MATH 829: Introduction to Data Mining and Analysis
The EM algorithm (part 2)

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

April 20, 2016

Recall

We are given independent observations $(x^{(i)}, z^{(i)})$ with missing values $z^{(i)}$.

- Let $\theta^{(0)}$ be an **initial guess** for θ.
- Given the current estimate $\theta^{(i)}$ of θ, compute

$$Q(\theta|\theta^{(i)}) := E_{z|x;\theta^{(i)}} \log p(x, z; \theta)$$
$$= \sum_{i=1}^{n} E_{z^{(i)}|x^{(i)};\theta^{(i)}} \left(\log p(x^{(i)}, z^{(i)}; \theta)\right) \qquad \text{(E step)}$$

(In other words, we average the missing values according to their distribution after observing the observed values.)

- We then optimize $Q(\theta|\theta^{(i)})$ with respect to θ:

$$\theta^{(i+1)} := \operatorname*{argmax}_{\theta} Q(\theta|\theta^{(i)}) \qquad \text{(M step)}.$$

Theorem: The sequence $\theta^{(i)}$ constructed by the EM algorithm satisfies:

$$l(\theta^{(i+1)}) \geq l(\theta^{(i)}).$$

Convergence of the EM algorithm - Jensen's inequality

Recall: if $\phi : \mathbb{R} \to \mathbb{R}$ is convex and X is a random variable, then

$$\phi(E(X)) \leq E(\phi(X)).$$

In other words, if μ is a probability measure on Ω, $g : \Omega \to \mathbb{R}$, and $\phi : \Omega \to \mathbb{R}$ is convex, then

$$\phi\left(\int_\Omega g \, d\mu\right) \leq \int_\Omega \phi \circ g \, d\mu.$$

Note:

1. The inequality is reversed if ϕ is concave instead of convex.
2. Equality holds iff g is constant or $\phi(x) = ax + b$.

Previously, to deal with missing values, our goal was to maximize

$$\sum_{i=1}^{n} \log p(x^{(i)}; \theta) = \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta).$$

Let $Q_i(z)$ be **any** probability distribution for $z^{(i)}$, i.e.,

1. $Q_i(z) \geq 0$
2. $\sum_z Q_i(z) = 1$.

Convergence of the EM algorithm (cont.)

Then, using Jensen's inequality:

$$l(\theta^{(i)}) = \sum_{i=1}^{n} \log p(x^{(i)}; \theta) = \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$
$$= \sum_{i=1}^{n} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$
$$\geq \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

Thinking of the inner sum as an expectation with respect to the distribution Q_i, we have shown:

$$\log p(x^{(i)}; \theta) \geq E_{z^{(i)} \sim Q_i} \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}.$$

How can we choose Q_i to get the best lower bound possible?

Convergence of the EM algorithm (cont.)

At every iteration of the EM algorithm, we choose Q_i to make the inequality

$$\log p(x^{(i)};\theta) \geq E_{z^{(i)}\sim Q_i} \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}.$$

tight at our "current" estimate $\theta = \theta^{(i)}$.

By the **equality case** in Jensen's inequality,

$$\log p(x^{(i)};\theta) = E_{z^{(i)}\sim Q_i} \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

if

$$\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} = c$$

for all $z^{(i)}$. In other words: $Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)};\theta)$.

Now, for Q_i to be a probability distribution, we need to choose:

$$Q_i(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)};\theta)}{\sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta)} = p(z^{(i)}|x^{(i)};\theta).$$

Convergence of the EM algorithm (cont.)

- The previous calculation motivates the E step

$$E_{z|x;\theta^{(i)}} \ \log p(x,z;\theta)$$

in the EM algorithm.

- We will now show that $l(\theta^{(i+1)}) \geq l(\theta^{(i)})$.
- With our choice of $Q_i^{(t)}(z^{(i)}) \propto p(x^{(i)}, z^{(i)};\theta^{(t)})$ at step t, we have:

$$\begin{aligned} l(\theta^{(t)}) = \sum_{i=1}^n \log p(x^{(i)};\theta^{(t)}) &= \sum_{i=1}^n \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta^{(t)}) \\ &= \sum_{i=1}^n \log \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})} \\ &= \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})}. \end{aligned}$$

Convergence of the EM algorithm (cont.)

Now,

$$\begin{aligned} l(\theta^{(t+1)}) &= \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t+1)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta^{(t+1)})}{Q_i^{(t+1)}(z^{(i)})} \\ &\geq \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} \\ &\geq \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})} \\ &= l(\theta^{(t)}). \end{aligned}$$

- First inequality holds by Jensen's inequality (our choice of Q_i gives equality in Jensen, but the inequality holds for **any** probability distribution).
- The second inequality holds by definition of $\theta^{(t+1)}$:

$$\theta^{(i+1)} := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i^{(t)}(z^{(i)})}.$$

Example - Univariate Gaussian

- We consider a simple example to illustrate the EM algorithm.
- Suppose $W \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$.
- Suppose w_i was observed for $i = 1, \dots, m$ and w_i is missing for $i = m+1, \dots, n$.
- Let $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ be the Gaussian density.
- The likelihood function for $\theta = (\mu, \sigma^2)$ is given by

$$\begin{aligned} L(\theta) = \prod_{i=1}^n f(w_i) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(w_i-\mu)^2}{2\sigma^2}} \\ &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(w_i-\mu)^2}{2\sigma}} \times \prod_{i=m+1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(w_i-\mu)^2}{2\sigma^2}} \end{aligned}$$

Marginalizing over the unobserved values, we get:

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(\theta)\ dw_{m+1} \dots dw_n = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(w_i-\mu)^2}{2\sigma}}.$$

Example - Univariate Gaussian (cont.)

Conclusion: The MLE for (μ, σ^2) is the usual MLE for the observed values:

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} w_i, \qquad \hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m} w_i^2 - \hat{\mu}^2.$$

We will now re-derive the same result using the EM algorithm.
The log-likelihood function is:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n}\left[-\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(w_i-\mu)^2 - \frac{1}{2}\log 2\pi\right] \\ &= -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\left[n\mu^2 + \sum_{i=1}^{n} w_i^2 - 2\mu\sum_{i=1}^{n} w_i\right] \end{aligned}$$

Remark: The likelihood is linear in $\sum_{i=1}^n w_i$ and $\sum_{i=1}^n w_i^2$.

Example - Univariate Gaussian (cont.)

- The E step of the EM algorithm at step t calculates:

$$E\left(\sum_{i=1}^{n} w_i | w_i^{\mathrm{obs}}; \theta^{(t)}\right) = \sum_{i=1}^{m} w_i + (n-m)\mu^{(t)}.$$

$$E\left(\sum_{i=1}^{n} w_i^2 | w_i^{\mathrm{obs}}; \theta^{(t)}\right) = \sum_{i=1}^{m} w_i^2 + (n-m)[(\mu^{(t)})^2 + (\sigma^2)^{(t)}].$$

Note: Replacing $\sum_{i=1}^n w_i$ and $\sum_{i=1}^n w_i^2$ in $l(\theta)$ by the above expressions, the resulting function has the same "functional form" as the usual log-likelihood.

We conclude that

$$\mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^{m} w_i + \frac{(n-m)}{n}\mu^{(t)},$$

$$(\hat{\sigma}^2)^{(t+1)} = \frac{1}{n}\sum_{i=1}^{m} w_i^2 + \frac{n-m}{n}(\mu^{(t)})^2 + (\sigma^2)^{(t)}) - (\mu^{(t+1)})^2.$$

Example - Univariate Gaussian (cont.)

Usually, one would iterate the following system until convergence:

$$\mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^{m} w_i + \frac{(n-m)}{n}\mu^{(t)},$$

$$(\hat{\sigma}^2)^{(t+1)} = \frac{1}{n}\sum_{i=1}^{m} w_i^2 + \frac{n-m}{n}(\mu^{(t)})^2 + (\sigma^2)^{(t)}) - (\mu^{(t+1)})^2.$$

In this simple case, we can directly compute the limit by letting $t \to \infty$ and solving:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{m} w_i + \frac{(n-m)}{n}\hat{\mu} \quad \Rightarrow \quad \hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} w_i,$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{m} w_i^2 + \frac{n-m}{n}(\hat{\mu}^2 + \hat{\sigma}^2) - \hat{\mu}^2 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m} w_i^2 - \hat{\mu}^2.$$

We obtain the same result as in the direct approach.

Example - summary

- Of course, one would not use the EM algorithm in the univariate Gaussian case.
- The important point here is that
  1. The E step was equivalent to computing the conditional expectation of the *sufficient statistics*.
  2. The M step was equivalent to a MLE problem with **complete data** (often available in closed form).

The same phenomenon occurs when working with *exponential family* distributions:

$$f(y|\theta) = b(y)\exp(\theta^T T(x) - a(\theta)),$$

where

- θ is a vector of parameters;
- $T(y)$ is a vector of *sufficient statistics*;
- $a(\theta)$ is a normalization constant (the log partition function).

Includes: Gaussian, Bernoulli, binomial, multinomial, geometric, exponential, Poisson, Dirichlet, gamma, chi-square, etc..

Example - fitting mixture models

- The EM algorithm is also useful to fitting models where there is no missing data, but where some *hidden* parameters make the estimation difficult.
- A *mixture model* is a probability model with density

$$f(x) = \sum_{i=1}^{K} p_i f_i(x),$$

where $p_i \geq 0$, $\sum_{i=1}^{K} p_i = 1$, and each f_i is a pdf.

- To sample from such a model:
  1. Choose a category C at random according to the distribution $\{p_i\}_{i=1}^{K}$.
  2. Choose $X|C = j \sim f_j$.
- The f_i are often taken from the same parametric family (e.g. Gaussian), but don't have to.

Example - mixture of Gaussians

- Consider a mixture of p-dimensional Gaussian distributions with
  - parameters $(\mu_i, \Sigma_i)_{i=1}^{K}$,
  - mixing probabilities $(p_i)_{i=1}^{K} \subset [0,1]$, $\sum_{i=1}^{K} p_i = 1$.
- Consider a sample $(x_i)_{i=1}^{n} \subset \mathbb{R}^p$ from this model.
- The category from which each sample was obtained is **unobserved**.
- The parameters of the model are $\theta := \{\mu_i, \Sigma_i, p_i : i = 1, \dots, K\}$. The density for that model is

$$f(x) = \sum_{i=1}^{K} p_i \cdot \phi(x; \mu_i, \Sigma_i),$$

where $\phi(x; \mu, \Sigma)$ denotes the Gaussian density with parameters (μ, Σ).

- The log-likelihood function is

$$l(\theta) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} p_j \cdot \phi(x_i; \mu_j, \Sigma_j)$$

Numerically optimizing $l(\theta)$ is known to be slow and unstable.

Example - mixture of Gaussians (cont.)

- The EM algorithm approach is simpler and faster.
- Suppose our observations are (x_i, c_i) where c_i is the (unobserved) category from which x_i was drawn.
- The log-likelihood function can be written as

$$l(\theta) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} \mathbf{1}_{\{C_i = j\}} p_j \phi(x_i; \mu_j, \Sigma_j).$$

- Using Bayes' rule:

$$\begin{aligned}\pi_{ij} := P(C_i = j | X_i = x_i) &= \frac{P(X_i = x_i | C_i = j) P(C_i = j)}{\sum_{k=1}^{K} P(X_i = x_i) | C_i = k) P(C_i = k)} \\ &= \frac{p_j \phi(x_i; \mu_j, \Sigma_j)}{\sum_{k=1}^{K} p_k \phi(x_i; \mu_k, \Sigma_k)}.\end{aligned}$$

Example - mixture of Gaussians (cont.)

The EM algorithm for a mixture of Gaussians:

- **E step:** Compute the "membership probabilities" (or "responsabilities")' using the current estimate of the parameters:

$$\pi_{ij}^{(t)} = \frac{p_j^{(t)} \phi(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{k=1}^{K} p_k^{(t)} \phi(x_i; \mu_k^{(t)}, \Sigma_k^{(t)})}.$$

- **M step:** Update parameters:

$$\begin{aligned}\mu_j^{(t+1)} &= \frac{1}{N_j} \sum_{i=1}^{n} \pi_{ij}^{(t)} x_i \\ \Sigma_j^{(t+1)} &= \frac{1}{N_j} \sum_{i=1}^{n} \pi_{ij}^{(t)} (x_i - \mu_i^{(t+1)})(x_i - \mu_i^{(t+1)})^T \\ p_j^{(t+1)} &= \frac{N_k}{n},\end{aligned}$$

where

$$N_k = \sum_{i=1}^{n} \pi_{ik}^{(t)}.$$