Machine Learning

The Expectation Maximization (EM) algorithm

Alessandro Chiuso

January 3, 2025

The Expectation Maximization (EM) algorithm

The EM algorithm (see the famous paper "Maximum Likelihood from incomplete data via the *EM* algorithm, by Dempster-Laird-Rubin, JRSS-B, 1977) has been developed to solve the Maximum Likelihood problem

$$\hat{\theta} := \arg\max_{\theta} \; p_\theta(x) = \arg\max_{\theta} \; \log(p_\theta(x))$$

It is often useful, to this purpose, to introduce some auxiliary (non-observable) variable z ("the missing data") so that the problem

$$\arg\max_{\theta} \; p_\theta(x, z) = \arg\max_{\theta} \; \log(p_\theta(x, z))$$

becomes "simple"

The Expectation Maximization (EM) algorithm

The EM algorithm is an alternating algorithm which provides a sequence $\hat{\theta}^{(k)}$, $k \in \mathbb{N}$, satisfying the following properties:

1. $\log(p_{\hat{\theta}^{(k+1)}}(x)) \geq \log(p_{\hat{\theta}^{(k)}}(x))$
2. $\hat{\theta}^{(k)} \to \theta^*$ where θ^* is a stationary point of $\log(p_\theta(x))$
   WARNING: limit cycles may exist.

To do so, the algorithm alternates between an *Expectation step* and a *Maximization step*

Expectation step

Since the variable z is not observed, one needs to "integrate it out". Intuitively, this can be done defining the following function:

$$Q(\theta, \theta') := \mathbf{E}_{p_{\theta'}(z|x)} \log(p_\theta(x, z))$$

The following result holds.
FACT:

$$Q(\theta, \theta') \leq \log(p_\theta(x)) + C$$

where C does NOT depend on θ, and equality holds for $\theta = \theta'$. This shows that the function $Q(\theta, \theta') - C$ provides a (tight at $\theta = \theta'$) lower bound for $\log(p_\theta(x))$.

Preliminary Definition (very useful in Statistics and Information Theory): the Kulback-Leibler (KL) divergence

Given two probability density functions, $p(x)$ and $q(x)$, such that $q(x) > 0 \; \forall x$ s.t. $p(x) > 0$, we define

$$KL(p||q) := \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx = \mathbf{E}_p \log\left(\frac{p(x)}{q(x)}\right)$$

is called the Kulback-Leibler divergenge between probability densities p and q, and it is a way to measure "how different" p and q are. In particular the following theorem holds:

Theorem

$$KL(p||q) \geq 0 \qquad KL(p||q) = 0 \Leftrightarrow p(x) = q(x) \quad a.e.\ w.r.t\ p(x)$$

Kulback-Leibler (KL) divergence: PROOF

Using the fact that the fucntion $\log(x)$ satisfies $-\log(x) \geq 1 - x$ we have:

$$\begin{aligned}
KL(p||q) &= \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx \\
&= -\int \log\left(\frac{q(x)}{p(x)}\right) p(x)\, dx \\
&\geq \int \left(1 - \frac{q(x)}{p(x)}\right) p(x)\, dx \\
&= \int \left(p(x) - q(x)\right)\, dx = 0
\end{aligned}$$

In addition, if $p(x) = q(x)$, $\forall\ x\ s.t. p(x) > 0$ we have

$$KL(p||q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx = \int \log(1)\, p(x)\, dx = 0$$

Kulback-Leibler (KL) divergence: PROOF 2 (a bit more difficult)

Conversely, assume by way of contradiction that $K(p||q) = 0$ while $p(x) \neq q(x)$ on a non-zero measure se $\mathcal{X}$, i.e. such that $P_x := \int_{\mathcal{X}} p(x)\, dx > 0$ (while $p(x) \neq q(x)$, $\forall x \in \mathcal{X}^c$ Then, $-\log\left(\frac{q(x)}{p(x)}\right) > 1 - \frac{q(x)}{p(x)}$, $\forall x \in \mathcal{X}$ so that

$$
\begin{aligned}
KL(p||q) &= -\int \log\left(\frac{q(x)}{p(x)}\right) p(x)\, dx \\
&> \int \left(1 - \frac{q(x)}{p(x)}\right) p(x)\, dx \\
&= 0
\end{aligned}
$$

against the assumption that $K(p||q) = 0$.

Expectation step: proof of the bound

The proof is based on properties of the Kullback-Leibler (KL) divergence:

$$
\begin{aligned}
Q(\theta, \theta') &= \mathbf{E}_{p_{\theta'}(z|x)} \log(p_\theta(x, z)) \\
&= \mathbf{E}_{p_{\theta'}(z|x)} \log\left(\frac{p_\theta(z|x) p_\theta(x)}{p_{\theta'}(z|x)}\right) + \mathbf{E}_{p_{\theta'}(z|x)} \log(p_{\theta'}(z|x)) \\
&= \underbrace{\mathbf{E}_{p_{\theta'}(z|x)} \log\left(\frac{p_\theta(z|x)}{p_{\theta'}(z|x)}\right)}_{\leq 0 \quad (=0 \ \ iff \ \ \theta = \theta')} \\
&\quad + \log\left(p_\theta(x)\right) + \mathbf{E}_{p_{\theta'}(z|x)} \log(p_{\theta'}(z|x)) \\
&\leq \log\left(p_\theta(x)\right) + \underbrace{\mathbf{E}_{p_{\theta'}(z|x)} \log(p_{\theta'}(z|x))}_{=C}
\end{aligned}
$$

Maximization step

Given a "current" estimate $\hat{\theta}^{(k)}$ and having performed the Expectation step to compute $Q(\theta, \hat{\theta}^{(k)})$, the Maximization step is as follows:

$$\hat{\theta}^{(k+1)} = \arg\max_{\theta} \; Q(\theta, \hat{\theta}^{(k)})$$

REMARK

This implies that $Q(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}) \geq Q(\hat{\theta}^{(k)}, \hat{\theta}^{(k)})$. Using now the fact that $Q(\theta, \theta') \leq \log(p_\theta(x)) + C$ and $Q(\theta', \theta') = \log(p'_\theta(x)) + C$ it is clear that

$$\begin{aligned} \log(p_{\hat{\theta}^{(k+1)}}(x)) &\geq Q(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}) - C \\ &\geq Q(\hat{\theta}^{(k)}, \hat{\theta}^{(k)}) - C \\ &= \log(p_{\hat{\theta}^{(k)}}(x)) \end{aligned}$$

proving that the likelihood increases along the sequence $\hat{\theta}^{(k)}$.

EM Algorithm for Gaussian Mixtures Models (GMM)

Let us now consider the Gaussian Mixture Model

$$x \sim p_\theta(x)$$

where

$$p_\theta(x) = \sum_{\ell=1}^{K} \pi_\ell p_\ell(x)$$

where $\pi_\ell \geq 0$, $\sum_{\ell=1}^{K} \pi_\ell = 1$ and $p_\ell(x)$ is the density of a Gaussian random vector with mean μ_ℓ and variance Σ_ℓ. The parameter vector θ contains all the means μ_ℓ, the variances Σ_ℓ as well as the mixing probabilities π_ℓ.

EM Algorithm for Gaussian Mixtures Models (GMM) (II)

Let us now introduce the indicator variable $z \in \{1, .., K\}$ which takes value ℓ if x comes from the $\ell - th$ Gaussian so that:

$$p_\theta(x|z = \ell) = p_\ell(x)$$

With this notation the density of x is of the form

$$p_\theta(x) = \sum_{\ell=1}^{K} \underbrace{p_\theta(x|z = \ell)}_{=p_\ell(x)} \underbrace{p_\theta(z = \ell)}_{=\pi_\ell}$$

EM Algorithm for Gaussian Mixtures Models (GMM) (II)

Now, given i.i.d. observations $\{x_i\}_{i=1,..,m}$ from the Gaussian Mixture Model, their joint density takes the form:

$$p_\theta(x_1, .., x_m) = \prod_{i=1}^{m} \sum_{\ell=1}^{K} \pi_\ell p_\ell(x_i)$$

Estimation of $\theta := (\mu_\ell, \Sigma_\ell, \pi_\ell, \ell = 1, .., K)$ in the GMM can be performed using the EM algorithm, using as "hidden variable" the indicator variables z_i, $i = 1, .., N$ alternating between the following steps:

- Given $\hat{\theta}^{(k)}$ compute $Q(\theta, \hat{\theta}^{(k)})$ as above
- Maximize $Q(\theta, \hat{\theta}^{(k)})$ over θ to obtain $\hat{\theta}^{(k+1)}$

Intuition behind the introduction of the variables z_i: *if one knew from which component of the mixture each observation x_i came from, then it would be simple to estimate the parameters of the corresponding component of the mixture*

EM Algorithm for Gaussian Mixtures Models (GMM) (E-Step)

Need to compute:

$$
\begin{aligned}
Q(\theta, \hat{\theta}^{(k)}) &:= \mathbf{E}_{p_{\hat{\theta}^{(k)}}(z|x)}[\log(p_\theta(x|z)p_\theta(z)] \\
&= \mathbf{E}_{p_{\hat{\theta}^{(k)}}(z|x)}\left[\sum_{i=1}^{m} \log(p_\theta(x_i|z_i)p_\theta(z_i))\right] \\
&= \sum_{i=1}^{m} \mathbf{E}_{p_{\hat{\theta}^{(k)}}(z_i|x_i)}\left[\log(p_\theta(x_i|z_i)p_\theta(z_i))\right] \\
&= \sum_{i=1}^{m} \left\{ \sum_{\ell=1}^{K} \log(p_\theta(x_i|z_i = \ell)\underbrace{p_\theta(z_i = \ell)}_{=\pi_\ell})\underbrace{p_{\hat{\theta}^{(k)}}(z_i = \ell|x_i)}_{w_{\ell i}:=} \right\} \\
&= \sum_{i=1}^{m} \left\{ \sum_{\ell=1}^{K} \log(p_\theta(x_i|z_i = \ell)\pi_\ell)w_{\ell i} \right\}
\end{aligned}
$$

where the second to last equation defines $w_{\ell i}$.

EM Algorithm for Gaussian Mixtures Models (GMM) (E-Step, II)

Now, using the fact that

$$\log(p_\theta(x_i|z_i=\ell)) = const - \frac{1}{2}\log(\det(\Sigma_\ell)) - \frac{1}{2}(x_i-\mu_\ell)^\top \Sigma_\ell^{-1}(x_i-\mu_\ell)$$

we obtain:

$$\begin{aligned} Q(\theta,\hat{\theta}^{(k)}) \quad := \quad & const - \tfrac{1}{2}\sum_{\ell=1}^{K}\log(\det(\Sigma_\ell))\sum_{i=1}^{m} w_{\ell i} - \\ & -\tfrac{1}{2}\sum_{\ell=1}^{K}\sum_{i=1}^{m}(x_i-\mu_\ell)^\top \Sigma_\ell^{-1}(x_i-\mu_\ell) w_{\ell i} \\ & + \left\{ \sum_{\ell=1}^{K}\log(\pi_\ell)\sum_{i=1}^{m} w_{\ell i} \right\} \end{aligned}$$

EM Algorithm for Gaussian Mixtures Models (GMM) (E-Step, III)

Observation:

$$w_{\ell i} := p_{\hat{\theta}^{(k)}}(z_i = \ell | x_i) = \frac{\overbrace{p_{\hat{\theta}^{(k)}}(x_i | z_i = \ell)}^{\mathcal{N}(\hat{\mu}_\ell^{(k)}, \hat{\Sigma}_\ell^{(k)})} \overbrace{p_{\hat{\theta}^{(k)}}(z_i = \ell)}^{=\hat{\pi}_\ell^{(k)}}}{\sum_{\ell=1}^{K} p_{\hat{\theta}^{(k)}}(x_i | z_i = \ell) p_{\hat{\theta}^{(k)}}(z_i = \ell)}$$

EM Algorithm for Gaussian Mixtures Models (GMM) (M-Step, I)

To maximise w.r.t. π_ℓ under the constraint $\sum_{\ell=1}^{K} \pi_\ell = 1$ we use Lagrange multipliers

$$\Lambda(\theta, \lambda) = Q(\theta, \hat{\theta}^{(k)}) + \lambda \left(\sum_{\ell=1}^{K} \pi_\ell - 1 \right)$$

setting to zero the partial derivatives

$$\frac{\partial \Lambda(\theta, \lambda)}{\partial \pi_\ell} = \frac{1}{\pi_\ell} \sum_{i=1}^{m} w_{\ell i} + \lambda = 0$$

which, under the condition $\sum_{\ell=1}^{K} \pi_\ell = 1$ has the unique solution

$$\hat{\pi}_\ell^{(k+1)} = \frac{\sum_{i=1}^{m} w_{\ell i}}{\sum_{j=1}^{K} \sum_{i=1}^{m} w_{ji}} = \frac{1}{m} \sum_{i=1}^{m} w_{\ell i}$$

EM Algorithm for Gaussian Mixtures Models (GMM) (M-Step, I)

Similarly, taking derivatives w.r.t. μ_ℓ we obtain:

$$\begin{aligned}\frac{\partial \Lambda(\theta,\lambda)}{\partial \mu_\ell} &= \frac{\partial Q(\theta,\hat{\theta}^{(k)})}{\partial \mu_\ell} \\ &= \Sigma_\ell^{-1} \sum_{i=1}^{m} (x_i - \mu_\ell) w_{\ell i} = 0\end{aligned}$$

which admits the unique solution

$$\hat{\mu}_\ell^{(k+1)} = \frac{\sum_{i=1}^{m} x_i w_{\ell i}}{\sum_{i=1}^{m} w_{\ell i}}$$

EM Algorithm for Gaussian Mixtures Models (GMM) (M-Step, I)

Last, it is possible to show (HOMEWORK) that the solution for Σ_ℓ is given by the equation:

$$\hat{\Sigma}_\ell^{(k+1)} = \frac{\sum_{i=1}^{m} (x_i - \hat{\mu}_\ell^{(k+1)})(x_i - \hat{\mu}_\ell^{(k+1)})^\top w_{\ell i}}{\sum_{i=1}^{m} w_{\ell i}}$$