A Tutorial on Clustering Algorithms

Introduction | K-means | Fuzzy C-means | Hierarchical | Mixture of Gaussians | Links

Clustering as a Mixture of Gaussians

Introduction to Model-Based Clustering

There's another way to deal with clustering problems: a *model-based* approach, which consists in using certain models for clusters and attempting to optimize the fit between the data and the model.
In practice, each cluster can be mathematically represented by a parametric distribution, like a Gaussian (continuous) or a Poisson (discrete). The entire data set is therefore modelled by a *mixture* of these distributions. An individual distribution used to model a specific cluster is often referred to as a *component* distribution.

A mixture model with high likelihood tends to have the following traits:

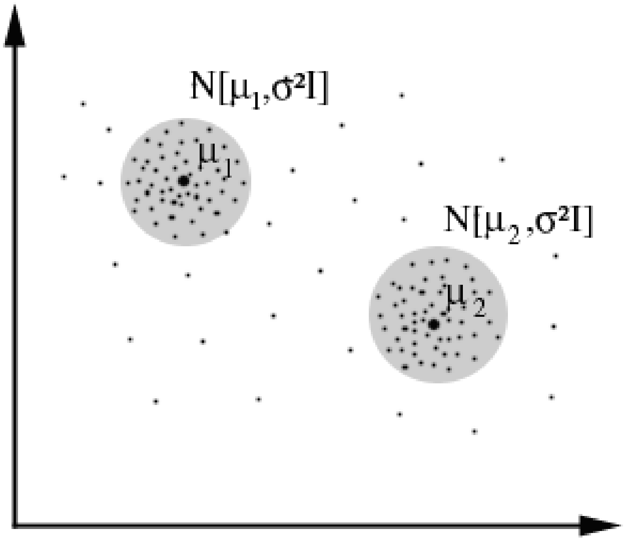
- component distributions have high "peaks" (data in one cluster are tight);
- the mixture model "covers" the data well (dominant patterns in the data are captured by component distributions).

Main advantages of model-based clustering:

- well-studied statistical inference techniques available;
- flexibility in choosing the component distribution;
- obtain a density estimation for each cluster;
- a "soft" classification is available.

Mixture of Gaussians

The most widely used clustering method of this kind is the one based on learning a *mixture of Gaussians*: we can actually consider clusters as Gaussian distributions centred on their barycentres, as we can see in this picture, where the grey circle represents the first variance of the distribution:



The algorithm works in this way:

- it chooses the component (the Gaussian) at random with probability $P(\omega_i)$;
- it samples a point $N(\mu_i, \sigma^2 I)$.

Let's suppose to have:

- $x_1, x_2, \ldots, x_N$
- $P(\omega_1), \ldots, P(\omega_K), \sigma$

We can obtain the likelihood of the sample: $P(x \mid \omega_i, \mu_1, \mu_2, \ldots, \mu_K)$.
What we really want to maximise is $P(x \mid \mu_1, \mu_2, \ldots, \mu_K)$ (probability of a datum given the centres of the Gaussians).

$$P(x \mid \mu_i) = \sum_i P(\omega_i) P(x \mid \omega_i, \mu_1, \mu_2, \ldots, \mu_K)$$

is the base to write the likelihood function:

$$P(\text{data} \mid \mu_i) = \prod_{i=1}^{N} \sum_i P(\omega_i) P(x \mid \omega_i, \mu_1, \mu_2, \ldots, \mu_K)$$

Now we should maximise the likelihood function by calculating $\frac{\partial L}{\partial \mu_i} = 0$, but it would be too difficult. That's why we use a simplified algorithm called EM (Expectation-Maximization).

The EM Algorithm

The algorithm which is used in practice to find the mixture of Gaussians that can model the data set is called EM (Expectation-Maximization) (Dempster, Laird and Rubin, 1977). Let's see how it works with an example.

Suppose x_k are the marks got by the students of a class, with these probabilities:

$x_1 = 30$ $\quad P(x_1) = \frac{1}{2}$

$x_2 = 18$ $\quad P(x_2) = \mu$

$x_3 = 0$ $\quad P(x_3) = 2\mu$

$x_4 = 23$ $\quad P(x_4) = \frac{1}{2} - 3\mu$

First case: we observe that the marks are so distributed among students:

x_1 : a students
x_2 : b students
x_3 : c students
x_4 : d students

$$P(a,b,c,d \mid \mu) \propto \left(\frac{1}{2}\right)^a * (\mu)^b * (2\mu)^c * \left(\frac{1}{2} - 3\mu\right)^d$$

We should maximise this function by calculating $\frac{\partial P}{\partial \mu} = 0$. Let's instead calculate the logarithm of the function and maximise it:

$$P_L = \log\left(\frac{1}{2}\right)^a + \log(\mu)^b + \log(2\mu)^c + \log\left(\frac{1}{2} - 3\mu\right)^d$$

$$\frac{\partial P_L}{\partial \mu} = \frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{\frac{1}{2} - 3\mu} = 0$$

$$\Rightarrow \mu = \frac{b+c}{6(b+c+d)}$$

Supposing a = 14, b = 6, c = 9 and d = 10 we can calculate that $\mu = \frac{1}{10}$.

Second case: we observe that marks are so distributed among students:

$x_1 + x_2$: h students
x_3 : c students
x_4 : d students

We have so obtained a circularity which is divided into two steps:

- expectation: $\mu \rightarrow a = \dfrac{\frac{1}{2}}{\frac{1}{2}+\mu}h,\ b = \dfrac{\mu}{\frac{1}{2}+\mu}h$
- maximization: $a, b \rightarrow \mu = \dfrac{b+c}{6(b+c+d)}$

This circularity can be solved in an iterative way.

Let's now see how the EM algorithm works for a mixture of Gaussians (parameters estimated at the *p*th iteration are marked by a superscript (*p*):

1. *Initialize parameters:*

$$\lambda_0 = \{\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_K^{(0)}, p_1^{(0)}, p_2^{(0)}, \ldots, p_K^{(0)}\}$$

2. *E-step:*

$$P(\omega_j \mid x_k, \lambda_t) = \frac{P(x_k \mid \omega_j, \lambda_t) P(\omega_j \mid \lambda_t)}{P(x_k \mid \lambda_t)} = \frac{P(x_k \mid \omega_i, \mu_i^{(t)}, \sigma^2) p_i^{(t)}}{\sum_k P(x_k \mid \omega_j, \mu_j^{(t)}, \sigma^2) p_j^{(t)}}$$

3. *M-step:*

$$\mu_i^{(t+1)} = \frac{\sum_k P(\omega_i \mid x_k, \lambda_t) x_k}{\sum_k P(\omega_i \mid x_k, \lambda_t)}$$

$$p_i^{(t+1)} = \frac{\sum_k P(\omega_i \mid x_k, \lambda_t)}{R}$$

where R is the number of records.

Bibliography

- A.P. Dempster, N.M. Laird, and D.B. Rubin (1977): "Maximum Likelihood from Incomplete Data via theEM algorithm", *Journal of the Royal Statistical Society*, Series B, vol. 39, 1:1-38
- Osmar R. Zaïane: "Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering" http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html
- Jia Li: "Data Mining - Clustering by Mixture Models" http://www.stat.psu.edu/~jiali/course/stat597e/notes/mix.pd

---

Previous page | Next page