

Cluster Analysis

Task: We want to classify objects (vectors) x_n , $n = 1, \dots, N$, into clusters S_k , $k = 1, \dots, K$ so that “close” objects are together.

Possible criterion:

$$\sum_k \sum_{x_n \in S_k} \sum_{x_j \in S_k} \|x_n - x_j\|^2.$$

Note: We consider the number of clusters as given, although its choice is a difficult problem.

1 The k-means algorithm [MacQueen, 1967]

1. Randomly select the centroids of the clusters c_k .
2. Assign each object x_n to the cluster S_k whose centroid is the closest,

$$k(n) \in \operatorname{argmin}_j \|x_n - c_j\|^2 .$$

3. In each cluster calculate the new centroid as the center of gravity of the cluster elements,

$$c_k := \frac{1}{|S_k|} \sum_{x_n \in S_k} x_n ,$$

where $|S_k|$ denotes the number of elements of cluster S_k .

4. Return to 2 if there was a significant change of the results.

Decreasing criterion

$$\sum_k \sum_{x_n \in S_k} \|x_n - c_k\|^2 .$$

Problems:

- If there are more nearest centers, we choose from them, for example, randomly.
- Converges to a **local** minimum.
- Choose initial cluster centroids different, preferably **much different**.
- We need non-empty clusters, preferably roughly equally big.
- It depends on the metric (you can normalize all coordinates first, but that may not be good either).

2 Fuzzy c-means (FCM) algorithm

[Dunn 1973, improved by Bezdek 1981]

Clusters S_k are fuzzy sets, characterized by a matrix of degrees of membership

$$\alpha_{n,k} = \mu_{S_k}(x_n),$$

which satisfy

$$\sum_k \alpha_{n,k} = 1, \quad \sum_n \alpha_{n,k} > 0.$$

1. We randomly choose the centroids of clusters, c_k .
2. The degree of membership of an object x_n to a cluster S_k is determined as

$$\alpha_{n,k} := \frac{\frac{1}{\|x_n - c_k\|^{\frac{2}{m-1}}}}{\sum_j \frac{1}{\|x_n - c_j\|^{\frac{2}{m-1}}}},$$

where $m > 1$ (e.g. 2) is a parameter of the method. (The denominator is a normalization factor; division by zero must be fixed.)

3. In each cluster, we calculate a new centroid as the center of gravity of the cluster elements weighted by the m -th power of their degree of membership,

$$c_k := \frac{\sum_n \alpha_{n,k}^m x_n}{\sum_n \alpha_{n,k}^m}.$$

4. Return to 2 if there was a significant change in the results.

Decreasing criterion

$$\sum_k \sum_{x_n \in S_k} \alpha_{n,k}^m \|x_n - c_k\|^2 .$$

If required, we defuzzify the final result (we classify each object into the cluster to which it has the highest degree of membership).

Similar problems, slightly smaller; a number of modifications, e.g., the distance of vectors (norm) can be more general.

3 Estimation of parameters of a mixture of normal distributions

Task: Based on the implementation of a random sample $(x_1, \dots, x_N)$ (here x_k are numbers), we search for a maximally plausible estimate of a mixture of K normal distributions with means c_k , variances σ_k^2 and mixture coefficients (weights) q_k , $k = 1, \dots, K$.

We optimize the parameter vector $\theta = (c_1, \dots, c_K, \sigma_1, \dots, \sigma_K, q_1, \dots, q_K)$.
The mixture has a density

$$f(t) = \sum_k q_k f_{N(c_k, \sigma_k^2)}(t),$$

the likelihood and its logarithm are

$$\begin{aligned} L(\theta) &= \prod_n \sum_k q_k f_{N(c_k, \sigma_k^2)}(x_n), \\ \ell(\theta) &= \sum_n \ln \sum_k q_k f_{N(c_k, \sigma_k^2)}(x_n). \end{aligned}$$

This is difficult to solve directly, but there is an iterative method:

4 EM algorithm

EM (Expectation-Maximization) [Dempster, Laird, and Rubin 1977], [M.I. Schlesinger 1968], [US Army ~1950]

The membership of x_n to the k -th component of the mixture (cluster) is described by the coefficient $\alpha_{n,k} \in [0, 1]$; we obtain a matrix whose coefficients satisfy

$$\sum_k \alpha_{n,k} = 1, \quad \sum_n \alpha_{n,k} > 0.$$

Here, for simplicity, we assume equal variances $\sigma_k^2 = \sigma^2$.

1. We randomly choose the mean values of the clusters c_k .

E. We determine the coefficients

$$\alpha_{n,k} := \frac{q_k f_{N(c_k, \sigma^2)}(x_n)}{\sum_j q_j f_{N(c_j, \sigma^2)}(x_n)}.$$

(The denominator is the normalization factor.)

M. We update the cluster weight

$$q_k := \frac{\sum_n \alpha_{n,k}}{\sum_j \sum_n \alpha_{n,j}} = \frac{1}{N} \sum_n \alpha_{n,k}$$

and its center as the centroid of the elements weighted by the degree of membership,

$$c_k := \frac{\sum_n \alpha_{n,k} x_n}{\sum_n \alpha_{n,k}} = \frac{\sum_n \alpha_{n,k} x_n}{N q_k}.$$

2. We repeat EM if there is a significant change in the results.

Theorem: During the EM algorithm, *the likelihood does not decrease*.

This is just a very special example of the EM algorithm; extending it to more dimensions is easy.

It can be used to find maximally plausible estimates of other distribution parameters.

But be careful: For the variance $\sigma_k^2 \rightarrow 0$ we get $\ell(\boldsymbol{\theta}) \rightarrow \infty$ and we would be looking for a maximum that does not exist $\implies$ the variance must be bounded from below!

The use for parameters of mixtures of distributions is typical, but not the only possible one.

Again, there are problems with getting stuck in a local extreme, etc., but the possibilities of using the maximum likelihood method are greatly expanded.