SYDE 372
Introduction to Pattern Recognition

Clustering

Alexander Wong

Department of Systems Design Engineering
University of Waterloo

Outline

Motivation

- All the approaches we have learned so far deal with problems where labeled samples are available to represent each class.
- However, there are many important problems where no such class-defining information is available!
- The data we have is just a collection of unlabeled samples and the problem is to find naturally occurring groups or clusters
- What are some practical clustering problems?
  - Identifying the number and variety of ground cover types in satellite photographs.
  - Identifying the number and variety of crowds based on the trajectories of individual persons.
  - Identifying distinct biological species based on physical attributes

Clusters

- Before we start discussing strategies for clustering unlabeled data, we need to think about what is meant by a cluster
- One way of defining a cluster is simply a set of samples that are similar to each other.
- If we use this definition, different similarity criterions can lead to different clustering results.
- Alternatively, we may define a cluster as a region in feature space containing a high density of samples.
- If we use this definition, peaks in the sample density function are associated with clusters.
- The way we define clusters influences the strategies that we use to perform clustering.

Intuitive Approaches

- Let's explore some possible clustering strategies that are relatively simple and intuitive in nature.
- Strategy 1: Mixture density strategy
  - Suppose we use the high sample density region definition of a cluster.
  - Given a set of N samples, we can estimate a combined PDF ($p(\underline{x})$) using density estimation (e.g., Parzen window estimation)
  - This density $p(\underline{x})$ is referred to as a mixture density since it is a mixture of the K class densities

$$p(\underline{x}) = \sum_{i=1}^{K} P(c_i) p(\underline{x}|c_i) \tag{1}$$

Intuitive Approaches

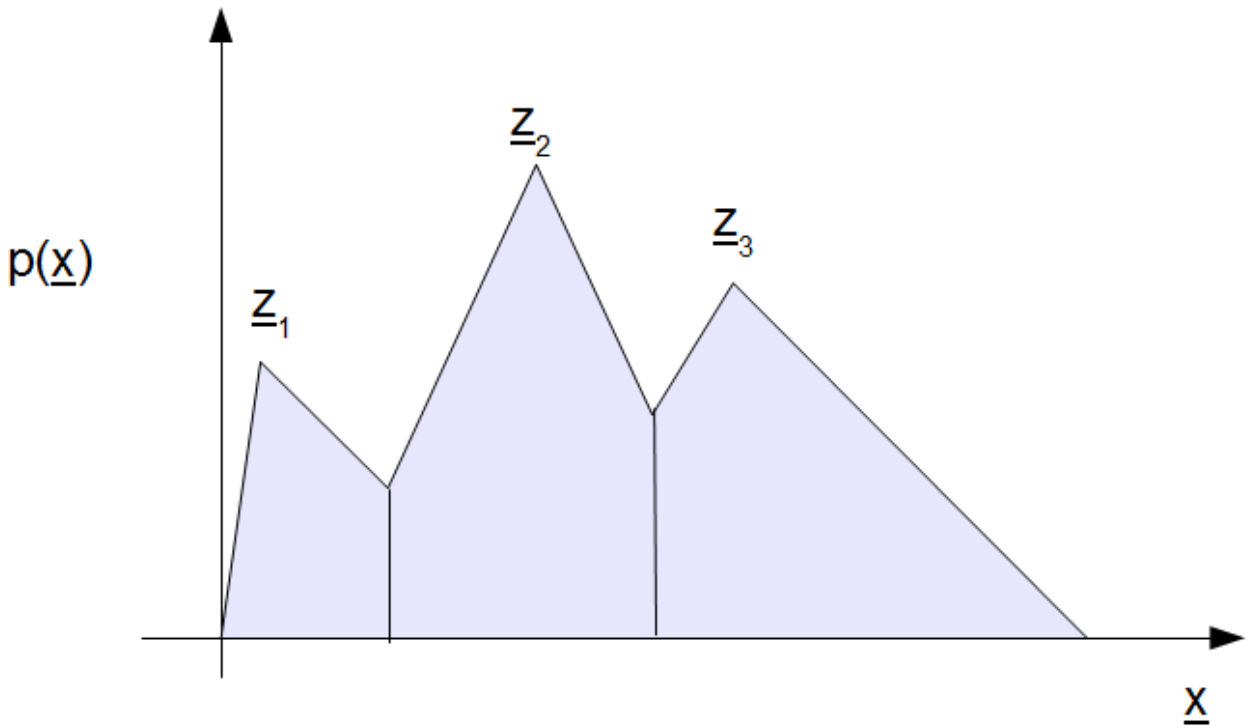
- Strategy 1: Mixture density strategy
  - If the K classes are reasonably distinct and nearly equally likely, $p(\underline{x})$ should have K peaks or local maxima, one for each class.
  - If that is the case, we can define the cluster prototype at each local maxima:

$$\underline{z}_i = \underline{x} \text{ such that } \underline{x} \text{ is the } i^{\text{th}} \text{ local maximum of } p(\underline{x}) \quad (2)$$

  - Based on the prototypes, we can determine our clusters using distance metrics (e.g., Euclidean distance) or cluster boundaries based on local minima.

Intuitive Approaches

- Strategy 1: Mixture density strategy example



Intuitive Approaches

- Strategy 1: Mixture density strategy drawbacks
  - In practice, too few samples will result in a large number of spurious local maxima
  - The individual sample densities of a particular class may be skewed (e.g., Gamma) and so a simple Euclidean distance metric may not define the cluster boundaries well.
  - Peaks of clusters may be too close (or even overlap) to distinguish properly.

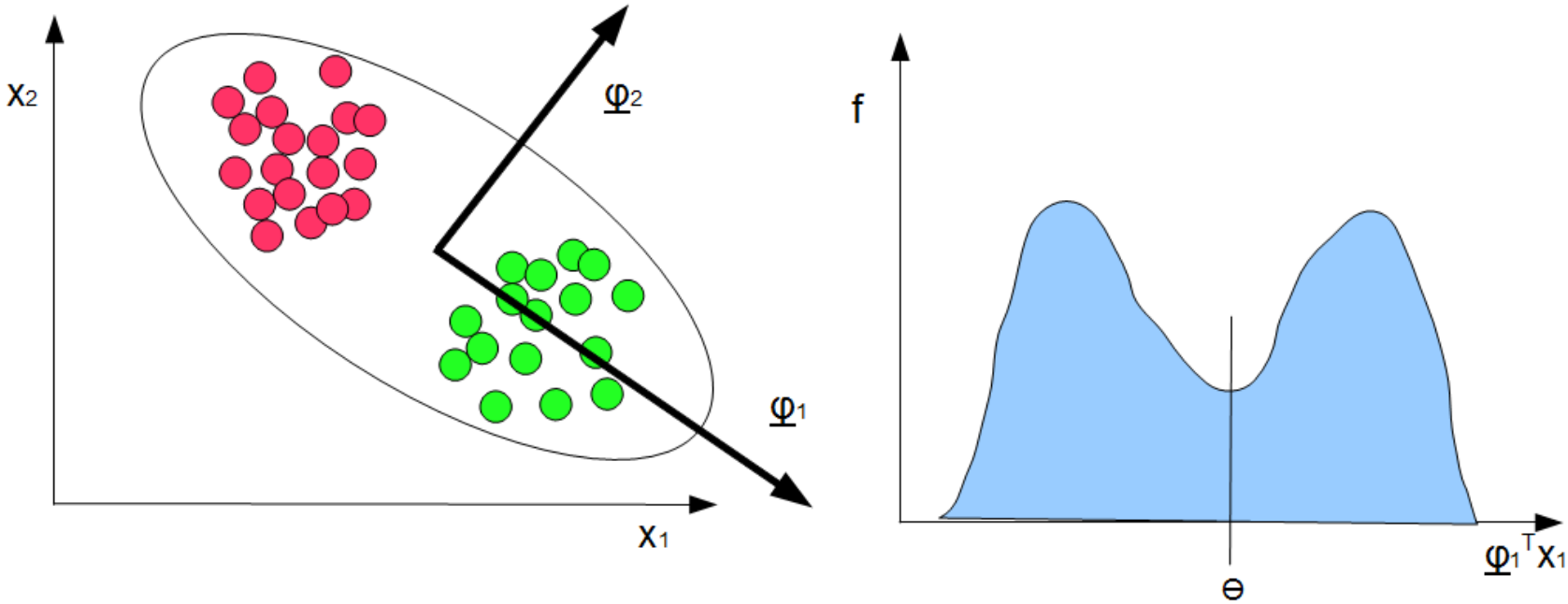
Intuitive Approaches

- Strategy 2: Principal component analysis strategy
  - Assume that we have two fairly compact and distinct clusters.
  - The combined covariance matrix describes the shape of the total sample distribution.
  - If we take the maximum eigenvalue eigenvector $\underline{\phi}_1$ as the maximum separation direction, we can do the following:
    - Project samples onto $\underline{\phi}_1$
    - Construct histogram
    - Choose threshold θ at minimum of distribution, and specify clusters based on the following rule:

$$\underline{x} \in c_1 \text{ if } \phi_1^T \underline{x} < \theta \tag{3}$$

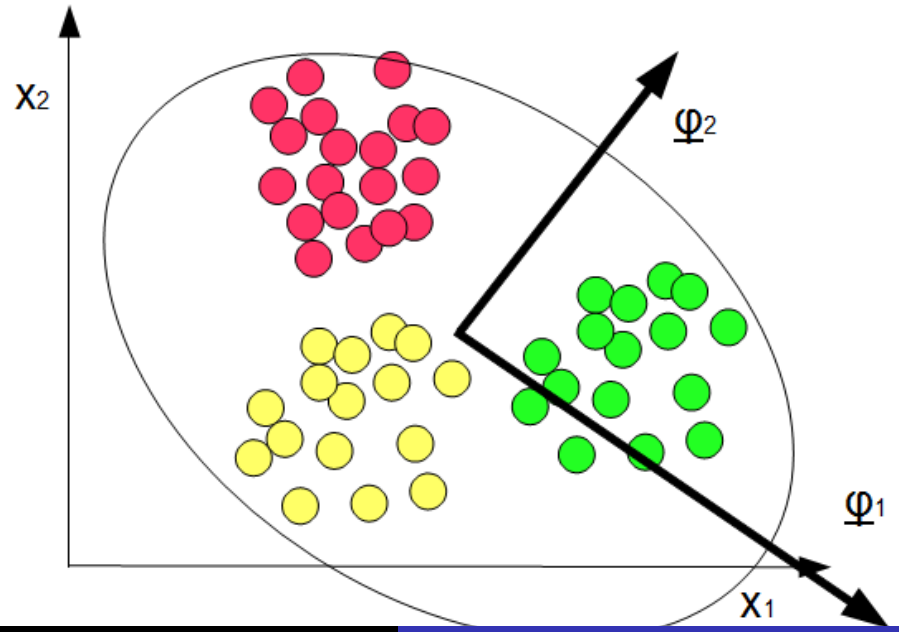
Intuitive Approaches

- Strategy 2: Principal component analysis strategy example



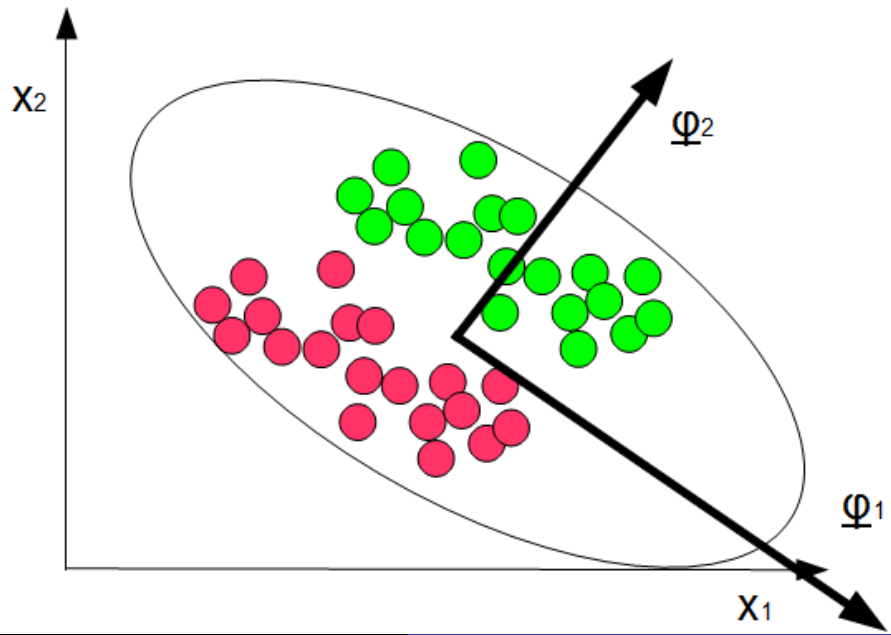
Intuitive Approaches

- Strategy 2: Principal component analysis strategy
- While this clustering scheme is simple and intuitive, it cannot handle more complicated but typical problems such as:
  - Having $K > 2$ clusters



Intuitive Approaches

- Strategy 2: Principal component analysis strategy
- While this clustering scheme is simple and intuitive, it cannot handle more complicated but typical problems such as:
  - Non-spherical distributions



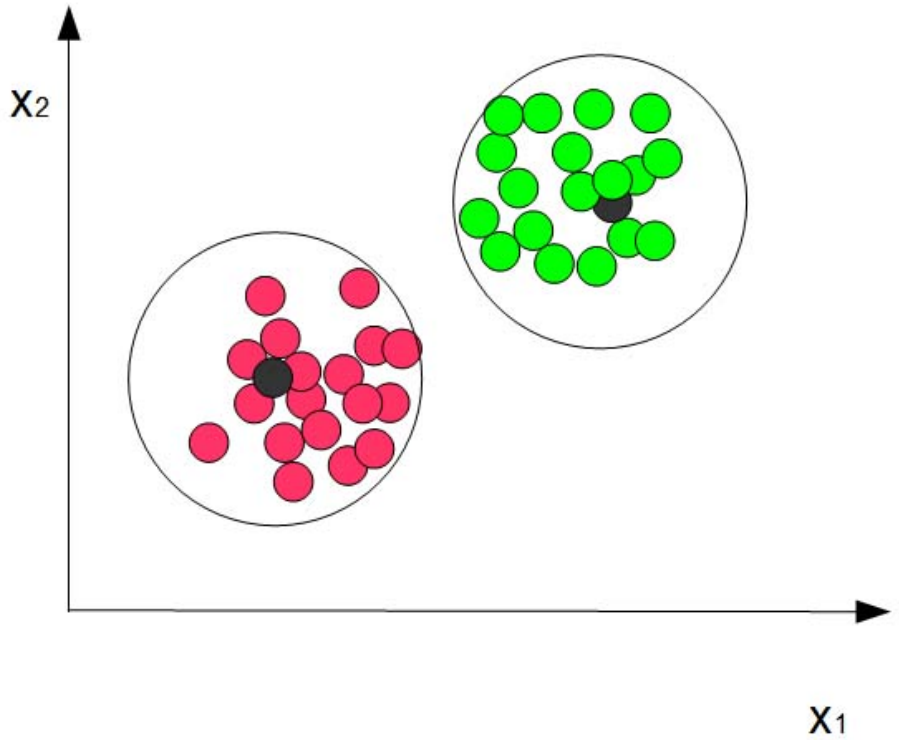
Intuitive Approaches

- Strategy 3: Euclidean distance threshold strategy

- Let's use the definition of clustering based on similarity with each other.
- Based on this notion, what we can do is assign a pattern to a cluster if the pattern to a cluster if the distance between the cluster and the prototype is less than some threshold T.
  - **Step 1**: Let $\underline{z}_1 = \underline{x}_1$, the first sample.
  - **Step 2**: For each sample, assign it to closest prototype for which $|\underline{x}_i - \underline{z}_j| < T$. If no such prototype exists, let $\underline{x}_i$ be a new prototype.
  - **Step 3**: Repeat Step 2 until all samples are assigned to clusters.

Intuitive Approaches

- Strategy 3: Euclidean distance threshold strategy



Intuitive Approaches

- Strategy 3: Euclidean distance threshold strategy drawbacks
- Threshold T is entirely arbitrary!
  - If T is small, many clusters will be found.
  - If T is large, few clusters will be found.
- Sensitivity to the order in which samples are considered (affects specific prototypes)
- As such, most of these intuitive approaches seem rather arbitrary!
- To help remedy such issues, we want to develop formal criteria for clustering that are less likely to impose some artificial structure on data, and more sensitive to any natural structure.

Criteria for Clustering

- A more formal approach to clustering is to define a criterion function
- This function serves as a quantitative measure of the clusters
- Thus, a given partitioning of the sample set is optimum when criterion function is maximized/minimized.
- Commonly used criterion functions:
  - Sum of squared errors
  - Scatter volume

Sum of Squared Error Criteria

- Suppose that we want K clusters, and a single cluster, c_i, has N_i samples.
- The sum of squared errors resulting when the N_i samples are represented by the cluster prototype, $\underline{z}_i$ can be expressed as:

$$J_i = \sum_{\underline{x} \in c_i} |\underline{x} - \underline{z}_i|^2 \tag{4}$$

- The prototype $\underline{z}_i$ which minimize the single class error is just the class mean $\underline{m}_i$

Sum of Squared Error Criteria

- Therefore, the total sum of squared error criterion is:

$$J_e = \sum_{i=1}^{K} J_i = \sum_{i=1}^{K} \sum_{\underline{x} \in c_i} |\underline{x} - \underline{m}_i|^2 \tag{5}$$

- A clustering or partitioning that minimizes this total error J_e is called **minimum variance partition**.

Scatter Volume Criteria

- Suppose we compute the scatter matrix for class c_i, which can be defined as:

$$S_i = \sum_{\underline{x} \in c_i} (\underline{x} - \underline{m}_i)(\underline{x} - \underline{m}_i)^T \tag{6}$$

- S_i is just N_i times the covariance matrix of class i.
- Summing over K classes, we have a measure of total within class scatter:

$$S_W = \sum_{i=1}^{K} S_i \tag{7}$$

Scatter Volume Criteria

- If we compute the trace of S_W, you will realize that it is just the sum of squared error criterion!

$$tr(S_W) = \sum_{i=1}^{K} tr(S_i) = \sum_{i=1}^{K} \sum_{\underline{x} \in c_i} |\underline{x} - \underline{m}_i|^2 = J_e \tag{8}$$

- Another good criterion is the scatter volume, which can be determined by taking the determinant of S_W:

$$J_v = |S_W| \tag{9}$$

Criteria for Clustering: Example

- Suppose we are given the following samples: $\underline{x}_1 = [4 \;\; 5]^T, \underline{x}_2 = [1 \;\; 4]^T, \underline{x}_3 = [0 \;\; 1]^T, \underline{x}_4 = [5 \;\; 0]^T$.
- Consider the following partitions:
  1. $C_1 = \{\underline{x}_1, \underline{x}_2\}, C_2 = \{\underline{x}_3, \underline{x}_4\}$
  2. $C_1 = \{\underline{x}_1, \underline{x}_4\}, C_2 = \{\underline{x}_2, \underline{x}_3\}$
  3. $C_1 = \{\underline{x}_1, \underline{x}_2, \underline{x}_3\}, C_2 = \{\underline{x}_4\}$

- (a) Which partition is favored by the sum of squared error criterion?
- (b) Which partition is favored by the scatter volume criterion?

Criteria for Clustering: Example

- Since the trace of S_w is the same as the sum of squared error, and the scatter volume is also based on S_w, let's first compute Sw for each case:
- For $C_1 = \{\underline{x}_1, \underline{x}_2\}, C_2 = \{\underline{x}_3, \underline{x}_4\}$:

$$\begin{aligned} \underline{m}_1 &= \frac{1}{2}\left\{[4 \;\; 5]^T + [1 \;\; 4]^T\right\} \\ \underline{m}_1 &= [2.5 \;\; 4.5]^T. \end{aligned} \tag{10}$$

$$\begin{aligned} \underline{m}_2 &= \frac{1}{2}\left\{[0 \;\; 1]^T + [5 \;\; 0]^T\right\} \\ \underline{m}_2 &= [2.5 \;\; 0.5]^T. \end{aligned} \tag{11}$$

Criteria for Clustering: Example

- The scatter matrices for each class are:

$$
\begin{aligned}
S_1 &= \{[4\ \ 5]^T[4\ \ 5] + [1\ \ 4]^T[1\ \ 4]\} - 2\{[2.5\ \ 4.5]^T[2.5\ \ 4.5]\} \\
S_1 &= \begin{bmatrix} 17 & 24 \\ 24 & 41 \end{bmatrix} - 2\begin{bmatrix} 12.5 & 22.5 \\ 22.5 & 40.5 \end{bmatrix} \\
S_1 &= \begin{bmatrix} 4.5 & 1.5 \\ 1.5 & 0.5 \end{bmatrix}
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
S_2 &= \{[0\ \ 1]^T[0\ \ 1] + [5\ \ 0]^T[5\ \ 0]\} - 2\{[2.5\ \ 0.5]^T[2.5\ \ 0.5]\} \\
S_2 &= \begin{bmatrix} 25 & 0 \\ 0 & 1 \end{bmatrix} - 2\begin{bmatrix} 12.5 & 2.5 \\ 2.5 & 0.5 \end{bmatrix} \\
S_2 &= \begin{bmatrix} 12.5 & -2.5 \\ -2.5 & 0.5 \end{bmatrix}
\end{aligned} \tag{13}
$$

Criteria for Clustering: Example

- Given S_1 and S_2, the total within scatter matrix S_W is:

$$S_W = S_1 + S_2$$

$$S_W = \begin{bmatrix} 4.5 & 1.5 \\ 1.5 & 0.5 \end{bmatrix} + \begin{bmatrix} 12.5 & -2.5 \\ -2.5 & 0.5 \end{bmatrix} \quad (14)$$

$$S_W = \begin{bmatrix} 17 & -1 \\ -1 & 1 \end{bmatrix}$$

- The trace and scatter volume of S_W are given by:

$$tr(S_W) = tr(\begin{bmatrix} 17 & -1 \\ -1 & 1 \end{bmatrix}) = 17 + 1 = 18 \quad (15)$$

$$|S_W| = 17(1) - (-1)(-1) = 16 \quad (16)$$

Criteria for Clustering: Example

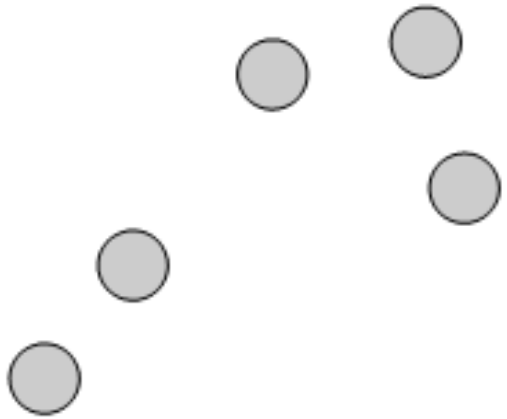
- Doing the same thing for case (a) and case (b) gives us the following:
  1. $tr(S_W) = 18, |S_W| = 16$
  2. $tr(S_W) = 18, |S_W| = 16$
  3. $tr(S_W) = 52/3, |S_W| = 64/3$
- (a) Since the sum of squared error criterion favors convergence towards the minimum trace of S_W, case (c) is the favored partition since it gives the lowest $tr(S_W)$.
- (a) Since the scatter volume criterion favors convergence towards the minimum scatter volume of S_W, cases (a) and (b) are the favored partitions since they give the lowest $|S_W|$.

Minimum Variance Algorithms

- Based on the criterion functions that we have discussed, we can now develop formal specific clustering algorithms to determine clusters that satisfies these criterion functions.
- A basic initial assumption for such methods is that the number of clusters is known (although more advanced methods have been developed for handling cases with unknown number of clusters by starting with a large number of clusters and then trying to converge to the ideal number of clusters)
- Here, we will look into **minimum variance algorithms**, where the goal is to minimize the sum of squared error criterion function.

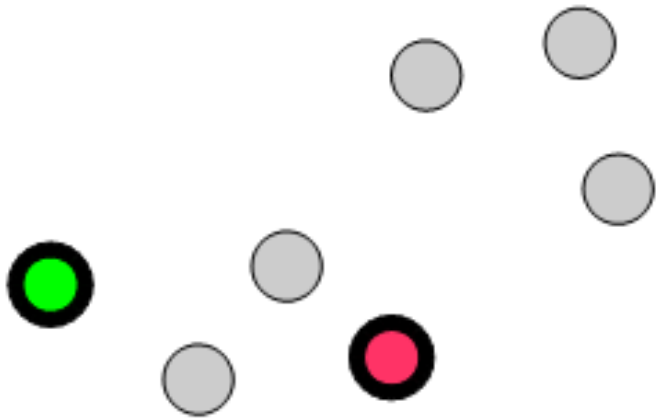
K-means algorithm

- Most famous of the minimum variance algorithms is the **K-means algorithm**.
- Suppose that we are given the following set of unlabeled samples:



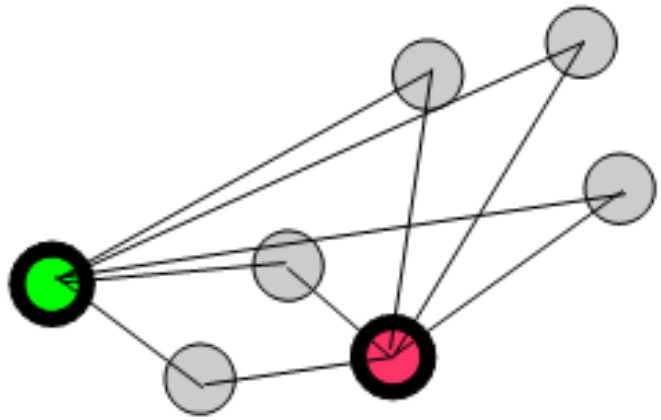
K-means algorithm

- **Step 1**: Choose prototypes $\{\underline{z}_1, \ldots, \underline{z}_k\}$ arbitrarily.



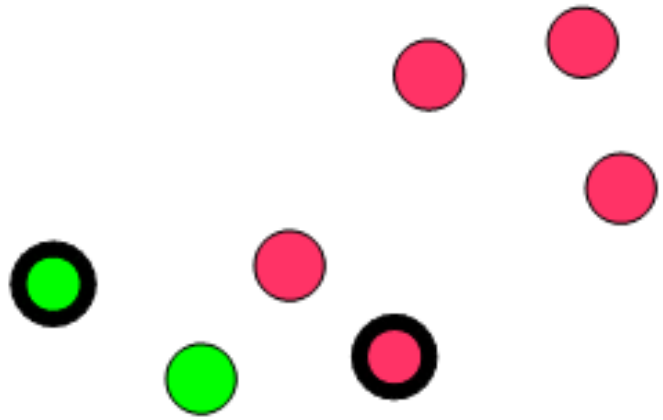
K-means algorithm

- **Step 2**: Compute the Euclidean distance between N samples to each cluster:



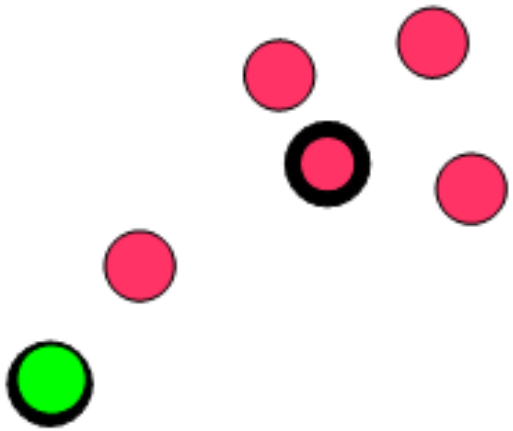
K-means algorithm

- **Step 3**: Assign the N samples to the K clusters based on the minimum Euclidean distance.
  $\underline{x} \in C_i$ if $|\underline{x} - \underline{z}_i| < |\underline{x} - \underline{z}_j|, j \neq i$



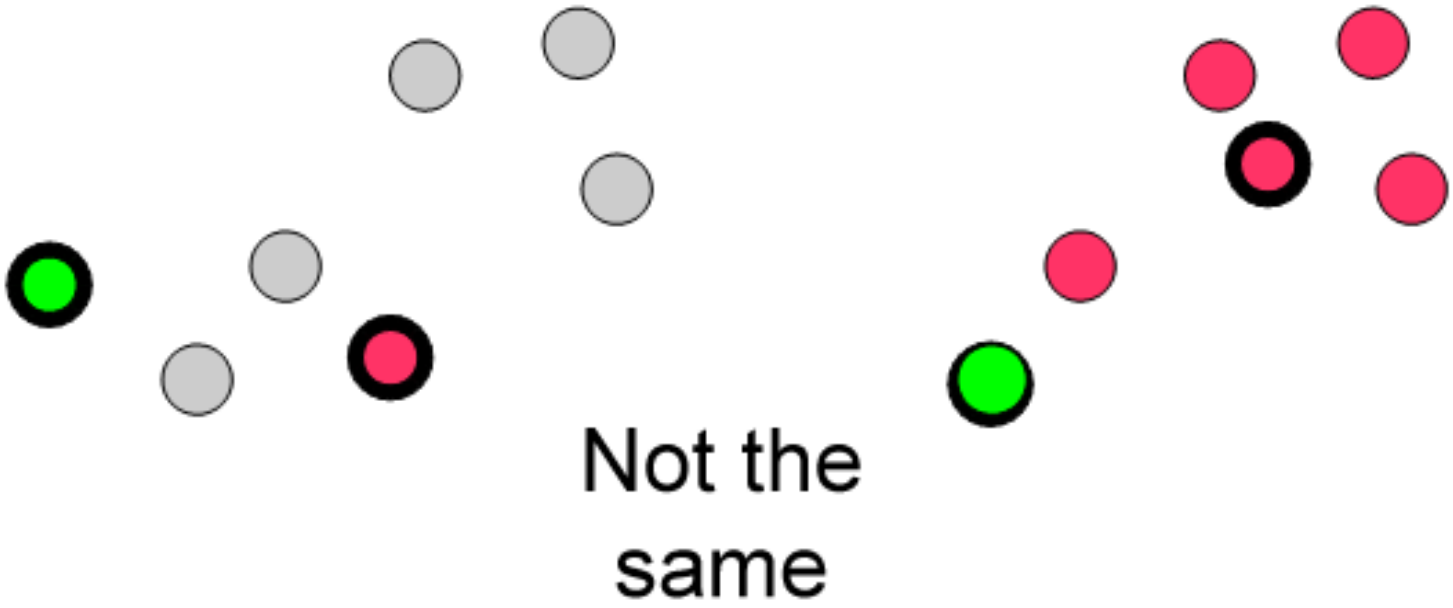
K-means algorithm

- **Step 4**: Compute new cluster prototypes as the cluster means:
  $z_{i,\text{new}} = \frac{1}{N_i} \sum_{i=1}^{N_i} \underline{x}_i, \quad \forall x \in C_i$



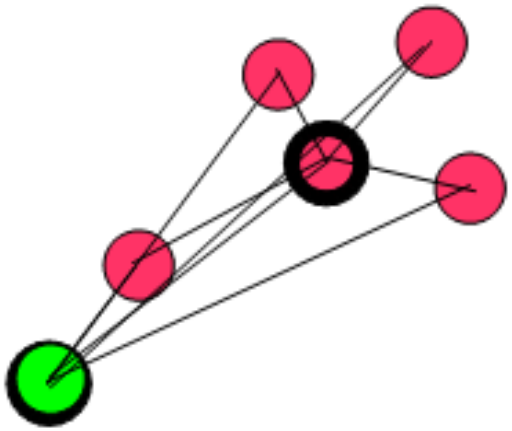
K-means algorithm

- **Step 5**: If any cluster prototypes change, go back to Step 2



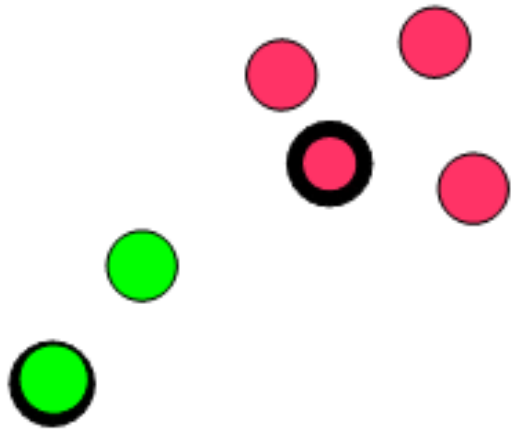
K-means algorithm

- **Step 2**: Compute the Euclidean distance between N samples to each cluster:



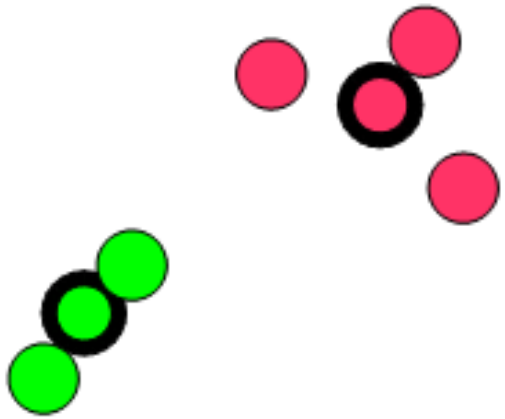
K-means algorithm

- **Step 3**: Assign the N samples to the K clusters based on the minimum Euclidean distance.
  $\underline{x} \in C_i$ if $|\underline{x} - \underline{z}_i| < |\underline{x} - \underline{z}_j|, j \neq i$



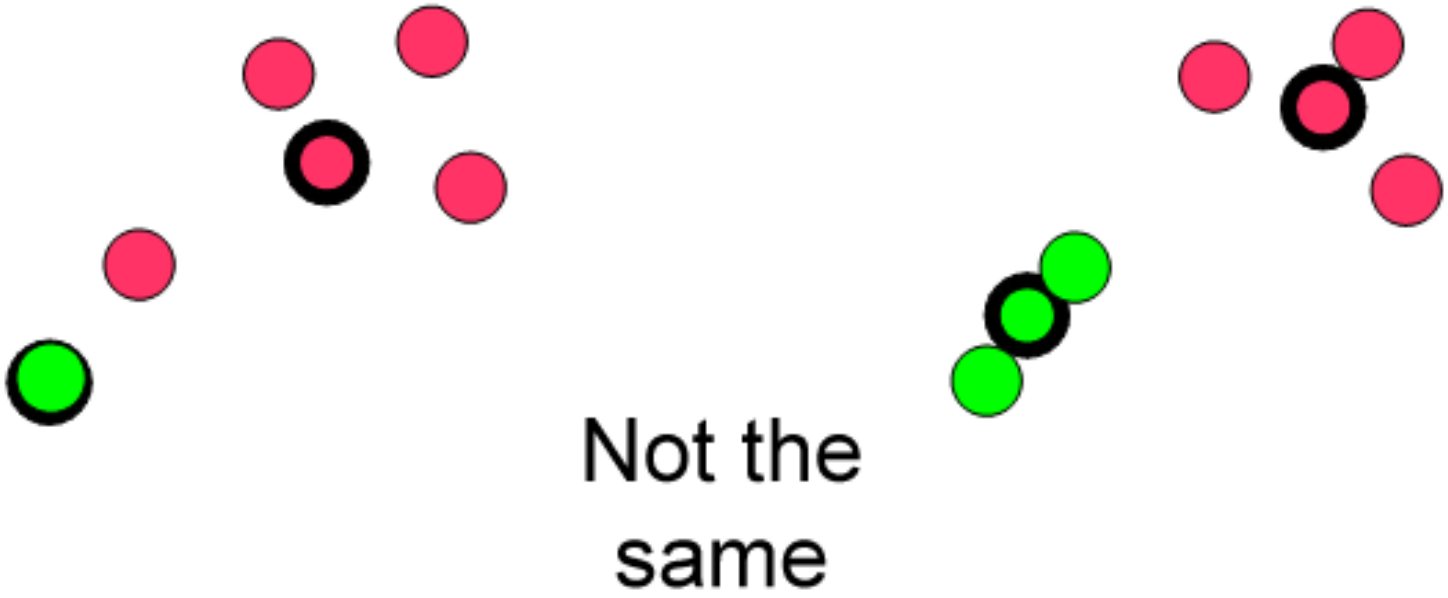
K-means algorithm

- **Step 4**: Compute new cluster prototypes as the cluster means:
  $z_{i,\text{new}} = \frac{1}{N_i} \sum_{i=1}^{N_i} \underline{x}_i, \quad \forall x \in C_i$



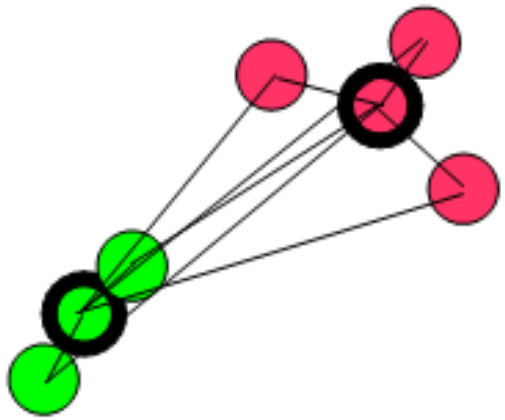
K-means algorithm

- **Step 5**: If any cluster prototypes change, go back to Step 2



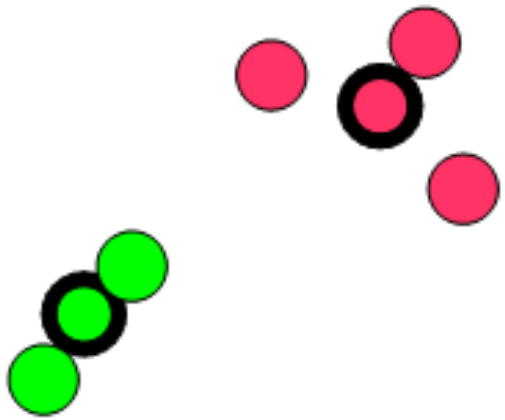
K-means algorithm

- **Step 2**: Compute the Euclidean distance between N samples to each cluster:



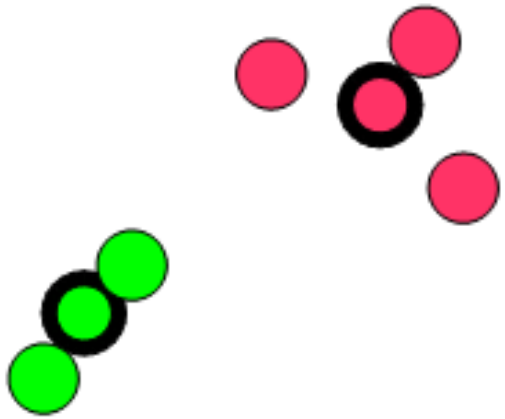
K-means algorithm

- **Step 3**: Assign the N samples to the K clusters based on the minimum Euclidean distance.
  $\underline{x} \in C_i$ if $|\underline{x} - \underline{z}_i| < |\underline{x} - \underline{z}_j|, j \neq i$



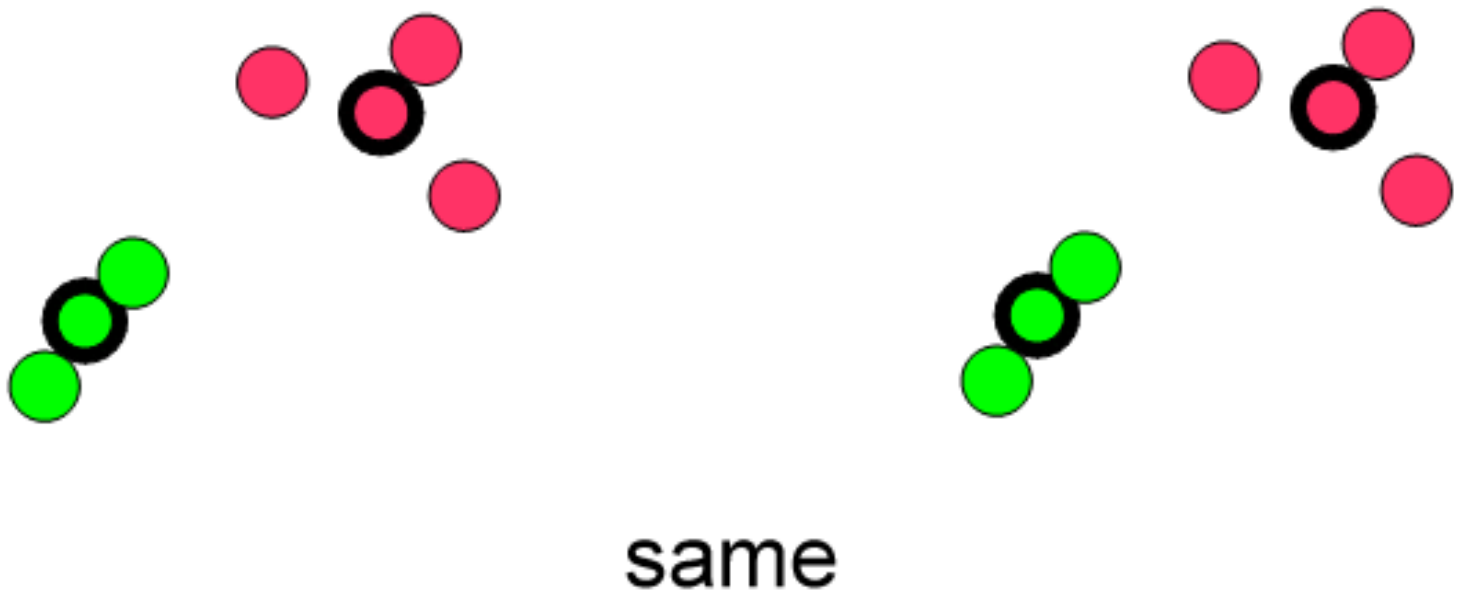
K-means algorithm

- **Step 4**: Compute new cluster prototypes as the cluster means:
  $z_{i,\text{new}} = \frac{1}{N_i} \sum_{i=1}^{N_i} \underline{x}_i, \quad \forall x \in C_i$



K-means algorithm

- **Step 5**: If any cluster prototypes change, go back to Step 2



- Therefore, stop!

K-means algorithm: Considerations and Possible solutions

- There are several considerations associated with K-means:
  - Sensitivity to initialization (can be reduced by running random initializations multiple times and choosing the best results)
  - Assumes number of clusters is known (can be reduced by starting with a large number of clusters and then consolidating them)
  - Sensitive to geometric properties of clusters

K-means algorithm: Variants

- A more complex variant of K-means is the ISODATA (Iterative Self Organizing Data Analysis Technique A):
  - If the number of samples in any cluster is less than some threshold, the cluster may be eliminated.
  - If the maximum variance feature for the cluster is larger than some threshold and there are sufficient samples in the cluster, split the cluster.
  - If the pairwise distance between clusters is below some threshold, the clusters are combined.
- ISODATA is quite flexible, but needs a number of thresholds to be defined.
- Therefore, ISODATA works best when used in an interactive, empirical environment.

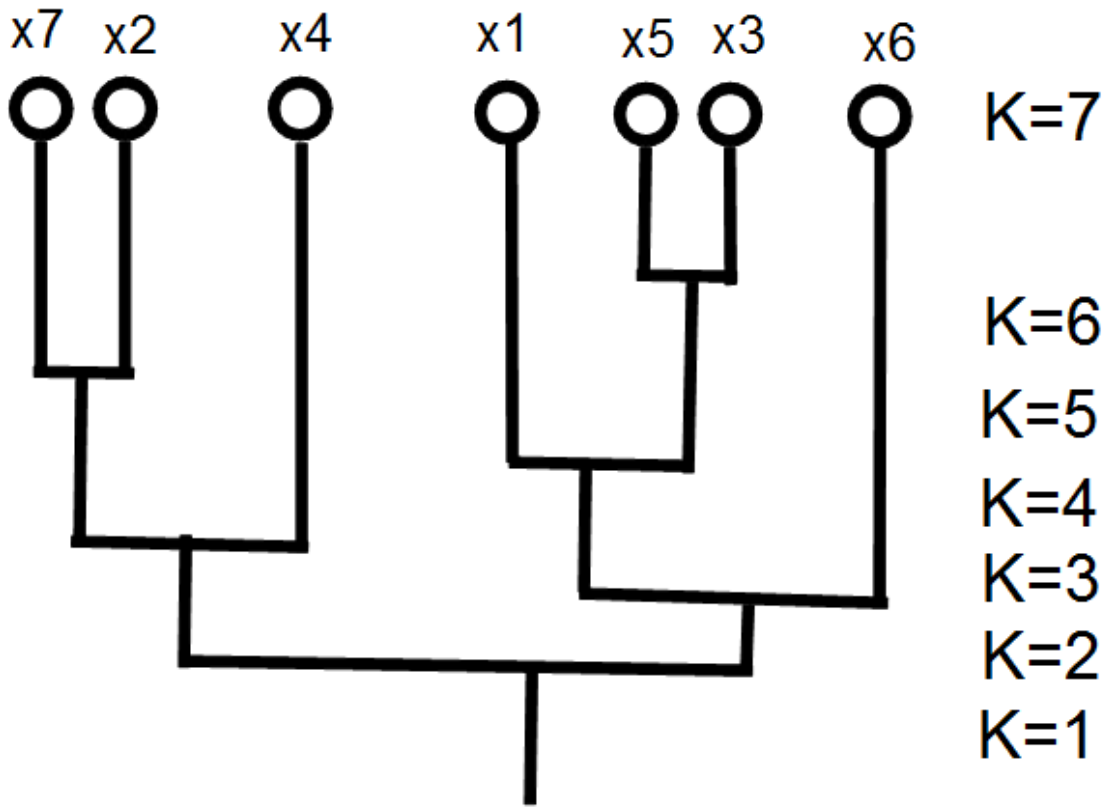
Hierarchical Clustering

- Both K means and ISODATA can be sensitive to initial conditions (e.g., number of clusters, choice of prototypes, sample ordering)
- A number of schemes can be devised to determine starting points for such approaches
- A popular approach for determining starting points (or even for clustering itself) is a hierarchical clustering approach,
- **Goal:** Develop a sequence of partitions what at each step sub-clusters are combined according to some criterion

Hierarchical Clustering: Simple Example

- Suppose we have *N* samples, and treat each as a cluster prototype (e.g., N one-sample clusters)
- We now find the two most similar clusters and merge them, resulting in N-1 clusters
- We continue combining the most similar clusters at each step until some criterion is satisfied
- Such a scheme can be conveniently represented using a dendogram

Dendogram



Measures of similarity for hierarchical clustering

- Popular similarity measures based on Euclidean distance (Each measure gives different clustering results):
  - Minimum distance (nearest neighbor)

$$d_{\min}(c_i, c_j) = \min_{\underline{x} \in c_i, \underline{x}' \in c_j} |\underline{x} - \underline{x}'| \tag{17}$$

  - Nearest neighbor measure well-suited for string-like clusters but highly sensitive to noise and outliers

Measures of similarity for hierarchical clustering

- Popular similarity measures based on Euclidean distance (Each measure gives different clustering results):
  - Maximum distance (furthest neighbor)

$$d_{\max}(c_i, c_j) = \max_{\underline{x} \in c_i, \underline{x}' \in c_j} |\underline{x} - \underline{x}'| \tag{18}$$

  - Furthest neighbor measure tend to discourage growth of elongated clusters since two sub-clusters are only merged if the least similar pair in the resulting cluster is sufficiently similar
  - This makes it less sensitive to outliers and noise and well-suited for compact, spherical clusters

Measures of similarity for hierarchical clustering

- Popular similarity measures based on Euclidean distance (Each measure gives different clustering results):
  - Average distance

$$d_{\text{avg}}(c_i, c_j) = \frac{1}{N_i N_j} \sum_{\underline{x} \in c_i} \sum_{\underline{x}' \in c_j} |\underline{x} - \underline{x}'| \tag{19}$$

  - Average distance also less sensitive to noise and outliers

Measures of similarity for hierarchical clustering

- Popular similarity measures based on Euclidean distance (Each measure gives different clustering results):
  - Distance between means

$$d_{\text{mean}}(c_i, c_j) = |\underline{m}_i - \underline{m}_j| \tag{20}$$

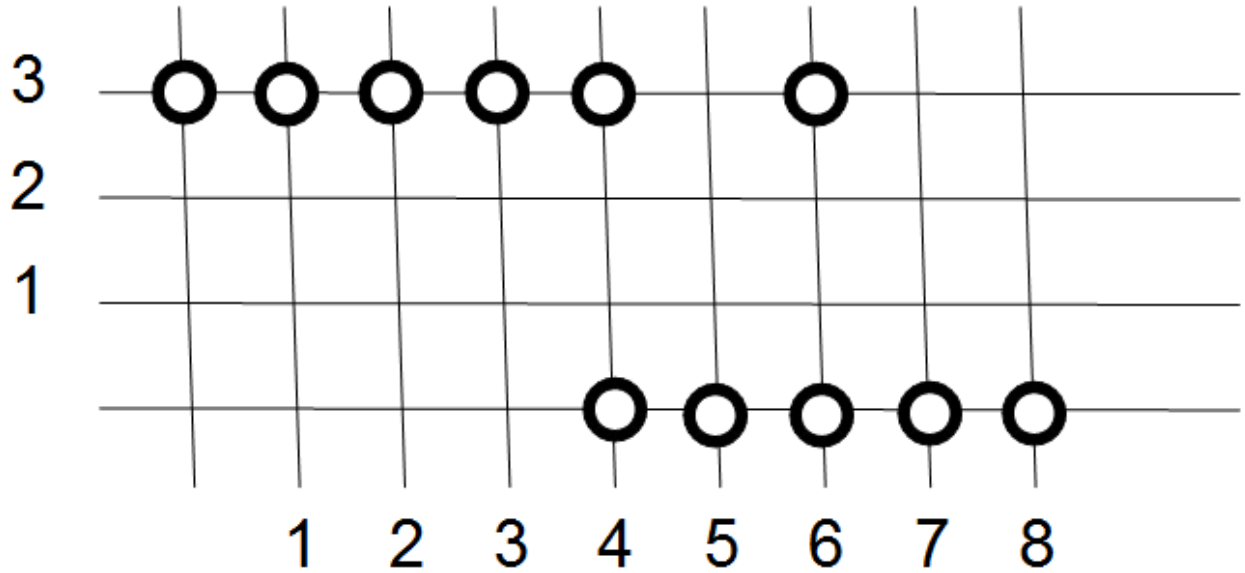
  - Inter-mean measure also less sensitive to noise and outliers
  - Most efficient to implement

Hierarchical clustering: Termination Criterion

- Termination occurs based on some criteria
  - Maximum number of levels
  - Number of clusters
  - Error threshold

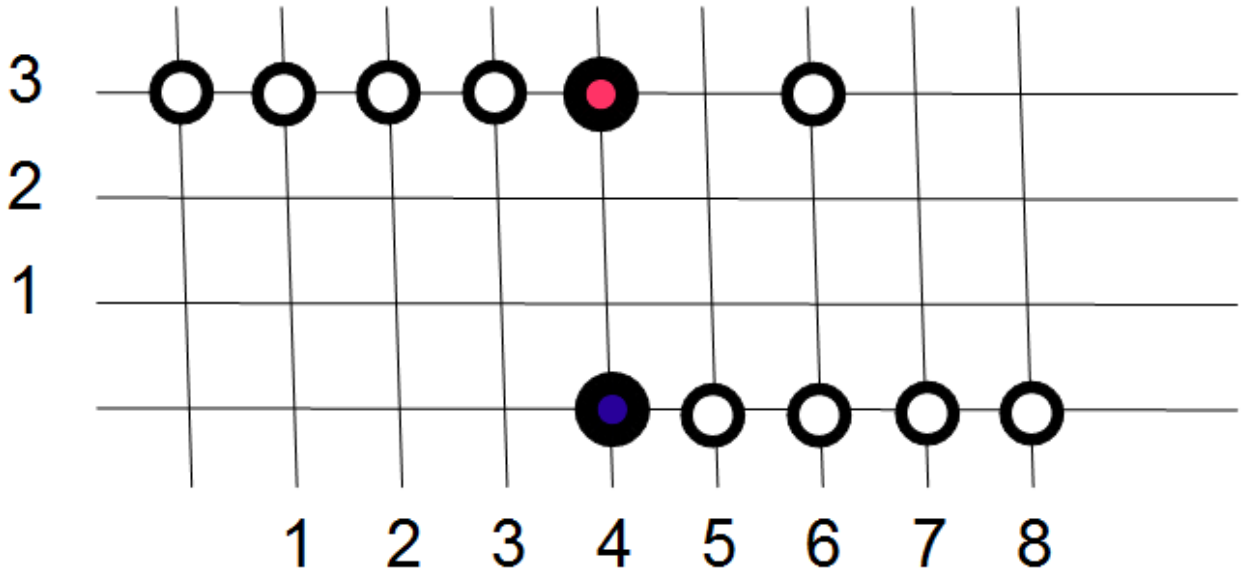
Clustering Example

- Suppose that we are given a set of unlabeled samples as shown in this scatter plot

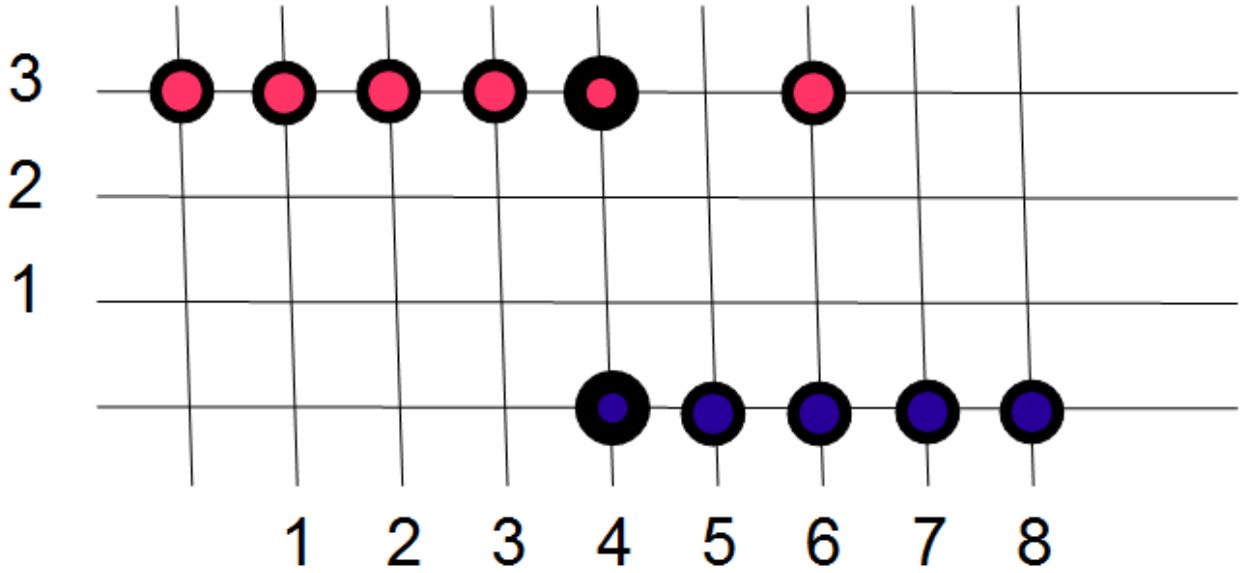


Clustering Example

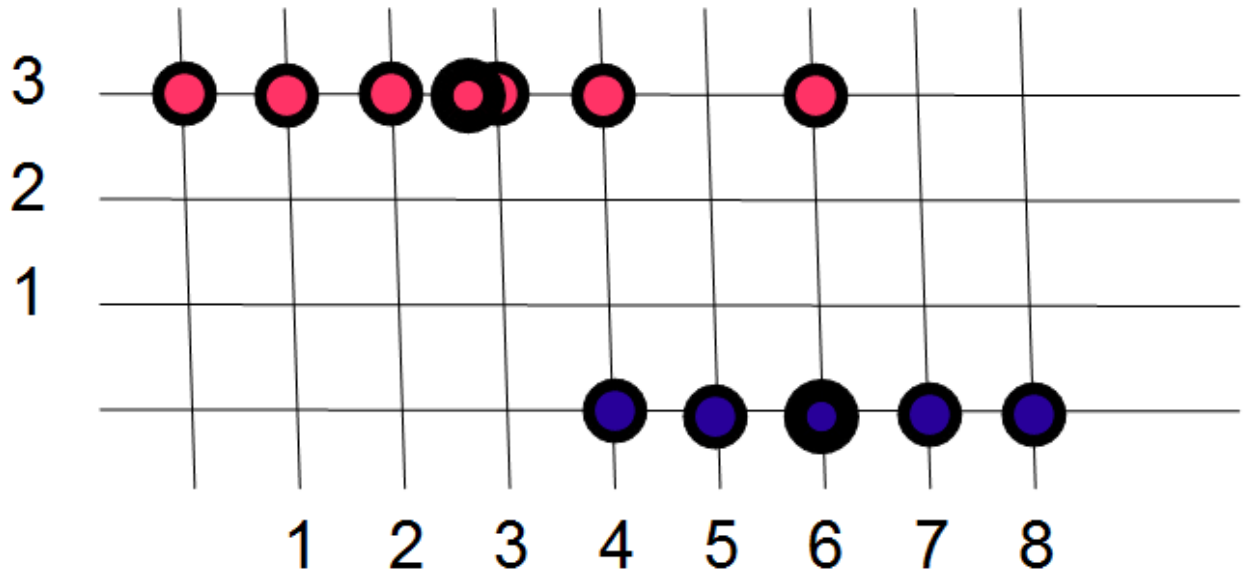
- Use K-means with K=2 and samples at (4,0) and (4,3) as initial prototypes to find clusters.



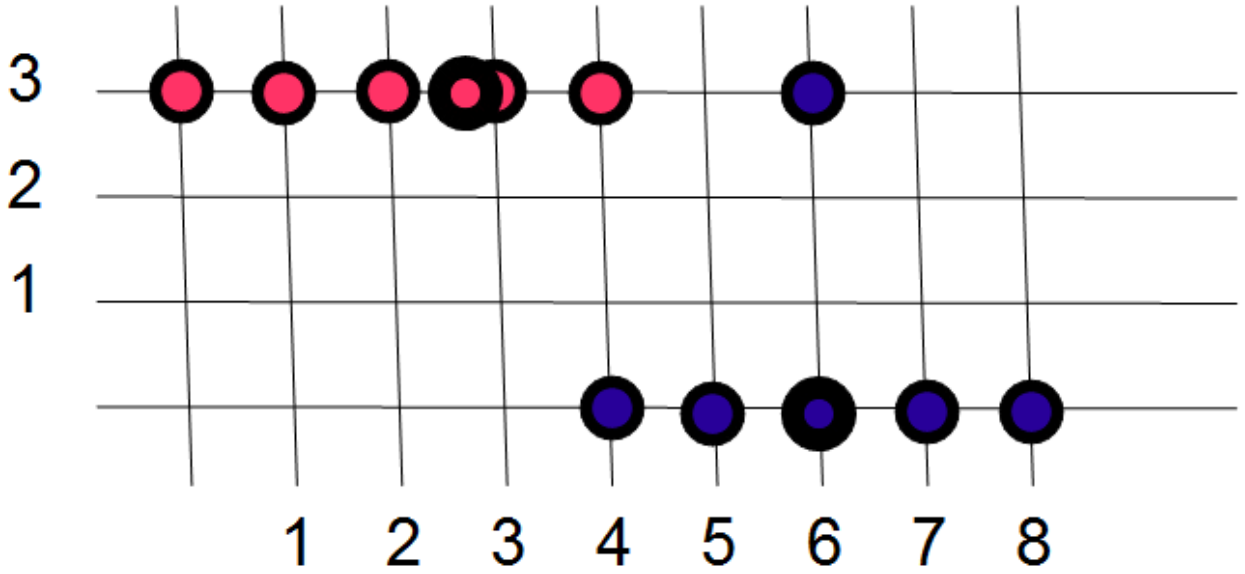
Clustering Example



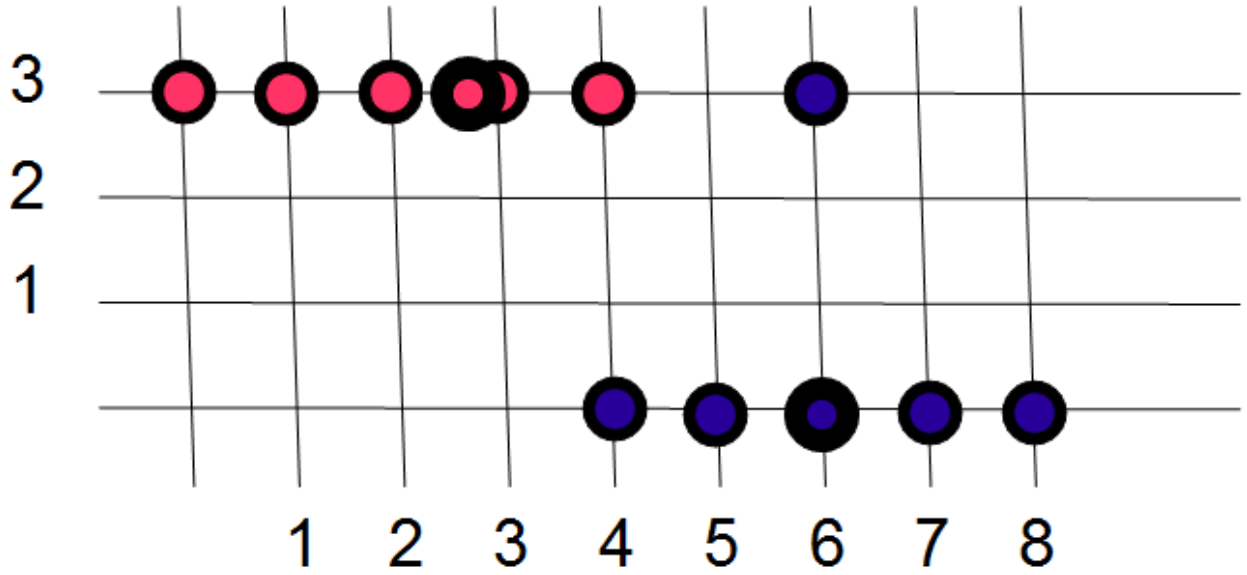
Clustering Example



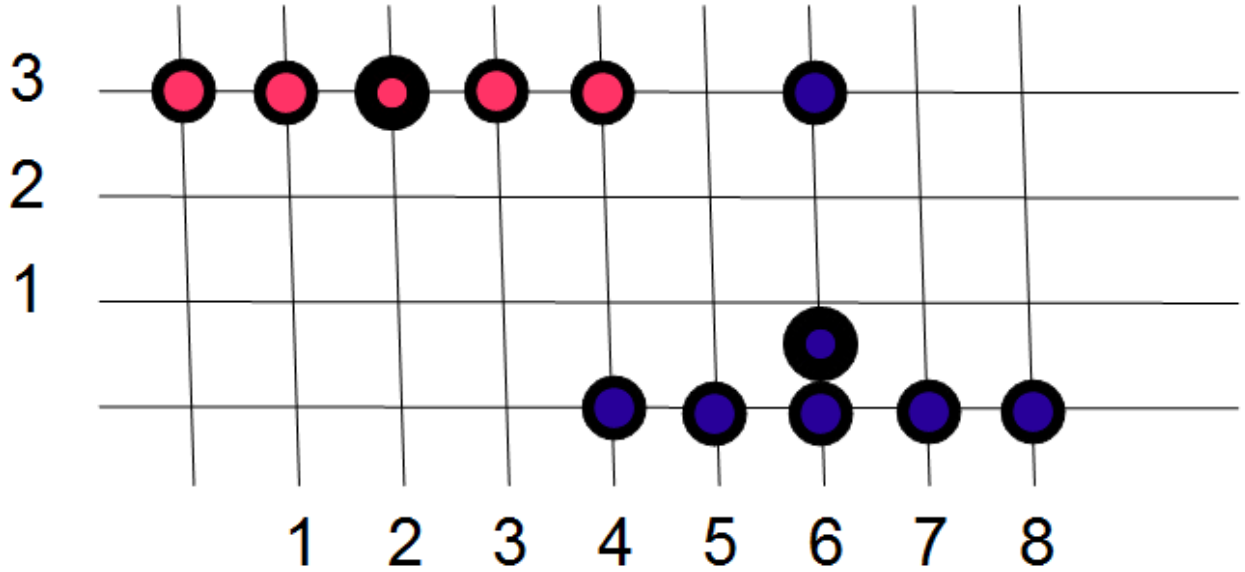
Clustering Example



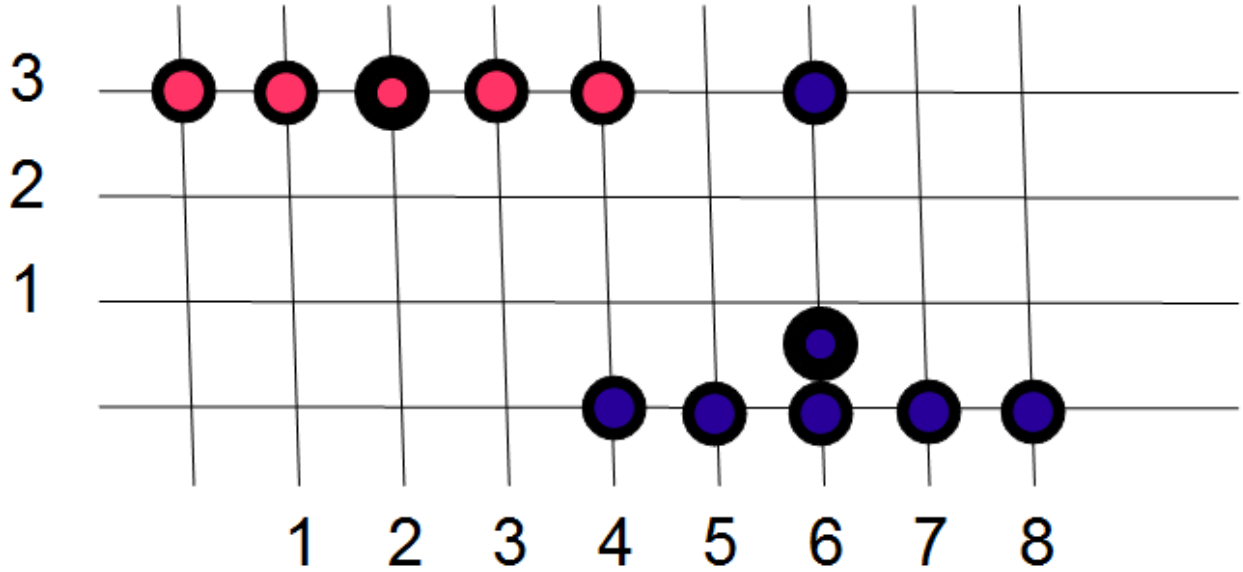
Clustering Example



Clustering Example



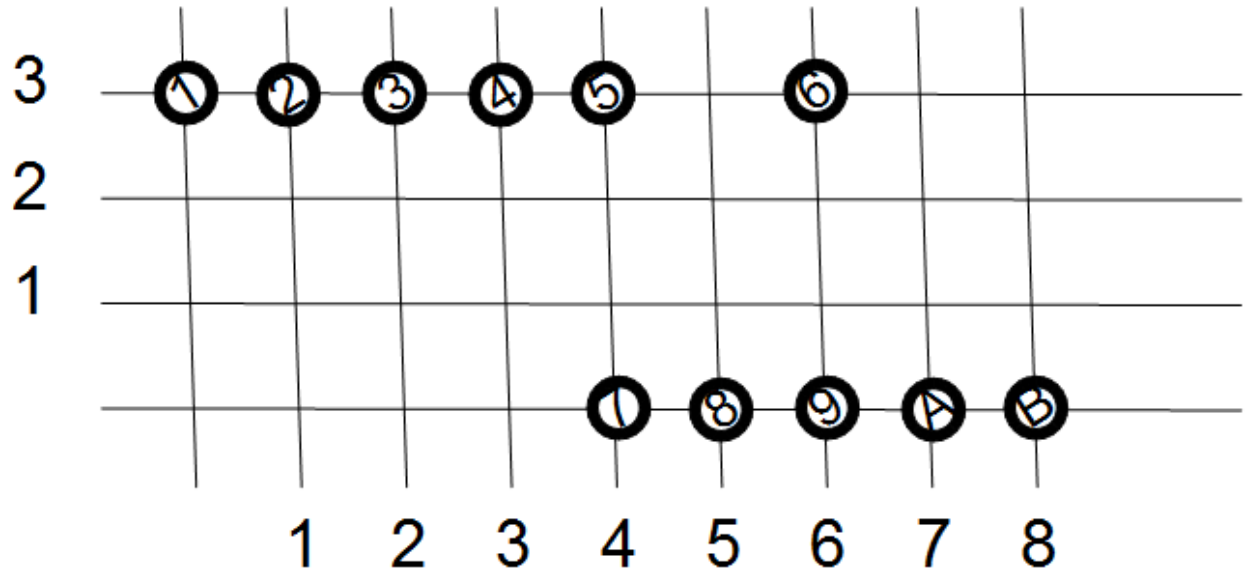
Clustering Example



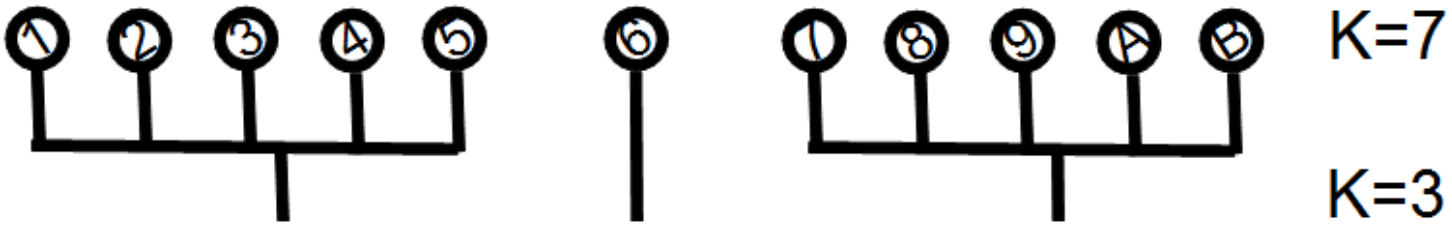
- Same as before, stop!

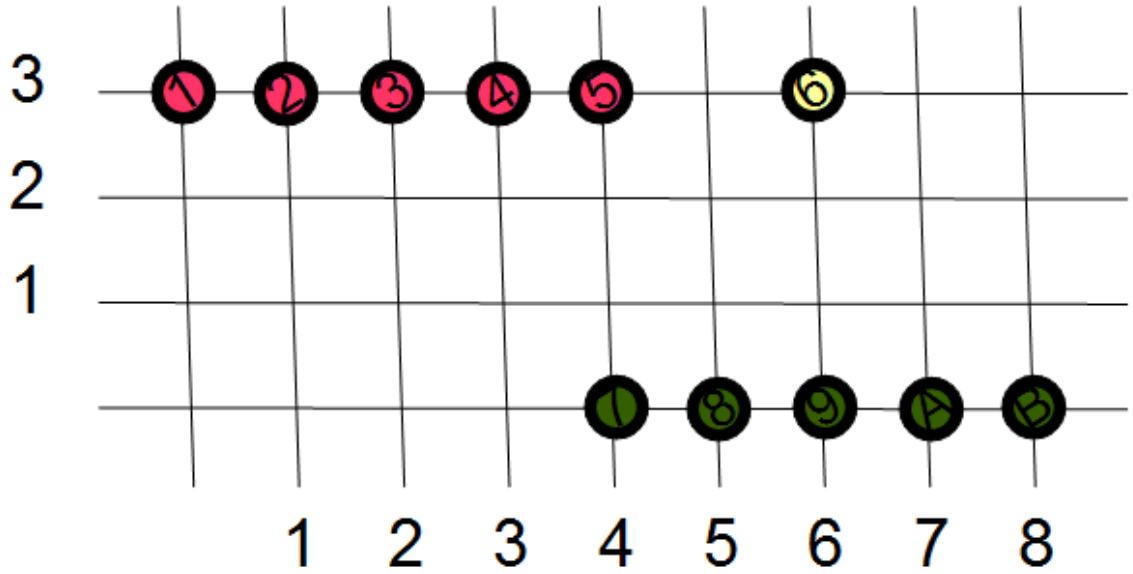
Clustering Example

- Use a hierarchical procedure with the nearest neighbor measure of inter-cluster similarity to find the clusters.



Clustering Example





Clustering Example