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 $(k = 1, \ldots, K)$)

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

Given an i.i.d data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $(\mathbf{x}_i \in \mathbb{R}^d)$, 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, namely 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

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

simultaneously w.r.t the cluster centres $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and the cluster labels $\mathbf{z}$. CEM for the GMM $p(\mathbf{x}_i; \mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\big(\mathbf{x}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\big)$ however maximizes the complete-data log-likelihood $\mathcal{L}_c(\mathbf{\Psi}, \mathbf{z})$ simultaneously w.r.t the Gaussian mixture parameters $\mathbf{\Psi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K)$ 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)
   - $\mathbf{\Sigma}_k = \sigma^2 \mathbf{I}$ $\forall k$ (identical isotropic covariance matrices)

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 with a variance σ_k^2:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_i \tag{2}$$

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.

- From (2), 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 $\boldsymbol{\Psi}$

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

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