

# **Lecture 21**

## **Clustering (2)**

# Outline

- Similarity (Distance) Measures
- Distortion Criteria  
Scattering Criterion
- Hierarchical Clustering and other clustering methods

# Distance Measure

- Distance Measure – What does it mean “Similar”?

- Norm:  $d(x, y) = \|x - y\|_m = \left[ \sum_{i=1}^N (x_i - y_i)^m \right]^{1/m}$

- Mahalanobis distance:

$$d(x, y) = |x - y|^T S_{xy}^{-1} |x - y|$$

- Angle:  $d(x, y) = x^T y / (|x| \cdot |y|)$

Binary and symbolic features (x, y contains 0, 1 only):

- Tanimoto coefficient:  $d(x, y) = \frac{x^T y}{x^T x + y^T y}$

# Clustering Criteria

- Is the current clustering assignment good enough?  
Most popular one is the mean-square error distortion measure

$$D = \sum_{i=1}^c \sum_{k=1}^n I(x_k, i) \|x_k - W(i)\|^2$$

$$= \sum_{i=1}^c \left( \frac{1}{N_i} \sum_{x, y \in C(i)} \|x - y\|^2 \right), \quad N_i = \sum_{k=1}^n I(x_k, i)$$

- Other distortion measures can also be used:

$$D = \sum_{i=1}^c \left( \frac{1}{N_i} \sum_{x, y \in C(i)} d(x, y) \right) \quad D = \sum_{i=1}^c \left( \frac{1}{N_i} \text{Min. } d(x, y) \right)$$

# Scatter Matrices

- Scatter matrices are defined in the context of analysis of variance in statistics.
- They are used in linear discriminant analysis.
- However, they can also be used to gauge the fitness of a particular clustering assignment.
- Mean vector for i-th cluster:

$$m_i = \frac{1}{N_i} \sum_{k=1}^N I(x_k, i) x_k$$

- Total mean vector

$$m = \frac{1}{N} \sum_{i=1}^c N_i m_i = \frac{1}{N} \sum_{k=1}^N x_k$$

- Scatter matrix for i-th cluster:

$$S_i = \sum_{k=1}^N I(x_k, i) [(x_k - m_i)(x_k - m_i)^T]$$

- Within-cluster scatter matrix

$$S_W = \sum_{i=1}^c S_i$$

- Between-cluster scatter matrix

$$S_B = \sum_{i=1}^c N_i [(m_i - m)(m_i - m)^T]$$

# Scattering Criteria

- Total scatter matrix:

$$S_T = \sum_{k=1}^N [(x_k - m)(x_k - m)^T]$$

$$= S_W + S_B$$

- Note that the total scatter matrix is independent of the assignment  $I(x_k, i)$ . But ...
- $S_W$  and  $S_B$  both depend on  $I(x_k, i)$ !
- Desired clustering property
  - $S_W$  small
  - $S_B$  large

- How to gauge  $S_W$  is small or  $S_B$  is large?  
There are several ways.

- $\text{Tr. } S_W$  (trace of  $S_W$ ): Let

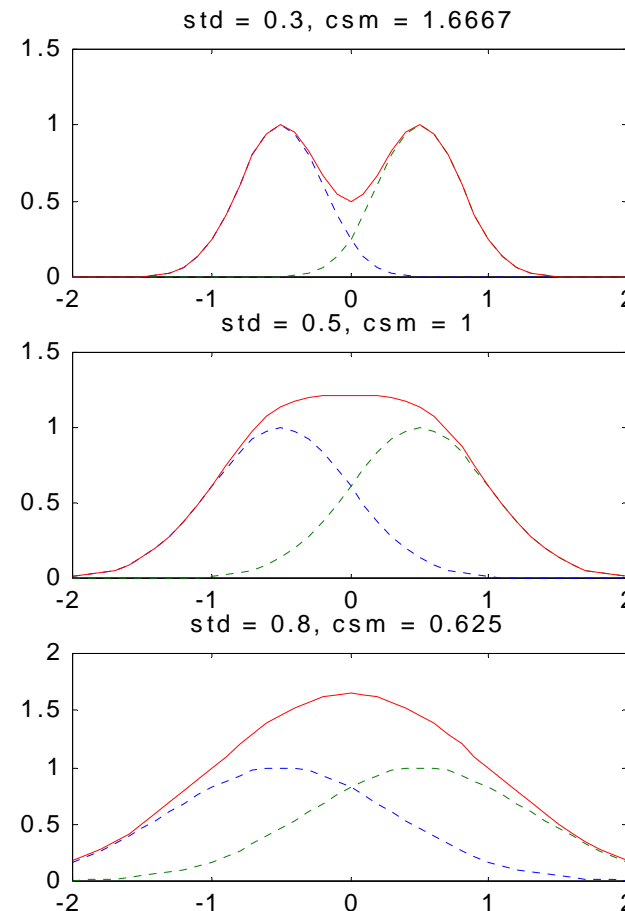
$$S_W = \sum_{m=1}^M \lambda_m v_m v_m^T$$

be the eigenvalue decomposition of  $S_W$ , then

$$\begin{aligned} \text{Tr. } S_W &= \sum_{m=1}^M \lambda_m = \sum_{i=1}^c \text{Tr. } S_i \\ &= \sum_{i=1}^c \sum_{k=1}^N I(x_k, i) \|x_k - m_i\|^2 = D \end{aligned}$$

# Cluster Separating Measure (CSM)

- Similar to scattering criteria.
- $csm = (m_i - m_j) / (\sigma_i + \sigma_j)$
- The larger its value, the more separable the two clusters.
- Assume underlying data distribution is Gaussian.



# Hierarchical Clustering

- Merge Method:

Initially, each  $x_k$  is a cluster. During each iteration, nearest pair of distinct clusters are merged until the number of clusters is reduced to 1.

- How to measure distance between two clusters:

$$d_{\min}(C(i), C(j)) = \min. d(x,y); \quad x \in C(i), y \in C(j)$$

– leads to *minimum spanning tree*

$$d_{\max}(C(i), C(j)) = \max. d(x,y); \quad x \in C(i), y \in C(j)$$

$$d_{\text{avg}}(C(i), C(j)) = \frac{1}{N_i N_j} \sum_{x \in C(i)} \sum_{y \in C(j)} d(x, y)$$

$$d_{\text{mean}}(C(i), C(j)) = m_i - m_j$$

# Hierarchical Clustering (II)

## Split method:

- Initially, only one cluster. Iteratively, a cluster is split into two or more clusters, until the total number of clusters reaches a predefined goal.
- The scattering criterion can be used to decide how to split a given cluster into two or more clusters.
- Another way is to perform a m-way clustering, using, say, k-means algorithm to split a cluster into m smaller clusters.