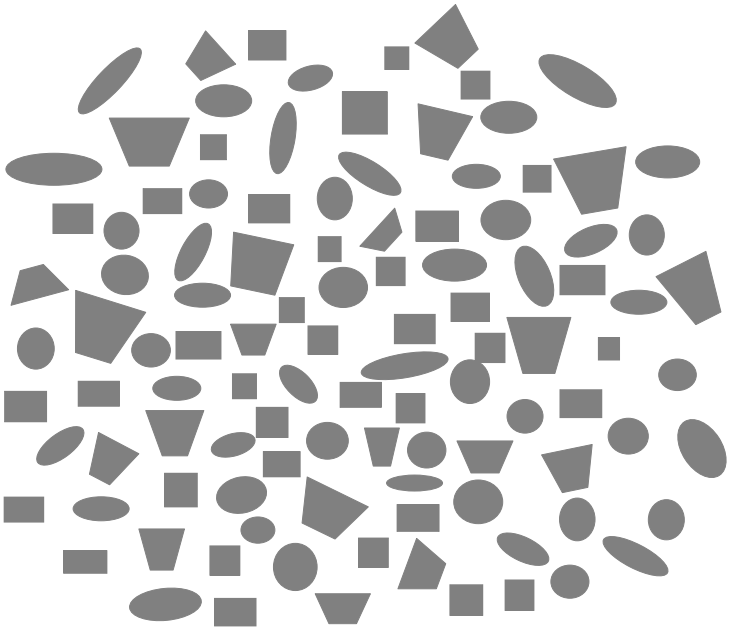
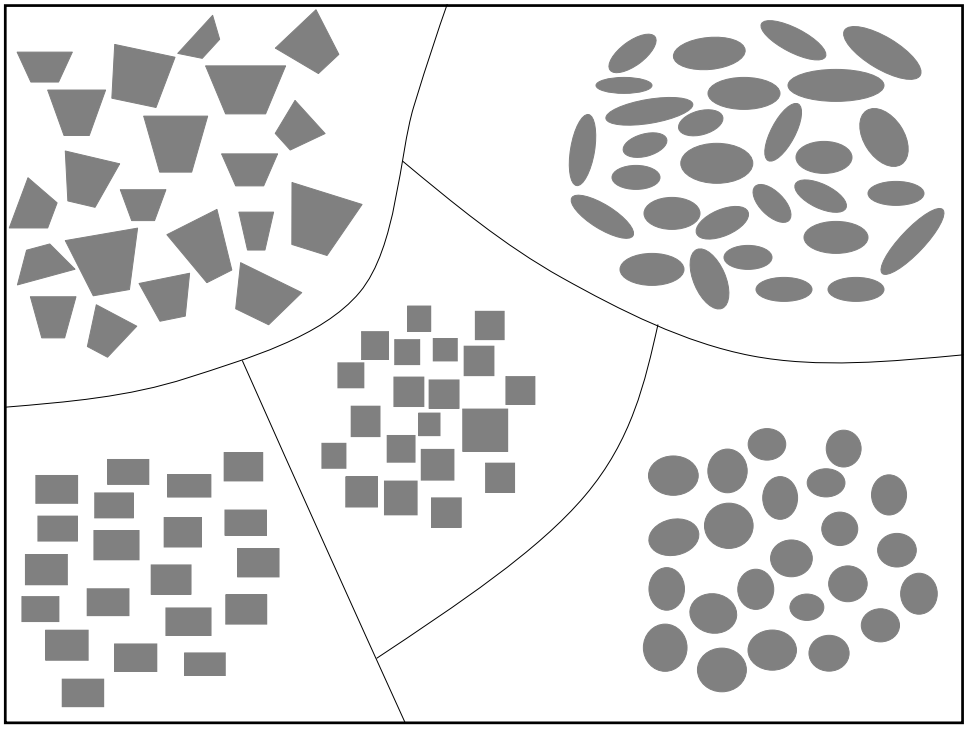
Unsupervised Learning

– A gentle introduction to clustering analysis

Pengyu Hong

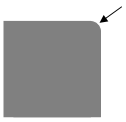
How Many Different Types of Shapes?



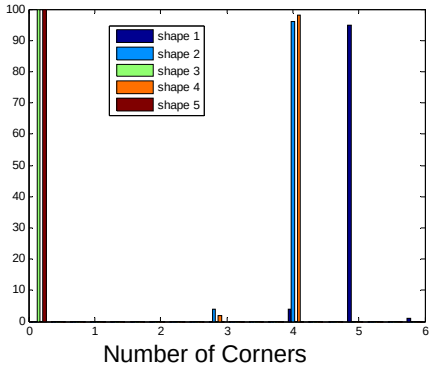


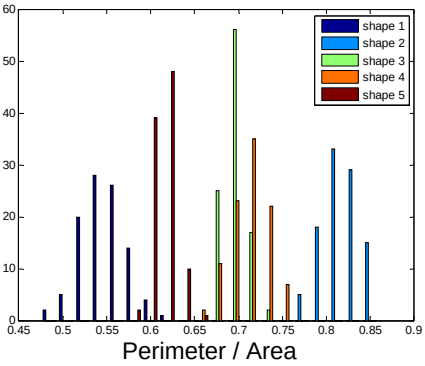
Features

- Extract features to represent objects
  - Corner
  - Perimeter / Area
- Noise

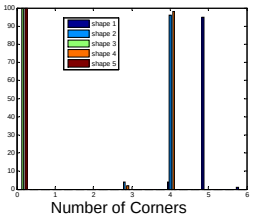


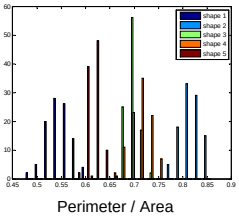
Data Distribution

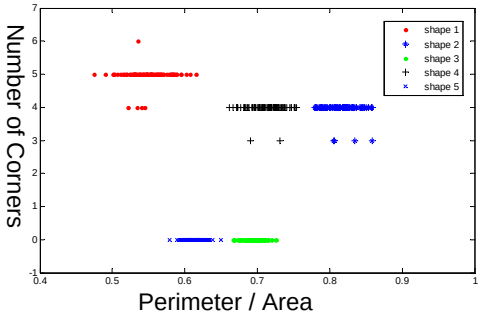




Data Distribution







Clustering Analysis

- Clustering is one of the most important *unsupervised learning* processes that organizing objects into groups whose members are similar in some way.
- Clustering finds *structures* in a collection of unlabeled data.
- A *cluster* is a collection of objects which are *similar* between them and are *dissimilar* to the objects belonging to other clusters.

Key Components of Clustering Analysis

- Data representation $\vec{x} = \begin{bmatrix} num_corner \\ perimeter / area \end{bmatrix}$
- Similarity measurement
- Clustering method

Similarity Measurements

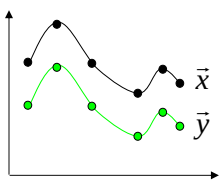
- Pearson Correlation

Two profiles (vectors) $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$ and $\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$

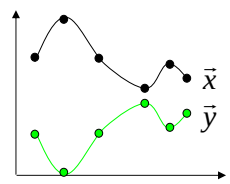
$$C_{pearson}(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{p}(x_i - m_x)(y_i - m_y)}{\sqrt{[\sum_{i=1}^{p}(x_i - m_x)^2][\sum_{i=1}^{p}(y_i - m_y)^2]}}$$

$$m_x = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$m_y = \frac{1}{N}\sum_{n=1}^{N} y_n$$

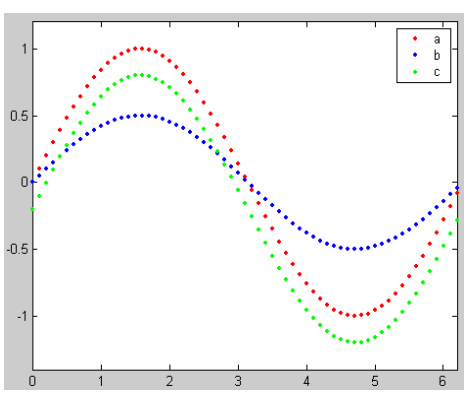


$+1 \geq$ Pearson Correlation ≥ -1



Similarity Measurements

- Pearson Correlation: Trend Similarity



$$\vec{b} = 0.5\vec{a}$$

$$\vec{c} = \vec{a} - 0.2$$

$$C_{pearson}(\vec{a},\vec{b}) = 1$$

$$C_{pearson}(\vec{a},\vec{c}) = 1$$

$$C_{pearson}(\vec{b},\vec{c}) = 1$$

Similarity Measurements

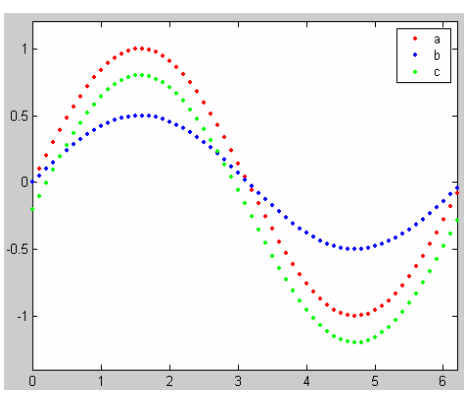
- Euclidean Distance

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$$

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{n=1}^{p} (x_n - y_n)^2}$$

Similarity Measurements

- Euclidean Distance: Absolute difference



$$\vec{b} = 0.5\vec{a}$$

$$\vec{c} = \vec{a} - 0.2$$

$$d(\vec{a}, \vec{b}) = 2.8025$$

$$d(\vec{a}, \vec{c}) = 1.5875$$

$$d(\vec{b}, \vec{c}) = 3.2211$$

Similarity Measurements

- Cosine Correlation

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$$

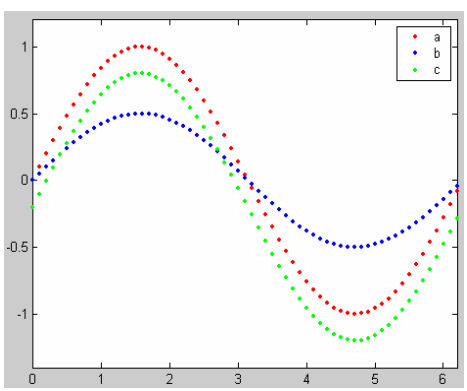
$$C_{\text{cosine}}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{p} x_i \times y_i}{\|\vec{x}\| \times \|\vec{y}\|}$$

the angle between two vectors

$\vec{x} = \vec{y}$ $\quad +1 \geq \text{Cosine Correlation} \geq -1 \quad$ $\vec{x} = -\vec{y}$

Similarity Measurements

- Cosine Correlation: Trend + Mean Distance



$$\vec{b} = 0.5\vec{a}$$

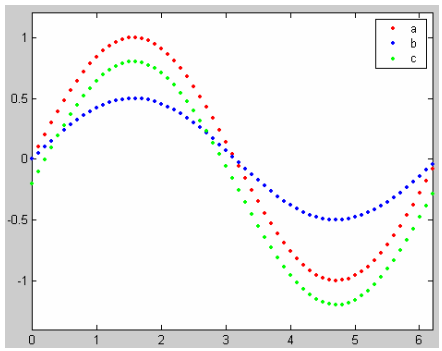
$$\vec{c} = \vec{a} - 0.2$$

$$C_{\text{cosine}}(\vec{a}, \vec{b}) = 1$$

$$C_{\text{cosine}}(\vec{a}, \vec{c}) = 0.9622$$

$$C_{\text{cosine}}(\vec{b}, \vec{c}) = 0.9622$$

Similarity Measurements



$\vec{b} = 0.5\vec{a}$

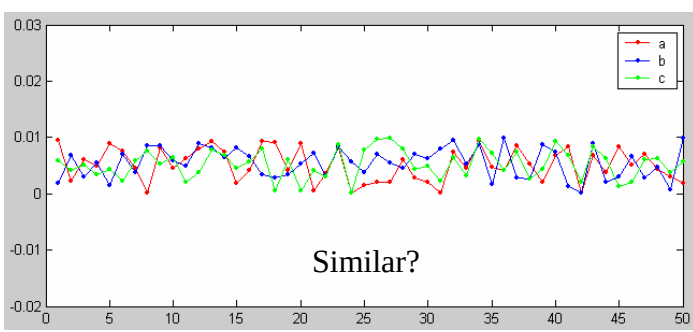
$\vec{c} = \vec{a} - 0.2$

$C_{pearson}(\vec{a},\vec{b}) = 1$ $\quad d(\vec{a},\vec{b}) = 2.8025$ $\quad C_{\cos ine}(\vec{a},\vec{b}) = 1$

$C_{pearson}(\vec{a},\vec{c}) = 1$ $\quad d(\vec{a},\vec{c}) = 1.5875$ $\quad C_{\cos ine}(\vec{a},\vec{c}) = 0.9622$

$C_{pearson}(\vec{b},\vec{c}) = 1$ $\quad d(\vec{b},\vec{c}) = 3.2211$ $\quad C_{\cos ine}(\vec{b},\vec{c}) = 0.9622$

Similarity Measurements

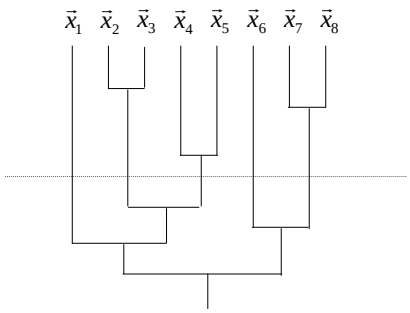


$C_{pearson}(\vec{a},\vec{b}) = -0.1175$ $\quad d(\vec{a},\vec{b}) = 0.0279$ $\quad C_{\cos ine}(\vec{a},\vec{b}) = 0.7544$

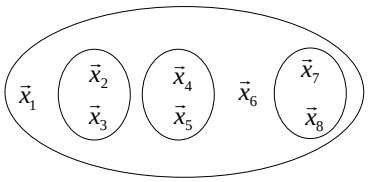
$C_{pearson}(\vec{a},\vec{c}) = 0.1244$ $\quad d(\vec{a},\vec{c}) = 0.0255$ $\quad C_{\cos ine}(\vec{a},\vec{c}) = 0.8092$

$C_{pearson}(\vec{b},\vec{c}) = 0.1779$ $\quad d(\vec{b},\vec{c}) = 0.0236$ $\quad C_{\cos ine}(\vec{b},\vec{c}) = 0.844$

Hierarchical Clustering



Dendrogram



Hierarchical Clustering (Cont.)

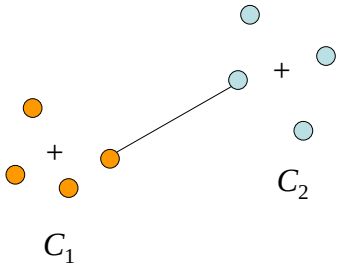
- Multilevel clustering: level 1 has *n* clusters → level *n* has one cluster.
- Agglomerative HC: starts with singleton and merge clusters.
- Divisive HC: starts with one sample and split clusters.

$\vec{x}_1 \quad \vec{x}_2 \quad \vec{x}_3 \quad \vec{x}_4 \quad \vec{x}_5 \quad \vec{x}_6 \quad \vec{x}_7 \quad \vec{x}_8$

Merge which pair of clusters?

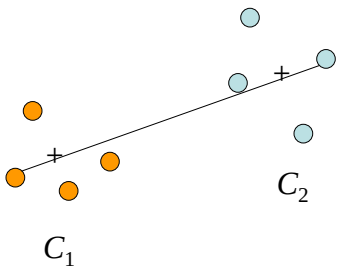
Nearest Neighbor Method: Starts with *n* nodes (*n* is the size of our sample), merges the 2 most similar nodes at each step, and stops when the desired number of clusters is reached.

Single Linkage



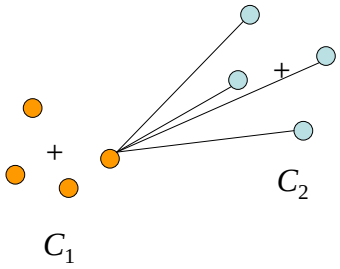
Dissimilarity between two clusters = Minimum dissimilarity between the members of two clusters

Complete Linkage



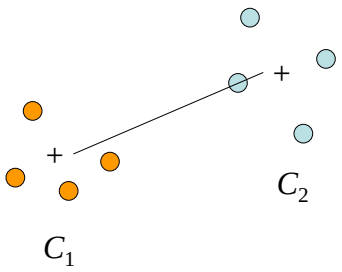
Dissimilarity between two clusters = Maximum dissimilarity between the members of two clusters

Average Linkage



Dissimilarity between two clusters = Averaged distances of all pairs of objects (one from each cluster).

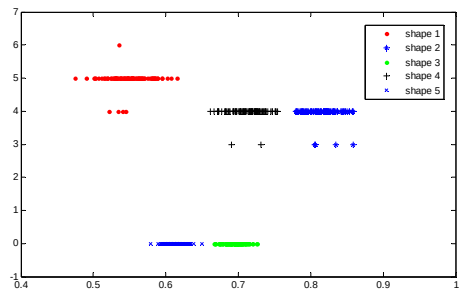
Centroid (Average Group) Linkage



Dissimilarity between two clusters = Distance between two cluster means.

Matlab Demo Clustering

Clustering Result → Classifier



$$P(shape \mid num_corner, perimeter_area)$$

$$= \frac{P(shape, num_corner, perimeter_area)}{P(num_corner, perimeter_area)}$$

$$= \frac{P(shape, num_corner, perimeter_area)}{\sum_{shape} P(shape, num_corner, perimeter_area)}$$

$$= \frac{P(num_corner, perimeter_area \mid shape) P(shape)}{\sum_{shape} P(shape, num_corner, perimeter_area)}$$

$$= \frac{P(num_corner \mid shape) P(perimeter_area \mid shape) P(shape)}{\sum_{shape} P(num_corner \mid shape) P(perimeter_area \mid shape) P(shape)}$$

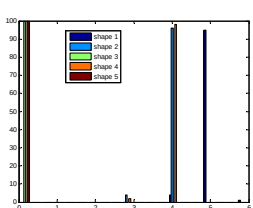
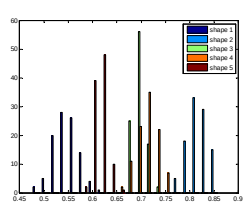
Clustering Result → Classifier

$$P(shape \mid num_corner, perimeter_area)$$

$$= \frac{P(num_corner \mid shape) P(perimeter_area \mid shape) P(shape)}{\sum_{shape} P(num_corner \mid shape) P(perimeter_area \mid shape) P(shape)}$$

$$P(shape)$$

$$P(num_corner \mid shape)$$



$$P(perimeter_area \mid shape)$$