Unsupervised Learning
Cluster Analysis

- What is Cluster Analysis?
- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - **As a stand-alone tool to get insight into data distribution**
  - **As a preprocessing step for other algorithms**

Slides modified from http://www.stat.columbia.edu › notes › clustering
and https://www.learndatasci.com/glossary/hierarchical-clustering/

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earthquake epicenters should be clustered along continent faults

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

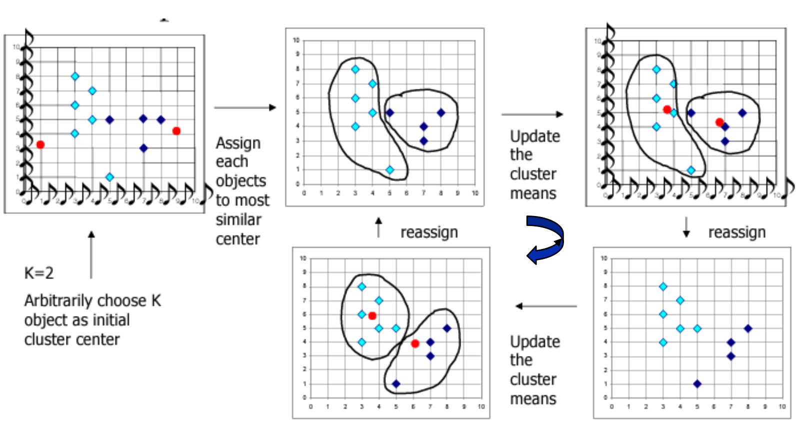
Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: d(i, j)
- There is a separate “quality” function that measures the “goodness” of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, and ordinal variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define “similar enough” or “good enough”
  - the answer is typically highly subjective.

Major Clustering Approaches

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Example: 2-Means



Partitioning Algorithm: K-Means

for $k = 1, \ldots, K$ let $\mathbf{r}(k)$ be a randomly chosen point from D;
while changes in clusters C_k happen **do**
  form clusters:
  for $k = 1, \ldots, K$ **do**
    $C_k = \{\mathbf{x} \in D \mid d(\mathbf{r}_k, \mathbf{x}) \leq d(\mathbf{r}_j, \mathbf{x})$ for all $j = 1, \ldots, K, j \neq k\}$
  end;
  compute new cluster centers:
  for $k = 1, \ldots, K$ **do**
    $\mathbf{r}_k$ = the vector mean of the points in C_k
  end;
end;

Comments on the K-Means Method

- Strength: Relatively efficient: O(t∗k∗n), where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.
  - Comparing: K-Medoids: O(k(n-k)2), O(ks2 + k(n-k))
- Comment: Often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms
- Weakness
  - Categorical data? Determining the distance measure
  - Need to specify k, the number of clusters, in advance
  - Unable to handle noisy data and outliers
  - Not suitable to discover clusters with non-convex shapes

Hierarchical Clustering: Types

- **Agglomerative**: Initially, each object is considered to be its own cluster. According to a particular procedure, the clusters are then merged step by step until a single cluster remains. At the end of the cluster merging process, a cluster containing all the elements will be formed.

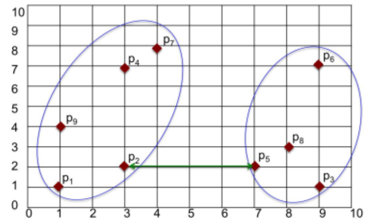
- **Divisive:** The Divisive method is the opposite of the Agglomerative method. Initially, all objects are considered in a single cluster. Then the division process is performed step by step until each object forms a different cluster. The cluster division or splitting procedure is carried out according to some principles that focus on maximum distance between neighboring objects in the cluster.

Hierarchical Clustering: Agglomerative

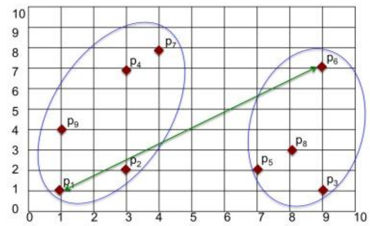
for $i = 1, \ldots, n$ let $C_i = \{\mathbf{x}(i)\}$;
while there is more than one cluster left **do**
  let C_i and C_j be the clusters
    minimizing the distance $\mathcal{D}(C_k, C_h)$ between any two clusters;
  $C_i = C_i \cup C_j$;
  remove cluster C_j;
end;

- Computing complexity $O(n^2)$

Computing the distance between clusters:
Linkage

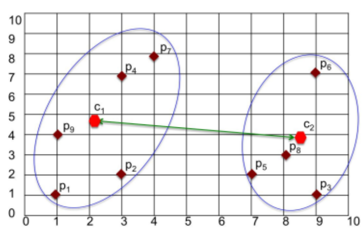


Single or Min Linkage:
Measure the distance between clusters by finding the minimum distance between points in those clusters.

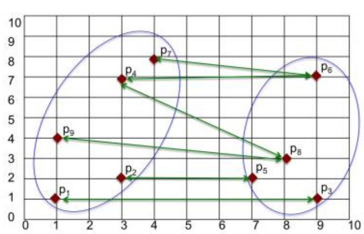


Complete or Max Linkage:
measure the distance between clusters by finding the maximum distance between points in two clusters

Computing the distance between clusters:
Linkage

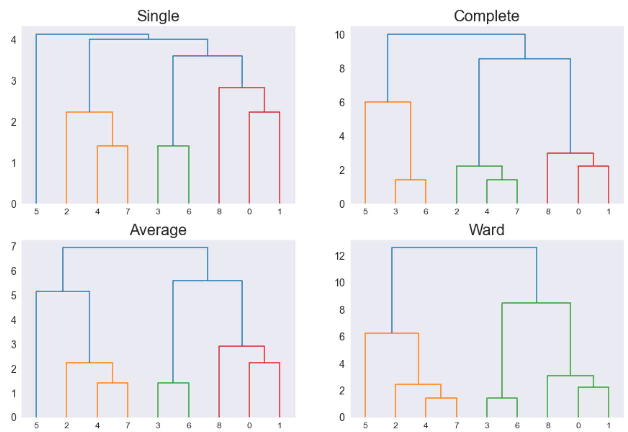


Centroid Linkage:
Measure the distance between clusters by measuring the distance between their centers/centroids.



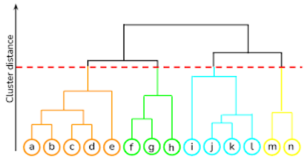
Average Linkage:
measure the distance between clusters as the average pairwise distance among all pairs of points in the clusters

Interpreting the results: Dendrograms



Interpreting Dendrograms

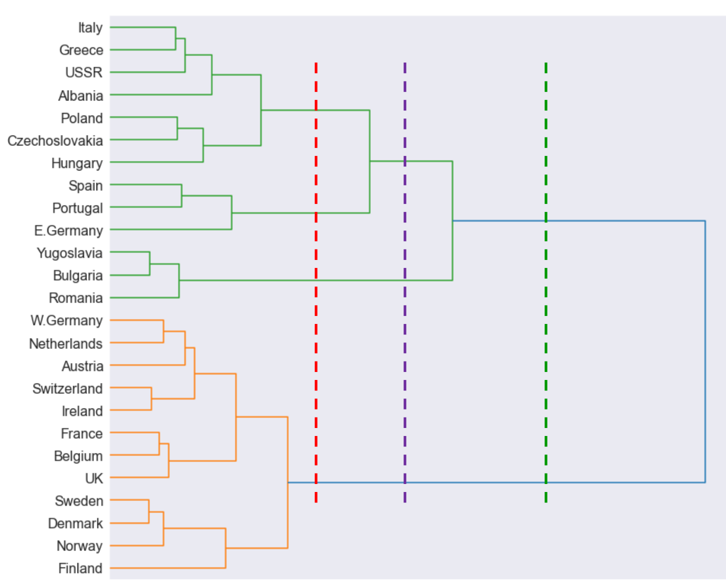
- The larger the length of the vertical lines in the dendrogram, more the distance between those clusters.
- The number of clusters will be the number of vertical lines intersected by the line drawn using the threshold.
- Horizontal line length is not necessarily the actual distance between them!



Protein Sources in Europe
From Biostatistics with R

	Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr.Veg
0	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
1	Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
2	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
3	Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
4	Czechoslovakia	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0

Real Data Set Results ???



Identifying the Number of Clusters

- No good way to identify true cluster structure!!!
  - Many methods proposed
- For many applications – look at the data
  - Are there reasonable explanations for the cluster?
  - Issue can be human bias brought to the data set.

Review Supervised Learning

- Training set of examples of input output (N)
  - $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$,
  - $y = f(x)$
- Function "h" is hypothesis about the world, approximates the true function f
  - drawn from a hypothesis space H of possible functions
  - h Model of the data, drawn from a model class H
- Consistent hypothesis: an h such that each x_i in the training set has $h(x_i) = y_i$.
- look for a best-fit function for which each $h(x_i)$ is close to y_i
- The true measure of a hypothesis, depends on how well it handles inputs it has not yet seen. E.g.: a second sample of (x_i, y_i)
- h generalizes well if it accurately predicts the outputs of the test set

Model Selection and Optimization

- Task of finding a good hypothesis as two subtasks:
  - Model selection: model selection chooses a good hypothesis space
  - Optimization (training) finds the best hypothesis within that space.
- A training set to create the hypothesis, and a test set to evaluate it.
- Error rate: the proportion of times that $h(x) \neq y$ for an (x, y)
- Ideally three data sets are needed:
  - A training set to train candidate models.
  - A validation set, also known as a development set or dev set, to evaluate the candidate models and choose the best one.
  - A test set to do a final unbiased evaluation of the best model.
- When insufficient amount of data to create three sets: k-fold cross-validation – see text

End