Fuzzy Clustering

Dr. David Mueller (Fall 2019 – ECPE 226 Computational Intelligence)

Fuzzy clustering falls into the category of unsupervised learning methods. It attempts to find order or structure in unlabeled data. Fuzzy clustering allows for data to be members of multiple clusters. In this sense, clusters may overlap.

What do we need?

First, we need **data**. Ideally, this data would contain natural clusters. There are methods that can be applied to the data to help determine if this is the case, but we will not discuss them here.

What does the data look like? Suppose we have a dataset X,

$$\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\mathrm{N}\}$$

where, $\mathrm{x_k} \in \mathbb{R}^d$ (the data points are real with dimension d).

In general, clustering involves determining **a distance metric** between data points. 'Close' data points are then clustered together. There are many measures of distance and a chosen measure has a strong influence on the result. One common distance measure is Euclidean distance,

$$\mathrm{d}^2(\mathbf{x}, \mathbf{v}) = (\mathbf{x} - \mathbf{v})^\mathrm{T}\mathbf{A}(\mathbf{x} - \mathbf{v})$$

the dot product of the difference between two vectors. Since this is Euclidean distance, the $d \times d$ matrix $\mathbf{A} = \boldsymbol{I}$, the identify matrix.

Clustering is based on the concept of creating **C-partitions** of the dataset. The partitioning of N data points into C clusters $A_1, \dots, A_C$ is defined by a $C \times N$ partition matrix,

$$\mathrm{U} = \begin{pmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & \ddots & \vdots \\ u_{C1} & \cdots & u_{CN} \end{pmatrix} = \{u_{ik}\}$$

where i is the cluster index and k is the data point index. The value u_{ik} represents the **degree of membership** for the k-th data point $\mathbf{x}_k$ in the i-th cluster A_i. As in most things fuzzy, the degree of membership is constrained by

$$0 \leq u_{ik} \leq 1.$$

Fuzzy C-Means

FCM seeks to minimize the following criterion function.

$$J(\mathrm{U},\mathrm{V}) = \sum_{k=1}^{N}\sum_{i=1}^{C}(u_{ik})^m d^2(\mathbf{x}_k,\mathbf{v}_i)$$

subject to the constraint that the sum of a data point's membership in all clusters equals 1,

$$\sum_{i=1}^{C} u_{ik} = 1$$

for all data points, k. Where m is the fuzzifier that must be larger than 1. It is typically set as $m = 2$.

For a given a data set, the fuzzy c-means (FCM) clustering algorithm determines the clusters by defining a set of **cluster centers**,

$$\mathrm{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_C\}$$

where $\mathbf{v}_i$ is the cluster center for cluster A_i. We compute/define $\mathbf{v}_i$ as

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N}(u_{ik})^m x_k}{\sum_{k=1}^{N}(u_{ik})^m}$$

Before we dive into the algorithm, we must consider the possibility that a vector center could coincide with one or more data points. To identify this, let's create a $C \times N$ matrix I. This matrix will contain the cluster index of the center(s) for which it coincides. For all data points define,

$$\mathrm{I}_k = \{i \mid 1 \le i \le C \; and \; d_{ik}^2 = 0\}.$$

If the data point and the cluster center have a distance metric between them of 0, then set

$$u_{ik} = \begin{cases} 0 & i \notin \mathrm{I}_k \\ \dfrac{1}{n} & i \in \mathrm{I}_k \end{cases}$$

where n is the number of non-empty elements (shared centers) in the vector I_k.

For the data points not coinciding with a cluster center(s), to minimize the function, we compute u_{ik} as

$$u_{ik} = \frac{1}{\sum_{j=1}^{C}\left(\dfrac{d^2(\mathbf{x}_k,\mathbf{v}_i)}{d^2(\mathbf{x}_k,\mathbf{v}_j)}\right)^{\frac{1}{m-1}}}$$

Fuzzy C-Mean Algorithm

Initialize
    C, the number of clusters
    m, the fuzzifier
    ϵ, the convergence threshold
    V, the cluster centers %initialize at random

t = 0

while $\sum_{i=1}^{C} d^2\left(\mathbf{v}_i^{(t)}, \mathbf{v}_i^{(t-1)}\right) > \epsilon$
    compute the memberships u_{ik} for each data point
    t = t+1
    estimate/compute $V^{(t)}$ using $U^{(t-1)}$
end