Latent Variable Models and Expectation Maximization

Piyush Rai

Introduction to Machine Learning (CS771A)

September 27, 2018
Recap: Latent Variable Models

- Assume each observation $x_n$ to be associated with a “local” latent variable $z_n$

- Parameters of $p(x|z, \theta)$ and $p(z|\phi)$ are collectively referred to as “global” parameters

- For brevity, we usually refer to the global parameters $\theta$ and $\phi$ as $\Theta = (\theta, \phi)$

- A Gaussian mixture model is an example of such a model
  - $z_n \in \{1, \ldots, K\}$ with $p(z_n|\phi) = \text{multinoulli}(\pi_1, \ldots, \pi_K)$
  - $x_n \in \mathbb{R}^D$ with $p(x_n|z_n, \theta) = \mathcal{N}(x|\mu_{z_n}, \Sigma_{z_n})$
  - Here $\Theta = (\phi, \theta) = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$

- Given data $X = \{x_1, \ldots, x_N\}$, the goal is to estimate the parameters $\Theta$ or latent variable $Z$ or both (note: we can usually estimate $\Theta$ given $Z$, and vice-versa)
Why Estimation is Difficult in LVMs?

- Suppose we want to estimate parameters $\Theta$. If we knew both $x_n$ and $z_n$ then we could do

$$\Theta_{MLE} = \arg \max_{\Theta} \sum_{n=1}^{N} \log p(x_n, z_n|\Theta) = \arg \max_{\Theta} \sum_{n=1}^{N} \left[ \log p(z_n|\phi) + \log p(x_n|z_n, \theta) \right]$$

- Simple to solve (usually closed form) if $p(z_n|\phi)$ and $p(x_n|z_n, \theta)$ are “simple” (e.g., exp-fam. dist.)

- However, in LVMs where $z_n$ is “hidden”, the MLE problem will be the following

$$\Theta_{MLE} = \arg \max_{\Theta} \sum_{n=1}^{N} \log p(x_n|\Theta) = \arg \max_{\Theta} \log p(X|\Theta)$$

- The form of $p(x_n|\Theta)$ may not be simple since we need to sum over unknown $z_n$’s possible values

$$p(x_n|\Theta) = \sum_{z_n} p(x_n, z_n|\Theta) \quad \text{... or if } z_n \text{ is continuous: } p(x_n|\Theta) = \int p(x_n, z_n|\Theta) \, dz_n$$

- The summation/integral may be intractable + may lead to complex expressions for $p(x_n|\Theta)$, in fact almost never an exponential family distribution. MLE for $\Theta$ won’t have closed form solutions!
An Important Identity

- Define $p_z = p(Z|X, \Theta)$ and let $q(Z)$ be some distribution over $Z$

- Assume discrete $Z$, the identity below holds for any choice of the distribution $q(Z)$

$$\log p(X|\Theta) = \mathcal{L}(q, \Theta) + KL(q||p_z)$$

$$\mathcal{L}(q, \Theta) = \sum_z q(Z) \log \left\{ \frac{p(X, Z|\Theta)}{q(Z)} \right\}$$

$$KL(q||p_z) = -\sum_z q(Z) \log \left\{ \frac{p(Z|X, \Theta)}{q(Z)} \right\}$$

(Exercise: Verify the above identity)

- Since $KL(q||p_z) \geq 0$, $\mathcal{L}(q, \Theta)$ is a lower-bound on $\log p(X|\Theta)$

$$\log p(X|\Theta) \geq \mathcal{L}(q, \Theta)$$

- Maximizing $\mathcal{L}(q, \Theta)$ will also improve $\log p(X|\Theta)$. Also, as we'll see, it's easier to maximize $\mathcal{L}(q, \Theta)$
Maximizing $\mathcal{L}(q, \Theta)$

- Note that $\mathcal{L}(q, \Theta)$ depends on two things $q(Z)$ and $\Theta$. Let’s do ALT-OPT for these.

- First recall the identity we had: $\log p(X|\Theta) = \mathcal{L}(q, \Theta) + \text{KL}(q||p_z)$ with

$$\mathcal{L}(q, \Theta) = \sum_Z q(Z) \log \left\{ \frac{p(X, Z|\Theta)}{q(Z)} \right\} \text{ and } \text{KL}(q||p_z) = -\sum_Z q(Z) \log \left\{ \frac{p(Z|X, \Theta)}{q(Z)} \right\}$$

- Maximize $\mathcal{L}$ w.r.t. $q$ with $\Theta$ fixed at $\Theta^{old}$: Since $\log p(X|\Theta)$ will be a constant in this case,

$$\hat{q} = \arg \max_q \mathcal{L}(q, \Theta^{old}) = \arg \min_q \text{KL}(q||p_z) = p_z = p(Z|X, \Theta^{old})$$

- Maximize $\mathcal{L}$ w.r.t. $\Theta$ with $q$ fixed at $\hat{q} = p(Z|X, \Theta^{old})$

$$\Theta^{new} = \arg \max_\Theta \mathcal{L}(\hat{q}, \Theta) = \arg \max_\Theta \sum_Z p(Z|X, \Theta^{old}) \log \frac{p(X, Z|\Theta)}{p(Z|X, \Theta^{old})} = \arg \max_\Theta \sum_Z p(Z|X, \Theta^{old}) \log p(X, Z|\Theta)$$

.. therefore, $\Theta^{new} = \arg \max_\Theta Q(\Theta, \Theta^{old})$ where $Q(\Theta, \Theta^{old}) = \mathbb{E}_{p(Z|X, \Theta^{old})}[\log p(X, Z|\Theta)]$

$Q(\Theta, \Theta^{old}) = \mathbb{E}_{p(Z|X, \Theta^{old})}[\log p(X, Z|\Theta)]$ is known as **expected complete data log-likelihood (CLL)**.
What’s Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(Z|X, \Theta^{old})$, $L(\hat{q}, \Theta)$ touches $\log p(X|\Theta)$ at $\Theta^{old}$
- Step 2: We maximize $L(\hat{q}, \Theta)$ w.r.t. $\Theta$ (equivalent to maximizing $Q(\Theta, \Theta^{old})$)
Step 1: We set \( \hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \), \( \mathcal{L}(\hat{q}, \Theta) \) touches \( \log p(\mathbf{X}|\Theta) \) at \( \Theta^{old} \).

Step 2: We maximize \( \mathcal{L}(\hat{q}, \Theta) \) w.r.t. \( \Theta \) (equivalent to maximizing \( Q(\Theta, \Theta^{old}) \)).
What’s Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(Z|X, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(X|\Theta)$ at $\Theta^{old}$
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. $\Theta$ (equivalent to maximizing $Q(\Theta, \Theta^{old})$)
Step 1: We set \( \hat{q} = p(Z|X, \Theta^{old}) \), \( L(\hat{q}, \Theta) \) touches \( \log p(X|\Theta) \) at \( \Theta^{old} \).

Step 2: We maximize \( L(\hat{q}, \Theta) \) w.r.t. \( \Theta \) (equivalent to maximizing \( Q(\Theta, \Theta^{old}) \)).
Step 1: We set $\hat{q} = p(Z|X, \Theta^{old})$, $L(\hat{q}, \Theta)$ touches $\log p(X|\Theta)$ at $\Theta^{old}$

Step 2: We maximize $L(\hat{q}, \Theta)$ w.r.t. $\Theta$ (equivalent to maximizing $Q(\Theta, \Theta^{old})$)
What’s Going On: A Visual Illustration..

- Step 1: We set \( \hat{q} = p(Z|X, \Theta^{old}) \), \( \mathcal{L}(\hat{q}, \Theta) \) touches \( \log p(X|\Theta) \) at \( \Theta^{old} \)
- Step 2: We maximize \( \mathcal{L}(\hat{q}, \Theta) \) w.r.t. \( \Theta \) (equivalent to maximizing \( Q(\Theta, \Theta^{old}) \))
Step 1: We set $\hat{q} = p(Z|X, \Theta^{old})$, $L(\hat{q}, \Theta)$ touches $\log p(X|\Theta)$ at $\Theta^{old}$

Step 2: We maximize $L(\hat{q}, \Theta)$ w.r.t. $\Theta$ (equivalent to maximizing $Q(\Theta, \Theta^{old})$)
What’s Going On: A Visual Illustration..

- Step 1: We set \( \hat{q} = p(Z|X, \Theta^{old}) \), \( \mathcal{L}(\hat{q}, \Theta) \) touches \( \log p(X|\Theta) \) at \( \Theta^{old} \)

- Step 2: We maximize \( \mathcal{L}(\hat{q}, \Theta) \) w.r.t. \( \Theta \) (equivalent to maximizing \( Q(\Theta, \Theta^{old}) \))

Local Maxima Found

Intro to Machine Learning (CS771A)

Latent Variable Models and Expectation Maximization
The two-step alternating optimization scheme we saw can never decrease $p(X|\Theta)$ (good thing).

To see this consider both steps: (1) Optimize $q$ given $\Theta = \Theta^{old}$; (2) Optimize $\Theta$ given this $q$.

- Step 1 keeps $\Theta$ fixed, so $p(X|\Theta)$ obviously can’t decrease (stays unchanged in this step).
- Step 2 maximizes the lower bound $\mathcal{L}(q, \Theta)$ w.r.t $\Theta$. Thus $p(X|\Theta)$ can’t decrease!
The Expectation Maximization (EM) Algorithm

The ALT-OPT of $\mathcal{L}(q, \Theta)$ that we saw leads to the EM algorithm (Dempster, Laird, Rubin, 1977).

**The EM Algorithm**

1. Initialize $\Theta$ as $\Theta^{(0)}$, set $t = 1$
2. Step 1: Compute posterior of latent variables given current parameters $\Theta^{(t-1)}$
   
   $$p(z_{n}^{(t)}|x_{n}, \Theta^{(t-1)}) = \frac{p(z_{n}^{(t)}|\Theta^{(t-1)})p(x_{n}|z_{n}^{(t)}, \Theta^{(t-1)})}{p(x_{n}|\Theta^{(t-1)})} \propto \text{prior} \times \text{likelihood}$$

3. Step 2: Now maximize the expected complete data log-likelihood w.r.t. $\Theta$
   
   $$\Theta^{(t)} = \arg \max_{\Theta} Q(\Theta, \Theta^{(t-1)}) = \arg \max_{\Theta} \sum_{n=1}^{N} \mathbb{E}_{p(z_{n}^{(t)}|x_{n}, \Theta^{(t-1)})} \left[ \log p(x_{n}, z_{n}^{(t)}|\Theta) \right]$$

4. If not yet converged, set $t = t + 1$ and go to step 2.

Note: If we can take the MAP estimate $\hat{z}_{n}$ of $z_{n}$ (not full posterior) in Step 1 and maximize the CLL in Step 2 using that estimate, i.e., do $\arg \max_{\Theta} \sum_{n=1}^{N} \log p(x_{n}, \hat{z}_{n}^{(t)}|\Theta)$, this will be identical to ALT-OPT.
Writing Down the Expected CLL

- Deriving the EM algorithm for any model requires finding the expression of the expected CLL

$$Q(\Theta, \Theta^{old}) = \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n, \Theta^{old})} [\log p(x_n, z_n|\Theta)]$$

$$= \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n, \Theta^{old})} [\log p(x_n|\Theta) + \log p(z_n|\Theta)]$$

- If $p(x_n|z_n, \Theta)$ and $p(z_n|\Theta)$ are exp-family distributions, expected CLL will have a simple form

- Finding the expression for the expected CLL in such cases is fairly straightforward
  - First write down the expressions for $p(x_n|z_n, \Theta)$ and $p(z_n|\Theta)$ and simplify as much as possible
  - In the resulting expressions, replace all terms containing $z_n$’s by their respective expectations, e.g.,
    - $z_n$ replaced by $\mathbb{E}_{p(z_n|x_n, \Theta^{old})}[z_n]$, i.e., the posterior mean of $z_n$
    - $z_n'z_n^\top$ replaced by $\mathbb{E}_{p(z_n|x_n, \Theta^{old})}[z_n'z_n^\top]$
    - .. and so on..

- The expected CLL may not always be computable and may need to be approximated
EM for Gaussian Mixture Model
EM for Gaussian Mixture Model

Let’s first look at the CLL. Similar to generative classification with Gaussian class-conditionals

$$\log p(X, Z|\Theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} [\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k)]$$  \hspace{1cm} (we’ve seen how we get this)

The expected CLL $Q(\Theta, \Theta^{old})$ will be

$$Q(\Theta, \Theta^{old}) = \mathbb{E}[\log p(X, Z|\Theta)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}[z_{nk}] [\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k)]$$

.. where the expectation is w.r.t. the current posterior of $z_n$, i.e., $p(z_n|x_n, \Theta^{old})$

In this case, we only need $\mathbb{E}[z_{nk}]$ which can be computed as

$$\mathbb{E}[z_{nk}] = \gamma_{nk} = 0 \times p(z_{nk} = 0|x_n, \Theta^{old}) + 1 \times p(z_{nk} = 1|x_n, \Theta^{old}) = p(z_{nk} = 1|x_n)$$

$$\propto p(z_{nk} = 1)p(x_n|z_{nk} = 1) \hspace{1cm} (from \ Bayes \ Rule)$$

Thus $\mathbb{E}[z_{nk}] \propto \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$  \hspace{1cm} (Posterior prob. that $x_n$ is generated by $k$-th Gaussian)

Note: We can finally normalize $\mathbb{E}[z_{nk}]$ as $\mathbb{E}[z_{nk}] = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{\ell=1}^{K} \pi_{\ell} \mathcal{N}(x_n|\mu_{\ell}, \Sigma_{\ell})}$ since $\sum_{k=1}^{K} \mathbb{E}[z_{nk}] = 1$
EM for Gaussian Mixture Model

1. Initialize $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ as $\Theta^{(0)}$, set $t = 1$

2. **E step:** compute the expectation of each $z_n$ (we need it in M step)

$$
E[z_{nk}^{(t)}] = \gamma_{nk}^{(t)} = \frac{\pi_k^{(t-1)} \mathcal{N}(x_n | \mu_k^{(t-1)}, \Sigma_k^{(t-1)})}{\sum_{\ell=1}^K \pi_\ell^{(t-1)} \mathcal{N}(x_n | \mu_\ell^{(t-1)}, \Sigma_\ell^{(t-1)})} \forall n, k
$$

3. Given “responsibilities” $\gamma_{nk} = E[z_{nk}]$, and $N_k = \sum_{n=1}^N \gamma_{nk}$, re-estimate $\Theta$ via MLE

$$
\mu_k^{(t)} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk}^{(t)} x_n
$$

$$
\Sigma_k^{(t)} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk}^{(t)} (x_n - \mu_k^{(t)})(x_n - \mu_k^{(t)})^T
$$

$$
\pi_k^{(t)} = \frac{N_k}{N}
$$

4. Set $t = t + 1$ and go to step 2 if not yet converged
Another Example: (Probabilistic) Dimensionality Reduction

Let’s consider a latent factor model for dimensionality reduction (will revisit this later)

\[ p(x_n|z_n, W, \sigma^2) = \mathcal{N}(Wz_n, \sigma^2 I_D) \quad p(z_n) = \mathcal{N}(0, I_K) \]

A low-dim \( z_n \in \mathbb{R}^K \) mapped to high-dim \( x_n \in \mathbb{R}^D \) via a projection matrix \( W \in \mathbb{R}^{D \times K} \)

The complete data log-likelihood for this model will be

\[
\log p(X, Z|W, \sigma^2) = \log \prod_{n=1}^{N} p(x_n, z_n|W, \sigma^2) = \log \prod_{n=1}^{N} p(x_n|z_n, W, \sigma^2) p(z_n) = \sum_{n=1}^{N} \{ \log p(x_n|z_n, W, \sigma^2) + \log p(z_n) \}
\]

Plugging in the expressions for \( p(x_n|z_n, W, \sigma^2) \) and \( p(z_n) \) and simplifying (exercise)

\[
\text{CLL} = - \sum_{n=1}^{N} \left\{ \frac{D}{2} \log \sigma^2 + \frac{1}{2\sigma^2}||x_n||^2 - \frac{1}{\sigma^2}z_n^TW^Tx_n + \frac{1}{2\sigma^2} \text{tr}(z_nz_n^TW^TW) + \frac{1}{2} \text{tr}(z_nz_n^T) \right\}
\]

Expected CLL will require replacing \( z_n \) by \( \mathbb{E}[z_n] \) and \( z_nz_n^T \) by \( \mathbb{E}[z_nz_n^T] \)

- These expectations can be obtained from the posterior \( p(z_n|x_n) \) (easy to compute due to conjugacy)
- The M step maximizes the expected CLL w.r.t. the parameters \((W, \sigma^2\) in this case)
The EM Algorithm: Some Comments

- The E and M steps may not always be possible to perform exactly. Some reasons
  - The posterior of latent variables $p(Z|X, \Theta)$ may not be easy to find
    - Would need to approximate $p(Z|X, \Theta)$ in such a case
  - Even if $p(Z|X, \Theta)$ is easy, the expected CLL, i.e., $\mathbb{E}[\log p(X, Z|\Theta)]$ may still not be tractable

$$
\mathbb{E}[\log p(X, Z|\Theta)] = \int \log p(X, Z|\Theta) p(Z|X, \Theta) dZ
$$

  - which can be approximated, e.g., using Monte-Carlo expectation (called Monte-Carlo EM)
  - Maximization of the expected CLL may not be possible in closed form

- EM works even if the M step is only solved approximately (Generalized EM)
- If M step has multiple parameters whose updates depend on each other, they are updated in an alternating fashion - called Expectation Conditional Maximization (ECM) algorithm

- Other advanced probabilistic inference algorithms are based on ideas similar to EM
  - E.g., Variational Bayesian (VB) inference