Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2015 : Lecture 19

Avinash Sharma

CVIT, IIIT Hyderabad

Lecture Plan

- Data Clustering
  - Introduction
  - Similarity Measures
  - Criterion Functions for Clustering
- Hierarchical Clustering
  - Agglomerative Clustering
- Kmeans Clustering (EM) & Variants (Next Class)

Introduction to Data Clustering

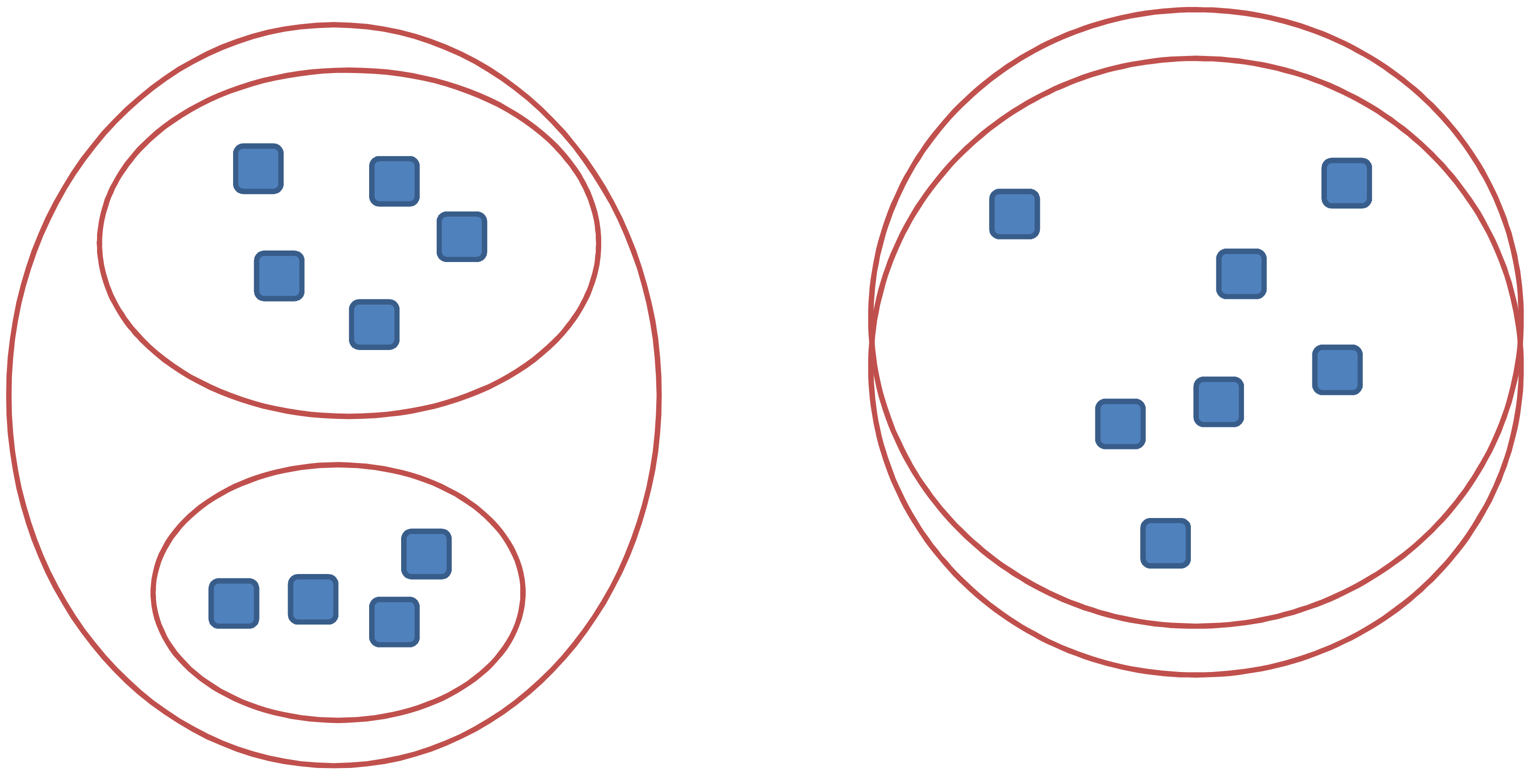
- Given a set of points, with a notion of distance between points, group the points into some number of clusters, so that
  - Members of a cluster are close/similar to each other.
  - Members of different clusters are dissimilar.

- Clustering is generally an ***unsupervised learning*** task as it attempts to recover the natural grouping of the data.

- Typically:
  - Points are sampled in a high dimensional space.
  - Generative Model assumption (with clusters having identical model parameters) rarely holds.

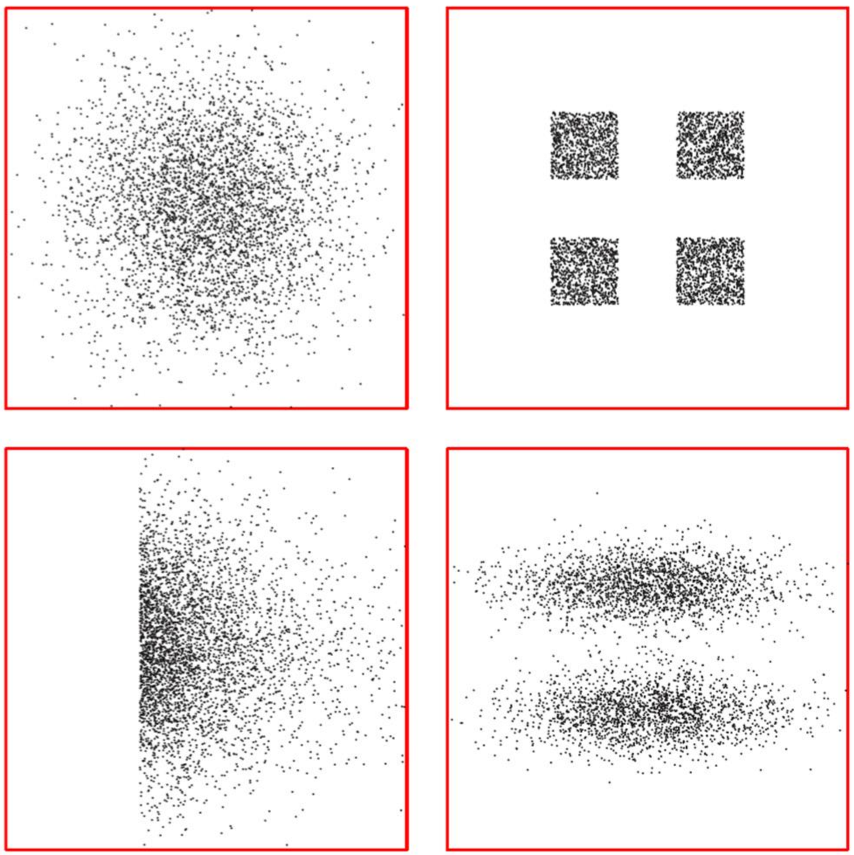
Introduction to Data Clustering

- How do we know what is the best clustering solution?



Introduction to Data Clustering

- Generative Model assumption (with clusters having identical model parameters) rarely holds !



Similarity Measures

- Vectors: Cosine distance.

$$s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^t \mathbf{x}'}{\|\mathbf{x}\| \; \|\mathbf{x}'\|}$$

- Sets: Jaccard distance.

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

(If A and B are both empty, we define $J(A,B) = 1$.)

$$0 \leq J(A,B) \leq 1.$$

- Points: Minkowski distance
  - q=2: Euclidean distance
  - q=1: City-block distance

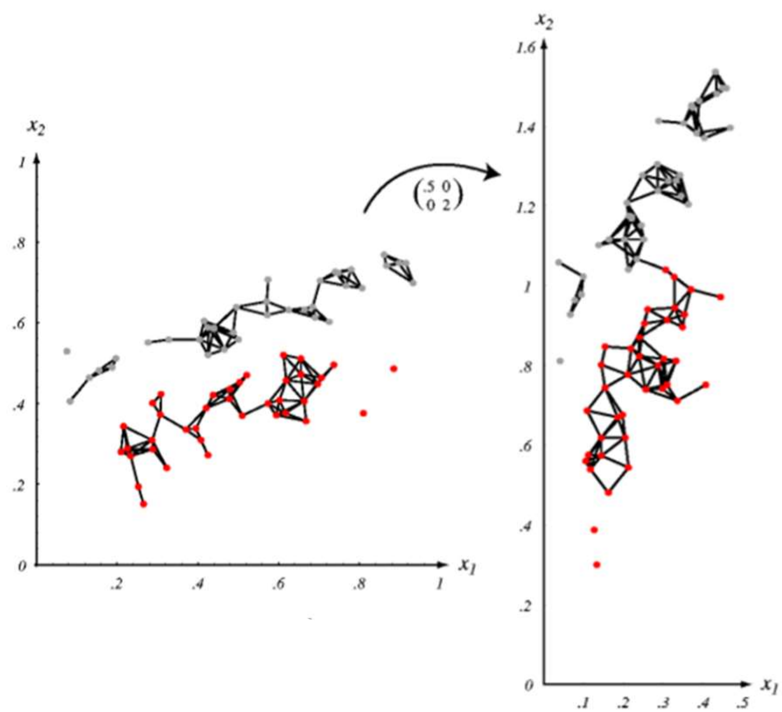
$$d(\mathbf{x}, \mathbf{x}') = \left(\sum_{k=1}^{d} |x_k - x'_k|^q \right)^{1/q},$$

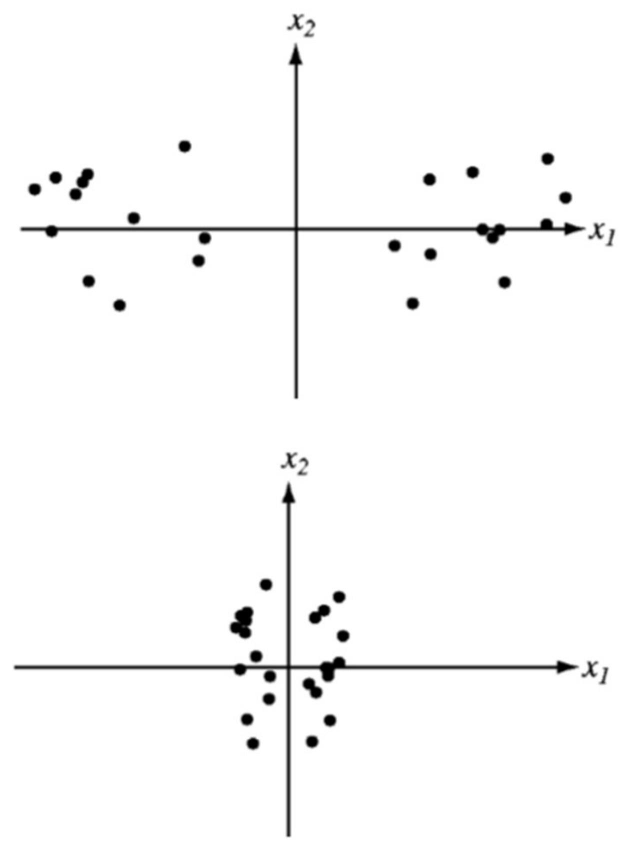
- Points: Mahalanobis metric
  - Data dependent

$$d(\mathbf{x}, \mathbf{y})^2 = (\mathbf{x} - \mathbf{y})^T S^{-1} (\mathbf{x} - \mathbf{y})$$

Similarity Measures

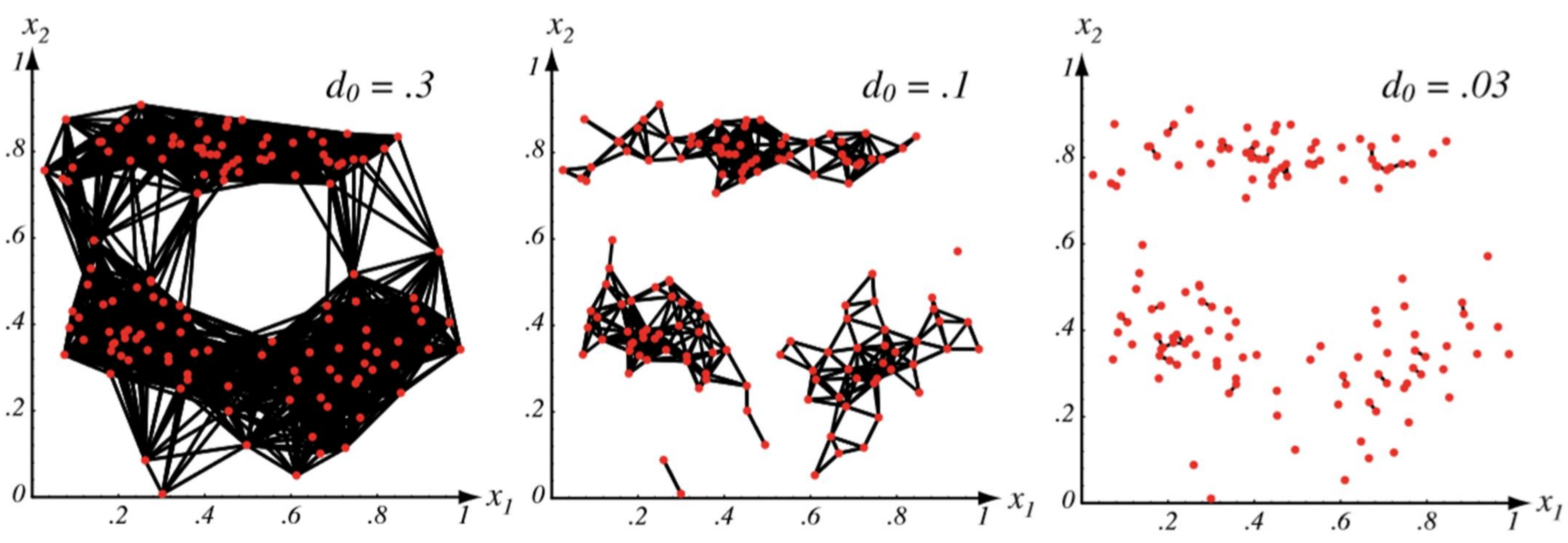
- Should we always normalize the data?
  - Not advisable when data is has clusters that are drawn from multiple distributions.





Similarity Measures

- Three clustering solutions with different parameter choices (distance thresholds).

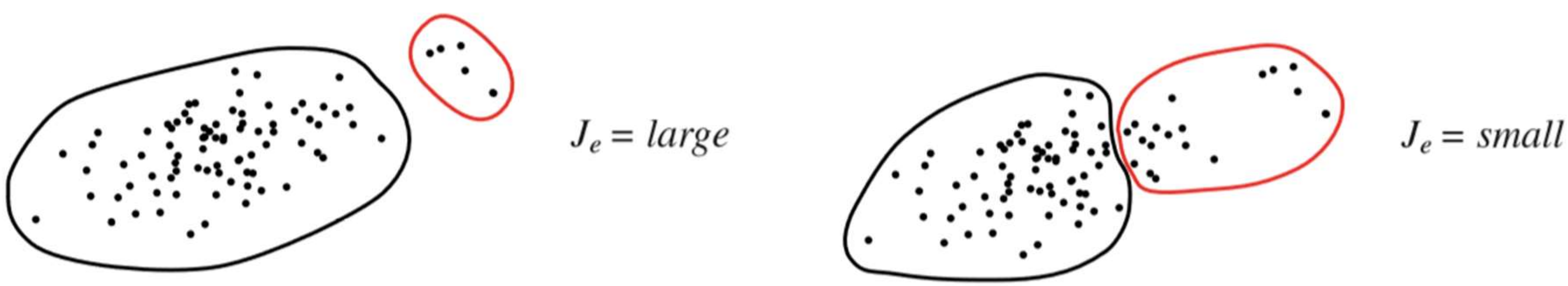


Criterion Functions for Clustering

- The Sum-of-Squared-Error Criterion:
  - Achieves minimum variance clustering

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in \mathcal{D}_i} \|\mathbf{x} - \mathbf{m}_i\|^2. \qquad \mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}.$$

  - Not always best criterion



Criterion Functions for Clustering

- Related Minimum Variance Criterion:

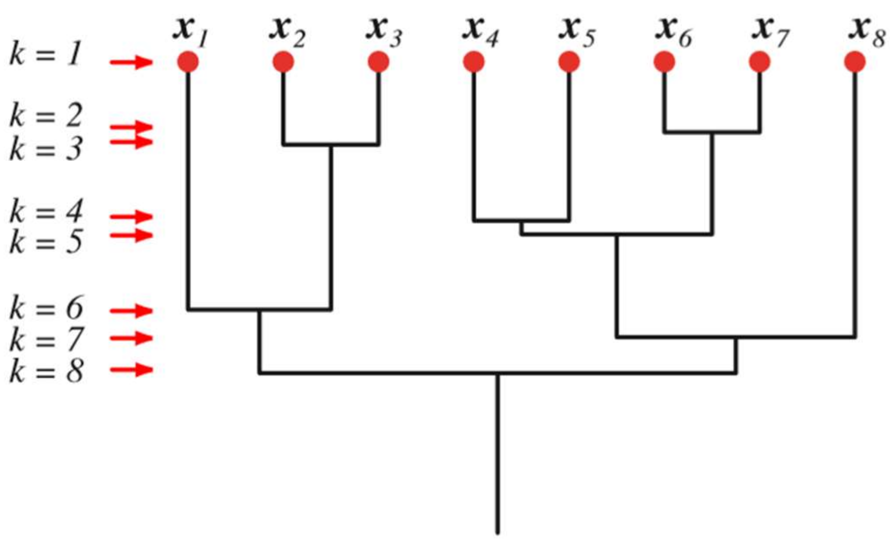
$$J_e = \frac{1}{2}\sum_{i=1}^{c} n_i \bar{s}_i, \qquad \bar{s}_i = \frac{1}{n^2}\sum_{\mathbf{x}\in\mathcal{D}_i}\sum_{\mathbf{x}'\in\mathcal{D}_i}\|\mathbf{x}-\mathbf{x}'\|^2.$$

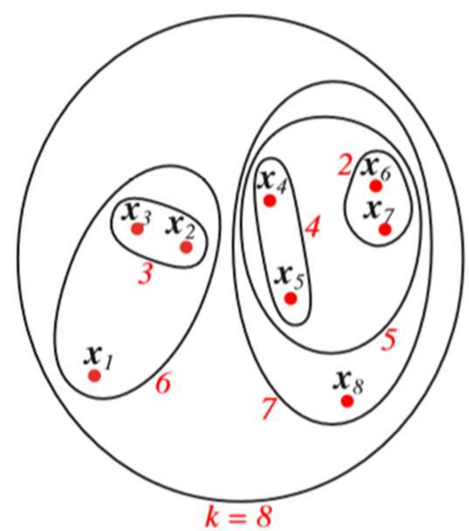
- Or, in a generalized manner:

$$\bar{s}_i = \frac{1}{n_i^2}\sum_{\mathbf{x}\in\mathcal{D}_i}\sum_{\mathbf{x}'\in\mathcal{D}_i} s(\mathbf{x},\mathbf{x}') \qquad \bar{s}_i = \min_{\mathbf{x},\mathbf{x}'\in\mathcal{D}_i} s(\mathbf{x},\mathbf{x}').$$

Hierarchical Clustering

- Combining two points/clusters at a time based on nearness of points/clusters until a fix number of clusters are remained as long as
  – *any two points put into a single cluster remains in the same cluster all the way till final solution.*





Agglomerative Clustering

- Agglomerative clustering is a bottom-up procedure that combines nearest cluster in each iteration until desired number of clusters are obtained.

Algorithm 4 (Agglomerative hierarchical clustering)

begin initialize $c, \hat{c} \leftarrow n, \mathcal{D}_i \leftarrow \{\mathbf{x}_i\}, i = 1, \ldots, n$
        do $\hat{c} \leftarrow \hat{c} - 1$
            Find nearest clusters, say, $\mathcal{D}_i$ and $\mathcal{D}_j$
            Merge $\mathcal{D}_i$ and $\mathcal{D}_j$
        until $c = \hat{c}$
    return c clusters
end

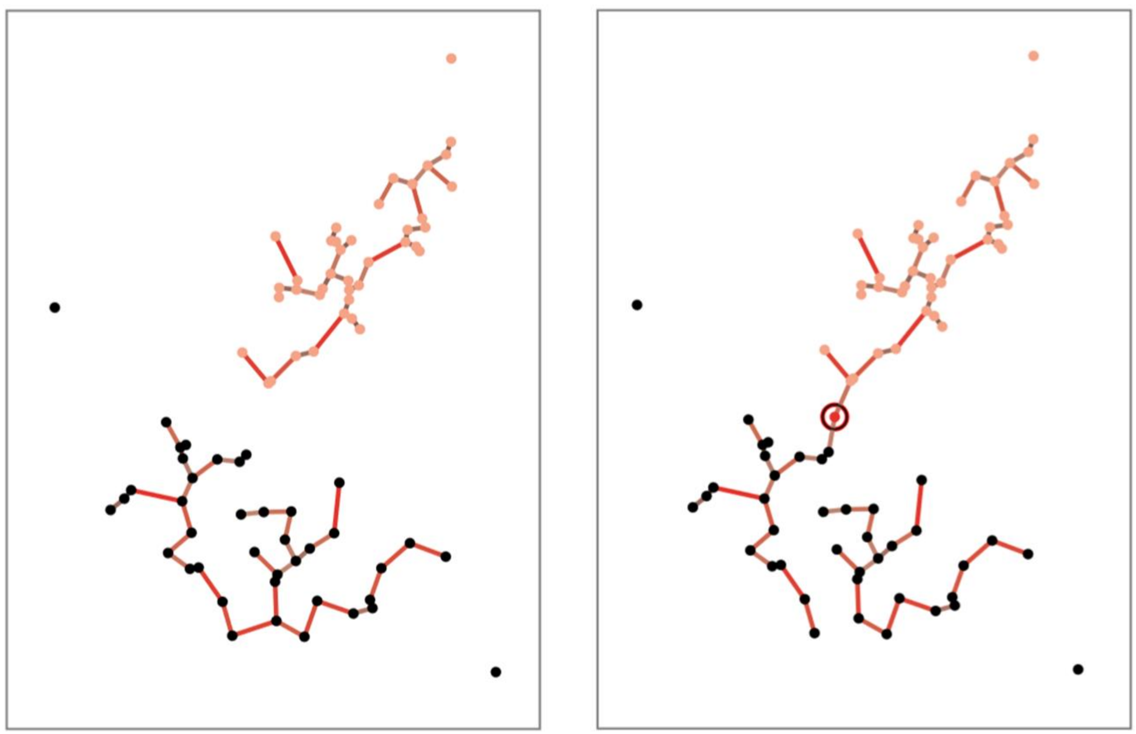
Agglomerative Clustering

- The measures of distance between clusters:

$$
\begin{aligned}
d_{min}(\mathcal{D}_i, \mathcal{D}_j) &= \min_{\substack{\mathbf{x}\in\mathcal{D}_i \\ \mathbf{x}'\in\mathcal{D}_j}} \|\mathbf{x} - \mathbf{x}'\| \\
d_{max}(\mathcal{D}_i, \mathcal{D}_j) &= \max_{\substack{\mathbf{x}\in\mathcal{D}_i \\ \mathbf{x}'\in\mathcal{D}_j}} \|\mathbf{x} - \mathbf{x}'\| \\
d_{avg}(\mathcal{D}_i, \mathcal{D}_j) &= \frac{1}{n_i n_j} \sum_{\mathbf{x}\in\mathcal{D}_i} \sum_{\mathbf{x}'\in\mathcal{D}_j} \|\mathbf{x} - \mathbf{x}'\| \\
d_{mean}(\mathcal{D}_i, \mathcal{D}_j) &= \|\mathbf{m}_i - \mathbf{m}_j\|.
\end{aligned}
$$

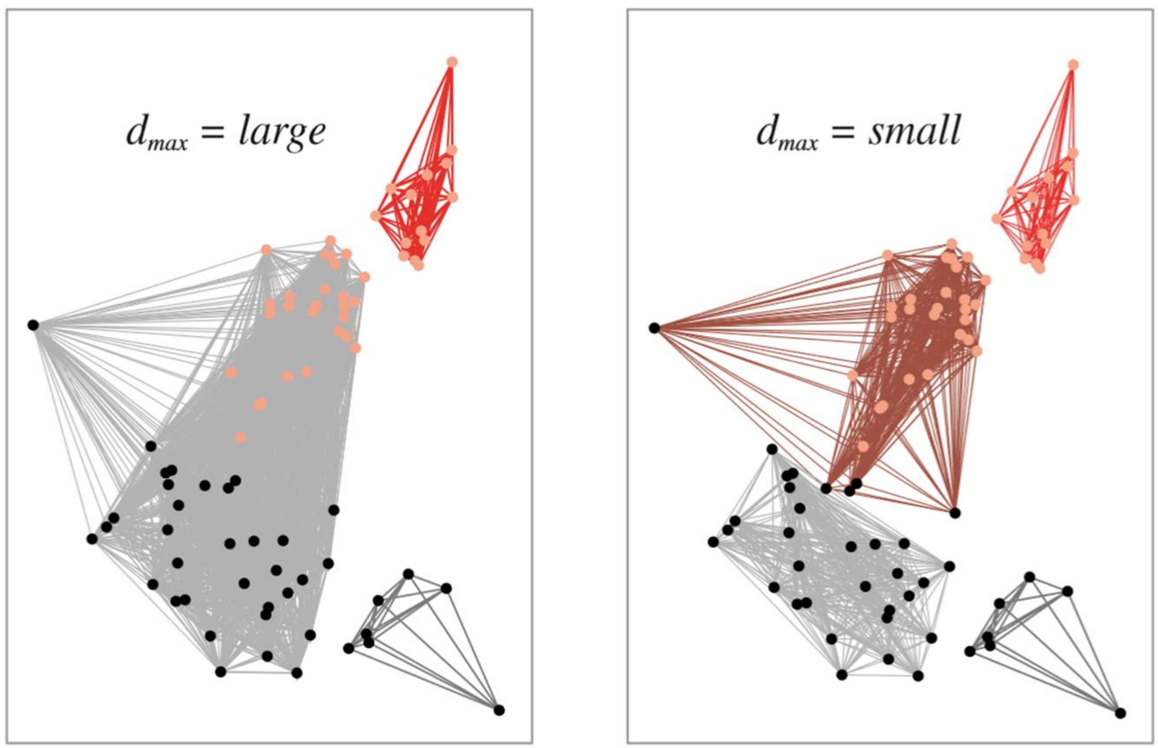
Agglomerative Clustering

- Nearest Neighbor strategy, also known as minimum algorithm or single-linkage algorithm, yields a minimum spanning tree solution.



Agglomerative Clustering

- Farthest neighbor clustering algorithm, also known as maximum algorithm or complete-linkage algorithm.



Agglomerative Clustering

- Compromises
  - Mean based distance is the simplest in terms of computational complexity
  - Average distance based algorithm is usable when distances are replaced with similarity measures.