Université du Sud Toulon - Var

Master 2 Informatique

Probabilistic Learning and Data Analysis

TD2: Model-based clustering

by Faicel CHAMROUKHI

Solution

The aim of this practical work is to show how the Classification EM algorithm for the well-known Gaussian mixture model (GMM), under some constraints, is exactly K-means so that CEM can be viewed as a probabilistic version of K-means. An additional feature of this work is to derive the EM algorithm to estimated a mixture of regressions. This can namely serve for curve clustering. In this case, the data are curves rather than vectorial data.

1 EM algorithm: updating the mixing proportions $\{\pi_k\}$

Consider the problem of finding the maximum of the function

$$Q_\pi(\pi_1, \ldots, \pi_K, \mathbf{\Psi}^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k$$

with respect to the mixing proportions $(\pi_1, \ldots, \pi_K)$ subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$, where $\tau_{ik}^{(q)}$ are the posterior probabilities at the qth iteration of EM.

- To perform this constrained maximization, introduce the Lagrange multiplier λ and derive the resulting unconstrained maximization problem (the Lagrangian function).
- To maximize the Lagrangian with respect to π_k $(k = 1, \ldots, K)$, first set the derivative of the Lagrangian with respect to π_k to zero, determine the Lagrange multiplier λ, and then the resulting value $\pi_k^{(q+1)}$ $(k = 1, \ldots, K)$ that corresponds to the maximum (the updating formula for the mixing proportions π_k $(k = 1, \ldots, K)$)

Solution

Consider the problem of finding the maximum of the function

$$Q_\pi(\pi_1, \ldots, \pi_K, \mathbf{\Psi}^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k$$

with respect to the mixing proportions $(\pi_1, \ldots, \pi_K)$ subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$. To perform this constrained maximization, we introduce the Lagrange multiplier λ such that the resulting Lagrangian function is given by:

$$L(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k + \lambda(1 - \sum_{k=1}^{K} \pi_k). \quad (1)$$

Taking the derivatives of the Lagrangian with respect to π_k for $k = 1, \ldots, K$ we obtain:

$$\frac{\partial L(\pi_1, \ldots, \pi_K)}{\partial \pi_k} = \frac{\sum_{i=1}^n \tau_{ik}^{(q)}}{\pi_k} - \lambda, \quad \forall k \in \{1, \ldots, K\}. \tag{2}$$

Then, setting these derivatives to zero yields:

$$\frac{\sum_{i=1}^n \tau_{ik}^{(q)}}{\pi_k} = \lambda, \quad \forall k \in \{1, \ldots, K\}. \tag{3}$$

By multiplying each hand side of (3) by π_k $(k = 1, \ldots, K)$ and summing over k we get

$$\sum_{k=1}^K \frac{\pi_k \times \sum_{i=1}^n \tau_{ik}^{(q)}}{\pi_k} = \sum_{k=1}^K \lambda \times \pi_k \tag{4}$$

which implies that $\lambda = n$. Finally, from (3) we get the updating formula for the mixing proportions π_k's, that is

$$\pi_k^{(q+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(q)}}{\lambda} = \frac{\sum_{i=1}^n \tau_{ik}^{(q)}}{n}, \quad \forall k \in \{1, \ldots, K\}. \tag{5}$$

2 CEM clustering as a probabilistic view for K-means clustering

Given an iid data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the aim is to automatically find a partition into K clusters. Lets us denote by $\mathbf{z} = (z_1, \ldots, z_n)$ the corresponding unknown cluster labels where $z_i \in \{1, \ldots, K\}$ denotes the cluster label of $\mathbf{x}_i$. To achieve this clustering task, we have seen that several algorithms can be used, namley K-means, EM or CEM with GMMs, etc.

Here we will consider the Classification EM (CEM) algorithm and the K-means algorithm. CEM is the classification version of EM.

Let us recall that K-means minimizes the following distortion measure

$$J(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z}) \quad = \quad \sum_{k=1}^K \sum_{i|z_i=k} \| \mathbf{x}_i - \boldsymbol{\mu}_k \|^2 \tag{6}$$

simultaneously w.r.t the cluster centres $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and the cluster labels $\mathbf{z}$. CEM for the GMM $p(\mathbf{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^K \pi_k \, \mathcal{N}\big(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\big)$ however maximizes the complete-data log-likelihood $\mathcal{L}_c(\boldsymbol{\Psi}, \mathbf{z})$ simultaneously w.r.t the Gaussian mixture parameters $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$ and the cluster labels $\mathbf{z}$.

1. Derive the expression of the optimized complete-data log-likelihood
2. Show that, under the following constraints, maximizing $\mathcal{L}_c$ by using CEM is equivalent to minimizing J by using K-means
   - $\pi_k = \frac{1}{K}$ $\forall k$ (same proportions)
   - $\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}$ $\forall k$ (identical isotropic covariance matrices)

We recall that the multivariate Gaussian density $\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is given by:

$$\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \Big)$$

Solution

The complete-data log-likelihood, for a GMM, which is the criterion optimized by the CEM algorithm, is given by:

$$\begin{aligned}
\mathcal{L}_c(\mathbf{\Psi}, \mathbf{z}) &= \log p((\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_n, z_n); \mathbf{\Psi}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i, z_i; \mathbf{\Psi}) \\
&= \sum_{i=1}^{n} \log \prod_{k=1}^{K} \Big[p(z_i = k) p(\mathbf{x} | z_i = k; \mathbf{\Psi}_k) \Big]^{z_{ik}} \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log [\pi_k p(\mathbf{x}_i | z_i = k; \mathbf{\Psi}_k)] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log [\pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)] \qquad (7)
\end{aligned}$$

To show that a particular case of CEM yields in K-means algorithm, consider the spherical Gaussian mixture model where the classes have the same diagonal covariance matrix: $\mathbf{\Sigma}_k = \sigma^2 \mathbf{I}\ \forall k$ and the same mixing proportions $\pi_k = \frac{1}{K}\ \forall k$. Then, the complete-data log-likelihood (7) takes the following form:

$$\begin{aligned}
\mathcal{L}_c(\mathbf{\Psi}, \mathbf{z}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log [\pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)] = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left[\pi_k \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}_k|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \Big) \right] \qquad (8) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \pi_k - \frac{d}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log |\mathbf{\Sigma}_k| - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \qquad (9) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \frac{1}{K} - \frac{nd}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log |\sigma^2 \mathbf{I}| - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\sigma^2 \mathbf{I})^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \qquad (10) \\
&= n \log \frac{1}{K} - \frac{nd}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\sigma^{2d}) - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\frac{1}{\sigma^2} \mathbf{I}) (\mathbf{x}_i - \boldsymbol{\mu}_k) \qquad (11) \\
&= n \log \frac{1}{K} - \frac{nd}{2} \log 2\pi - nd \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\mathbf{x}_i - \boldsymbol{\mu}_k) \qquad (12)
\end{aligned}$$

where we used:

$\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} = n$ (from (9) to (10) and from (10) to (11))

$|\sigma^2 \mathbf{I}_d| = \sigma^{2d}$ (from (10) to (11))

$(\sigma^2 \mathbf{I}_d)^{-1} = (\sigma^2)^{-1} \mathbf{I}_d$ (from (10) to (11))

Since σ is constant (we are not maximizing w.r.t it), maximizing the complete-data log-likelihood (12) w.r.t the means and the clusters $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z})$ is therefore equivalent to maximizing

$$-\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\mathbf{x}_i - \boldsymbol{\mu}_k) = -\sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \parallel \mathbf{x}_i - \boldsymbol{\mu}_k \parallel^2 = -\sum_{k=1}^{K} \sum_{i | z_i = k} \parallel \mathbf{x}_i - \boldsymbol{\mu}_k \parallel^2 . \qquad (13)$$

which is also equivalent minimizing w.r.t the means and the cluster labels $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z})$ the criterion

$$J(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z}) = \sum_{k=1}^{K} \sum_{i | z_i = k} \parallel \mathbf{x}_i - \boldsymbol{\mu}_k \parallel^2 \qquad (14)$$

which is none other than the distortion criterion (1) minimized by the K-means algorithm.

3 EM for mixture of polynomial regressions

The aim here is to cluster n iid curves $((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$ into K clusters using mixture of regressions and EM. Each curve consists of m observations $\mathbf{y} = (y_{i1}, \ldots, y_{im})$ regularly observed at the inputs $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ for all $i = 1, \ldots, n$ (e.g., $\mathbf{x}$ may represent the sampling time in a temporal context).

Model definition

The polynomial regression mixture model arises when we assume that, each class of curves has a prior probability α_k and generate a curve according to polynomial function (with polynomial coefficients $\boldsymbol{\beta}_k$) corrupted by a zero-mean Gaussian noise witha variance σ_k^2:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_i \tag{15}$$

where $\mathbf{y}_i$ is an $n \times 1$ curve, $\mathbf{X}_i$ is the $n \times (p+1)$ regression matrix (Vandermonde matrix) with rows $(1, x_{ij}, x_{ij}^2 \ldots, x_{ij}^p)$, p being the order of the polynomial, $\boldsymbol{\beta}_k = (\beta_{k0}, \ldots, \beta_{kp})^T$ is the vector of regression coefficients for class k and $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma_k^2\mathbf{I}_m)$ is its corresponding Gaussian noise.

- From (15), derive the corresponding density for the observed curve $\mathbf{y}_i$ given $\mathbf{x}_i$ (mixture of polynomial regressions)
- Derive the EM algorithm for estimating the model parameters $\boldsymbol{\Psi} = (\alpha_1, \ldots, \alpha_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$
- provide the updating formula for $\boldsymbol{\Psi}$

Solution:

The conditional mixture density of a curve $\mathbf{y}_i$ $(i = 1, \ldots, n)$ can be written as:

$$p(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} p(z_i = k)p(\mathbf{y}_i|\mathbf{x}_i, z_i = k; \boldsymbol{\beta}_k, \sigma_k^2) = \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_m), \tag{16}$$

$\boldsymbol{\Psi} = (\alpha_1, \ldots, \alpha_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ with $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2)$, σ_k^2 being the noise variance for the cluster k. The unknown parameter vector $\boldsymbol{\Psi}$ is estimated by maximum likelihood via the EM algorithm.

Parameter estimation via the EM algorithm

Given an i.i.d training set of n curves $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ regularly observed for the inputs $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the log-likelihood of $\boldsymbol{\Psi}$ is given by:

$$\mathcal{L}(\boldsymbol{\Psi}) = \log \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_m). \tag{17}$$

The log-likelihood is maximized by the EM algorithm. Before giving the EM steps, the complete-data log-likelihood is given by:

$$\mathcal{L}_c(\boldsymbol{\Psi}) \quad = \quad \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \alpha_k + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \mathcal{N}(\mathbf{y}_i; \boldsymbol{\beta}_k^T\mathbf{X}_j, \sigma_k^2\mathbf{I}_m) \tag{18}$$

where $\mathbf{z} = (z_1, \ldots, z_n)$ is the vector of cluster labels for the n curves and z_{ik} is an indicator binary-valued variable such that $z_{ik} = 1$ if $z_i = k$ (i.e., if $\mathbf{y}_i$ is generated by the cluster r). The EM algorithm for PRMs and PSRMs starts with an initial model parameters $\boldsymbol{\Psi}^{(0)}$ and alternates between the two following steps until convergence:

E-step: Compute the expected complete-data log-likelihood given the curves $\mathbf{Y}$, the inputs $\mathbf{X}$ and the current value of the parameter $\boldsymbol{\Psi}$ denoted by $\boldsymbol{\Psi}^{(q)}$:

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) &= \mathbb{E}\big[\mathcal{L}_c(\boldsymbol{\Psi})|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Psi}^{(q)}\big] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}\big[z_{ik}|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Psi}^{(q)}\big] \log \alpha_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}\big[z_{ik}|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Psi}^{(q)}\big] \log \mathcal{N}(\mathbf{y}_i; \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_m) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log \mathcal{N}(\mathbf{y}_i; \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_m) \end{aligned} \tag{19}$$

where

$$\tau_{ik}^{(q)} = p(z_i = k|\mathbf{y}_i, \mathbf{t}; \mathbf{\Psi}^{(q)}) = \frac{\alpha_k^{(q)} \mathcal{N}\big(\mathbf{y}_i; \mathbf{X}\boldsymbol{\beta}_k^{T(q)}, \sigma_k^{2(q)}\mathbf{I}_m\big)}{\sum_{k'=1}^{K} \alpha_{k'}^{(q)} \mathcal{N}(\mathbf{y}_i; \mathbf{X}\boldsymbol{\beta}_{k'}^{(q)T}, \sigma_{k'}^{2(q)}\mathbf{I}_m)} \tag{20}$$

is the posterior probability that the curve $\mathbf{y}_i$ is generated by the cluster r. This step therefore only requires the computation of the posterior cluster probabilities $\tau_{ik}^{(q)}$ $(i = 1, \ldots, n)$ for each of the R clusters.

M-step: Compute the update $\mathbf{\Psi}^{(q+1)}$ fo $\mathbf{\Psi}$ by maximizing the Q-function (19) with respect to $\mathbf{\Psi}$. The two terms of the Q-function are maximized separately. The first term, that is the function $\sum_{i=1}^{n}\sum_{k=1}^{R} \tau_{ik}^{(q)} \log \alpha_k$ is maximized with respect to $(\alpha_1, \ldots, \alpha_R)$ subject to the constraint $\sum_{k=1}^{K} \alpha_k = 1$ using Lagrange multipliers which gives the following updates:

$$\alpha_k^{(q+1)} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(q)} \quad (r = 1, \ldots, R). \tag{21}$$

The second term of (19) can also be decomposed independently as a sum of R functions of $(\boldsymbol{\beta}_k, \sigma_k^2)$ to perform R separate maximizations. The maximization of each of the R functions, that is $\sum_{i=1}^{n} \tau_{ik}^{(q)} \log \mathcal{N}(\mathbf{y}_i; \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_m)$ corresponds therefore to solving a weighted least-squares problem. The solution of this problem is straightforward and is given by:

$$\boldsymbol{\beta}_k^{(q+1)} \quad = \quad \big(\mathbf{X}^{*T}\mathbf{W}_k^{(q)}\mathbf{X}^*\big)^{-1}\mathbf{X}^{*T}\mathbf{W}_k^{(q)}\mathbf{y}^* \tag{22}$$

$$\sigma_k^{2(q+1)} \quad = \quad \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(q)}}(\mathbf{y}^* - \mathbf{X}^{*T}\boldsymbol{\beta}_k^{(q+1)})^T\mathbf{W}_k^{(q)}(\mathbf{y}^* - \mathbf{X}^{*T}\boldsymbol{\beta}_k^{(q+1)}) \tag{23}$$

where $\mathbf{X}^* = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$, the vector $\mathbf{y}^*$ is an $nm \times 1$ vector composed of the n curves by stacking them one curve after another, that is $\mathbf{y}^* = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$ and $\mathbf{W}_k^{(q)}$ is the $nm \times nm$ diagonal matrix whose diagonal elements are $(\underbrace{\tau_{1k}^{(q)}, \ldots, \tau_{1k}^{(q)}}_{m \text{ times}}, \ldots, \underbrace{\tau_{nk}^{(q)}, \ldots, \tau_{nk}^{(q)}}_{m \text{ times}})$.