Data Analysis 3

Feature Selection, Representative-based Clustering, Hierarchical Clustering

Jan Platoš

September 22, 2019

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Unsupervised learning

Unsupervised learning

Unsupervised machine learning is the machine learning task of inferring a function that describes the structure of ”unlabeled” data (i.e. data that has not been classified or categorized).

Clustering

Clustering

Given a set of data points, partition them into groups containing very similar points.

- Possible applications:
  - Data summarization
  - Customer Segmentation
  - Social network analysis
  - Preprocessing data for other algorithms (classification, outliers detection, etc.)

Feature Selection for Clustering

- How to select only the features that are important?
- How to measure the ability of the feature to cluster objects?
- Two main approaches exists:
  - Filter models
  - Wrapper models

Feature Selection for Clustering - Filter models

Filter models

- Definition of measures that may evaluate the quality of a feature or a feature combination

Term Strength

- Suitable for Text data or other sparse documents.
- A conditional probability that a selected term appear in a document Y when it appear in a document X.
- The documents pairs are randomly sampled from similar documents.

$$Term\ strength = P(t \in Y | t \in X)$$

Feature Selection for Clustering - Filter models

Predictive Attribute Dependence

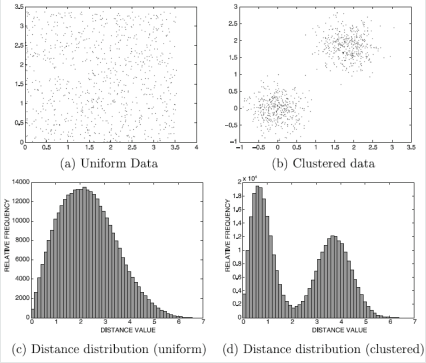
- Correlated features should end in better results than uncorrelated.
- Set of features should predict the value of another correlated feature.
- A regression or classification algorithm is used as a predictor.

Predictive Attribute Dependence - Principle

- For each feature:
  - Use other features to predict the value of the selected feature.
  - Compute the accuracy of prediction and use it as relevance measure.
- All relevant features are used for clustering.

Feature Selection for Clustering - Filter models

Entropy



- Highly clustered data reflects some characteristics in the underlying distance distribution.
- The goal is to quantify the shape of the distance distribution on a given subset of features and to pick the most suitable subset.
- A systematic way to search possible combination of features.
- A algorithm for entropy measure have to be defined (greedy, ...).

Feature Selection for Clustering - Filter models

Entropy

- For k-features a k-dimensional space is defined.
- Define m multidimensional grid regions over this space.
- The p_i is a fraction of points in a region i.

$$E = \sum_{i=1}^{m} p_i \log(p_i)$$

- Distribution with poor clustering behavior results in high entropy.
- Alternatively an entropy over a distance distribution may be computed.

Feature Selection for Clustering - Filter models

Hopkins Statistics

- Evaluates the clustering tendency of the whole datasets/selected features (greedy approach may be used).
- D is a dataset, R is a random sample of r points from D, sample S of r randomly generated points.
- The $a_1, a_2, \ldots, a_r$ are distances from points from R to the nearest neighbors from D and $b_1, \ldots, b_r$ are distances from points from S to the nearest neighbors from D.

$$H = \frac{\sum_{i=1}^{r} b_i}{\sum_{i=1}^{r} (a_i + b_i)}$$

- H is in the range $(0, 1)$.
- Uniformly distributed data will have $H = 0.5$, H close to 1 means clustered data.
- Due to random sampling, each tests will produce different results.
- Confidence test have to be used for statistical importance.

Feature Selection for Clustering - Wrapper models

Wrapper models

- Uses a cluster validity criterion for feature subset evaluation.
- The are highly imperfect.
- Search space of features is exponentially related to the dimensionality.
- The principle is sensitive to the choice of the validity criterion.
- Simpler methodology:
  - Cluster points according the selected feature subset.
  - Assign labels to the points according to the cluster the points belongs to.
  - Use supervised criterion to measure the quality of each feature.

Representative-based Algorithms

- Simplest clustering algorithm.
- Based on the distance/similarity measure between points.
- Clusters are created in a single step.
- Hierarchical relationships do not exist among different clusters.
- The representatives are computed or selected from the cluster.
- A distance function is used for defined representatives to assign points into closest representative/cluster.

Representative-based Algorithms

- A number of clusters k is usually defined by the user.
- A dataset D contains n data points $X_1, \ldots, X_n$.
- The goal is to determine k representatives $Y_1, \ldots, Y_k$ that minimizes function O:

$$O = \sum_{i=1}^{n} \left[min_j Dist(X_i, Y_i) \right]$$

- i.e. the sum of the distances of the different data points to their closest representatives needs to be minimized.

Representative-based Algorithms

- The position of representatives and points assignment is not known a priori.
- An Iterative approach is used to solve the problem of representative/point assignment.
- General approach for representative-based algorithms:
  1. Initialize k representatives (random sampling or other method)
  2. Assign each data point to its closest representative using distance function $Dist(\cdot, \cdot)$
  3. Form a clusters from points assigned to each of the representative.
  4. Determine the optimal representative for each cluster C_j using minimization of a local objective function $\sum_{X_i \in C_j} [Dist(X_i, Y_j)]$

Representative-based Algorithms

Algorithm 1: GenericRepresentative(Database: D, Number of Rpresentatives: k)

1 **begin**
2 Initialize representative set S;
3 **repeat**
4 Create clusters $(C_1, \ldots, C_k)$ by assigning each point in D to closest representative in S using the distance function $Dist(\cdot, \cdot)$;
5 Recreate set S by determining one representative Y_j for each C_j that minimizes $\sum_{X_i \in C_j} Dist(X_i, Y_j)$;
6 **until** *convergence*;
7 **return** $(C_1, \ldots, C_k)$
8 **end**

Definition of the distance function

- The distance function $Dist(X, Y)$ defines the behavior of the algorithm.
- The general definition is a L_p-norm.

$$Dist(X, Y) = \|X - Y\|_p^p = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

- Another possibility is a cosine measure.

$$\cos(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^{d} (x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}}$$

Definition of the distance function

- Mahalanobis distance is a measure that takes into account a statistical distribution along each dimension

$$Maha(X, Y) = \sqrt{(X - Y)\Sigma^{-1}(X - Y)^T}$$

- Where Σ is a covariance matrix where Σ_{ij} is a variance between i-th and j-th dimension.

$$\Sigma = \begin{bmatrix} E\left[(X_1 - \mu_1)(X_1 - \mu_1)\right] & E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] & \cdots & E\left[(X_1 - \mu_1)(X_d - \mu_d)\right] \\ \vdots & & \ddots & \vdots \\ E\left[(X_d - \mu_d)(X_1 - \mu_1)\right] & E\left[(X_d - \mu_d)(X_2 - \mu_2)\right] & \cdots & E\left[(X_d - \mu_d)(X_d - \mu_d)\right] \end{bmatrix}$$

- μ_i is expected value of the dimension i.

Representative-based Algorithms - K-medians Algorithm

K-medians algorithm

- Uses an Manhattan distance for measuring distance between data points and representatives.

$$Dist(X, Y) = \|X - Y\|_1^1 = \sum_{i=1}^{d} |x_i - y_i|$$

- The optimal representatives for each cluster is a **median along each dimension in a cluster**.
- The k-medians algorithm is more robust than k-means: median is not as sensitive to outliers as the mean.

Representative-based Algorithms - K-means Algorithm

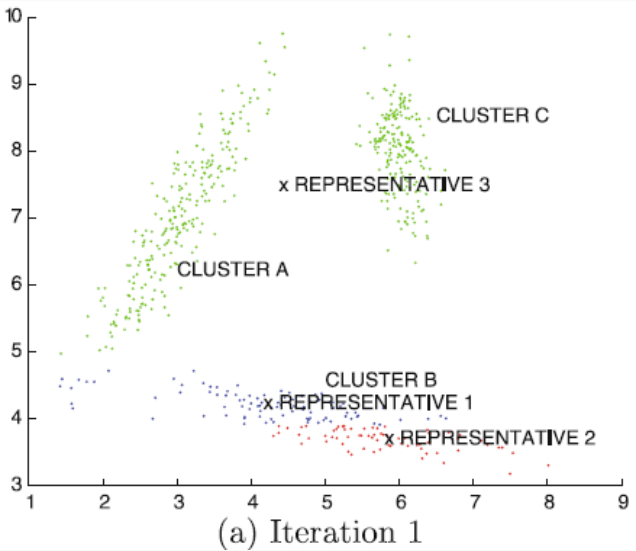
K-means algorithm

- Uses an Euclidean distance for measuring distance between data points and representatives.

$$Dist(X, Y) = \|X - Y\|_2^2 = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

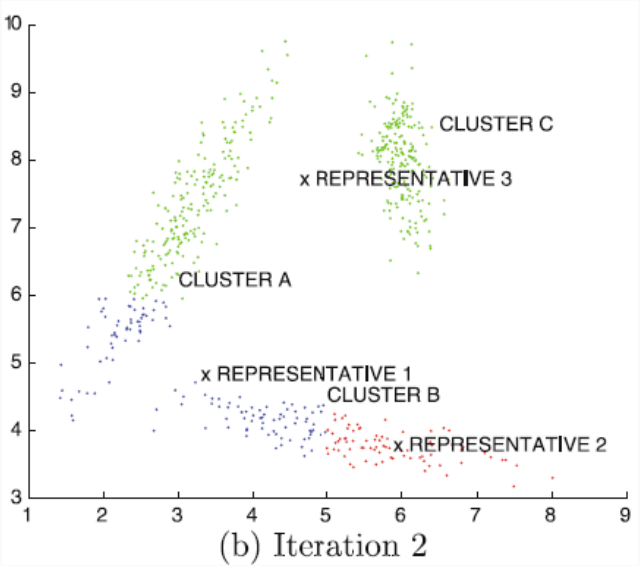
- The optimal representatives for each cluster is a mean.
- The k-means does not work well when the clusters are not spherical.
- A kernel variant of this algorithm is possible.

Representative-based Algorithms - K-means Algorithm



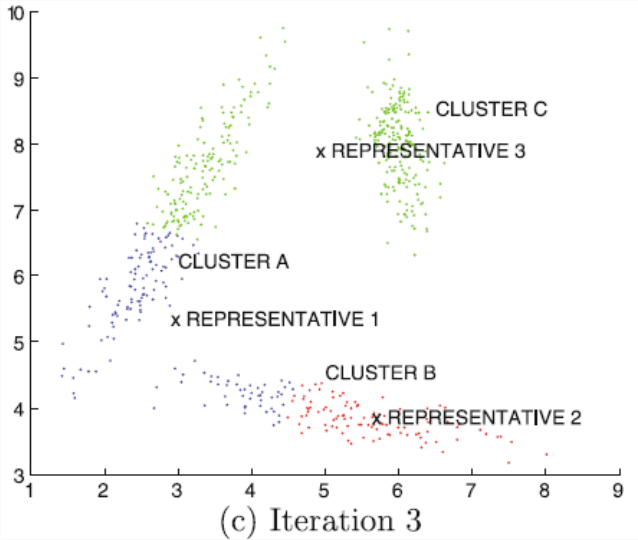
(a) Iteration 1

Representative-based Algorithms - K-means Algorithm



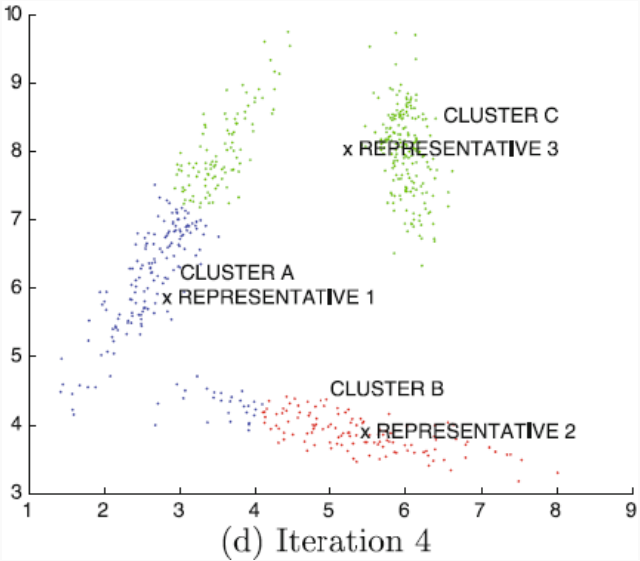
(b) Iteration 2

Representative-based Algorithms - K-means Algorithm



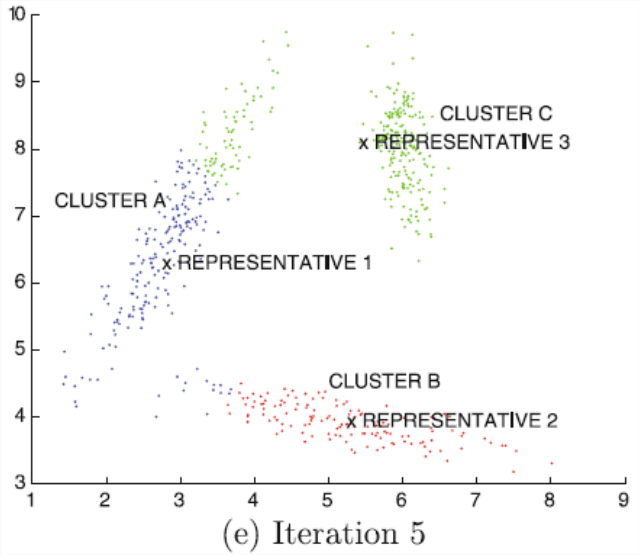
(c) Iteration 3

Representative-based Algorithms - K-means Algorithm



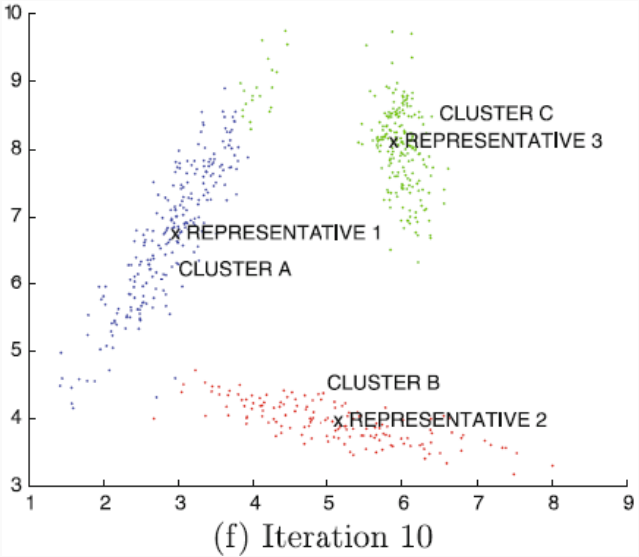
(d) Iteration 4

Representative-based Algorithms - K-means Algorithm



(e) Iteration 5

Representative-based Algorithms - K-means Algorithm



(f) Iteration 10

Representative-based Algorithms - K-medoids Algorithm

K-medoids algorithm

- The representatives are always selected from the dataset.
- Why:
  - For means: the representatives may lie outside the cluster in an empty region due to presence of outliers.
  - Sometimes it is difficult to compute the optimal central representative of a set of data points of a complex data type, e.g. set of time series of varying lengths.
- The *k*-medoid approach may be defined for any data type as long as a distance/similarity function is known.

Representative-based Algorithms - K-medoids Algorithm

K-medoids algorithm

- A set of representatives Y is randomly selected from dataset.
- Iterative improvement is performed on the set of representatives:
  - Exhaustive search (extremely expensive).
  - A set of pairs (X, Y) is randomly generated, X from dataset and Y from representatives, and the best pair is used for exchange.
- The k-medoids is slower than k-means.
- May be scalable implemented.

Representative-based Algorithms - Implementation Issues

Selection of k

- Very difficult automatically.
- Usually a large value is chosen.
- A post processing step may reduce number of clusters.

Initialization

- Algorithms are robust to the choice of the initialization step.
- Random generation of the representatives.
- Random sampling of the dataset.
- A centroids of m randomly chosen samples.

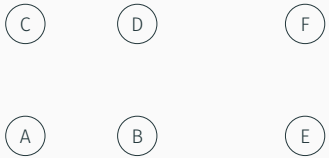
Hierarchical Clustering

- Creates a hierarchical structure above the objects from the dataset.
- The different levels of clustering granularity provide different application-specific insights to the data.
- Hierarchical organization of the data allows even better flat cluster

Hierarchical Clustering - Algorithm types

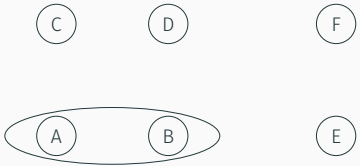
- Bottom-up (agglomerative) methods
  - Individual data objects are agglomerated into higher level clusters.
  - Objective function is used for computing similarity.
- Top-down (divisive) methods
  - Partitioning of the data objects into tree-like structure.
  - A flat clustering algorithm may be used for the partitioning in a given step.
  - A trade-off in balance of the tree between number of clusters and the number of objects in each cluster/leaf.

Hierarchical Clustering - Bottom-Up Agglomerative Methods

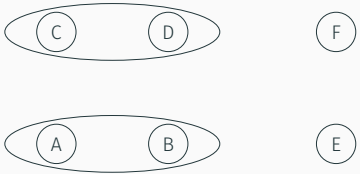


A B C D E F

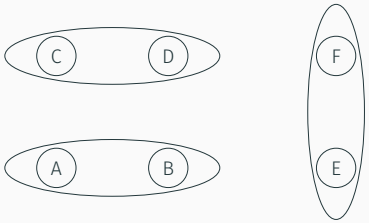
Hierarchical Clustering - Bottom-Up Agglomerative Methods



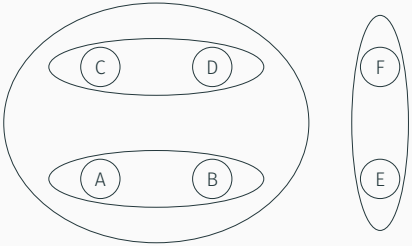
Hierarchical Clustering - Bottom-Up Agglomerative Methods

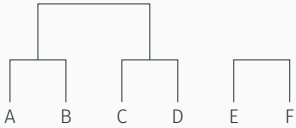


Hierarchical Clustering - Bottom-Up Agglomerative Methods

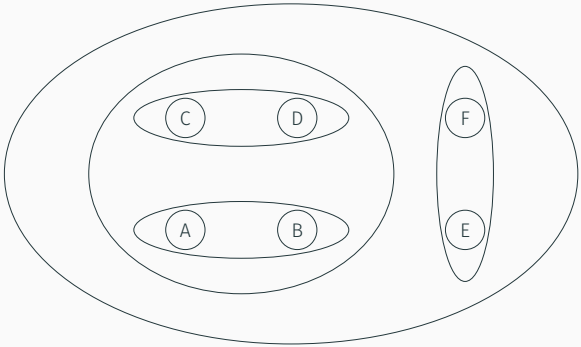


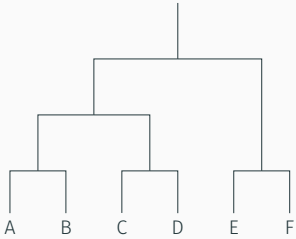
Hierarchical Clustering - Bottom-Up Agglomerative Methods





Hierarchical Clustering - Bottom-Up Agglomerative Methods





Hierarchical Clustering - Bottom-Up Agglomerative Methods

- Iterative approach starting with individual data object.
- Two clusters are merged in each iteration.
- Each merging step reduces the number of clusters by one.
- A carefully selected measure for computation of the distance between individual objects need to be defined.
- A proper strategy for measuring the distance between clusters need to be defined also.
- A distance matrix should be stored in a memory, the computational complexity increases when not.

Hierarchical Clustering - Bottom-Up Agglomerative Methods

Algorithm 2: AgglomerativeMerge(Dataset: D)

```
1  begin
2      Initialize n × n distance matrix M using D;
3      repeat
4          Pick the closest pair of clusters i and j using M;
5          Merge clusters i and j;
6          Delete rows/columns i and j from M and create a new row and column for
             newly merged cluster;
7          Update the entries of the new row and column of M;
8      until termination criterion;
9      return current merged cluster set
10 end
```

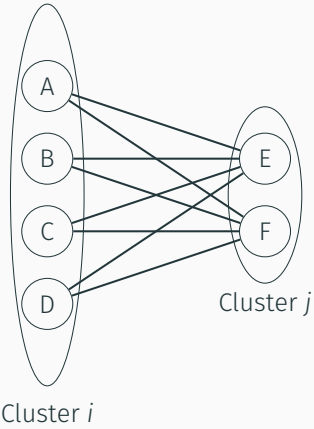
Hierarchical Clustering - Bottom-Up Agglomerative Methods

Group Similarity Computation

- Distance between two groups of objects need to be computed.
- The distance is a function of the distances between all pairs of objects from different cluster.

$$D(C_i, C_j) = func_{\forall x \in C_i, \forall y \in C_j}(d(x, y))$$

- Different strategies exists.
- Each criterion has different advantages and disadvantages.



- Bottom-Up Agglomerative Methods - Group Similarity Computation

- Best (single) linkage.
  - The distance is equal to the minimum distance between all pairs of objects.
  - Its corresponds to the closes pair of objects between the two groups.
  - Very efficient approach in discovering clusters of arbitrary shape.
  - Very sensitive to noise that connects different clusters.

$$D(C_i, C_j) = \min_{\forall x \in C_i, \forall y \in C_j} \{d(x, y)\}$$

- Worst (complete) linkage:
  - The distance is equal to the maximum distance between all pairs of objects.
  - Its corresponds to the farthest pair of objects between the two groups.
  - This criterion attempts to minimize the maximum diameter of a cluster.

$$D(C_i, C_j) = \max_{\forall x \in C_i, \forall y \in C_j} \{d(x, y)\}$$

- Bottom-Up Agglomerative Methods - Group Similarity Computation

- Group-average linkage
  - The distance is equal to the average distance between all pairs of objects.
  - A weighted average is used for computation.

$$D(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

- Closest centroid
  - The closest centroid are merged in each iteration.
  - The centroids lose information about the relative spreads of the clusters.
  - The method will not discriminate between clusters of varying sizes.
  - Typically larger clusters has statistically more likely centroid closer to each other than smaller clusters.

- Variance-based criterion
  - This criterion minimizes the cluster variance objective function during merging.
  - Merging always results into worsening of the clustering objective function due to loss of granularity.
  - Each cluster maintain 0^{th}, 1^{st} and 2^{nd} order moment statistics.
  - The average squared error of a cluster i is defined as:

$$SE_i = \sum_{r=1}^{d} \left(\frac{S_{ir}}{m_i} - \frac{F_{ir}^2}{m_i^2} \right)$$

  - m_i is the number of points, S_{ir} squared sum of the data points in a cluster across each direction r, F_{ir} is a sum of data points along each direction.

- Variance-based criterion
  - The moment statistics of a merge of two clusters is the sum of the clusters moment statistics.
  - The change of a variance on executing merge on clusters i and j is defined as:

$$\Delta SE_{i \cup j} = SE_{i \cup j} - SE_i - SE_j$$

  - This change is always positive.
  - The pair of clusters with the smallest variance change is selected for merge.

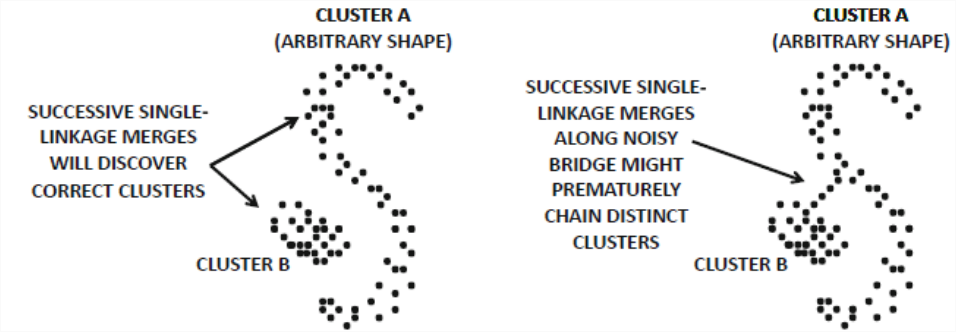
- Ward's method
  - Slightly different approach than Variance-based criterion.
  - It is unscaled sum of squared error as the merging criterion.

$$SE = \sum_{r=1}^{d} \left(m_i S_{ir} - F_{ir}^2 \right)$$

  - It's a variant of the centroid method.
  - It's the equivalent of the multiplication of the squared Euclidean distance between centroid and the harmonic mean of the number of points in each pair.
  - Larger clusters are penalized.

- Bottom-Up Agglomerative Methods - Group Similarity Computation

- Single linkage method is able to discover clusters of arbitrary shape, but it may merge clusters connected by noise points.



- Bottom-Up Agglomerative Methods - Group Similarity Computation

- Complete linkage method
  - Focus on the minimization of the maximum distance between pair of points.
  - This may be viewed as the approximation of the cluster diameter.
  - The method tries to create clusters with the similar diameter.
  - Larger natural cluster may be broken into small clusters.
  - The created clusters tends to be spherical.
- Group-average, variance and Ward's method are robust to the noise.

- Bottom-Up Agglomerative Methods - Implementation notes

- A heap of sorted distances need to be maintained for efficient minimal distance determination.
- Initial matrix computation requires $O(n^2 \cdot d)$ time.
- The maintenance of the sort heap requires $O(n^2 \cdot \log n)$. The required space for distance matrix is $O(n^2)$.
- When a distance matrix didn't fit into memory the computation time is $O(n^3 \cdot d)$.

- Bottom-Up Agglomerative Methods - Implementation notes

- The results is a binary tree of clusters.
- It is very difficult to control the structure of the tree.
- Difficult to use when the specific structure is desired.
- The method is sensitive to a small number of mistakes during merging process, therefore, its very sensitive to the noise.
- These methods are impractical for large dataset.
- Frequently combined with sampling and partitioning methods.

Hierarchical Clustering - Top-Down Divisive Methods

- Uses a general-purpose flat-clustering algorithms as a subroutine.
- The algorithm initializes the tree at the root with all points.
- The particular node is split into multiple nodes in each iteration.
- The strategy for selection of the node to split may affect the balanced tree by height or number of clusters.
- When the clustering subroutine is stochastic, several trials are tested the best is selected.

Hierarchical Clustering - Top-Down Divisive Methods

Algorithm 3: GenericTopDownClustering(Dataset: D, Flat Algorithm: A)

```
1 begin
2     Initialize tree T to root containing D;
3     repeat
4         Select a leaf node L in T based on pre-defined criterion;
5         Use algorithm A to split L into L_1, ..., L_k;
6         Add L_1, ..., L_k as children of L in T;
7     until termination criterion;
8     return tree T
9 end
```

Hierarchical Clustering - Top-Down Divisive Methods - Example

- Bisecting k-Means
  - Each node in split into two children with 2-means algorithm.
  - Several runs of randomized initialization are used.
  - The one with the best impact on the overall clustering objective is used.
  - Several strategies for node selection exists.

Questions?