Fuzzy c-means algorithm

Fernando Lobo

Data mining

Fuzzy c-means algorithm

- Uses concepts from the field of fuzzy logic and fuzzy set theory.
- Objects are allowed to belong to more than one cluster.
- Each object belongs to every cluster with some weight.

Fuzzy c-means algorithm

- When clusters are well separated, a crisp classification of objects into clusters makes sense.
- But in many cases, clusters are not well separated.
  - in a crisp classification, a borderline object ends up being assigned to a cluster in an arbitrary manner.

Fuzzy sets

- Introduced by Lotfi Zadeh in 1965 as a way of dealing with imprecision and uncertainty.
- Fuzzy set theory allows an object to belong to a set with a degree of membership between 0 and 1.
- Traditional set theory can be seen as a special case that restrict membership values to be either 0 or 1.

Fuzzy clusters

- Assume a set of n objects $X = \{x_1, x_2, \dots, x_n\}$, where x_i is a d-dimensional point.
- A fuzzy clustering is a collection of k clusters, $C_1, C_2, \dots, C_k$, and a partition matrix $W = w_{i,j} \in [0, 1]$, for $i = 1 \dots n$ and $j = 1 \dots k$, where each element $w_{i,j}$ is a weight that represents the degree of membership of object i in cluster C_j.

Restrictions
(to have what is called a fuzzy pseudo-partition)

1. All weights for a given point, x_i, must add up to 1.

$$\sum_{j=1}^{k} w_{i,j} = 1$$

2. Each cluster C_j contains, with non-zero weight, at least one point, but does not contain, with a weight of one, all the points.

$$0 < \sum_{i=1}^{n} w_{i,j} < n$$

Fuzzy c-means (FCM) is a fuzzy version of k-means

Fuzzy c-means algorithm:

1. Select an initial fuzzy pseudo-partition, i.e., assign values to all $w_{i,j}$
2. Repeat
3. compute the centroid of each cluster using the fuzzy partition
4. update the fuzzy partition, i.e, the $w_{i,j}$
5. Until the centroids don't change

There's alternative stopping criteria. Ex: "change in the error is below a specified threshold", or "absolute change in any $w_{i,j}$ is below a given threshold".

Fuzzy c-means

- As with k-means, FCM also attempts to minimize the sum of the squared error (SSE).

- In k-means:

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_i} dist(c_i, x)^2$$

- In FCM:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{i,j}^{p} dist(x_i, c_j)^2$$

p is a parameter that determines the influence of the weights.
$p \in [1..\infty[$

Computing centroids

- For a cluster C_j, the corresponding centroid c_j is defined as:

$$c_j = \frac{\sum_{i=1}^{n} w_{ij}^{p} x_i}{\sum_{i=1}^{n} w_{ij}^{p}}$$

- This is just an extension of the definition of centroid that we have seen for k-means.

- The difference is that all points are considered and the contribution of each point to the centroid is weighted by its membership degree.

Updating the fuzzy pseudo-partition

- Formula can be obtained by minimizing the SSE subject to the constraint that the weights sum to 1.

$$w_{ij} = \frac{(1/dist(x_i, c_j)^2)^{\frac{1}{p-1}}}{\sum_{q=1}^{k}(1/dist(x_i, c_q)^2)^{\frac{1}{p-1}}}$$

- Intuition: w_{ij} should be high if x_i is close to the centroid c_j, i.e., if $dist(x_i, c_j)$ is low.

- Denominator (sum of all weights) is needed to normalize weights for a point.

Effect of parameter p

- If $p > 2$, then the exponent $1/(p-1)$ decrease the weight assigned to clusters that are close to the point.
- If $p \rightarrow \infty$, then the exponent $\rightarrow 0$. This implies that the weights $\rightarrow 1/k$.
- If $p \rightarrow 1$, the exponent increases the membership weights of points to which the cluster is close. As $p \rightarrow 1$, membership $\rightarrow 1$ for the closest cluster and membership $\rightarrow 0$ for all the other clusters (this corresponds to k-means).