Probabilistic Model-Based Clustering

Probabilistic Clustering (why)

- In all the cluster analysis methods we have discussed so far, each data object can be assigned to at most one clusters.
- In what situations may a data object belong to more than one cluster?
  - Clustering product reviews: a customer review might relate to multiple products/services. If we want to cluster reviews per product/service, we must allow that a review can belong to many clusters.
  - Clustering to study user search intent: a user of an online store would typically perform some search. It is important to understand the search intent: searching for product, for customer support, for offers, etc. In one session however, the user may search with multiple intents.

Fuzzy Set and Fuzzy Cluster

Fuzzy cluster: A fuzzy set $F_S : X \rightarrow [0, 1]$ (value between 0 and 1)

Example: Popularity of cameras is defined as a fuzzy mapping

Camera	Sales (units)
A	50
B	1320
C	860
D	270

$$\text{Pop}(o) = \begin{cases} 1 & \text{if } 1,000 \text{ or more units of } o \text{ are sold} \\ \frac{i}{1000} & \text{if } i \ (i < 1000) \text{ units of } o \text{ are sold} \end{cases}$$

Function *pop()* defines a fuzzy set of popular digital cameras. The fuzzy set of digital cameras according to Pop() is {*A*(0.05), *B*(1), *C*(0.86), *D*(0.27)}, where the degree of membership is written in parentheses.

In fuzzy clustering, a cluster is a fuzzy set of objects that belong to this cluster. The degree of membership of every object indicates how strong this object is related to this cluster.

Complexity of Fuzzy Clustering Algorithm

Fuzzy (Soft) Clustering

Formally, given a set of objects, $o_1, \ldots, o_n$, a **fuzzy clustering** of k **fuzzy clusters**, $C_1, \ldots, C_k$, can be represented using a **partition matrix**, $M = [w_{ij}]$. where w_{ij} is the membership degree of o_i in fuzzy cluster C_j. The partition matrix should satisfy the following three requirements:

P1: for each object o_i and cluster C_j, $0 \le w_{ij} \le 1$ (fuzzy set).

P2: for each object o_i, $\sum_{j=1}^{k} w_{ij} = 1$. equal participation in the clustering

P3: for each cluster C_j, $0 < \sum_{i=1}^{n} w_{ij} < n$ ensures there is no empty cluster.

Fuzzy (Soft) Clustering

Example: Let cluster features be

C_1 :“digital camera” and “lens”

C_2: “computer“

$$w_{ij} = \frac{|R_i \cap C_j|}{|R_i \cap (C_1 \cup C_2)|} = \frac{|R_i \cap C_j|}{|R_i \cap \{digital\ camera, lens, computer\}|}.$$

Review-id	Keywords
R_1	digital camera, lens
R_2	digital camera
R_3	lens
R_4	digital camera, lens, computer
R_5	computer, CPU
R_6	computer, computer game

$$M = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \frac{2}{3} & \frac{1}{3} \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Fuzzy (Soft) Clustering

How to evaluate how well a fuzzy clustering describes a data set ?

Let c_1, …, c_k as the center of the k clusters

For an object o_i, sum of the squared error (SSE), p is a parameter:

$$\mathrm{SSE}(o_i) = \sum_{j=1}^{k} w_{ij}^p dist(o_i, c_j)^2$$

For a cluster C_j, SSE:

$$\mathrm{SSE}(C_j) = \sum_{i=1}^{n} w_{ij}^p dist(o_i, c_j)^2$$

For the whole clustering:

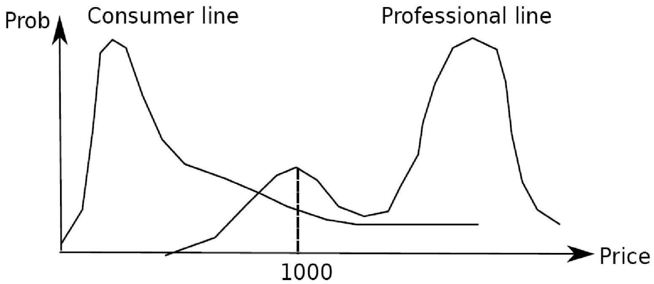
$$\mathrm{SSE}(\mathcal{C}) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^p dist(o_i, c_j)^2$$

Probabilistic Model-Based Clustering

- Is there Algorithm to detect probabilistic clusters in the data ?
- A data set that is the subject of cluster analysis can be regarded as a sample of the possible instances of the hidden clusters, but without any cluster labels.
- Statistically, we can assume that a hidden cluster is a distribution over the data space, which can be mathematically represented using a probability density function (or distribution function).
- For a probabilistic cluster, C, its probability density function, f, and a point, o, in the data space, $f(o)$ is the relative likelihood that an instance of C appears at o.

Probabilistic Model-Based Clustering

- Example: Suppose that there are 2 categories for digital cameras sold
  - consumer line vs. professional line
  - density functions f_1, f_2 for C_1, C_2
  - obtained by probabilistic clustering



- For a price value of, say, \$1000, $f_1(1000)$ is the relative likelihood that the price of a consumer-line camera is \$1000. Similarly, $f_2(1000)$ is the relative likelihood that the price of a professional-line camera is \$1000.
- The starting point of our analysis is that we don’t know the probability density function of the two clusters.

Probabilistic Model-Based Clustering

- Suppose we want to find *k* probabilistic clusters, C_1, … ,C_k, through cluster analysis of *D*.
- Conceptually, we can assume that *D* is formed as follows. Each cluster C_j is associated with a probability, w_j. We then run the following two steps n times to generate $D = \{o_1, \ldots, o_n\}$:
  - Choose a cluster, C_j , according to probabilities w_j.
  - Choose an instance of C_j according to its probability density function, f_j.
- This is called the **mixture model.** It assumes that a set of observed objects is a mixture of instances from multiple probabilistic clusters, and conceptually each observed object is generated independently.
- The task of *probabilistic model-based cluster analysis* is to infer infer a set of k probabilistic clusters that is mostly likely to generate D using the above data generation process.

Probabilistic Model-Based Clustering

- Consider a set *C* of *k* probabilistic clusters $C_1, \ldots, C_k$ with probability density functions $f_1, \ldots, f_k$, respectively, and their probabilities $\omega_1, \ldots, \omega_k$.
- The probability of an object *o* generated by cluster C_j is

$$P(o|C_j) = \omega_j f_j(o)$$

- The probability of *o* generated by the set of cluster ***C*** is

$$P(o|\boldsymbol{C}) = \sum_{j=1}^{k} \omega_j f_j(o)$$

- Since objects are assumed to be generated independently, for a data set D = $\{o_1, \ldots, o_n\}$, we have,

$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

Probabilistic Model-Based Clustering

- Since objects are assumed to be generated independently, for a data set D = {o_1, ..., o_n}, we have,

$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

- Task: Find a set *C* of *k* probabilistic clusters s.t. *P*(*D*|***C***) is maximized.
- However, maximizing *P*(*D*|***C***) is often intractable since the probability density function of a cluster can take an arbitrarily complicated form.
- To make it computationally feasible (as a compromise), assume the probability density functions being some parameterized distributions

Univariate Gaussian Mixture Model

- Assume that the probability density function of each cluster follows a 1-d Gaussian distribution. Suppose that there are k clusters.
- The probability density function of each cluster are centered at μ_j with standard deviation σ_j, θ_j, = (μ_j, σ_j) is:

$$P(o_i|\Theta_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(o_i-\mu_j)^2}{2\sigma^2}}$$

- Assuming that each cluster has the same probability w_j:

$$P(o_i|\mathbf{\Theta}) = \sum_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(o_i-\mu_j)^2}{2\sigma^2}}$$

- The task is then to minimize:

$$P(\mathbf{O}|\mathbf{\Theta}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(o_i-\mu_j)^2}{2\sigma^2}}$$

The EM (Expectation Maximization) Algorithm

The k-means algorithm has two steps at each iteration:

- **Expectation Step** (E-step): Given the current cluster centers, each object is assigned to the cluster whose center is closest to the object: An object is *expected to belong to the closest cluster*
- **Maximization Step** (M-step): Given the cluster assignment, for each cluster, the algorithm *adjusts the center* so that *the sum of distance* from the objects assigned to this cluster and the new center is minimized

The (EM) algorithm: A framework to approach maximum likelihood or maximum a posteriori estimates of parameters in statistical models.

- **E-step** assigns objects to clusters according to the current fuzzy clustering or parameters of probabilistic clusters
- **M-step** finds the new clustering or parameters that minimize the sum of squared error (SSE) or the expected likelihood

Quality: What Is Good Clustering?

A good clustering method will produce high quality clusters

- high intra-class similarity: cohesive within clusters
- low inter-class similarity: distinctive between clusters

The quality of a clustering method depends on

- the similarity measure used by the method
- its implementation, and
- Its ability to discover some or all of the hidden patterns

Examples of Quality Measures

- *Sum of square distances to centroids:*
  - The squared distance between from the representative to every other point in the cluster is calculated, then summed over all points.
  - Suitable for representative based methods.
  - Favors spherical clusters.
- *Intracluster to intercluster distance ratio:*
  - Sample pairs of points in D.
  - Let P denote the pairs in the same cluster, and Q denote the pairs in different clusters.
  - Compute:
    - Intra/Inter
  - Smaller values are better.

$$Intra = \sum_{(\overline{X_i}, \overline{X_j}) \in P} dist(\overline{X_i}, \overline{X_j})/|P|$$

$$Inter = \sum_{(\overline{X_i}, \overline{X_j}) \in Q} dist(\overline{X_i}, \overline{X_j})/|Q|$$

Examples of Quality Measures

- Silhouette coefficient:
  - Let $D_{avg\text{-}in}$ denote the average distance between a point in the cluster and every other point in the same cluster.
  - Let $D_{avg\text{-}out\_i}$ denote the average distance between a point in the cluster and every other point in the cluster i.
  - Let $D_{avg\text{-}out\text{-}min}$ be the minimum Davg-out_i.

$$S_i = \frac{Dmin_i^{out} - Davg_i^{in}}{\max\{Dmin_i^{out}, Davg_i^{in}\}}$$

- The silhouette coefficient will be drawn from the range (−1, 1). Large positive values indicate highly separated clustering. Negative values are indicative of some level of “mixing” of data points from different clusters.

Credits and Readings

- These slides, except when explicitly stated, use material from:
  - Charu C.Aggarwal. Data Mining The Textbook, Springer
  - Han, J., Kamber, M. and Pei, J. Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, Burlington.