Model selection for Gaussian mixtures

Pierre-Alexandre Mattei

http://pamattei.github.io – @pamattei
pierre-alexandre.mattei@inria.fr

MSc Data Science

Recap on GMMs

Model selection for GMMs

Recap on GMMs

Model selection for GMMs

Gaussian mixture models

We have some data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$. A Gaussian mixture model with K clusters is a statistical model $(p_\theta, \theta \in \Theta_K)$ with the following density

$$p_\theta(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

with $\theta = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$.

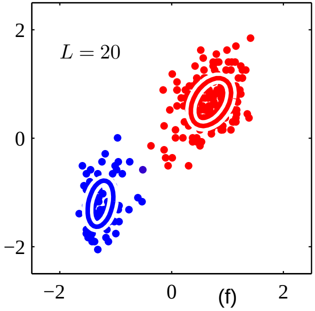

Figure: **Figure 9.8 from Bishop's book.**

The parameters of a Gaussian mixture model

Consider the model

$$p_\theta(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

with $\theta = (\pi_1, ..., \pi_K, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K)$.

What are the constraints on the parameters? In other words, in which space the proportions, means, and covariances live?

The proportions and means

The proportions must sum to one! So

$$(\pi_1, ..., \pi_K) \in \Delta_K,$$

with $\Delta_K = \{\mathbf{t} \in \mathbb{R}^K, t_1 + ... + t_K = 1\}$. This set is usually called the simplex.

The means $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$ are simply vectors in $\mathbb{R}^D$.

The covariances $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K$ must be symmetric positive definite matrices. This means that they are symmetric and all their eigenvalues are strictly positive.

Recap on GMMs

Model selection for GMMs

Choosing the number of clusters as model selection

The goal is to choose the number of clusters K. This is a model selection problem. Each number of clusters corresponds indeed to a model (i.e. a parametric family of densities)

$$\mathcal{M}_k = (p_\theta, \theta \in \Theta_K).$$

What model selection techniques do you know?

Model selection

The goal is to choose the number of clusters K. This is a model selection problem. Each number of clusters corresponds indeed to a model (i.e. a parametric family of densities)

$$\mathcal{M}_k = (p_\theta, \theta \in \Theta_K).$$

What model selection techniques do you know?

Penalised model selection

A first important school of techniques for model selection is to maximise a penalised likelihood criterion:

$$\mathcal{L}(\mathcal{M}_k) = \sum_{i=1}^{n} \log p_{\hat{\theta}_k}(x_i) - \text{penalty}(\mathcal{M}_k),$$

where $\hat{\theta}_k$ is the maximum likelihood estimate for model $\mathcal{M}_k$.
At the end, we choose the model with the largest $\mathcal{L}(\mathcal{M}_k)$.
The role of the penalty is to discourage overly complex models, and avoid overfitting. What does "overly complex" mean in a clustering context?

Penalised model selection

$$\mathcal{L}(\mathcal{M}_k) = \sum_{i=1}^{n} \log p_{\hat{\theta}_k}(x_i) - \text{penalty}(\mathcal{M}_k),$$

Overly complex can (for example) mean " with too many clusters". One way to formalise this is by counting the number of free parameters $q_k = \dim(\Theta_k)$. This leads to the following (famous?) penalties:

- $\text{penalty}_{\text{AIC}} = q_k$
- $\text{penalty}_{\text{BIC}} = q_k \log(n)/2$.

The BIC also has a Bayesian interpretation, but we won't talk about it today.

Validation using the likelihood

Another way of doing model selection is to compute the likelihood on a validation set, and choose the model with the largest validation likelihood.

A more advanced approach would be to use cross-validation.