KHU-BME
Pattern Classification
Lecture 16

Unsupervised Classification (DHS Ch. 10)

Outline

- **Similarity and Similarity Measures**
- **Chain Method of Clustering**
- **Clustering Criterion Functions**
  - **Sum of Squared Error Criterion**
  - **Related Minimum-variance Criteria**
  - **Scattering Criteria**
    - **Trace criterion**
    - **Determinant criterion**
    - **Another possible criterion**

- **Iterative Optimization**
    - **For sum-of-squared-error criterion**
- **Clustering procedure – basic min. squared error**

- **K-means Clustering**

- **Hierarchical Clustering**
  - **Agglomerative Hierarchical Clustering**
  - **Nearest Neighbor Algorithm (Single Linkage)**
  - **Furthest Neighbor Algorithm (Complete Linkage)**

- **Mixture Densities: Gaussian Mixtures***

- **Component Analysis: PCA, ICA***

***special topics**

Unsupervised Classification (DHS 10.1)

Don't know which class each prototype belongs to.

Why?

1. Collection and labeling of prototypes is often expensive and time-consuming
2. Feature characteristics may not be stationary in time
3. It may be desirable to study structures of the data. New Information or discovery of subclasses may alter the decision
4. Perform exploratory data analysis
5. Classification in the reverse direction

Ex: Classifying unknown targets

1. Criterion function methods
   - Sum of squared error
   - Min. variance
   - Scatter matrices
   - Optimization problem
2. Heuristic methods
   - Chain
   - Hierarchical
   - Min. spanning tree
3. Unmixing methods
   - PCA
   - Gaussian mixture
   - ICA

Data Description and Clustering (DHS 10.6)

Similarity Measures (DHS 10.6.1)

Figure 10.6 shows four different data sets. All have the same mean and covariance matrix, yet their distributions are different. => mixture density ideas

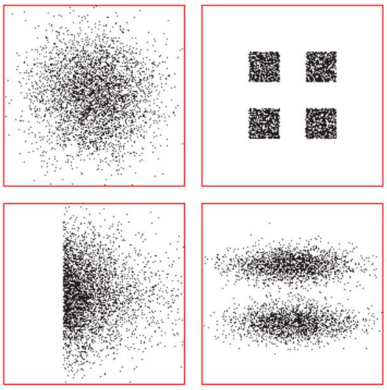


FIGURE 10.6. These four data sets have identical statistics up to second-order—that is, the same mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In such cases it is important to include in the model more parameters to represent the structure more completely. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Major Questions

1. How to measure similarity between samples?
2. How to evaluate partitioning into clusters?

Similarity

Samples: $x_1, x_2, \ldots, x_J$

Ex: Distance measures

$$d(x_i,x_j) = \{ \sum_{n=1}^{N}[x_i(n)-x_j(n)]^2 \}^{1/2} \qquad \text{(Euclidean distance)}$$

$$d'(x_i,x_j) = \sum_{n=1}^{N}\left|x_i(n)-x_j(n)\right| \qquad \text{(Manhattan distance or absolute distance)}$$

Clustering example: compare distance to a threshold, d_o

Figure 10.7: Effect of d_o as the distance threshold for clustering.

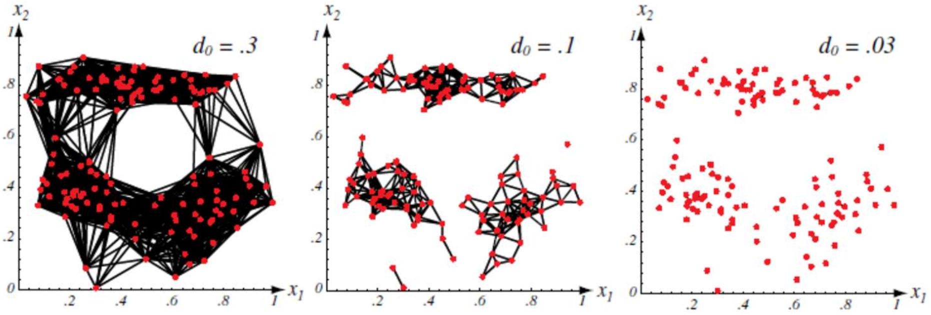


FIGURE 10.7. The distance threshold affects the number and size of clusters in similarity based clustering methods. For three different values of distance d_0, lines are drawn between points closer than d_0—the smaller the value of d_0, the smaller and more numerous the clusters. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Clustering by Euclidian distance is invariant to translations and rotations in feature space, but variant to linear transformations in general.

Scaling

Scaling problems can often be avoided using normalized distance measures, e.g. Mahalanobis distance.

Figure 10.8. Effect of scaling for clustering.

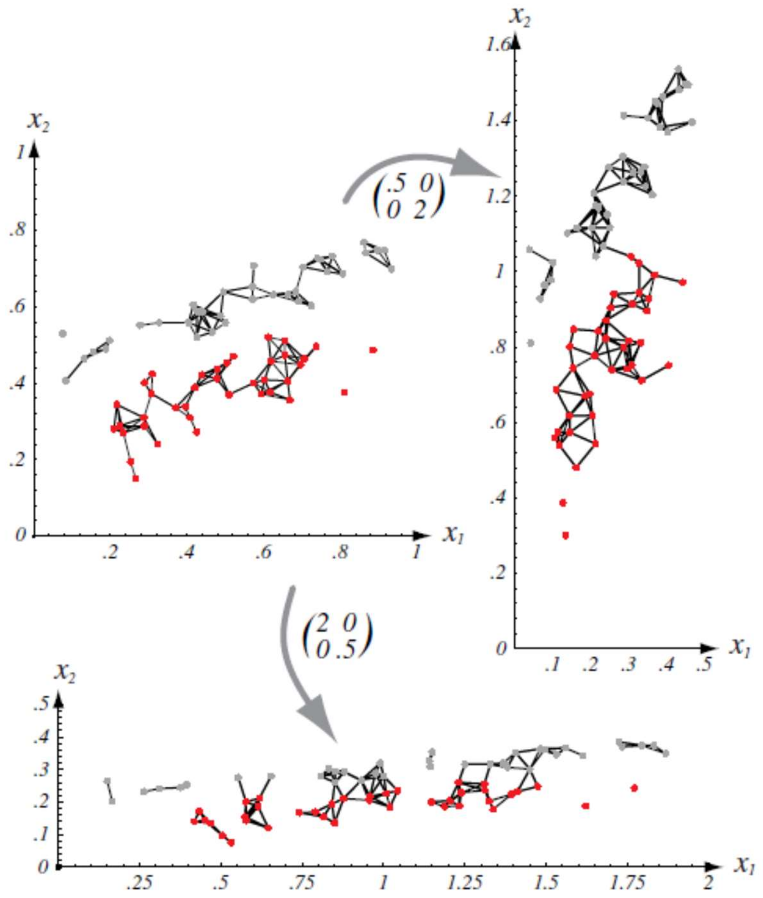


FIGURE 10.8. Scaling axes affects the clusters in a minimum distance cluster method. The original data and minimum-distance clusters are shown in the upper left; points in one cluster are shown in red, while the others are shown in gray. When the vertical axis is expanded by a factor of 2.0 and the horizontal axis shrunk by a factor of 0.5, the clustering is altered (as shown at the right). Alternatively, if the vertical axis is shrunk by a factor of 0.5 and the horizontal axis is expanded by a factor of 2.0, smaller more numerous clusters result (shown at the bottom). In both these scaled cases, the assignment of points to clusters differ from that in the original space. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Figure 10.9. Effect of normalization

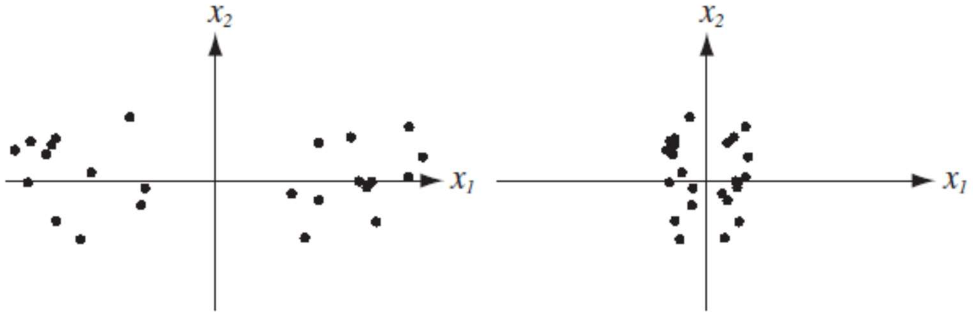


FIGURE 10.9. If the data fall into well-separated clusters (left), normalization by scaling for unit variance for the full data may reduce the separation, and hence be undesirable (right). Such a normalization may in fact be appropriate if the full data set arises from a single fundamental process (with noise), but inappropriate if there are several different processes, as shown here. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Other Similarity Measures

Angular Similarity

Introduce a nonmetric similarity function S to compare x_i and x_j

Similarity, $S(x_i,x_j)= x_i^T x_j \ / \ ||x_i||\ ||x_j||$ (normalized inner product, cosine of the angle between x_i and x_j)

Useful if angle between two vectors is meaningful.

Binary Similarity

For binary features

S is a measure of the relative possession of common attributes

$S(x_i,x_j)= x_i^T x_j \ / \ x_i^T x_i + x_j^T x_j – x_i^T x_j$

$x_i^T x_j$=No. of shared attributes

S=Percentage of attributes that are shared.

Chain Method of Clustering

1. First sample assigned to cluster #1
2. Compute distance d of the next sample to the previous sample and compare to d_o (a threshold pre-specified). If $d<d_o$ assign sample to the same (first) cluster; otherwise form a new cluster for the sample.
3. Compute distance d of the next sample to all existing clusters. Find min. distance d_{min} if $d_{min}<d_o$, assign to corresponding cluster. Otherwise form a new cluster. (1-pass algorithm)

Distance d to cluster = distance to the first sample of cluster.
or variation = distance to the sample mean of cluster.

However, it is (1) sensitive to d_o and (2) order of samples.

Variations:

- Could perform multiple passes, each with different d_o
- Could use cluster means for d

So far, how to measure "similarity," now the criterion functions.

Clustering Criterion Functions (DHS 10.7)

A set z of J samples x_1, …, x_j.

Divide z into K subsets z_1, z_2, …., z_k.

Define a criterion function to measure clustering quality of any partition of the J samples into K subsets. Find extremum (min. or max.) of the criterion function.

1. Sum of Squared Error Criterion (DHS 10.7.1)

Simplest and most widely used.

Let J_i = No. of samples in z_i

m_i = mean of samples in z_i

$$m_i = (1/J_i) \sum_{x_j \in z_i} x_j$$

Sum of Squared Error:

$$J_e = \sum_{i=1}^{K} \sum_{x_j \in z_i} \|x_j - m_i\|^2$$

Smallest J_e gives the minimum variance clustering. So it is good for the clusters that for compact clouds and well separated from one another.

Good for the clusters from compact clouds that are rather well-separated from one another

But look at Figure 10.10

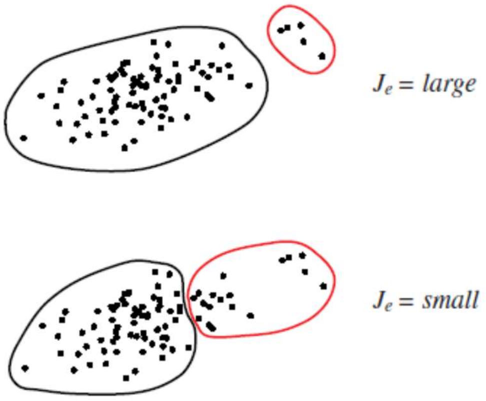


FIGURE 10.10. When two natural groupings have very different numbers of points, the clusters minimizing a sum-squared-error criterion J_e of Eq. 54 may not reveal the true underlying structure. Here the criterion is smaller for the two clusters at the bottom than for the more natural clustering at the top. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

2. Minimum-Variance Criteria (DHS 10.7.2)

Can re-write J_e as

$$J_e=(1/2)\sum_{i=1}^{K} J_i\bar{s}_i$$

where $\bar{s}_i = (1/J_i^2)\sum_{x_j\in z_i}\sum_{x_l\in z_i}\|x_j-x_l\|^2$, the average squared distance between any two points in the i-th cluster

Similarly, $\bar{s}_i$ could be replaced by

Different choices:

$$\bar{s}_i = (1/J_i^2)\sum_{x_j\in z_i}\sum_{x_l\in z_i} s(x_j, x_l)$$

where s=a similarity function.

or

$$\bar{s}_i = \min_{x_j,x_l\in z_i}[s(x_j,x_l)]$$

Objective: find extremum of criterion function

3. Scattering Criteria (DHS 10.7.3)

Form matrices S_W and S_B to measure scattering of samples. Form a criterion function from S_W and/or S_B.

Mean of i-th cluster: $m_i= (1/J_i)\sum_{x_j\in z_i} x_j$

Total mean vector: $m=(1/J)\sum x_j = (1/J)\sum_{i=1}^{K} J_i m_i$

Scatter matrix for i-th cluster: $S_i= \sum_{x_j\in z_i}[x_j-m_i]\,[x_j-m_i]^T$

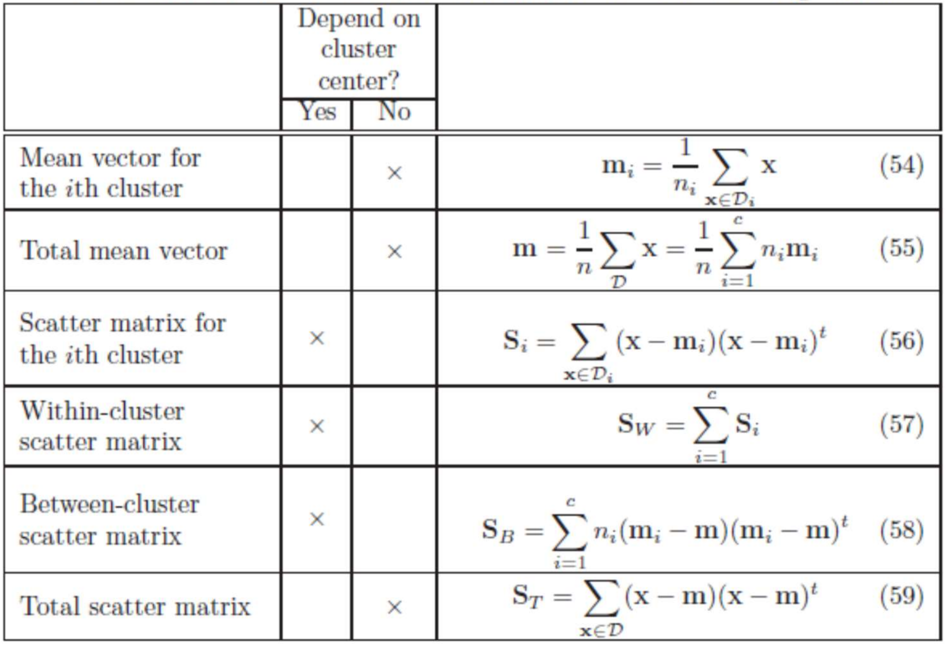
Within-class scatter matrix: $S_W=\sum_{i=1}^{K} S_i$, NxN matrix

Between-cluster scatter matrix: $S_B=\sum_{i=1}^{K} J_i[m_i-m]\,[m_i-m]^T$, NxN matrix

Possible pitfall: if no. of clusters = no. of samples, $S_i = S_W = 0$

Total scatter matrix: $S_T = \sum_{x_j \in z_i} [x_j - m]\,[x_j - m]^T$

Table 10.1: Mean vectors and scatter matrices used in clustering criteria.



	Depend on cluster center?		
	Yes	No	
Mean vector for the ith cluster		×	$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}$ (54)
Total mean vector		×	$\mathbf{m} = \frac{1}{n} \sum_{\mathcal{D}} \mathbf{x} = \frac{1}{n} \sum_{i=1}^{c} n_i \mathbf{m}_i$ (55)
Scatter matrix for the ith cluster	×		$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$ (56)
Within-cluster scatter matrix	×		$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i$ (57)
Between-cluster scatter matrix	×		$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$ (58)
Total scatter matrix		×	$\mathbf{S}_T = \sum_{\mathbf{x} \in \mathcal{D}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$ (59)

Can show

$S_T = S_W + S_B$

1. S_T is independent of clustering or partition of data.
2. As S_W increases, S_B decreases.
3. Need a scalar measure of S_B or S_W to maximize or minimize

Scalar Measure Option 1: Trace Criterion

$$\mathrm{Tr}\{S_W\} = \sum_{i=1}^{K} \mathrm{Tr}\{S_i\} = \sum_{i=1}^{K} \sum_{x_j \in z_i} \|x_j - m_i\|^2$$

$\mathrm{Tr}\{S_W\} = J_e$ = sum of squared errors.

$\mathrm{Tr}\{S_T\} = \mathrm{Tr}\{S_W\} + \mathrm{Tr}\{S_B\}$

Min. $\mathrm{Tr}\{S_W\} \Leftrightarrow$ Max. $\mathrm{Tr}\{S_B\}$

$$\mathrm{Tr}\{S_B\} = \sum_{i=1}^{K} J_i \|m_i - m\|^2$$

Scalar Measure Option 2: Determinant Criterion

$J_d=|S_W|=\sum_{i=1}^{K}|S_i|$

$|S_B|$ could maximize this

$S_B=\sum_{i=1}^{K}J_i[m_i-m]\,[m_i-m]^T$

Exception: singular if $K\leq N$, $\therefore |S_B|$ not useful, poor choice

$S_i=\sum_{x_j\in z_i}[x_j-m_i]\,[x_j-m_i]^T$

Max. Rank of $S_i=J_i-1$

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

Max. rank of S_W is $(\sum_{i=1}^{K}J_i)-K=J-K$

If $N>J-K$, S_W is singular

J_d is invariant to a linear transformation of coordinate (J_e is not)

Let $x_j' = Tx_j$ T = linear transformation

Then $S_i' = TS_iT^T$

$S_W' = TS_WT^T$

$m_i' = Tm_i$

$J_d' = |S_W'| = |TS_WT^T| = |T||S_W||T^T|$

$|T|$ = same for all partitions.

Scalar Measure Option 3: Another possibility (Invariant Criteria)

$J_f=Tr\{S_W^{-1}S_B\}=\sum_{n=1}^{N}\lambda_n$ (since the trace of a matrix is the sum of its eigenvalues)

where λ_n = n-th eigenvalue of $S_W^{-1}S_B$

is also invariant to linear transformations.

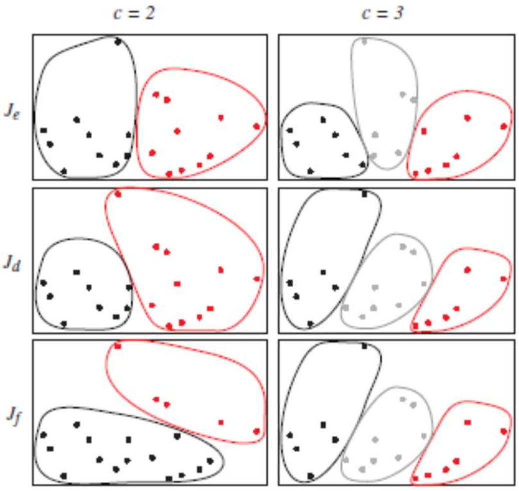
Example 3: Clustering criteria

We can gain some intuition by considering these criteria applied to the following data set.

sample	x_1	x_2
1	-1.82	0.24
2	-0.38	-0.39
3	-0.13	0.16
4	-1.17	0.44
5	-0.92	0.16
6	-1.69	-0.01
7	0.33	-0.17
8	-0.71	-0.21
9	1.27	-0.39
10	-0.16	-0.23

sample	x_1	x_2
11	0.41	0.91
12	1.70	0.48
13	0.92	-0.49
14	2.41	0.32
15	1.48	-0.23
16	-0.34	1.88
17	0.83	0.23
18	0.62	0.81
19	-1.42	-0.51
20	0.67	-0.55

All of the clusterings seem reasonable, and there is no strong argument to favor one over the others. For the case $c = 2$, the clusters minimizing the J_e indeed tend to favor clusters of roughly equal numbers of points, as illustrated in Fig. 10.9; in contrast, J_d favors one large and one fairly small cluster. Since the full data set happens to be spread horizontally more than vertically, the eigenvalue in the horizontal direction is greater than that in the vertical direction. As such, the clusters are “stretched” horizontally somewhat. In general, the differences between the cluster criteria become less pronounced for large numbers of clusters. For the $c = 3$ case, for instance, the clusters depend only mildly upon the cluster criterion — indeed, two of the clusterings are identical.



The clusters found by minimizing a criterion depends upon the criterion function as well as the assumed number of clusters. The sum-of-squared-error criterion J_e (Eq. 49), the determinant criterion J_d (Eq. 63) and the more subtle trace criterion J_f (Eq. 65) were applied to the 20 points in the table with the assumption of $c = 2$ and $c = 3$ clusters. (Each point in the table is shown, with bounding boxes defined by $-1.8 < x < 2.5$ and $-0.6 < y < 1.9$.)