Clustering

COMP4211

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

Supervised Learning vs Unsupervised Learning

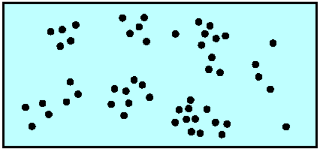
Supervised learning

- The learner is provided with a set of inputs together with the corresponding desired outputs
- Given training set: $(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \ldots, (\mathrm{x}_N; y_N)$
- Find a general function $y = h(\mathrm{x})$
- An approximation to a target (true) function $y = f(\mathrm{x})$
  - h: hypothesis

Unsupervised learning

- training examples as input patterns, with no associated output patterns
- Given training set $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N$
- unlabeled training examples
- no teacher

- find clusters



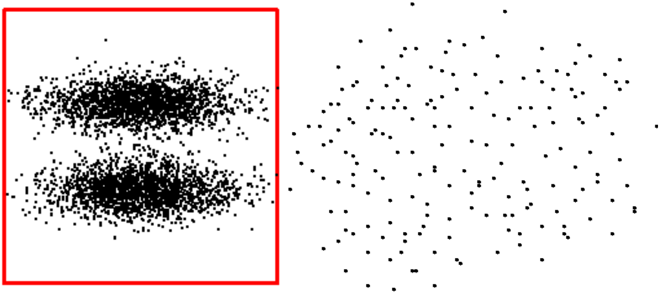
- in the early stages of an investigation, it may be helpful to perform exploratory data analysis to gain some insight into the nature or structure of the data

Problem

Given:

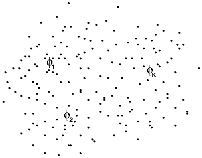
- $x_1, x_2, \ldots, x_n$
- they fall into k clusters

Determine: the cluster centers (centroids) $m_1, m_2, \ldots, m_k$

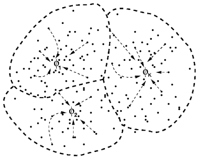


k-Means Clustering

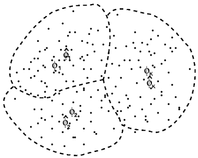
1. Make initial guesses for $m_1, m_2, \ldots, m_k$
    - usually, just randomly choose k of the examples



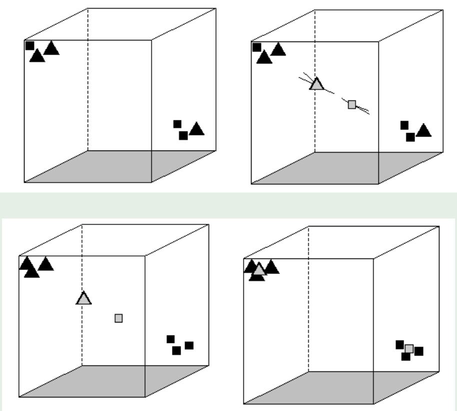
2. Use the estimated cluster centers to put the patterns into clusters
    - put x_j into cluster i if $\|x_j - m_i\|$ is the minimum of all the k distances
    - the feature space is partitioned into k clusters



3. for $i = 1$ to k, replace m_i with the mean of all examples for cluster i
4. Go back to step 2 until there are no changes in the m_i's



Example



(demo)

Distance Measures

- Euclidean distance: $d(\mathrm{x}, \mathrm{z}) = \sqrt{\sum_{i=1}^{n} (x_i - z_i)^2}$
- scaled Euclidean distance: $d(\mathrm{x}, \mathrm{z}) = \sqrt{\sum_{i=1}^{n} w_i (x_i - z_i)^2}$
- L_1 distance: $d(\mathrm{x}, \mathrm{z}) = \sum_{i=1}^{n} |x_i - z_i|$
- L_∞ distance: $d(\mathrm{x}, \mathrm{z}) = \max(|x_i - z_i|)$

Similarity Measures

similarity functions

- gives a large value when two feature vectors are similar

Example

Normalized inner product

$$s(x_1, x_2) = \frac{x_1' x_2}{\|x_1\| \cdot \|x_2\|}$$

- cosine of angle between vectors
- for binary-valued (0/1) features, the normalized inner product gives a relative count of features shared by the two vectors
- a simple variation is the fraction of features shared:

$$s(x_1, x_2) = \frac{x_1' x_2}{d}$$

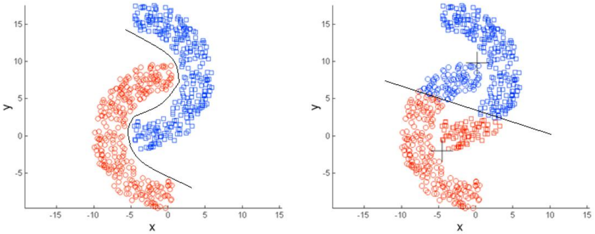
Issues

Different initialization means that you may get different clusters each time

- multiple runs
- pick the solution with minimum sum of squared error $\sum_{i=1}^{K} \sum_{\mathrm{x} \in C_i} \|\mathrm{x} - \mathrm{m}_i\|^2$

Implicit assumptions about the shapes of clusters

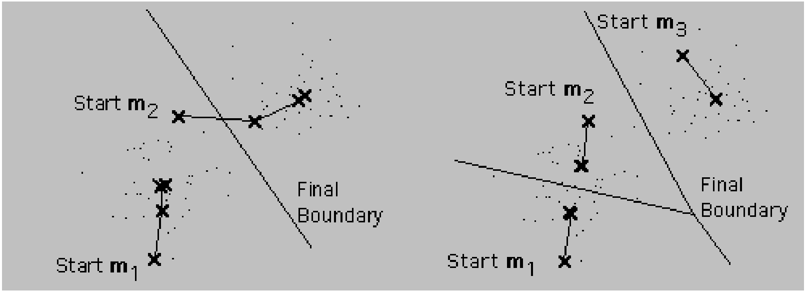
- can get wrong results when clusters have other shapes



Data points are assigned to only one cluster (hard assignment)

You have to pick the number of clusters

- in general, clustering result depends on k



(demo)