

Unsupervised Learning and Clustering

Selim Aksoy

Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr

CS 551, Spring 2009



Introduction

- ▶ Until now we have assumed that the training examples were labeled by their class membership.
- ▶ Procedures that use labeled samples are said to be *supervised*.
- ▶ In this chapter, we will study *clustering* as an *unsupervised* procedure that uses unlabeled samples.



Introduction

- ▶ Unsupervised procedures are used for several reasons:
 - ▶ Collecting and labeling a large set of sample patterns can be costly or may not be feasible.
 - ▶ One can train with large amount of unlabeled data, and then use supervision to label the groupings found.
 - ▶ Unsupervised methods can be used for feature extraction.
 - ▶ Exploratory data analysis can provide insight into the nature or structure of the data.



Data Description

- ▶ Assume that we have a set of unlabeled multi-dimensional patterns.
- ▶ One way of describing this set of patterns is to compute their sample mean and covariance.
- ▶ This description uses the assumption that the patterns form a cloud that can be modeled with a hyperellipsoidal shape.
- ▶ However, we must be careful about any assumptions we make about the structure of the data.



Data Description

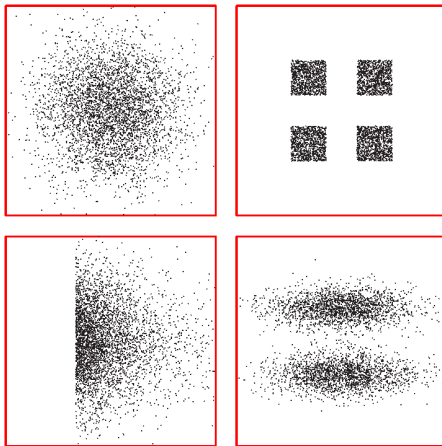


Figure 1: These four data sets have identical first-order and second-order statistics. We need to find other ways of modeling their structure. Clustering is an alternative way of describing the data in terms of groups of patterns.

Clusters

- ▶ A *cluster* is comprised of a number of similar objects collected or grouped together.
- ▶ Other definitions of clusters (from Jain and Dubes, 1988):
 - ▶ A cluster is a set of entities which are alike, and entities from different clusters are not alike.
 - ▶ A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it.
 - ▶ Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points.



Clustering

- ▶ *Cluster analysis* organizes data by abstracting the underlying structure either as a grouping of individuals or as a hierarchy of groups.
- ▶ These groupings are based on measured or perceived similarities among the patterns.
- ▶ Clustering is unsupervised. Category labels and other information about the source of data influence the interpretation of the clusters, not their formation.



Clustering

- Clustering is a very difficult problem because data can reveal clusters with different shapes and sizes.



Figure 2: The number of clusters in the data often depend on the resolution (fine vs. coarse) with which we view the data. How many clusters do you see in this figure? 5, 8, 10, more?

Clustering

- ▶ Clustering algorithms can be divided into several groups:
 - ▶ *Exclusive* (each pattern belongs to only one cluster) vs. *nonexclusive* (each pattern can be assigned to several clusters).
 - ▶ *Hierarchical* (nested sequence of partitions) vs. *partitional* (a single partition).



Clustering

- ▶ Implementations of clustering algorithms can also be grouped:
 - ▶ *Agglomerative* (merging atomic clusters into larger clusters) vs. *divisive* (subdividing large clusters into smaller ones).
 - ▶ *Serial* (processing patterns one by one) vs. *simultaneous* (processing all patterns at once).
 - ▶ *Graph-theoretic* (based on connectedness) vs. *algebraic* (based on error criteria).



Clustering

- ▶ Hundreds of clustering algorithms have been proposed in the literature.
- ▶ Most of these algorithms are based on the following two popular techniques:
 - ▶ Iterative squared-error partitioning,
 - ▶ Agglomerative hierarchical clustering.
- ▶ One of the main challenges is to select an appropriate measure of similarity to define clusters that is often both data (cluster shape) and context dependent.



Similarity Measures

- ▶ The most obvious measure of *similarity* (or dissimilarity) between two patterns is the distance between them.
- ▶ If distance is a good measure of dissimilarity, then we can expect the distance between patterns in the same cluster to be significantly less than the distance between patterns in different clusters.
- ▶ Then, a very simple way of doing clustering would be to choose a threshold on distance and group the patterns that are closer than this threshold.



Similarity Measures

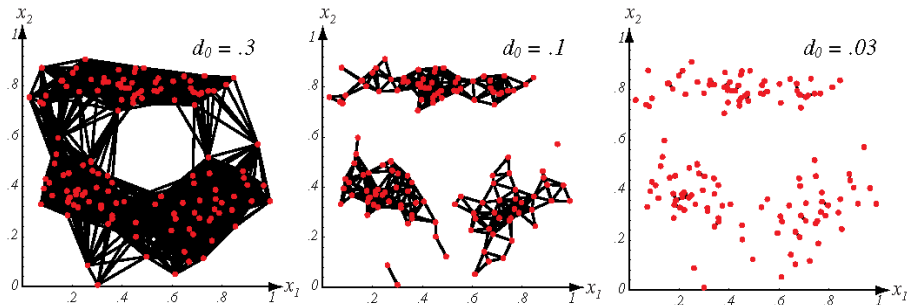


Figure 3: The distance threshold affects the number and size of clusters that are shown by lines drawn between points closer than the threshold.

Criterion Functions

- ▶ The next challenge after selecting the similarity measure is the choice of the criterion function to be optimized.
- ▶ Suppose that we have a set $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of n samples that we want to partition into exactly k disjoint subsets $\mathcal{D}_1, \dots, \mathcal{D}_k$.
- ▶ Each subset is to represent a cluster, with samples in the same cluster being somehow more similar to each other than they are to samples in other clusters.
- ▶ The simplest and most widely used criterion function for clustering is the *sum-of-squared-error* criterion.



Squared-error Partitioning

- ▶ Suppose that the given set of n patterns has somehow been partitioned into k clusters $\mathcal{D}_1, \dots, \mathcal{D}_k$.
- ▶ Then, the sum of squared errors is defined by

$$J_e = \sum_{i=1}^k \sum_{\mathbf{x} \in \mathcal{D}_i} \|\mathbf{x} - \mathbf{c}_i\|^2$$

where $\mathbf{c}_i$ is the point representative of cluster i .

- ▶ For a given cluster $\mathcal{D}_i$ with n_i samples, the mean vector

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}$$

is the best representative ($\mathbf{c}_i$) of the samples in $\mathcal{D}_i$.

- ▶ The mean vector corresponds to the centroid of the samples.



Squared-error Partitioning

- ▶ A general algorithm for iterative squared-error partitioning:
 1. Select an initial partition with k clusters. Repeat steps 2 through 5 until the cluster membership stabilizes.
 2. Generate a new partition by assigning each pattern to its closest cluster center.
 3. Compute new cluster centers as the centroids of the clusters.
 4. Repeat steps 2 and 3 until an optimum value of the criterion function is found (e.g., when a local minimum is found or a predefined number of iterations are completed).
 5. Adjust the number of clusters by merging and splitting existing clusters or by removing small or outlier clusters.
- ▶ This algorithm, without step 5, is also known as the *k-means* algorithm.



Squared-error Partitioning

- ▶ k -means is computationally efficient and gives good results if the clusters are compact, hyperspherical in shape, and well-separated in the feature space.
- ▶ However, choosing k and choosing the initial partition are the main drawbacks of this algorithm.
- ▶ The value of k is often chosen empirically or by prior knowledge about the data.
- ▶ The initial partition is often chosen by generating k random points uniformly distributed within the range of the data, or by randomly selecting k points from the data.

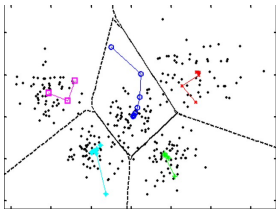


Squared-error Partitioning

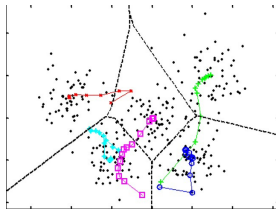
- ▶ Numerous attempts have been made to improve the performance of the basic k -means algorithm:
 - ▶ incorporating a fuzzy criterion resulting in fuzzy k -means,
 - ▶ using genetic algorithms, simulated annealing, deterministic annealing to optimize the resulting partition,
 - ▶ using iterative splitting to find the initial partition.
- ▶ Another alternative is to use model-based clustering using Gaussian mixtures to allow more flexible shapes for individual clusters (k -means with Euclidean distance assumes spherical shapes).
- ▶ In model-based clustering, the value of k corresponds to the number of components in the mixture.



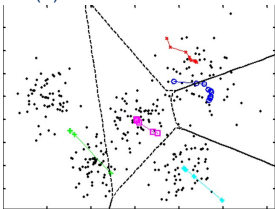
Examples



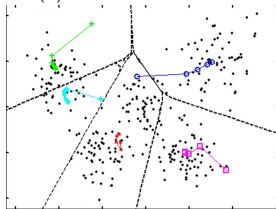
(a) Good initialization.



(b) Good initialization.



(c) Bad initialization.



(d) Bad initialization.

Figure 4: Examples for k -means with different initializations of five clusters for the same data.

Hierarchical Clustering

- ▶ The k -means algorithm produces a *flat* data description where the clusters are disjoint and are at the same level.
- ▶ In some applications, groups of patterns share some characteristics when looked at a particular level.
- ▶ Hierarchical clustering tries to capture these multi-level groupings using *hierarchical* representations rather than flat partitions.



Hierarchical Clustering

- ▶ In hierarchical clustering, for a set of n samples,
 - ▶ the first level consists of n clusters (each cluster containing exactly one sample),
 - ▶ the second level contains $n - 1$ clusters,
 - ▶ the third level contains $n - 2$ clusters,
 - ▶ and so on until the last (n 'th) level at which all samples form a single cluster.
- ▶ Given any two samples, at some level they will be grouped together in the same cluster and remain together at all higher levels.



Hierarchical Clustering

- ▶ A natural representation of hierarchical clustering is a tree, also called a *dendrogram*, which shows how the samples are grouped.
- ▶ If there is an unusually large gap between the similarity values for two particular levels, one can argue that the level with fewer number of clusters represents a more natural grouping.



Hierarchical Clustering

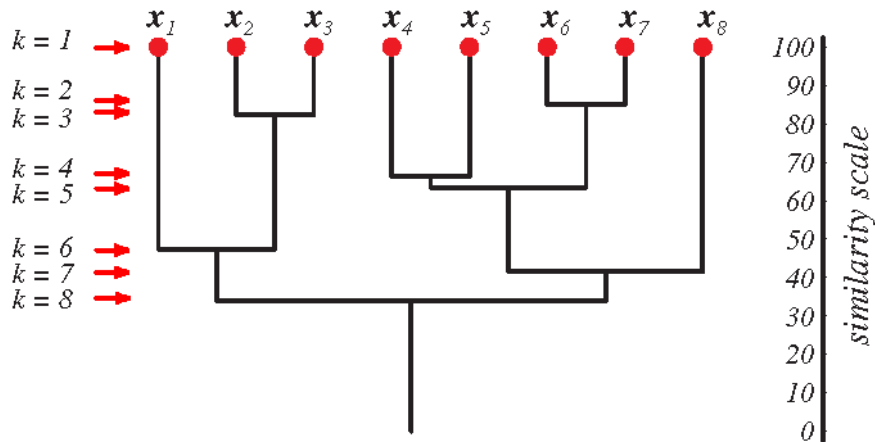


Figure 5: A dendrogram can represent the results of hierarchical clustering algorithms. The vertical axis shows a generalized measure of similarity among clusters.

Hierarchical Clustering

► *Agglomerative Hierarchical Clustering:*

1. Specify the number of clusters. Place every pattern in a unique cluster and repeat steps 2 and 3 until a partition with the required number of clusters is obtained.
2. Find the closest clusters according to a distance measure.
3. Merge these two clusters.
4. Return the resulting clusters.



Hierarchical Clustering

- Popular distance measures (for two clusters $\mathcal{D}_i$ and $\mathcal{D}_j$):

$$d_{\min}(\mathcal{D}_i, \mathcal{D}_j) = \min_{\substack{\mathbf{x} \in \mathcal{D}_i \\ \mathbf{x}' \in \mathcal{D}_j}} \|\mathbf{x} - \mathbf{x}'\|$$

$$d_{\max}(\mathcal{D}_i, \mathcal{D}_j) = \max_{\substack{\mathbf{x} \in \mathcal{D}_i \\ \mathbf{x}' \in \mathcal{D}_j}} \|\mathbf{x} - \mathbf{x}'\|$$

$$d_{\text{avg}}(\mathcal{D}_i, \mathcal{D}_j) = \frac{1}{\#\mathcal{D}_i \#\mathcal{D}_j} \sum_{\mathbf{x} \in \mathcal{D}_i} \sum_{\mathbf{x}' \in \mathcal{D}_j} \|\mathbf{x} - \mathbf{x}'\|$$

$$d_{\text{mean}}(\mathcal{D}_i, \mathcal{D}_j) = \|\mathbf{m}_i - \mathbf{m}_j\|$$



Hierarchical Clustering

- ▶ When $d_{\min}$ is used to measure the distance between clusters, the algorithm is called the nearest neighbor clustering algorithm.
- ▶ Moreover, if the algorithm is terminated when the distance between nearest clusters exceeds a threshold, it is called the *single linkage algorithm* where
 - ▶ patterns represent the nodes of a graph,
 - ▶ edges connect patterns belonging to the same cluster,
 - ▶ merging two clusters corresponds to adding an edge between the nearest pair of nodes in these clusters.



Hierarchical Clustering

- ▶ When $d_{\max}$ is used to measure the distance between clusters, the algorithm is called the farthest neighbor clustering algorithm.
- ▶ Moreover, if the algorithm is terminated when the distance between nearest clusters exceeds a threshold, it is called the *complete linkage algorithm* where
 - ▶ patterns represent the nodes of a graph,
 - ▶ edges connect all patterns belonging to the same cluster,
 - ▶ merging two clusters corresponds to adding edges between every pair of nodes in these clusters.



Hierarchical Clustering

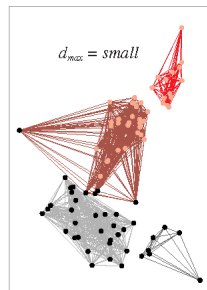
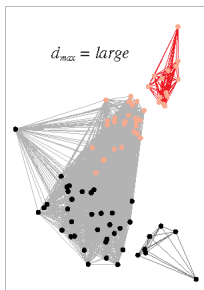
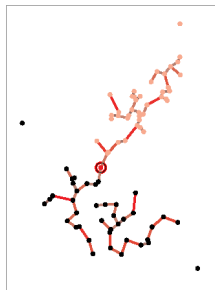
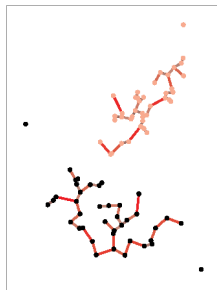


Figure 6: Examples for single linkage clustering. Figure 7: Examples for complete linkage clustering.

Hierarchical Clustering

► *Stepwise-Optimal Hierarchical Clustering:*

1. Specify the number of clusters. Place every pattern in a unique cluster and repeat steps 2 and 3 until a partition with the required number of clusters is obtained.
 2. Find the clusters whose merger increases an error criterion the least.
 3. Merge these two clusters.
- When the sum-of-squared-error criterion J_e is used, the pair of clusters whose merger increases J_e as little as possible is the pair for which the distance

$$d_e(\mathcal{D}_i, \mathcal{D}_j) = \frac{\#\mathcal{D}_i \#\mathcal{D}_j}{\#\mathcal{D}_i + \#\mathcal{D}_j} \|\mathbf{m}_i - \mathbf{m}_j\|^2$$

is minimum.



Graph Clustering

- ▶ A set of patterns can be represented as a graph where
 - ▶ the nodes represent the patterns, and
 - ▶ the edge weights represent the similarity between each pair of patterns.
- ▶ The goal of graph clustering is to find a partition of the graph such that
 - ▶ the edges between different groups have low weights (which means that patterns in different clusters are dissimilar from each other), and
 - ▶ the edges within a group have high weights (which means that patterns within the same cluster are similar to each other).



Graph Clustering

- ▶ The edge weights are typically computed using distances in the feature space.
 - ▶ Algorithms for graph clustering include
 - ▶ Clique finding algorithms
 - ▶ Minimum spanning tree algorithms
 - ▶ Graph cut algorithms
 - ▶ Min cut
 - ▶ Ratio cut
 - ▶ Normalized cut
- and many others.



Cluster Validity

- ▶ The procedures we have considered so far either assume that the number of clusters is known or use some thresholds to decide how the clusters are formed.
- ▶ These may be reasonable assumptions for some applications but are usually unjustified if we are exploring a data set whose properties and structure are unknown.
- ▶ Furthermore, most of the iterative algorithms that we use may find a local extremum and may not give the best result.



Cluster Validity

- ▶ Methods for validating the results of a clustering algorithm include:
 - ▶ Repeating the clustering procedure for different values of the parameters, and examining the resulting values of the criterion function for large jumps or stable ranges.
 - ▶ Evaluating the goodness-of-fit using measures such as the chi-squared or Kolmogorov-Smirnov statistics.
 - ▶ Formulating hypothesis tests that check whether multiple clusters found have been formed by chance, and whether the observed change in the error criterion has any significance.



Cluster Validity

- ▶ The groupings by the unsupervised clustering can also be compared to the known labels if ground truth is available.
- ▶ In an optimal result, the patterns with the same class labels in the ground truth must be assigned to the same cluster and the patterns corresponding to different classes must appear in different clusters.
- ▶ The following measures quantify how well the results of the unsupervised clustering algorithm reflect the groupings in the ground truth:
 - ▶ Entropy,
 - ▶ Rand index.



Cluster Validity

- ▶ *Entropy* is an information theoretic criterion that measures the homogeneity of the distribution of the clusters with respect to different classes.
- ▶ Given K as the number of clusters resulting from the clustering algorithm and C as the number of classes in the ground truth, let
 - ▶ h_{ck} denote the number of patterns assigned to cluster k with a ground truth class label c .
 - ▶ $h_{c.} = \sum_{k=1}^K h_{ck}$ denote the number of patterns with a ground truth class label c .
 - ▶ $h_{.k} = \sum_{c=1}^C h_{ck}$ denote the number of patterns assigned to cluster k .



Cluster Validity

- ▶ The quality of individual clusters is measured in terms of the homogeneity of the class labels within each cluster.
- ▶ For each cluster k , the cluster entropy E_k is given by

$$E_k = - \sum_{c=1}^C \frac{h_{ck}}{h_{.k}} \log \frac{h_{ck}}{h_{.k}}.$$

- ▶ Then, the overall cluster entropy E_{cluster} is given by a weighted sum of individual cluster entropies as

$$E_{\text{cluster}} = \frac{1}{\sum_{k=1}^K h_{.k}} \sum_{k=1}^K h_{.k} E_k.$$



Cluster Validity

- ▶ A smaller cluster entropy value indicates a higher homogeneity.
- ▶ However, the cluster entropy continues to decrease as the number of clusters increases.
- ▶ To overcome this problem, another entropy criterion that measures how patterns of the same class are distributed among the clusters can be defined.



Cluster Validity

- For each class c , the class entropy E_c is given by

$$E_c = - \sum_{k=1}^K \frac{h_{ck}}{h_c} \log \frac{h_{ck}}{h_c}.$$

- Then, the overall class entropy E_{class} is given by a weighted sum of individual class entropies as

$$E_{\text{class}} = \frac{1}{\sum_{c=1}^C h_c} \sum_{c=1}^C h_c E_c.$$



Cluster Validity

- ▶ Unlike the cluster entropy, the class entropy increases when the number of clusters increases.
- ▶ Therefore, the two measures can be combined for an overall entropy measure as

$$E = \beta E_{\text{cluster}} + (1 - \beta) E_{\text{class}}$$

where $\beta \in [0, 1]$ is a weight that balances the two measures.



Cluster Validity

- ▶ The *Rand index* can also be used to measure the agreement of every pair of patterns according to both unsupervised and ground truth labelings.
- ▶ The agreement occurs if
 - ▶ two patterns that belong to the same class are put into the same cluster, or
 - ▶ two patterns that belong to different classes are put into different clusters.



Cluster Validity

- ▶ The Rand index is computed as the proportion of all pattern pairs that agree in their labels.
- ▶ The index has a value between 0 and 1, where 0 indicates that the two labelings do not agree on any pair of patterns and 1 indicates that the two labelings are exactly the same.

