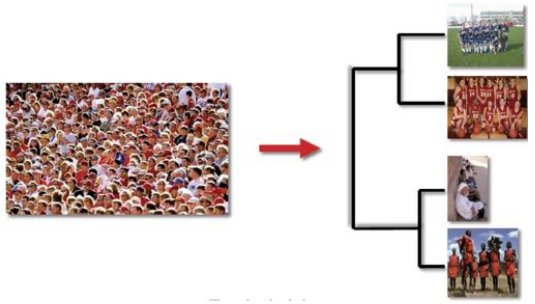
Hierarchical Clustering

Mikhail Dozmorov
Fall 2016

What is clustering

- Partitioning of a data set into subsets.
- A cluster is a group of relatively homogeneous cases or observations



What is clustering

Given n objects, assign them to k groups (clusters) based on their similarity

- Unsupervised Machine Learning
- Class Discovery
- Difficult, and maybe ill-posed problem!

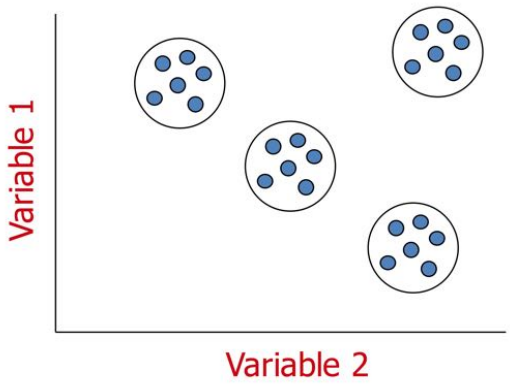
Clustering impossible

- Scale-invariance - meters vs inches
- Richness - all partitions as possible solutions
- Consistency - increasing distances between clusters and decreasing distances within clusters should yield the same solution

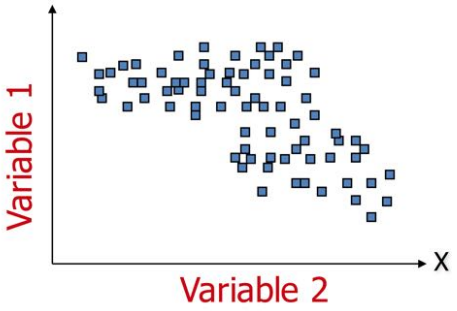
No function exists that satisfies all three.

J. Kleinberg. "An Impossibility Theorem for Clustering. Advances in Neural Information Processing Systems" (NIPS) 15, 2002.
https://www.cs.cornell.edu/home/kleinber/nips15.pdf

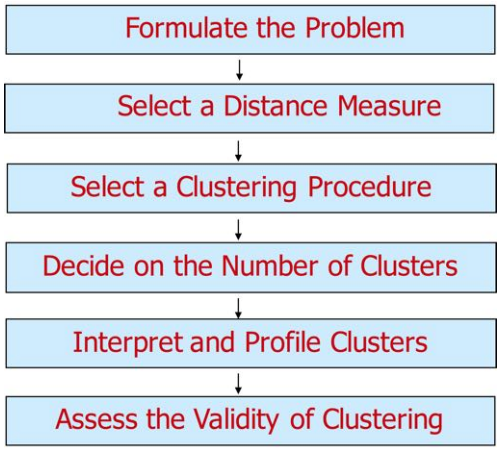
Clustering utopia



Clustering reality



Conducting Cluster Analysis



Clustering gene expression

Gene expression matrix

Samples

Genes

$$\begin{bmatrix} x_{11} & x_{12} & \mathsf{L} & x_{1n} \\ x_{21} & x_{22} & \mathsf{L} & x_{2n} \\ \mathsf{M} & \mathsf{M} & \mathsf{L} & \mathsf{M} \\ x_{g1} & x_{g2} & \mathsf{L} & x_{gn} \end{bmatrix}$$

Formulating the Problem

- Most important is selecting the variables on which the clustering is based.
- Inclusion of even one or two irrelevant variables may distort a clustering solution.
- Variables selected should describe the similarity between objects in terms that are relevant to the marketing research problem.
- Should be selected based on past research, theory, or a consideration of the hypotheses being tested.

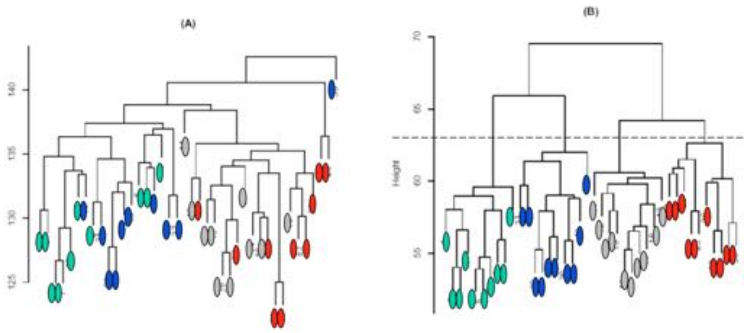
Filtering

- Non-informative genes contribute random terms in the calculation of distances
- The resulting effect is that they hide the useful information provided by other genes
- Therefore, assign non-informative genes zero weight, i.e., exclude them from the cluster analysis

Filtering examples

- % Present >= X - remove all genes that have missing values in greater than (100-X) percent of the columns
- SD (Gene Vector) >= X - remove all genes that have standard deviations of observed values less than X
- At least X Observations with abs(Val) >= Y - remove all genes that do not have at least X observations with absolute values greater than Y
- MaxVal-MinVal >= X - remove all genes whose maximum minus minimum values are less than X

Clustering noise



A: 450 relevant genes plus 450 “noise” genes.

B: 450 relevant genes.

Cluster the right data

Clustering works as expected when the data to be clustered is processed correctly

- Log Transform Data - replace all data values x by log2 (x). Why?
- Center genes [mean or median] - subtract the row-wise mean or median from the values in each row of data, so that the mean or median value of each row is 0.
- Center arrays [mean or median] - subtract the column-wise mean or median from the values in each column of data, so that the mean or median value of each column is 0.

Cluster the right data

Clustering works as expected when the data to be clustered is processed correctly

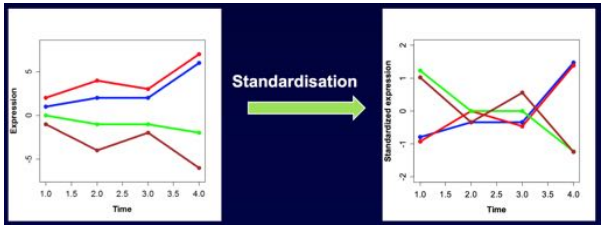
- Normalize genes - multiply all values in each row of data by a scale factor S so that the sum of the squares of the values in each row is 1.0 (a separate S is computed for each row).
- Normalize arrays - multiply all values in each column of data by a scale factor S so that the sum of the squares of the values in each column is 1.0 (a separate S is computed for each column).
- These operations are not associative, so the order in which these operations is applied is very important
- Log transforming centered genes are not the same as centering log transformed genes.

Standartization

In many cases, we are not interested in the absolute amplitudes of gene expression, but in the relative changes. Then, we standardize:

$$g_s = (g - \hat{g})/\sigma(g)$$

Standardized gene expression vectors have a mean value of zero and a standard deviation of one.



Distance

- Clustering organizes things that are close into groups
- What does it mean for two genes to be close?
- What does it mean for two samples to be close?
- Once we know this, how do we define groups?

How to define (dis)similarity among objects

Distance

- We need a mathematical definition of distance between two points
- What are points?
- If each gene is a point, what is the mathematical definition of a point?

Points

$Gene_1 = (E_{11}, E12, \ldots, E1N)$

$Gene_2 = (E_{21}, E22, \ldots, E2N)$

$Sample_1 = (E_{11}, E_{21}, \ldots, E_{G1})$

$Sample_2 = (E_{12}, E_{22}, \ldots, E_{G2})$

$E_{gi} = expression\ gene\ g,\ sample\ i$

Distance definition

For all objects i, j, and h

$$d(i,j) \geq 0$$
$$d(i,i) = 0$$
$$d(i,j) = d(j,i)$$
$$d(i,j) \leq d(i,h) + d(h,j)$$

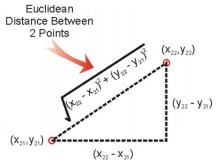
Most famous distance

Euclidean distance

Example distance between gene 1 and 2:

– Sqrt of Sum of $(E_{1i} - E_{2i})^2$, $i = 1, \ldots, N$

- When N is 2, this is distance as we know it:



- When N is 20,000 you have to think abstractly

Distance measures

- Euclidean distance

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$$

- Manhattan distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$$

- Minkowski distance (L_q metric)

$$d(i,j) = \left(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{in} - x_{jn}|^q \right)^{1/q}$$

- Disadvantages: not scale invariant, not for negative correlations

Distance measures

- When deciding on an appropriate value of q, the investigator must decide whether emphasis should be placed on large differences.
- Larger values of q give relatively more emphasis to larger differences.

Distance measures

- Canberra distance

$$d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

- Binary (0/1 vectors), aka Jaccard distance

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- Maximum distance between two components of x and y

Similarity definition

- For all objects i, j

$$0 \leq sim(i,j) \leq 1$$
$$sim(i,i) = 1$$
$$sim(i,j) = sim(j,i)$$

Similarity measures

- Gene expression profiles represent comparative expression measures
- Euclidean distance may not be meaningful
- Need distance measure that score based on similarity
- The more objects i and j are alike (or close) the larger $s(i,j)$ becomes

Similarity measures

Cosine similarity. From Euclidean dot product between two non-zero vectors:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

The cosine similarity is

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\left[\sum_{k=1}^{n} x_{ik}^2 \sum_{k=1}^{n} x_{jk}^2\right]^{1/2}}$$

Similarity measures

Pearson correlation coefficient [-1, 1]

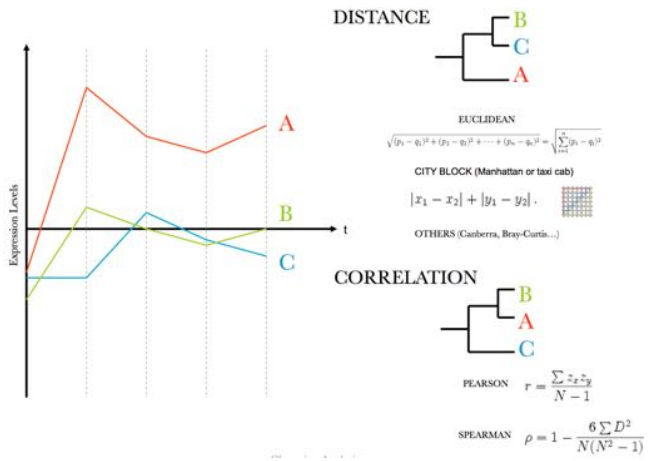
Vectors are normalized to the vector's means

$$s(i,j)=\frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_{i.})(x_{jk}-\bar{x}_{j.})}{\left[\sum_{k=1}^{n}(x_{ik}-x_{i.})^2\sum_{k=1}^{n}(x_{jk}-\bar{x}_{j.})^2\right]^{1/2}}$$

Convert to dissimilarity [0, 1]

$d(i,j) = (1 - s(i,j))/2$

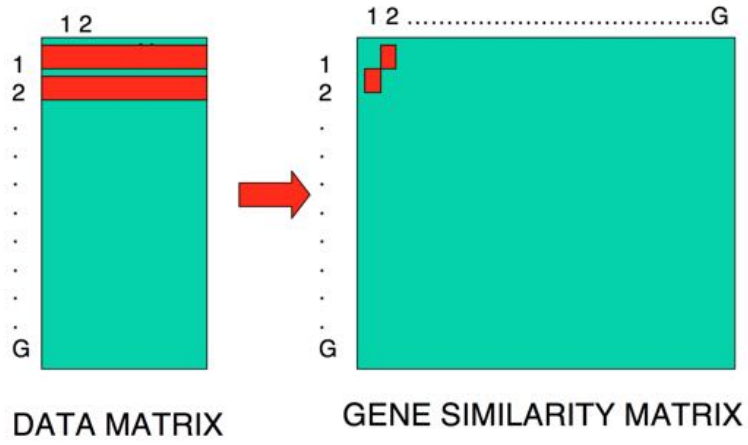
Distances between gene expression profiles



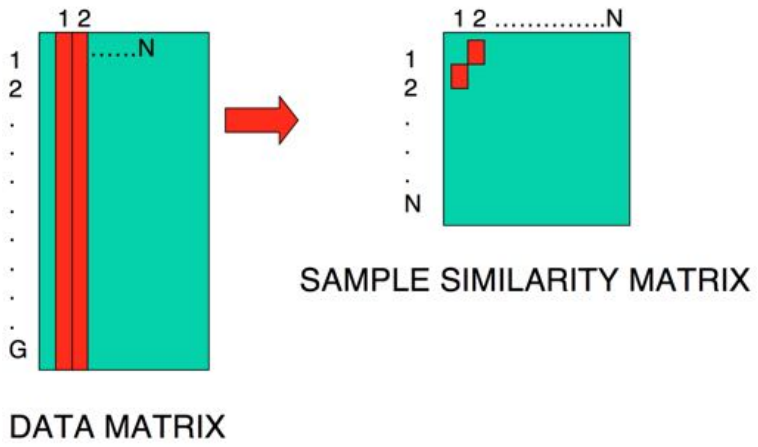
Convert correlation to dissmilarity

$$d(X_i, X_j) = \frac{1 - Cor(X_i, X_j)}{2}$$

The (dis-)similarity matrixes



The (dis-)similarity matrixes



Clustering binary data

- Two columns with binary data, encoded 0 and 1
- a - number of rows where both columns are 1
- b - number of rows where this and not the other column is 1
- c - number of rows where the other and not this column is 1
- d - number of rows where both columns are 0

Jaccard distance

$$\frac{a}{a+b+c}$$

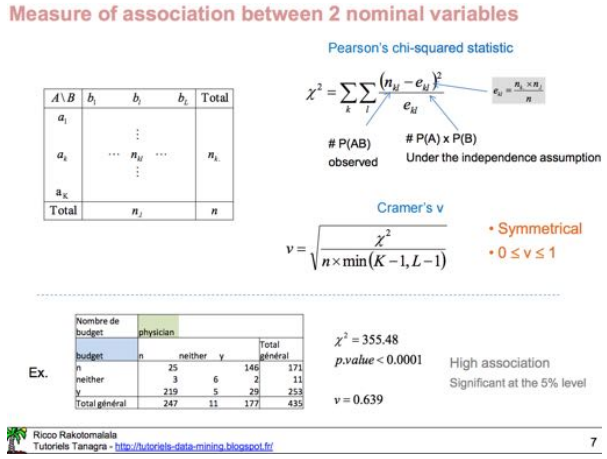
Clustering binary data

- Two columns with binary data, encoded 0 and 1
- a - number of rows where both columns are 1
- b - number of rows where this and not the other column is 1
- c - number of rows where the other and not this column is 1
- d - number of rows where both columns are 0

Tanimoto distance

$$\frac{a+d}{a+d+2(b+c)}$$

Clustering categorical data



Clustering mixed data

Gower distance

J. C. Gower "A General Coefficient of Similarity and Some of Its Properties" Biometrics 1971
http://venus.unive.it/romanaz/modstat_ba/gowdis.pdf

- Idea: Use distance measure between 0 and 1 for each pair of variables: $d_{ij}^{(f)}$
- Aggregate: $d(i,j) = \frac{1}{p} \sum_{i=1}^{p} d_{ij}^{(f)}$

Gower distance

How to calculate distance measure for each pair of variables

- Quantitative: interval-scaled distance $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$, where x_{if} is the value for object i in variable f, and R_f is the range of variable f for all objects
- Categorical: use "1" when x_{if} and x_{jf} agree, and "0" otherwise
- Ordinal: Use normalized ranks, then like interval-scaled based on range

Choose (dis-)similarity metric

- Think hard about this step!
- Remember: garbage in - garbage out
- The metric that you pick should be a valid measure of the distance/similarity of genes.

Examples

- Applying correlation to highly skewed data will provide misleading results.
- Applying Euclidean distance to data measured on categorical scale will be invalid.

Distances in R

Function	Package	Distances
dist	stats	Euclidean, Manhattan, Canberra, max, binary
daisy	cluster, bioDist	Euclidean, Manhattan
distancematrix, distancevector	hopach	Euclidean, cor, cosine-angle (abs versions)
vegdist	vegan	Jaccard, Gower, many others

Other packages: `cclust`, `e1071`, `flexmix`, `fpc`, `mclust`, `Mfuzz`, `class`

Assembling objects into clusters

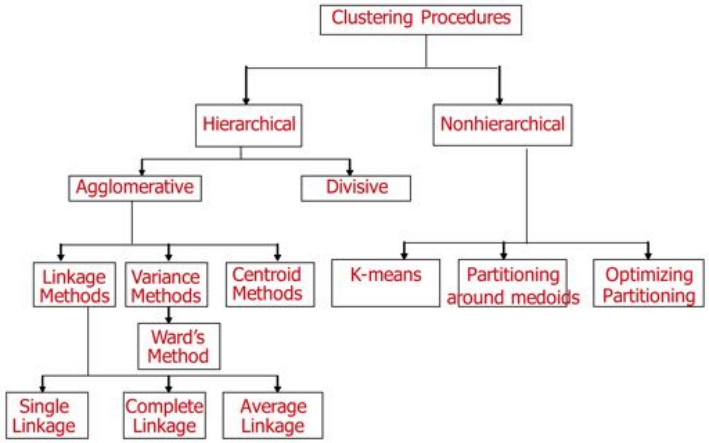
Assembling objects into clusters

- The number of ways to partition a set of n objects into k non-empty classes

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k-1}(-1)^i\binom{k}{i}(k-i)^n$$

- $S(n,1) = 1$ - one way to partition n object in to 1 group, or n disjoint groups
- $S(n,2) = 2^{n-1} - 1$ ways to partition n objects into two non-empty groups

Classification of Clustering Procedures



Hierarchical Clustering

- Allows organization of the clustering data to be represented in a tree (dendrogram)
- Agglomerative (Bottom Up): each observation starts as own cluster. Clusters are merged based on similarities
- Divisive (Top Down): all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Agglomerative clustering (bottom-up)

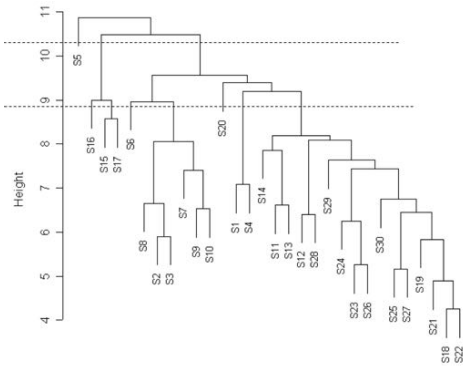
- Idea: ensure nearby points end up in the same cluster
- Starts with as each gene in its own cluster
- Joins the two most similar clusters
- Then, joins next two most similar clusters
- Continues until all genes are in one cluster

Divisive clustering (top-down)

- Starts with all genes in one cluster
- Choose split so that genes in the two clusters are most similar (maximize “distance” between clusters)
- Find next split in same manner
- Continue until all genes are in single gene clusters

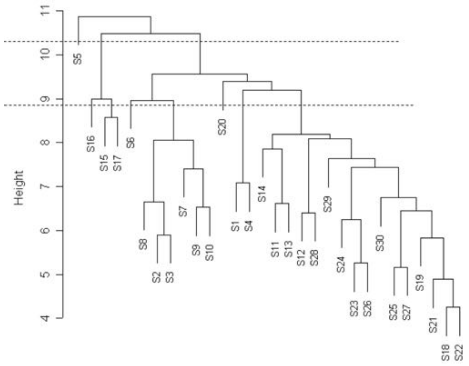
Dendrograms

- We can then make dendrograms showing divisions
- The y-axis represents the distance between the groups divided at that point



Dendrograms

- Note: Left and right is assigned arbitrarily. Vertical distance is what's matter
- Look at the height of division to find out distance. For example, S5 and S16 are very far.



Which to use?

- Both agglomerative and divisive are only ‘step-wise’ optimal: at each step the optimal split or merge is performed
- Outliers will irreversibly change clustering structure

Which to use?

Agglomerative/Bottom-Up

– Computationally simpler, and more available.

– More "precision" at bottom of tree

– When looking for small clusters and/or many clusters, use agglomerative

Which to use?

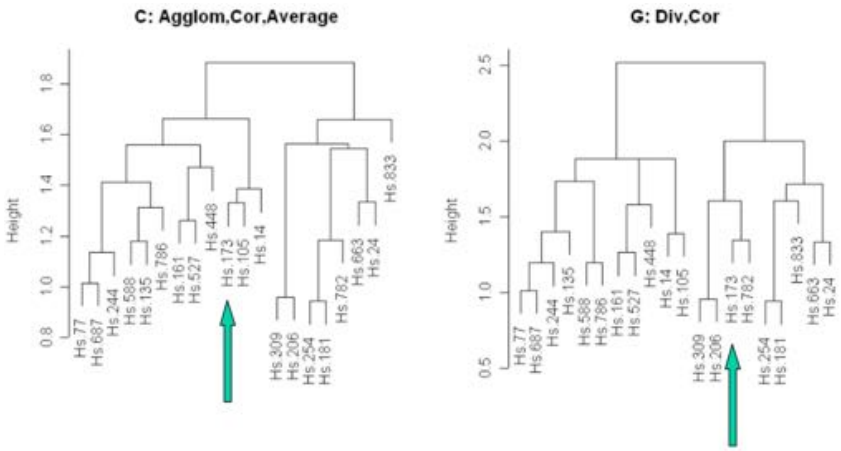
Divisive/Top-Down

– More "precision" at top of tree.

– When looking for large and/or few clusters, use divisive

Results ARE sensitive to choice!

Which to use?

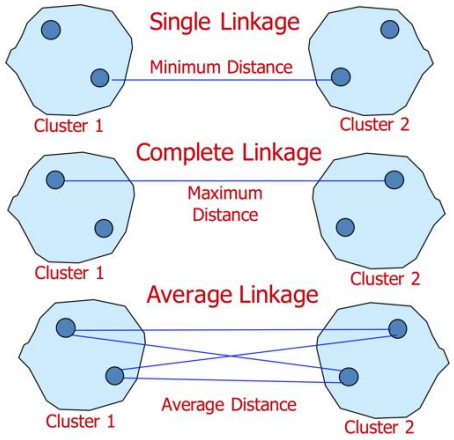


Linkage between clusters

- Single Linkage - join clusters whose distance between closest genes is smallest (elliptical)
- Complete Linkage - join clusters whose distance between furthest genes is smallest (spherical)
- Average Linkage - join clusters whose average distance is the smallest.

Linking objects based on the distance between them

Linkage between clusters

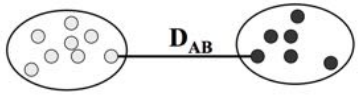


Single linkage

Cluster-to-cluster distance is defined as the *minimum distance* between members of one cluster and members of the another cluster. Single linkage tends to create 'elongated' clusters with individual genes chained onto clusters.

$$D_{AB} = \min (d(u_i, v_j))$$

where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B



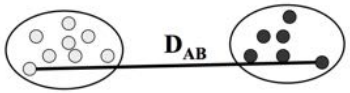
5

Complete linkage

Cluster-to-cluster distance is defined as the *maximum distance* between members of one cluster and members of the another cluster. Complete linkage tends to create clusters of similar size and variability.

$$D_{AB} = \max (d(u_i, v_j))$$

where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B



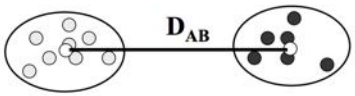
7

Average linkage

Cluster-to-cluster distance is defined as the *average distance* between all members of one cluster and all members of another cluster. Average linkage has a slight tendency to produce clusters of similar variance.

$$D_{AB} = 1/(N_A N_B) \Sigma \Sigma (d(u_i, v_j))$$

where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B



Ward's method

- Ward's procedure is commonly used. For each cluster, the sum of squares is calculated. The two clusters with the smallest increase in the overall sum of squares within cluster distances are combined.

$$\begin{aligned}\Delta(A,B) &= \sum_{i\in A\cup B}\|\vec{x}_i - \vec{m}_{A\cup B}\|^2 - \sum_{i\in A}\|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i\in B}\|\vec{x}_i - \vec{m}_B\|^2 \\ &= \frac{n_A n_B}{n_A + n_B}\|\vec{m}_A - \vec{m}_B\|^2\end{aligned}$$

- Δ - Merging cost of combining the clusters A and B. m_j is the center of cluster j, and n_j is the number of points in it.
- The sum of squares starts at 0 (each point is in its own cluster), and grows as clusters are merged. Ward's method keep this growth to minimum.

Ward, J. H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association
http://iv.slis.indiana.edu/sw/data/ward.pdf

Ward's method

- The distance d between two clusters C_i and C_j is defined as the loss of information (or: the increase in error) in merging two clusters.
- The error of a cluster C is measured as the sum of distances between the objects in the cluster and the cluster centroid $cenC$.
- When merging two clusters, the error of the merged cluster is larger than the sum or errors of the two individual clusters, and therefore represents a loss of information.
- The merging is performed on those clusters which are most homogeneous, to unify clusters such that the variation inside the merged clusters increases as little as possible.
- Ward's method tends to create compact clusters of small size. It is a least squares method, so implicitly assumes a Gaussian model.

Ward's method

An important issue though is the form of input that is necessary to give Ward's method. For an input data matrix, x, in R's `hclust` function the following command is required: `hclust(dist(x)^2, method="ward")` although this is not mentioned in the function's documentation file.

Fionn Murtagh "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" Journal of Classification 2014 https://link.springer.com/article/10.1007/s00357-014-9161-z