Machine Learning

(Soft) k-means

Irving Gómez Méndez

UNIVERSIDAD
IBEROAMERICANA
CIUDAD DE MÉXICO ®

EM Algorithm

To fit the statistical models we hold a sample $(x_1, \ldots, x_n)$. However, sometimes we do not know all the observations. For example, if X is the survival time of a component and τ is the experiment's duration, it might happened that the component would not broke before τ, i.e. $X > \tau$ and we would not observed its value (we only would knew that $X > \tau$) and the observation would be censored.

Assume that the **complete data have a joint density** $f(x;\phi)$, while the **joint density of the observed data** is denoted by $g(y;\phi)$, and the **joint density of the complete data given the observed** is given by $k(x|y;\phi)$.

Thus, we want to find the value of the parameters ϕ that maximize $f(x;\phi)$. Note that if we have access to the complete observations, then the observed data do not add information so

$$f(x;\phi) = f(x,y;\phi) = k(x|y;\phi)g(y;\phi)$$

and

$$\log f(x;\phi) = \log g(y;\phi) + \log k(x|y;\phi)$$

But we do not have access to x! We only have the observed data y. To solve this problem, we are going to substitute the unobserved values by their expected value **fixing the value of the parameters to be** ϕ', i.e.

Step E:

$$\begin{aligned} Q(\phi|\phi') &= \mathbb{E}_{x\sim f(\cdot;\phi')} \log g(y;\phi) + \mathbb{E}_{x\sim f(\cdot;\phi')} \log k(x|y;\phi) \\ &= \log g(y;\phi) + \mathbb{E}_{x\sim f(\cdot;\phi')} \log k(x|y;\phi) \end{aligned}$$

Step M: We find the parameters ϕ that maximize $Q(\phi|\phi')$.

Finally the EM algorithm consists in iterating the previous steps.

Gaussian Mixture

A gaussian mixture is a powerful and useful model to estimate densities, where the density function is modeled as

$$f(x) = \sum_{\ell=1}^{k} \pi_\ell \phi(x; \mu_\ell, \Sigma_\ell)$$

Let be $Y \in \{1, \dots, k\}$ the random variable which determines the gaussian distribution that generated x, i.e.

$$\begin{aligned} \mathbb{P}(X = x) &= \sum_{\ell=1}^{k} \mathbb{P}(X = x | Y = \ell) \mathbb{P}(Y = \ell) \\ &= \sum_{\ell=1}^{k} \pi_\ell \phi_\ell(x) \end{aligned}$$

In the context of density estimation, we only have observations $(x_1, \ldots, x_n)$ that come from a unknown distribution, whose density we want to estimate thus, there is no $(y_1, \ldots, y_n)$. But in order to estimate the real density function through our gaussian mixture, we can consider the "complete" data as $((x_1, y_1), \ldots, (x_n, y_n))$ and the "observed" data as $(x_1, \ldots, x_n)$.

Moreover, we are going to denote

$$y_i = \begin{cases} (1,0,0,\ldots,0) & \text{with prob. } \pi_1 \\ (0,1,0,\ldots,0) & \text{with prob. } \pi_2 \\ \vdots & \vdots \\ (0,0,0,\ldots,1) & \text{with prob. } \pi_k \end{cases}$$

and the density of y_i can be written as

$$f(y_i|\theta) = \pi_1^{w_{i1}} \pi_2^{w_{i2}} \cdots \pi_k^{w_{ik}},$$

where $w_{ij} \in \{0,1\}$ for $j = 1, \ldots, k$ and $\sum_{j=1}^{k} w_{ij} = 1$

With this notation, we can write the conditional density of x_i given y_i as

$$f(x_i|y_i,\theta) = \phi_1^{w_{i1}}(x_i;\mu_1,\Sigma_1)\phi_2^{w_{i2}}(x_i;\mu_2,\Sigma_2)\cdots\phi_k^{w_{ik}}(x_i;\mu_k,\Sigma_k).$$

Thus, the joint distribution of the "complete" data can be written as

$$\begin{aligned} f(\mathbf{x},\mathbf{y}|\theta) &= \prod_{i=1}^{n} f(x_i,y_i|\theta) \\ &= \prod_{i=1}^{n} f(x_i|y_i,\theta)f(y_i|\theta), \end{aligned}$$

And

$$\begin{aligned}
\log f(\mathbf{x}, \mathbf{y}|\theta) &= \sum_{i=1}^{n} \log f(y_i|\theta) + \log f(x_i|y_i, \theta) \\
&= \sum_{i=1}^{n} \left(\sum_{\ell=1}^{k} w_{i\ell} \log \pi_\ell + \sum_{\ell=1}^{k} w_{i\ell} \log \phi_\ell(x_i; \mu_\ell, \Sigma_\ell) \right) \\
&= \sum_{i=1}^{n} \sum_{\ell=1}^{k} w_{i\ell} \Big(\log \pi_\ell + \log \phi_\ell(x_i; \mu_\ell, \Sigma_\ell) \Big)
\end{aligned}$$

But, the w_{ij} are not observed! To solve this problem we are going to use the EM algorithm.

Step E:

$$Q\left(\theta|\theta^{(t)}\right) = \sum_{i=1}^{n}\sum_{\ell=1}^{k}\mathbb{E}\left[w_{i\ell}|x_i,\theta^{(t)}\right]\left(\log\pi_\ell + \log\phi_\ell(x_i;\mu_\ell,\Sigma_\ell)\right),$$

let be $w_{i\ell}^{\star(t)} = \mathbb{E}\left[w_{i\ell}|x_i,\theta^{(t)}\right]$

Step M:

$$\underset{\{\pi_\ell,\mu_\ell,\Sigma_\ell\}}{\text{maximize}}\sum_{i=1}^{n}\sum_{\ell=1}^{k} w_{i\ell}^{\star(t)}\left(\log\pi_\ell + \log\phi_\ell(x_i;\mu_\ell,\Sigma_\ell)\right)$$

$$\text{subject to } \sum_{\ell=1}^{k}\pi_\ell = 1$$

The optimization problem can be solved using Lagrange multipliers. The part of the Lagrangian associated to π_j is given by

$$L(\pi_j) = \sum_{i=1}^{n} w_{ij}^{\star(t)} \log \pi_j + \lambda \left(1 - \sum_{\ell=1}^{k} \pi_\ell\right)$$

Deriving L with respect to π_j and evaluating at $(\boldsymbol{\pi}^\star, \lambda^\star)$. We have that

$$\left.\frac{\partial L}{\partial \pi_j}\right|_{(\boldsymbol{\pi}^\star, \lambda^\star)} = 0$$

$$\Leftrightarrow \frac{\sum_{i=1}^{n} w_{ij}^{\star(t)}}{\pi_j^\star} - \lambda^\star = 0$$

It follows that

$$\lambda^\star = \frac{\sum_{i=1}^n w_{ij}^{\star(t)}}{\pi_j^\star},$$

and

$$\begin{aligned}
\pi_j^\star &= \frac{1}{\lambda^\star}\sum_{i=1}^n w_{ij}^{\star(t)} \\
\Rightarrow \sum_{j=1}^k \pi_j^\star &= \frac{1}{\lambda^\star}\sum_{i=1}^n \sum_{j=1}^k w_{ij}^{\star(t)} \\
\Rightarrow 1 &= \frac{n}{\lambda^\star} \Rightarrow \lambda^\star = n.
\end{aligned}$$

Thus,

$$\pi_j^\star = \frac{1}{n}\sum_{i=1}^n w_{ij}^{\star(t)}$$

Similarly, considering the part of the Lagrangian associated to μ_j and Σ_j, it can be shown that

$$\mu_j^\star = \frac{1}{\sum_{i=1}^n w_{ij}^{\star(t)}} \sum_{i=1}^n w_{ij}^{\star(t)} x_i,$$

and

$$\Sigma_j^\star = \frac{1}{\sum_{i=1}^n w_{ij}^{\star(t)}} \sum_{i=1}^n w_{ij}^{\star(t)} (x_i - \mu_j^\star)(x_i - \mu_j^\star)^T$$

All that rest is to calculate $w_{ij}^{\star(t)}$.

First, use Bayes' theorem to show that

$$\mathbb{P}(y_i|x_i,\theta) = \frac{\mathbb{P}(x_i|y_i,\theta)\mathbb{P}(y_i|\theta)}{\mathbb{P}(x_i|\theta)}$$

$$= \frac{[\pi_1\phi_1(x_i;\mu_1,\Sigma_1)]^{w_{i1}}\cdots[\pi_k\phi_k(x_i;\mu_k,\Sigma_k)]^{w_{ik}}}{\sum_{\ell=1}^{k}\pi_\ell\phi_\ell(x_i,\mu_\ell,\Sigma_\ell)}$$

Note that $w_{ij} \in \{0,1\}$, i.e. $w_{ij} \sim$ Bernoulli and $\{w_{ij}=1\} \equiv \{y_i = (0,\ldots,1,\ldots,0)\}$, where the 1 is at the j-th position.

Thus,

$$\begin{aligned} w_{ij}^{\star(t)} &= \mathbb{E}\left[w_{ij}|x_i, \theta^{(t)}\right] \\ &= \mathbb{P}\left(w_{ij} = 1|x_i, \theta^{(t)}\right) \\ &= \mathbb{P}\left(y_i = (0, \ldots, 1, \ldots, 0)|x_i, \theta^{(t)}\right) \end{aligned}$$

$$= \frac{\pi_j^{(t)} \phi_j\left(x_i; \mu_j^{(t)}, \Sigma_j^{(t)}\right)}{\sum_{\ell=1}^k \pi_\ell^{(t)} \phi_\ell\left(x_i, \mu_\ell^{(t)}, \Sigma_\ell^{(t)}\right)}$$

Here is the final algorithm to estimate the gaussian mixture:

1. Initialize $\hat{\pi}_1^{(1)}, \ldots, \hat{\pi}_k^{(1)}, \hat{\mu}_1^{(1)}, \ldots, \hat{\mu}_k^{(1)}, \widehat{\Sigma}_1^{(1)}, \ldots, \widehat{\Sigma}_k^{(1)}$.
2. **E-step:** Compute

$$\widehat{w}_{ij}^{(t)} = \frac{\hat{\pi}_j^{(t)} \phi_j \left(x_i; \hat{\mu}_j^{(t)}, \widehat{\Sigma}_j^{(t)} \right)}{\sum_{\ell=1}^k \hat{\pi}_\ell^{(t)} \phi_\ell \left(x_i, \hat{\mu}_\ell^{(t)}, \widehat{\Sigma}_\ell^{(t)} \right)}, \quad j = 1, \ldots, k.$$

3. **M-step:** Update

$$\hat{\pi}_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \widehat{w}_{ij}^{(t)}$$

$$\hat{\mu}_j^{(t+1)} = \frac{1}{\sum_{i=1}^n \widehat{w}_{ij}^{(t)}} \sum_{i=1}^n \widehat{w}_{ij}^{(t)} x_i$$

$$\widehat{\Sigma}_j^{(t+1)} = \frac{1}{\sum_{i=1}^n \widehat{w}_{ij}^{(t)}} \sum_{i=1}^n \widehat{w}_{ij}^{(t)} \left(x_i - \hat{\mu}_j^{(t)} \right) \left(x_i - \hat{\mu}_j^{(t)} \right)^T$$

k-means

Ina clustering problem we intend to split the data into k groups. To do so we look for a point that represents the group and assign the observations to the closest cluster.

Let be Y the (unobserved) random variable that indicate to which group X belongs, then

$$\begin{aligned}\mathbb{P}(X = x) &= \sum_{\ell=1}^{k} \mathbb{P}(X = x|Y = \ell)\mathbb{P}(Y = \ell)\\ &= \sum_{\ell=1}^{k} \mathbb{P}_\ell(x;\theta_\ell)\pi_\ell\end{aligned}$$

We can make $\mathbb{P}_\ell(\cdot;\theta_\ell) = \phi(\cdot;\mu_\ell,\Sigma_\ell)$ and take μ_ℓ as the group representative. Then

$$\mathbb{P}(X = x) = \sum_{\ell=1}^{k} \pi_\ell \phi(x;\mu_\ell,\Sigma_\ell),$$

and estimate $\pi_\ell, \mu_\ell, \Sigma_\ell$.

But this is exactly the problem of estimating a gaussian mixture!

Now, two questions arise, how many groups/clusters we need and to evaluate the classification of the observations to the clusters. There are at least two ways to answer these questions.

The first way to answer the questions is through the use of the Square Sum of Residuals (SSR). If x_i was assigned to the j-th cluster, then since μ_j is represents the group, we can take $\hat{x}_i = \mu_j$ as an estimate of x_i, this the SSR is given by

$$SSR = \sum_{i=1}^{n} (x_i - \hat{x}_i)^T (x_i - \hat{x}_i)$$

However, the SSR decreases as we increase the number of clusters, then an overfitting problem can arise if we minimize SSR. Therefore, instead of minimizing the SSR, we estimate the number of clusters taking the knee of the SSR considering it as a function of the number of clusters.

The second way to answer the questions is through the use of the silhouette graphs. Let be

- a_i : the average distance of the i-th observation to the rest of the members of its cluster.
- $b_i = \min_c d(c_i, c)$: the minimum distance of x_i to the rest of the groups, and $d(x_i, c)$ is the average distance of x_i to cluster c, for all c such that $x_i \notin c$.

We define the silhouette of i as:

$$s_i = \frac{b_i - a_i}{\max\{b_i, a_i\}}$$

If

$$s_i \begin{cases} \simeq 1 & \text{then } i \text{ is well-classified,} \\ \simeq 0 & \text{then } i \text{ is in the frontier of its cluster,} \\ < 0 & \text{then } i \text{might be classified in the wring cluster.} \end{cases}$$

To determine the number of clusters, we can take $\bar{s} = \frac{1}{n} \sum_{i=1}^{n} s_i$ and maximize it as a function of the number of clusters.