the homogeneity or heterogeneity within clusters as the number of clusters change (i.e. K is changed). One way to measure is use sum of square errors in each cluster



Machine Learning (K-means Clustering)

Visual Analysis for Clustering

1. Visual analysis of the attributes selected for the clustering may given an idea of the range of values that K should be evaluated in

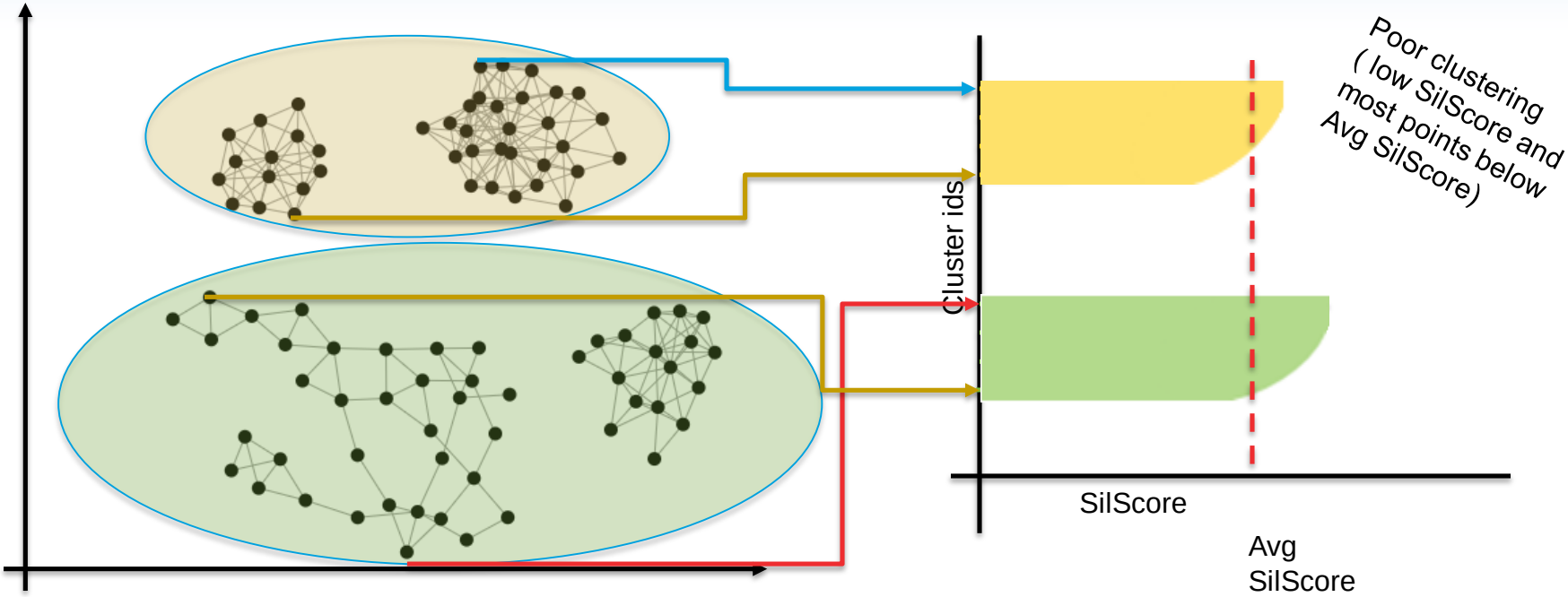


2. Identifying the attributes on which clusters are clearly demarcated and using them in incremental order to build the multi-dimensional clusters likely to give much better clusters than using all the attributes at one go

Unsupervised Machine Learning

greatlearning
Learning for Life

Silhouette Score



SilScore = B – A / max (A,B)

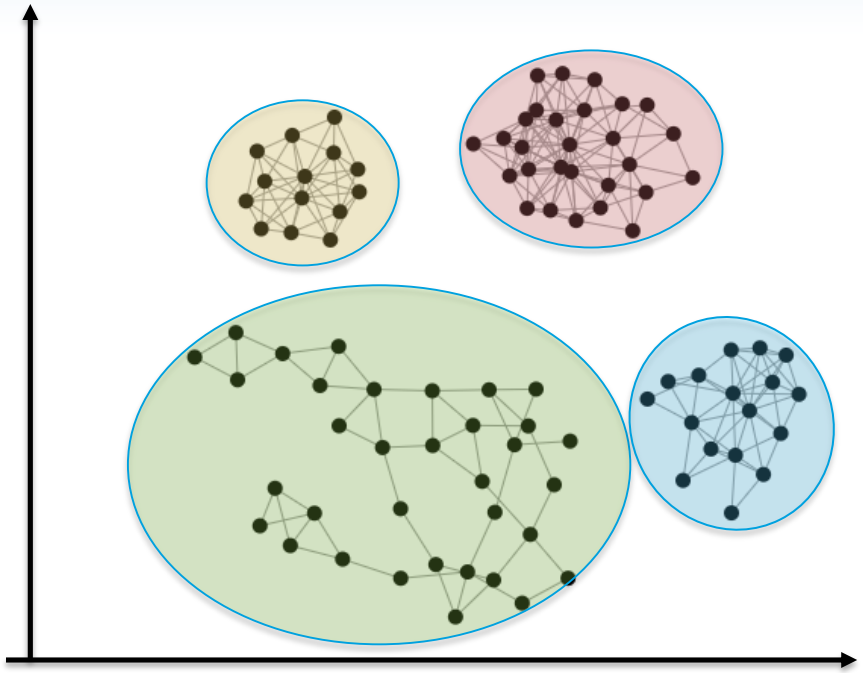
A = avg distance of a point from other points in its cluster

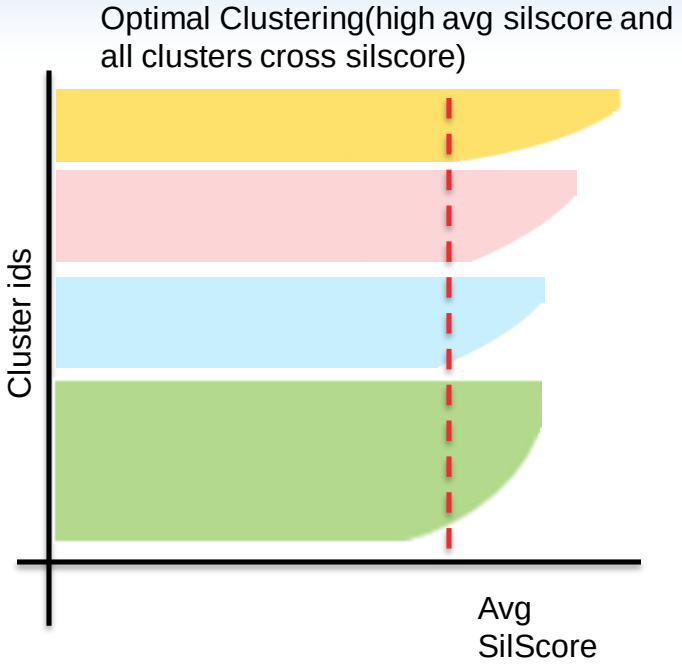
B = avg distance of a point from the points in the nearest cluster

Note: Every data point has a SilScore

Avg SilScore = sum of all SilScore / Number of data points

Unsupervised Machine Learning





SilScore = B – A / max (A,B)

A = avg distance of a point from other points in its cluster

B = avg distance of a point from the points in the nearest cluster

Note: Every data point has a SilScore

Avg SilScore = sum of all SilScore / Number of data points

Machine Learning (K-means Clustering)

Dynamic Clustering

1. Clustering on correct attributes is the key to good clustering results.

2. We can also consider those attributes who's value changes with time. For e.g. age, income category, years of work experience etc.

3. We can use sequential k means clustering over time to track individual clusters (how they change in size, shape and position

4. We can also understand how individual data points move across clusters, form new clusters etc.

5. Analyzing the changes in the clusters over time using metrics such as

6. Cluster size, new entries and exits one can analyze the impact of strategies designed based on earlier clustering analysis

Machine Learning (K-Means Clustering)

Strengths	Weakness
Use simple principles without the need for any complex statistical terms	Computationally intensive How to fix K?
Once clusters and their associated centroids are identified, it is easy to assign new objects (for example, new customers) to a cluster based on the object's distance from the closest centroid	The k-means algorithm is sensitive to the starting positions of the initial centroid. Thus, it is important to rerun the k-means analysis several times for a particular value of k to ensure the cluster results provide the overall minimum WSS
Because the method is unsupervised, using k-means helps to eliminate subjectivity from the analysis.	Susceptible to curse of dimensionality

Machine Learning (K-means Clustering)

Lab- 1 Analyze auto mpg data set using K means to explore the data in terms of the various attributes

Sol: KMeansClustering_Auto_Mpg.ipynb

Machine Learning (Hierarchical (Agglomerative) Clustering)

1. The agglomerative clustering starts with each cluster comprising exactly one data point in the feature space

2. It progressively agglomerates / combines the two nearest clusters until there is one grand cluster left in the feature space

3. For the closest cluster analysis, each of the inter cluster distance measurement techniques (single link, complete link, average link, centroid distance) can be implemented
   a. In single linkage method, the minimum distance between nearest points from the two clusters is used to consolidate clusters
   b. In complete linkage, distance between two farthest points form each cluster is considered
   c. Group average clustering is based on the average distance between clusters

4. Prior domain knowledge helps in deciding the inter cluster distance metric selection. If the clusters are likely to be in long chain or sausage like, minimum distance (single linkage) would be a good choice

5. Complete and average linkage are better choice if the clusters are likely to be spherical

Machine Learning (Clustering – Measuring distance between clusters)

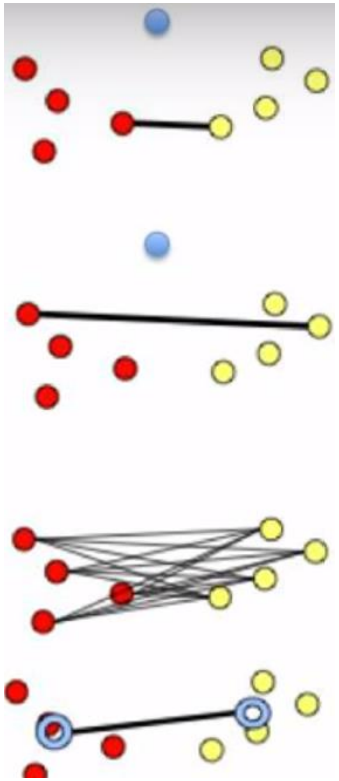
1. Ideally, a good clustering should result in compact clusters separated from one another by maximal distance. This calls for measuring the distance between cluster. The most widely used methods include :

   a. Minimum distance(single linkage) – is the distance between pair of records Ai and Bj that belong to clusters A and B respectively and are closest

   b. Maximum distance(complete linkage) – is the largest distance between the pair of records Ai and Bj that belong to cluster A and B respectively

   c. Average distance (average linkage) - average distance of all possible distances between records in one cluster to records in other cluster

   d. Centroid distance - the distance between centroids of the different clusters.

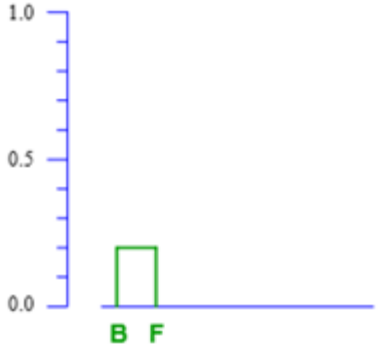


2. Distance between clusters is used in hierarchical clustering

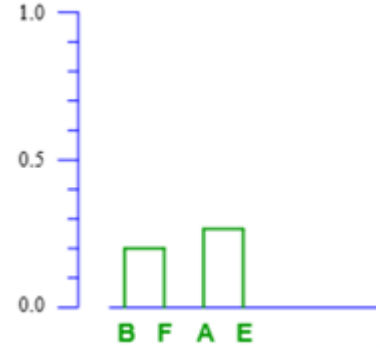
Complete Linkage

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

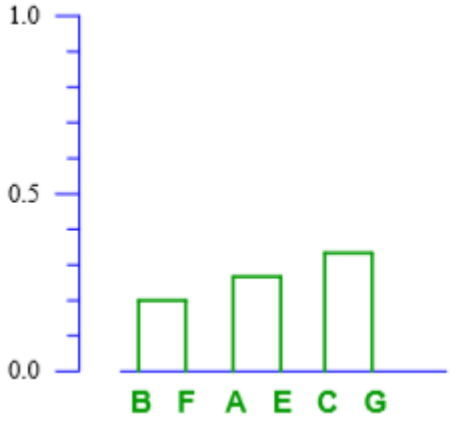
samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0

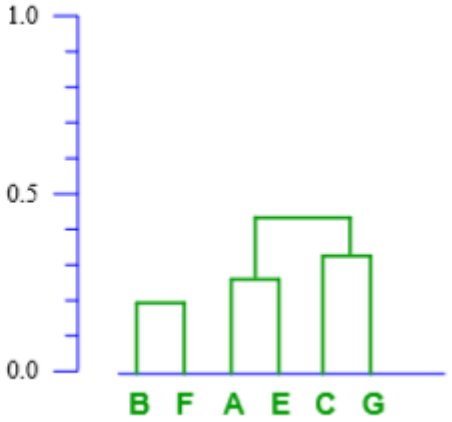


samples	(A,E)	(B,F)	C	D	G
(A,E)	0	0.7778	0.4286	1.0000	0.3750
(B,F)	0.7778	0	0.7143	0.8333	0.7778
C	0.4286	0.7143	0	1.0000	0.3333
D	1.0000	0.8333	1.0000	0	0.8571
G	0.3750	0.7778	0.3333	0.8571	0



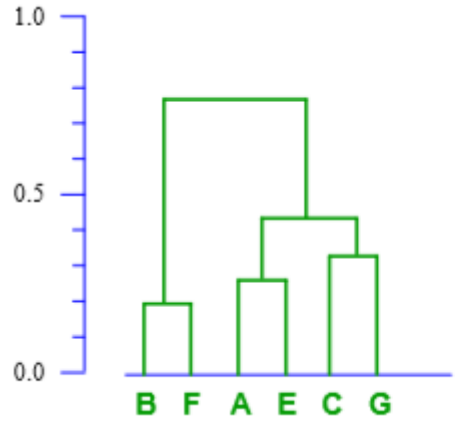
samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0

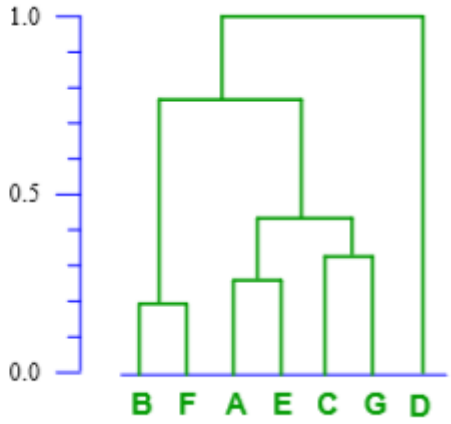




samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0

samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0



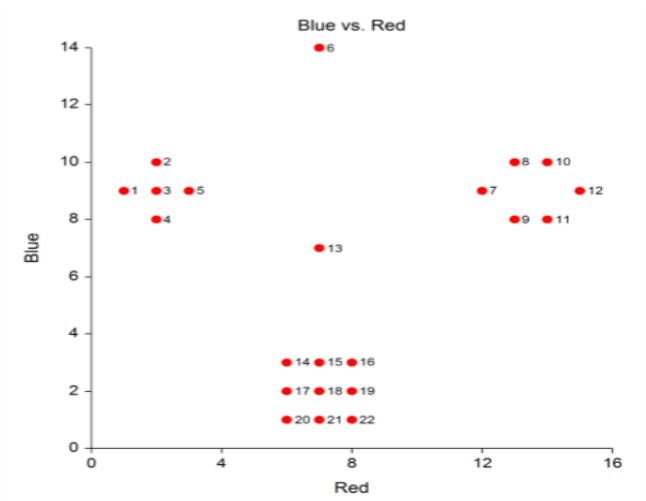


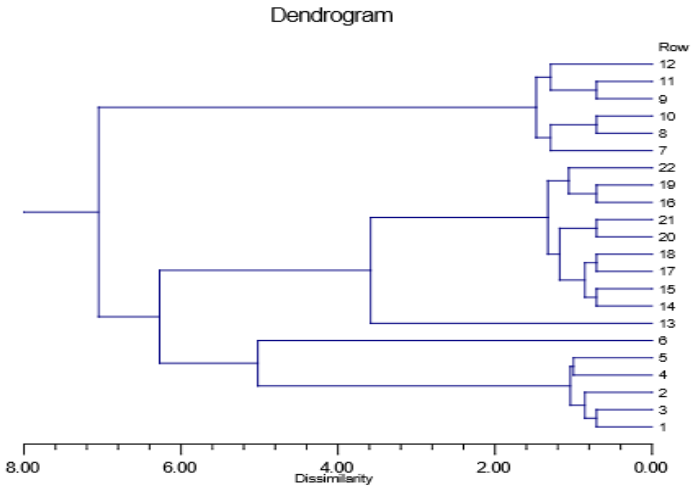
Machine Learning (Hierarchical (Agglomerative) Clustering)

1. Dendrogram is a tree like diagram that summarizes the process of clustering. At the leaf are the records representing the data points while root node is the entire data. The intermediate nodes has two daughter nodes representing sub groups

2. Similar records are joined by lines who's vertical length reflects the relative distance between the data points

3. When viewed bottom up, the tree posses a monotonicity property. Dissimilarity between the merged clusters is monotone increasing with the level of merger





Source: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf

Machine Learning (Hierarchical (Agglomerative) Clustering)

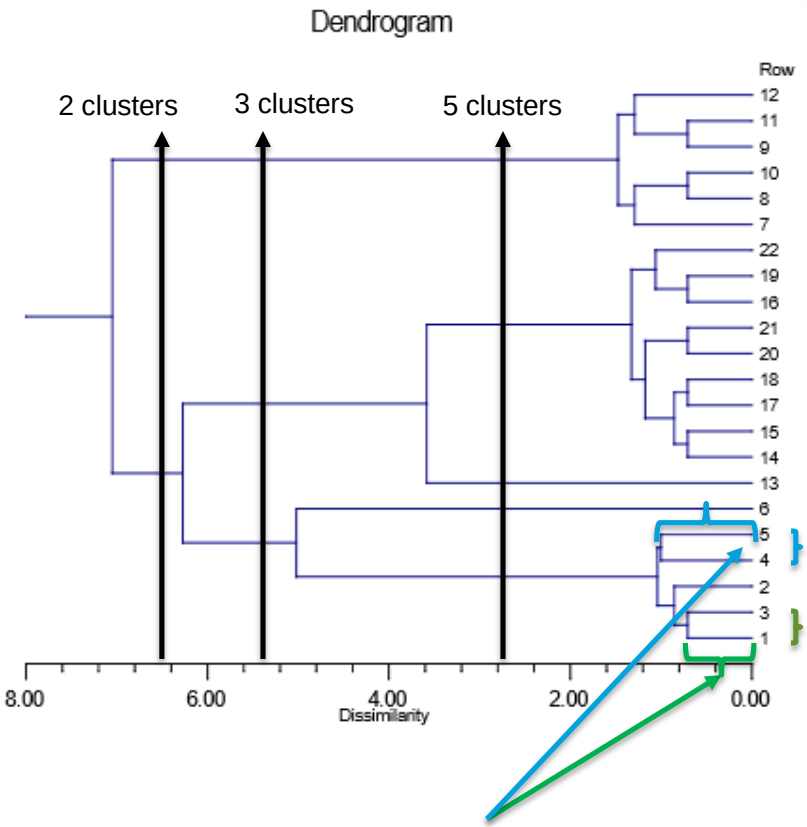
1. The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters (the scale is in reverse order)

2. The vertical axis represents the objects and clusters.

3. Each fusion of two clusters is represented on the graph by the splitting of a horizontal line

4. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters

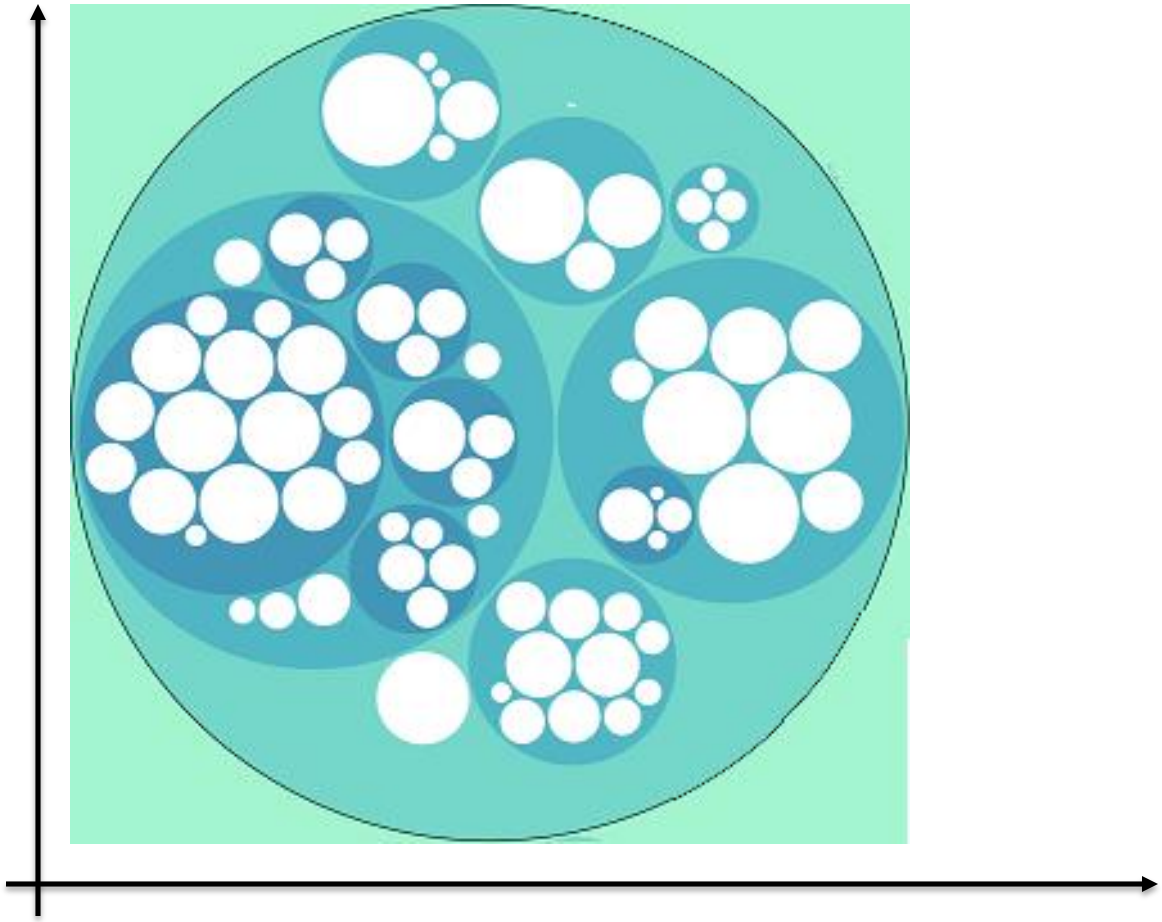
5. When we draw a vertical at any point on the X axis, the number of lines it cuts indicates number of clusters at that value of dissimilarity



Distance/ dissimilarity between 1,3 is less than between 4 and 5. This is reflected in the length of the horizontal bar which is longer for 4,5 compared to 1,3

Source: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf

Clusters within clusters



Machine Learning (Hierarchical (Agglomerative) Clustering)

6. Dendogram is often viewed as a graphical summary of the data and it's clusters. Such interpretation has to be taken with caution as the dendograms change with change in the linkage methods and distance functions!

7. The hierarchical structure created is the result of the linkage and distance methods. Whether such structure exists? Is a question that needs to be answered before interpreting the hierarchy

8. The extent to which the hierarchical structure produced by the dendogram actually represents the data itself can be judged by the *cophenetic correlation coefficient*

9. This is the correlation between the N(N-1)/2 pairwise observation dissimilarities Dii' input to the algorithm and their corresponding cophenetic dissimilarity Cii'

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

samples	(A,E)	(B,F)	C	D	G
(A,E)	0	0.7778	0.4286	1.0000	0.3750
(B,F)	0.7778	0	0.7143	0.8333	0.7778
C	0.4286	0.7143	0	1.0000	0.3333
D	1.0000	0.8333	1.0000	0	0.8571
G	0.3750	0.7778	0.3333	0.8571	0

Distance between	Pairwise Dist	Dendogramic
A – B	.5000	.7778
A – F	.6250	.7778
E – B	.6667	.7778
E – F	.7778	.7778

Machine Learning (Agglomerative Clustering)

Lab- 2 Analyze the wines data set using agglomerative clustering

Description – This data is about red wines. The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). We have to use K-means clustering to understand if 10 clusters exist and what their characteristics are

Sol: HierarchialClustering_Wine.ipynb

Machine Learning (Clustering effectiveness Cophenetic correlation)

1. Suppose that the original data {Xi} have been modeled using a cluster method to produce a dendrogram {Ti}

2. Define the following distance measures :
    1. x(i,j)=|Xi−Xj| , the ordinary Euclidean distance between the i th and j th observations

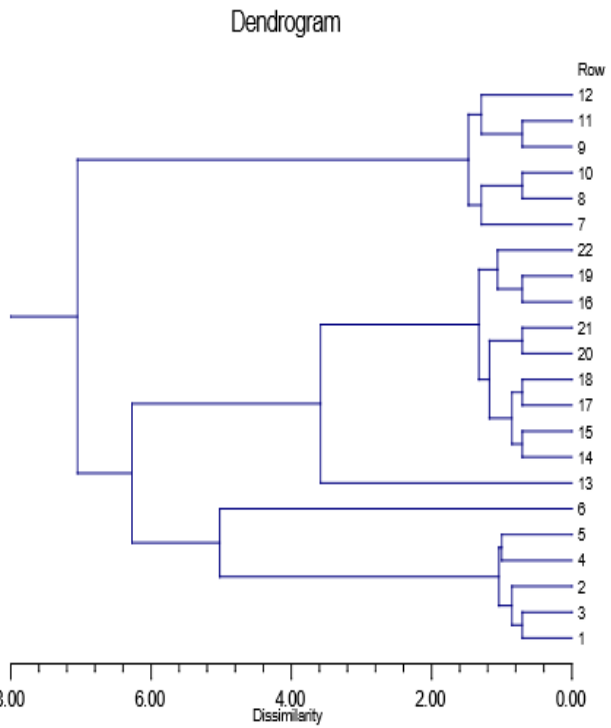
    2. t(i,j)= the dendrogrammatic distance between the model points Ti and Tj. This distance is the height of the node a which these two points are first joined together

    3. Then, letting x be the average of the x(i,j) , and letting t be the average of the t(i,j),

3. Cophenetic correlation coefficient c is defined as

$$c = \frac{\sum_{i<j}(x(i,j)-x)(t(i,j)-t)}{\sqrt{[\sum_{i<j}(x(i,j)-x)^2][\sum_{i<j}(t(i,j)-t)^2]}}$$

4. Values close to 1 is preferred



For comparison of distance methods, clustering techniques and Cophenetic, read

https://journalofinequalitiesandapplications.springeropen.com/track/pdf/10.1186/1029-242X-2013-203?site=journalofinequalitiesandapplications.springeropen.com

Machine Learning (Clustering effectiveness Silhouette coefficient)

1. Silhouette analysis can be used to study the separation distance between the resulting clusters.

2. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

3. If the silhouette plot shows values close to one for each observation, the fit was good; if there are many observations closer to zero, it's an indication that the fit was not good.

Thank You