EM Algorithm for Latent Variable Models

David Rosenberg

New York University

October 29, 2016
EM Algorithm for Latent Variable Models
General Latent Variable Model

- Two sets of random variables: $z$ and $x$.
- $z$ consists of unobserved **hidden variables**.
- $x$ consists of **observed variables**.
- Joint probability model parameterized by $\theta \in \Theta$:
  \[
p(x, z \mid \theta)
  \]
Complete and Incomplete Data

- An observation of \( x \) is called an **incomplete data set**.
- An observation \((x, z)\) is called a **complete data set**.

Suppose we have an incomplete data set \( D = (x_1, \ldots, x_n) \).

To simplify notation, take \( x \) to represent the entire dataset

\[
x = (x_1, \ldots, x_n),
\]

and \( z \) to represent the corresponding unobserved variables

\[
z = (z_1, \ldots, z_n).
\]
Complete and Incomplete Data

- An observation of $x$ is called an incomplete data set.
- An observation $(x, z)$ is called a complete data set.

Suppose we have an incomplete data set $D = (x_1, \ldots, x_n)$.

To simplify notation, take $x$ to represent the entire dataset

$$x = (x_1, \ldots, x_n),$$

and $z$ to represent the corresponding unobserved variables

$$z = (z_1, \ldots, z_n).$$
Our Objectives

- Given incomplete dataset $\mathcal{D} = x = (x_1, \ldots, x_n)$, find MLE
  $$\hat{\theta} = \arg \max_{\theta} p(\mathcal{D} | \theta).$$

- Given $x_i$, find conditional distribution over $z_i$:
  $$p(z_i | x_i, \theta).$$

- For Gaussian mixture model, this is easy once we have $\theta$:
  $$p(z_i | x_i, \theta) \propto p(z_i, x_i | \theta_i) = \pi(z_i) N(x_i | \mu_{z_i}, \Sigma_{z_i}).$$
Let \( q(z) \) be any PMF on \( Z \), the support of \( Z \):

\[
\log p(x \mid \theta) = \log \left[ \sum_z p(x, z \mid \theta) \right]
\]

\[
= \log \left[ \sum_z q(z) \left( \frac{p(x, z \mid \theta)}{q(z)} \right) \right] \quad \text{(log of an expectation)}
\]

\[
\geq \sum_z q(z) \log \left( \frac{p(x, z \mid \theta)}{q(z)} \right) \quad \text{(expectation of log)}
\]

\( \mathcal{L}(q, \theta) \)

We’ll maximize the lower \( \mathcal{L}(q, \theta) \) w.r.t. \( \theta \) and \( q \).
Consider maximizing the lower bound $\mathcal{L}(q, \theta)$:

$$\mathcal{L}(q, \theta) = \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right)$$

$$= \sum_z q(z) \log p(x, z | \theta) - \sum_z q(z) \log q(z)$$

Lower Bound and Expected Complete Log-Likelihood

Maximizing $\mathcal{L}(q, \theta)$ equivalent to maximizing $\mathbb{E}[\text{complete data log-likelihood}]$.
A Family of Lower Bounds

- Each $q$ gives a different lower bound: $\log p(x | \theta) \geq \mathcal{L}(q, \theta)$
- Two lower bounds, as functions of $\theta$:

From Bishop's *Pattern recognition and machine learning*, Figure 9.14.
Choose sequence of $q$’s and $\theta$’s by “coordinate ascent”.

EM Algorithm (high level):

1. Choose initial $\theta^{\text{old}}$.
2. Let $q^* = \arg \max_q \mathcal{L}(q, \theta^{\text{old}})$
3. Let $\theta^{\text{new}} = \arg \max_\theta \mathcal{L}(q^*, \theta^{\text{old}})$.
4. Go to step 2, until converged.

Will show: $p(x | \theta^{\text{new}}) \geq p(x | \theta^{\text{old}})$

Get sequence of $\theta$’s with monotonically increasing likelihood.
EM: Coordinate Ascent on Lower Bound

1. Start at $\theta^{old}$.
2. Find $q$ giving best lower bound at $\theta^{old} \implies \mathcal{L}(q, \theta)$.
3. $\theta^{new} = \arg \max_{\theta} \mathcal{L}(q, \theta)$.

From Bishop's *Pattern recognition and machine learning*, Figure 9.14.
The Lower Bound

- Let’s investigate the lower bound:

\[ \mathcal{L}(q, \theta) = \sum_z q(z) \log \left( \frac{p(x, z | \theta)}{q(z)} \right) \]

\[ = \sum_z q(z) \log \left( \frac{p(z | x, \theta) p(x | \theta)}{q(z)} \right) \]

\[ = \sum_z q(z) \log \left( \frac{p(z | x, \theta)}{q(z)} \right) + \sum_z q(z) \log p(x | \theta) \]

\[ = -\text{KL}[q(z), p(z | x, \theta)] + \log p(x | \theta) \]

- Amazing! We get back an equality for the marginal likelihood:

\[ \log p(x | \theta) = \mathcal{L}(q, \theta) + \text{KL}[q(z), p(z | x, \theta)] \]
The Best Lower Bound

- Find $q$ maximizing

$$\mathcal{L}(q, \theta^{\text{old}}) = -\text{KL}[q(z), p(z \mid x, \theta^{\text{old}})] + \log p(x \mid \theta^{\text{old}})$$

- Recall $\text{KL}(p\|q) \geq 0$, and $\text{KL}(p\|p) = 0$.
- Best $q$ is $q^*(z) = p(z \mid x, \theta^{\text{old}})$ and

$$\mathcal{L}(q^*, \theta^{\text{old}}) = -\text{KL}[p(z \mid x, \theta^{\text{old}}), p(z \mid x, \theta^{\text{old}})] + \log p(x \mid \theta^{\text{old}}) = 0$$

- Summary:

$$\log p(x \mid \theta^{\text{old}}) = \mathcal{L}(q^*, \theta^{\text{old}}) \quad \text{tangent at } \theta^{\text{old}}.$$  
$$\log p(x \mid \theta) \geq \mathcal{L}(q^*, \theta) \quad \forall \theta$$
Tight lower bound for any chosen $\theta$

Fix any $\theta'$ and take $q'(z) = p(z \mid x, \theta')$. Then

1. $\log p(x \mid \theta) \geq \mathcal{L}(q', \theta) \ \forall \theta$. [Global lower bound].
2. $\log p(x \mid \theta') = \mathcal{L}(q', \theta')$. [Lower bound is tight at $\theta'$.]

From Bishop's *Pattern recognition and machine learning*, Figure 9.14.
**General EM Algorithm**

1. **Choose initial** $\theta^{\text{old}}$.

2. **Expectation Step**
   - Let $q^*(z) = p(z \mid x, \theta^{\text{old}})$. [$q^*$ gives best lower bound at $\theta^{\text{old}}$]
   - Let
     \[
     J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left( \frac{p(x, z \mid \theta)}{q^*(z)} \right)
     \]
     expectation w.r.t. $z \sim q^*(z)$

3. **Maximization Step**
   \[
   \theta^{\text{new}} = \arg \max_\theta J(\theta).
   \]
   [Equivalent to maximizing expected complete log-likelihood.]

4. **Go to step 2, until converged.**
EM Monotonically Increases Likelihood
EM Gives Monotonically Increasing Likelihood: By Picture

From Bishop's *Pattern recognition and machine learning*, Figure 9.14.
EM Gives Monotonically Increasing Likelihood: By Math

1. Start at $\theta^{\text{old}}$.
2. Choose $q^*(z) = \arg \max_q \mathcal{L}(q, \theta^{\text{old}})$. We’ve shown
   \[
   \log p(x \mid \theta^{\text{old}}) = \mathcal{L}(q^*, \theta^{\text{old}})
   \]
3. Choose $\theta^{\text{new}} = \arg \max_{\theta} \mathcal{L}(q^*, \theta^{\text{old}})$. So
   \[
   \mathcal{L}(q^*, \theta^{\text{new}}) \geq \mathcal{L}(q^*, \theta^{\text{old}}).
   \]

Putting it together, we get

\[
\begin{align*}
\log p(x \mid \theta^{\text{new}}) &\geq \mathcal{L}(q^*, \theta^{\text{new}}) & \mathcal{L} \text{ is a lower bound} \\
&\geq \mathcal{L}(q^*, \theta^{\text{old}}) & \text{By definition of } \theta^{\text{new}} \\
&= \log p(x \mid \theta^{\text{old}}) & \text{Bound is tight at } \theta^{\text{old}}.
\end{align*}
\]
Suppose We Maximize the Lower Bound...

- Suppose we have found a **global maximum** of $\mathcal{L}(q, \theta)$:

\[
L(q^*, \theta^*) \geq L(q, \theta) \quad \forall q, \theta,
\]

where of course

\[
q^*(z) = p(z \mid x, \theta^*).
\]

- Claim: $\theta^*$ is a global maximum of $\log p(x \mid \theta^*)$.

- Proof: For any $\theta'$, we showed that for $q'(z) = p(z \mid x, \theta')$ we have

\[
\log p(x \mid \theta') = \mathcal{L}(q', \theta') + \text{KL}[q', p(z \mid x, \theta')]
\]

\[
= \mathcal{L}(q', \theta')
\]

\[
\leq \mathcal{L}(q^*, \theta^*)
\]

\[
= \log p(x \mid \theta^*)
\]
Convergence of EM

- Let $\theta_n$ be value of EM algorithm after $n$ steps.
- Define “transition function” $M(\cdot)$ such that $\theta_{n+1} = M(\theta_n)$.
- Suppose log-likelihood function $\ell(\theta) = \log p(x | \theta)$ is differentiable.
- Let $S$ be the set of stationary points of $\ell(\theta)$. (i.e. $\nabla_\theta \ell(\theta) = 0$)

**Theorem**

*Under mild regularity conditions*\(^a\), for any starting point $\theta_0$,

- $\lim_{n \to \infty} \theta_n = \theta^*$ for some stationary point $\theta^* \in S$ and
- $\theta^*$ is a fixed point of the EM algorithm, i.e. $M(\theta^*) = \theta^*$. Moreover,
- $\ell(\theta_n)$ strictly increases to $\ell(\theta^*)$ as $n \to \infty$, unless $\theta_n \equiv \theta^*$.

\(^a\)For details, see “Parameter Convergence for EM and MM Algorithms” by Florin Vaida in *Statistica Sinica* (2005).
Variations on EM
EM Gives Us Two New Problems

- The “E” Step: Computing

\[ J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left( \frac{p(x, z | \theta)}{q^*(z)} \right) \]

- The “M” Step: Computing

\[ \theta^{\text{new}} = \arg \max_{\theta} J(\theta). \]

- Either of these can be too hard to do in practice.
Generalized EM (GEM)

- Addresses the problem of a difficult "M" step.
- Rather than finding
  \[ \theta^{\text{new}} = \arg\max_{\theta} J(\theta), \]
  find any \( \theta^{\text{new}} \) for which
  \[ J(\theta^{\text{new}}) > J(\theta^{\text{old}}). \]
- Can use a standard nonlinear optimization strategy
  - e.g. take a gradient step on \( J \).
- We still get monotonically increasing likelihood.
Suppose “E” step is difficult:
- Hard to take expectation w.r.t. $q^*(z) = p(z | x, \theta^{old})$.

Solution: Restrict to distributions $Q$ that are easy to work with.

Lower bound now looser:

$$q^* = \arg\min_{q \in Q} KL[q(z), p(z | x, \theta^{old})]$$
EM in Bayesian Setting

- Suppose we have a prior $p(\theta)$.
- Want to find MAP estimate: $\hat{\theta}_{\text{MAP}} = \arg\max_\theta p(\theta \mid x)$:

  $$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)}$$

  $$\log p(\theta \mid x) = \log p(x \mid \theta) + \log p(\theta) - \log p(x)$$

- Still can use our lower bound on $\log p(x, \theta)$.

  $$J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left( \frac{p(x, z \mid \theta)}{q^*(z)} \right)$$

- Maximization step becomes

  $$\theta^{\text{new}} = \arg\max_\theta [J(\theta) + \log p(\theta)]$$

- Homework: Convince yourself our lower bound is still tight at $\theta$. 
Homework: Gaussian Mixture Model (Hints)
Homework: Derive EM for GMM from General EM Algorithm

- Subsequent slides may help set things up.
- Key skills:
  - MLE for multivariate Gaussian distributions.
  - Lagrange multipliers
Gaussian Mixture Model (\(k\) Components)

- **GMM Parameters**

  - Cluster probabilities: \(\pi = (\pi_1, \ldots, \pi_k)\)
  - Cluster means: \(\mu = (\mu_1, \ldots, \mu_k)\)
  - Cluster covariance matrices: \(\Sigma = (\Sigma_1, \ldots \Sigma_k)\)

- Let \(\theta = (\pi, \mu, \Sigma)\).

- **Marginal log-likelihood**

  \[
  \log p(x \mid \theta) = \log \left\{ \sum_{z=1}^{k} \pi_z \mathcal{N}(x \mid \mu_z, \Sigma_z) \right\}
  \]
Homework: Gaussian Mixture Model (Hints)

$q^*(z)$ are “Soft Assignments”

- Suppose we observe $n$ points: $X = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$.
- Let $z_1, \ldots, z_n \in \{1, \ldots, k\}$ be corresponding hidden variables.
- Optimal distribution $q^*$ is:
  \[
  q^*(z) = p(z \mid x, \theta).
  \]
- Convenient to define the conditional distribution for $z_i$ given $x_i$ as
  \[
  \gamma_i^j := p(z = j \mid x_i) = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)}
  \]
Expectation Step

- The complete log-likelihood is

\[
\log p(x, z \mid \theta) = \sum_{i=1}^{n} \log [\pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z)]
\]

\[
= \sum_{i=1}^{n} \left( \log \pi_z + \log \mathcal{N}(x_i \mid \mu_z, \Sigma_z) \right)
\]

simplifies nicely

- Take the expected complete log-likelihood w.r.t. \(q^*\):

\[
J(\theta) = \sum_z q^*(z) \log p(x, z \mid \theta)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma^j_i [\log \pi_j + \log \mathcal{N}(x_i \mid \mu_j, \Sigma_j)]
\]
Maximization Step

Find $\theta^*$ maximizing $J(\theta)$:

$$\mu_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c x_i$$

$$\Sigma_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c (x_i - \mu_{\text{MLE}}) (x_i - \mu_{\text{MLE}})^T$$

$$\pi_c^{\text{new}} = \frac{n_c}{n},$$

for each $c = 1, \ldots, k$. 