Expectation-Maximization Algorithm

Mohammad Emtiyaz Khan
EPFL

Nov 5, 2015

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

Computing maximum likelihood for Gaussian mixture model is difficult due to the log outside the sum.

$\sum_{n=1}^{N} (y_n - \beta^T \tilde{x}_n)^2$

$\sum_n \log \sum_k \pi_k e^{-\frac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1}(\cdot)}$

$\sum_n \sum_k \left(\log \pi_k - \frac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1} (x_n-\mu_k)\right)$

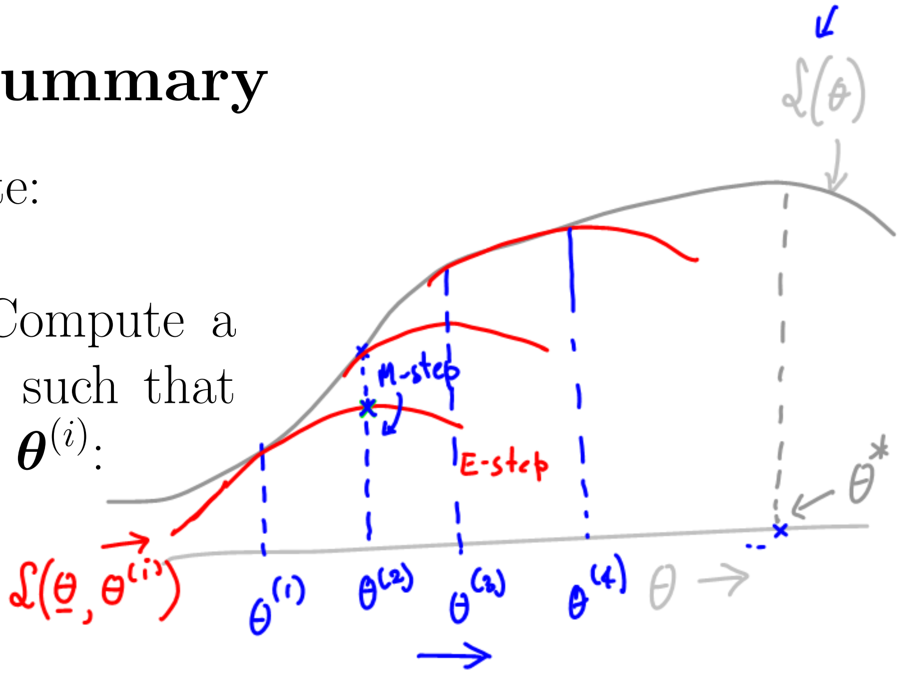
$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) := \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$\theta = \{\mu_{1:K}, \Sigma_{1:K}, \pi_{1:K}\}$

Expectation-Maximization (EM) algorithm provides an elegant and general method to optimize such optimization problems. It uses an iterative two-step procedure where individual steps usually involve problems that are easy to optimize.

EM algorithm: Summary

Start with $\boldsymbol{\theta}^{(1)}$ and iterate:



1. (Expectation step) Compute a lower bound to the cost such that it is tight at the previous $\boldsymbol{\theta}^{(i)}$:

$\mathcal{L}(\boldsymbol{\theta}) \geq \underline{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)})$ and
$\mathcal{L}(\boldsymbol{\theta}^{(i)}) = \underline{\mathcal{L}}(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i)})$.

2. (Maximization step) Update $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\theta} \underline{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}).$$

Concavity of log

Given a vector $\mathbf{p} = [p_1, p_2, \ldots, p_K]^T$ s.t. $0 < p_k < 1$, $\forall k$, $\sum_k p_k = 1$, the following holds for any $t_k > 0$:

$$\log \sum_{k=1}^{K} \pi_k \mathcal{N}(\underline{x}_n | \underline{\mu}_k, \Sigma_k) \quad p_{kn} \quad t_{kn}$$

s.t. $0 < p_{kn} < 1$, $\sum_{k=1}^{K} p_{kn} = 1$

$$\log\left(\sum_{k=1}^{K} p_k t_k\right) \geq \sum_{k=1}^{K} p_k \log t_k$$

The expectation step

A
$$\log \underbrace{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{=\ \mathcal{L}(\theta)} \geq \sum_{k=1}^{K} p_{kn}^{(i)} \log \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{p_{kn}^{(i)}}$$

$\theta \quad \theta^{(i)}$

$:= \underline{\mathcal{L}}(\theta, \theta^{(i)})$

with equality when,

To show equality at $\theta^{(i)}$

$$\mathcal{L}(\theta^{(i)}) = \underline{\mathcal{L}}(\theta^{(i)}, \theta^{(i)})$$

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

This is not a coincidence.

$\underline{\theta}^{(i)} = \{\underline{\pi}^{(i)}, \underline{\mu}^{(i)}, \Sigma^{(i)}\}$

We substitute $p_{kn}^{(i)}$ from B in A to show equality $\mathcal{L}(\theta^{(i)}) = \underline{\mathcal{L}}(\theta^{(i)}, \theta^{(i)})$

$\Downarrow$

Prior Likelihood

Let $\tilde{p}_{kn}^{(i)} = \pi_k^{(i)} \mathcal{N}(\underline{x}_n | \underline{\mu}_k^{(i)}, \Sigma_k^{(i)})$

and $p_{kn}^{(i)} = \dfrac{\tilde{p}_{kn}^{(i)}}{\sum_j \tilde{p}_{jn}^{(i)}}$

$$\sum_k p_{kn}^{(i)} \log \frac{\pi_k^{(i)} \mathcal{N}(\ )}{p_{kn}^{(i)}} = \sum_k \frac{\pi_k^{(i)} \mathcal{N}(\ )}{\sum_j \pi_j^{(i)} \mathcal{N}(\ )} \log \sum_j \pi_j^{(i)} \mathcal{N}(\ )$$

$$= \frac{\sum_k \pi_k^{(i)} \mathcal{N}(\ )}{\sum_j \pi_j^{(i)} \mathcal{N}(\ )} \log \sum_j \pi_j^{(i)} \mathcal{N}(\ )$$

$$= \log \sum_j \pi_j^{(i)} \mathcal{N}(\ )$$

$p_{kn}^{(i)}$ is the posterior (defined later)

The lower bound is $= \sum_n \sum_k p_{kn}^{(i)} \log \frac{\pi_k \mathcal{N}(\quad)}{\cancel{p_{kn}^{(i)}}}$ is constant so we can ignore it.

The maximization step

Maximize the lower bound w.r.t. $\boldsymbol{\theta}$.

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \sum_{k=1}^{K} p_{kn}^{(i)} [\log \pi_k + \underline{\log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}]$$

$$\left[\log \pi_k \quad - \frac{1}{2} (\underline{x}_n - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{x}_n - \underline{\mu}_k) \quad - \frac{1}{2} \log |\Sigma_k| \right]$$

Differentiating w.r.t. $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^{-1}$, we can get the updates for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

$$\boldsymbol{\mu}_k^{(i+1)} = \frac{\sum_{n=1}^{N} p_{kn}^{(i)} \mathbf{x}_n}{\sum_n p_{kn}^{(i)}} \qquad \frac{\sum_n r_{kn} \underline{x}_n}{\sum_{n=} r_{kn}}$$

$$\boldsymbol{\Sigma}_k^{(i+1)} = \frac{\sum_{n=1}^{N} p_{kn}^{(i)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{(i+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(i+1)})^T}{\sum_n p_{kn}^{(i)}}$$

For π_k, we use the fact that they sum to 1. Therefore, we add a Lagrangian term, differentiate w.r.t. π_k and set to 0, to get the following update:

<u>Homework</u>

$$\pi_k^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} p_{kn}^{(i)}$$

Summary of EM for GMM

Intialize $\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}, \boldsymbol{\pi}^{(1)}$ and iterate between the E and M step, until $\mathcal{L}(\boldsymbol{\theta})$ stabilizes.

1. E-step: Compute responsibilities $p_{kn}^{(i)}$:

   Prior x likelihood

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

2. Compute the marginal likelihood (cost).

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

3. M-step: Update $\boldsymbol{\mu}_k^{(i+1)}, \boldsymbol{\Sigma}_k^{(i+1)}, \pi_k^{(i+1)}$.

$$\boldsymbol{\mu}_k^{(i+1)} = \frac{\sum_{n=1}^N p_{kn}^{(i)} \mathbf{x}_n}{p_{kn}^{(i)}}$$

$$\boldsymbol{\Sigma}_k^{(i+1)} = \frac{\sum_{n=1}^N p_{kn}^{(i)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{(i+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(i+1)})^T}{p_{kn}^{(i)}}$$

$$\pi_k^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} p_{kn}^{(i)}$$

If we let, covariance be diagonal i.e. $\boldsymbol{\Sigma}_k := \sigma^2 \mathbf{I}$, then EM algorithm is same as K-means as $\sigma^2 \to 0$.

E-step $\quad p_{kn}^{(i)} \quad$ M-step

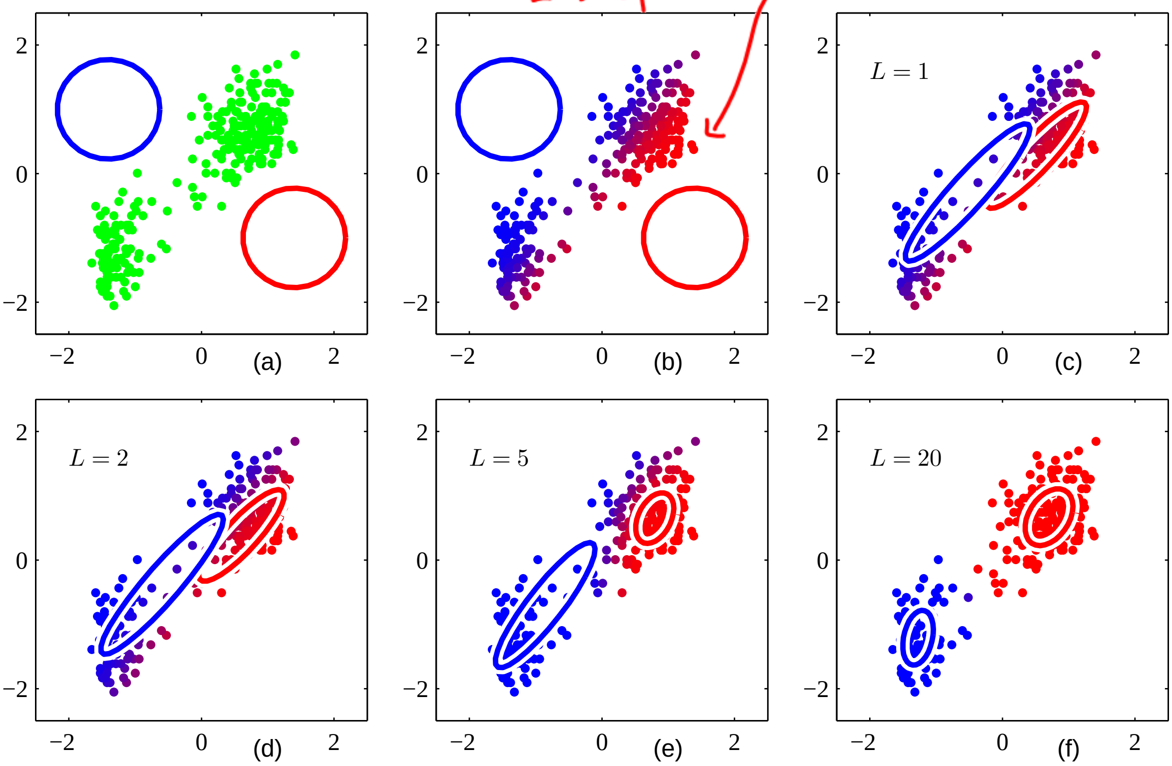


Figure 1: EM algorithm for GMM

Bayes rule enables probabistic inversion.

Forward model $\quad A\underline{x} = \underline{b}$

Inverse model $\quad \underline{x} = A^{-1}\underline{b}$

Posterior distribution

We now show that $p_{kn}^{(i)}$ is the posterior distribution of the latent variable, i.e. $p_{kn}^{(i)} = p(r_n{=}k|\mathbf{x}_n, \boldsymbol{\theta}^{(i)})$

Joint $\quad$ Likelihood $\times$ Prior $\quad$ Posterior $\times$ Marginal likelihood

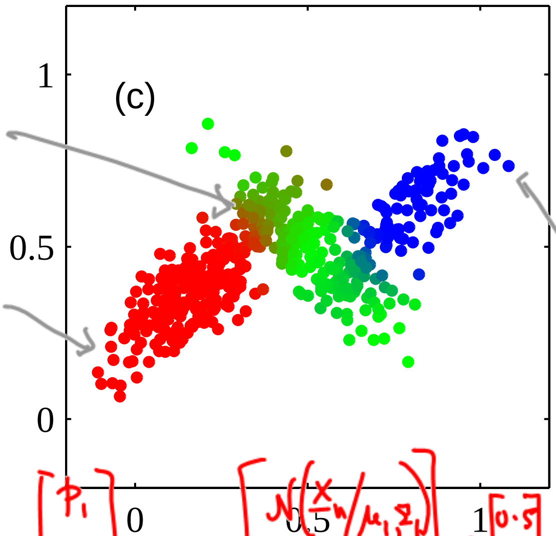
$$p(\mathbf{x}_n, r_n|\boldsymbol{\theta}) = p(\mathbf{x}_n|r_n, \boldsymbol{\theta})p(r_n|\boldsymbol{\theta}) = p(r_n|\mathbf{x}_n, \boldsymbol{\theta})p(\mathbf{x}_n|\boldsymbol{\theta})$$

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes' rule

$$p(r_n|\underline{x}_n,\theta) = \frac{p(\underline{x}_n|r_n,\theta)\,p(r_n|\theta)}{\sum_{j=1}^{K} p(\underline{x}_n, r_n{=}j|\theta)} = \frac{p(\underline{x}_n|r_n,\theta)\,p(r_n|\theta)}{\sum_{j=1}^{K} p(\underline{x}_n|r_n{=}j,\theta)\,p(r_n{=}j|\theta)}$$

$\begin{bmatrix}0.5\\0.5\\0\end{bmatrix}$ $\quad \begin{bmatrix}1\\0\\0\end{bmatrix}$



$\underline{p}_n^{(i)} = \begin{bmatrix}0 & \cdots\\0 & \cdots\\1 & \end{bmatrix}$

$$\underline{p} = \begin{bmatrix}p_1\\p_2\\p_3\end{bmatrix} = \begin{bmatrix}N(\underline{x}_n|\mu_1,\Sigma_1)\\N(\underline{x}_n|\mu_2,\Sigma_2)\\N(\underline{x}_n|\mu_3,\Sigma_3)\end{bmatrix} .* \begin{bmatrix}0.5\\0.4\\0.1\end{bmatrix}$$

$\underline{p} = \underline{p}./\mathrm{sum}(\underline{p})$

EM in general

Data $\leftarrow$ Likelihood $\times$ Prior

$$\log p(X/\theta) = \log \sum \frac{p(X, r/\theta) \times p(r/\theta^{(i)}, x)}{p(r/\theta^{(i)}, x)}$$

$$\geq \sum \left[\log p(x, r/\theta)\right] p(r/\theta^{(i)}, x) + H(\cancel{p(r/\theta^{(i)}}, x))$$

Given a generic joint distribution $p(\mathbf{x}_n, r_n|\boldsymbol{\theta})$, marginal likelihood can be lower bounded in a similar way.

EM algorithm can be compactly written as follows:

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{p(r_n|\mathbf{x}_n, \boldsymbol{\theta}^{(i)})} [\log p(\mathbf{x}_n, r_n|\boldsymbol{\theta})]$$

Another interpreation is that part of the data is missing, i.e. $(\mathbf{x}_n, r_n)$ is the "complete" data and r_n is missing. EM algorithm averages over the "unobserved" part of the data.

To do

1. Identify the joint, likelihood, prior, and marginal distributions respectively. Understand the use of Bayes rule that relates all these distributions together.
2. Derive the posterior distribution for GMM.
3. Understand the relation between EM and K-means.
4. Relate the lower bound to EM for probabilistic models in general.
5. Read the Wikipedia page on how to find a good K.
6. Read Bishop Section 14.5 to learn about conditional mixture models and mixture of experts.
7. Read about other mixture models in KPM book.