Tutorial: Gaussian mixture model and the EM algorithm

MAP573 – Introduction to unsupervised learning

École Polytechnique - Autumn 2019

Preliminaries

Goals.

1. Gaussian mixture models
2. Expectation-Maximization algorithm for mixture models

Instructions. Each student *may* send an `R markdown` report generated via `R studio` to julien.chiquet@inra.fr at the end of the tutorial. This report should answer the questions by commentaries and codes generating appropriate graphical outputs. A cheat sheet of the markdown syntax can be found here.

Required packages. Check that the following packages are available on your computer:

```
library(aricode)
library(mixtools)
```

You also need `Rstudio`, LaTeX and packages for markdown:

```
library(knitr)
library(rmarkdown)
```

1 Gaussian Mixture Models

We consider a collection of random variables $(X_1, \dots, X_n)$ associated with n individuals drawn from Q populations. The label of each individual describes the population (or class) to which it belongs and is unobserved. The Q classes have *a priori* distribution $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_Q)$ with $\alpha_q = \mathbb{P}(i \in q)$. The hidden random indicator variables $(Z_{iq})_{i \in \mathcal{P}, q \in \mathcal{Q}}$ describe the label of each individuals, that is,

$$\alpha_q = \mathbb{P}(Z_{iq} = 1) = \mathbb{P}(i \in q), \quad \text{such that} \sum_{q=1}^{Q} \alpha_q = 1.$$

Remark that we have $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iQ}) \sim \mathcal{M}(1, \boldsymbol{\alpha})$. The distribution of X_i conditional on the label of i is assumed to be a univariate gaussian distribution with unknown parameters, that is, $X_i | Z_{iq} = 1 \sim \mathcal{N}(\mu_q, \sigma_q^2)$.

2 Questions

- *Likelihood.* Write the model complete-data loglikelihood.
- *E-step.* For fixed values of $\hat{\mu}_q, \hat{\sigma}_q^2$ and $\hat{\alpha}_q$, give the expression of the estimates of the posterior probabilities $\tau_{iq} = \mathbb{P}(Z_{iq} = 1 | X_i)$.
- *M-step.* For fixed values of $\hat{\tau}_{iq}$, show that the maximization step leads to the following estimator for the model parameters:

$$\hat{\alpha}_q = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_{iq}, \quad \hat{\mu}_q = \frac{\sum_i \hat{\tau}_{iq} x_i}{\sum_i \hat{\tau}_{iq}}, \quad \hat{\sigma}_q^2 = \frac{\sum_i \hat{\tau}_{iq} (x_i - \hat{\mu}_q)^2}{\sum_i \hat{\tau}_{iq}}$$

- *Implementation.* Test your EM algorithm on simulate data. Try different values for μ_q, σ_q. Also consider different initialization. Compare with the output of the function `normalmixEM` in the package **mixtools**.
- *Model Selection.* Compute the ICL criterion and test it on your simulated data.

$$ICL(Q) = -2 \log L(X, \hat{Z}; \hat{\alpha}, \hat{\mu}, \hat{\sigma^2}) + \log(n) \mathrm{df}(Q).$$