Inference in Latent Variable Models: The EM Algorithm

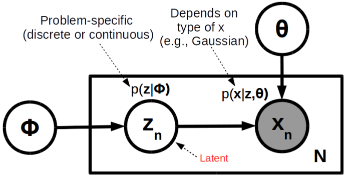
Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Feb 8, 2018

Parameter Estimation in Latent Variable Models

- Assume each observation $\boldsymbol{x}_n$ to be associated with a "local" latent variable $\boldsymbol{z}_n$



- Although we can do fully Bayesian inference for all the unknowns, suppose we only want a point estimate of the "global" parameters $\Theta = (\theta, \phi)$ via MLE/MAP
- The MLE would be $\hat{\Theta} = \arg\max_{\Theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\Theta)$ where

$$\text{if } \boldsymbol{z}_n \text{ is discrete:} \quad \log p(\boldsymbol{x}_n|\Theta) = \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) = \log \sum_{k=1}^{K} p(\boldsymbol{x}_n|\boldsymbol{z}_n = k, \Theta) p(\boldsymbol{z}_n = k|\Theta)$$

$$\text{if } \boldsymbol{z}_n \text{ is continuous:} \quad \log p(\boldsymbol{x}_n|\Theta) = \log \int p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) d\boldsymbol{z}_n = \log \int p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta) p(\boldsymbol{z}_n|\Theta) d\boldsymbol{z}_n$$

- MLE difficult: $p(\boldsymbol{x}_n|\Theta)$ is usually not in exp-family $\Rightarrow \log p(\boldsymbol{x}_n|\Theta)$ won't have a simple form

Parameter Estimation in Latent Variable Models

- Rather than MLE on $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\Theta)$, how about doing MLE on $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta)$?
  - One reason: $\log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) = \log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)p(\boldsymbol{z}_n|\Theta)$ usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ and $p(\boldsymbol{z}_n|\Theta)$ are exponential family distributions
- $p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta)$ known as complete data likelihood, $p(\boldsymbol{x}_n|\Theta)$ known as the incomplete data likelihood
- Is MLE on $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta)$ instead of our original objective $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\Theta)$ right thing?
  - Yes :-) Will see the justification shortly
- Denoting $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$, consider the new MLE problem

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

- However since $\mathbf{Z}$ is latent, we will actually maximize the expectation $\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$

$$\boxed{\hat{\Theta} = \arg\max_{\Theta} \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]}$$

.. where the expectation is w.r.t. to an "optimal" distribution over $\mathbf{Z}$ (will see shortly what it is)

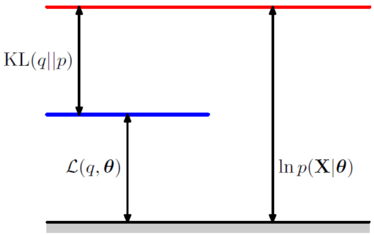
An Important Identity

- Define $p_z = p(\mathbf{Z}|\mathbf{X}, \Theta)$ and let $q(\mathbf{Z})$ be some distribution over $\mathbf{Z}$
- Assume discrete $\mathbf{Z}$, the identity below holds for any choice of the distribution $q(\mathbf{Z})$

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

$$\begin{aligned} \mathcal{L}(q, \Theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \right\} \\ \text{KL}(q||p_z) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \Theta)}{q(\mathbf{Z})} \right\} \end{aligned}$$

(Exercise: Verify the above identity)



- Since $\text{KL}(q||p_z) \geq 0$, $\mathcal{L}(q, \Theta)$ is a lower-bound on $\log p(\mathbf{X}|\Theta)$
- Maximizing $\mathcal{L}(q, \Theta)$ won't decrease $\log p(\mathbf{X}|\Theta)$. Can maximize $\log p(\mathbf{X}|\Theta)$ by maximizing $\mathcal{L}(q, \Theta)$
- Note: $\mathcal{L}(q, \Theta)$ depends on q and Θ. Consider alternating maximization w.r.t each fixing the other

An Alternating Optimization Scheme for $\mathcal{L}(q, \Theta)$

- The identity we had was $\log p(\mathbf{X}|\Theta) = \mathcal{L}(q, \Theta) + \mathrm{KL}(q||p_z)$ with

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

- With Θ fixed at Θ^{old}, $\log p(\mathbf{X}|\Theta)$ won't change. The q that maximizes $\mathcal{L}(q, \Theta)$ will be

$$\hat{q} = \arg\max_q \mathcal{L}(q, \Theta^{old})$$

.. which is attained when $\mathrm{KL}(q||p_z) = 0$. Therefore $\boxed{\hat{q} = p_z = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})}$

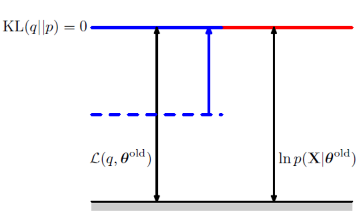
- With q fixed at $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, the Θ that maximizes $\mathcal{L}(q, \Theta)$ will be

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

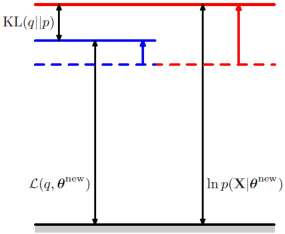
.. therefore, $\boxed{\Theta^{new} = \arg\max_\theta \mathcal{Q}(\Theta, \Theta^{old})}$ where $\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$

Why Alternating Optimization Works Here?

- The two-step alternating optimziation scheme we saw never decreases $p(\mathbf{X}|\Theta)$
- To see this consider both steps: (1) Optimize q with $\Theta = \Theta^{old}$; (2) Optimize Θ using this q



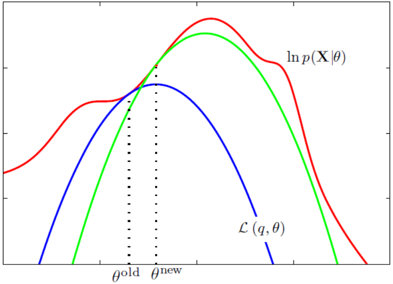
(Step 1)



(Step 2)

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

Convergence: A View in the Parameter (Θ) Space



- E-step: Update of q makes the $\mathcal{L}(q, \Theta)$ curve touch the $\log p(\mathbf{X}|\Theta)$ curve at Θ^{old}
- M-step gives the maxima Θ^{new} of $\mathcal{L}(q, \Theta^{old})$
- Next E-step readjusts $\mathcal{L}(q, \Theta^{old})$ curve (green) to meet $\log p(\mathbf{X}|\Theta)$ curve again, now at Θ^{new}
- This continues until a local maxima of $\log p(\mathbf{X}|\Theta)$ is reached

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior distribution $p(\mathbf{Z}|\mathbf{X},\Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior distribution

$$
\begin{aligned}
\mathcal{Q}(\Theta,\Theta^{old}) &= \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X},\mathbf{Z}|\Theta)] = \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\log p(\boldsymbol{x}_n,\boldsymbol{z}_n|\Theta)] \\
&= \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n,\Theta) + \log p(\boldsymbol{z}_n|\Theta)]
\end{aligned}
$$

- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. Θ

$$
\Theta^{new} = \arg\max_{\Theta} \mathcal{Q}(\Theta,\Theta^{old})
$$

- If the incomplete log-lik $p(\mathbf{X}|\Theta)$ not yet converged then set $\Theta^{old} = \Theta^{new}$ and go to the E step.

The Expected CLL

- Deriving the EM algorithm requires finding the expression of the expected CLL

$$\mathcal{Q}(\Theta, \Theta^{old}) = \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \Theta^{old})}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta) + \log p(\boldsymbol{z}_n|\Theta)]$$

- If $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ and $p(\boldsymbol{z}_n|\Theta)$ 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 $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ and $p(\boldsymbol{z}_n|\Theta)$ and simplify as much as possible
  - In the resulting expressions, replace all terms containing $\boldsymbol{z}_n$'s by their respective expectations, e.g.,
    - $\boldsymbol{z}_n$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \Theta^{old})}[\boldsymbol{z}_n]$, i.e., the posterior mean of $\boldsymbol{z}_n$
    - $\boldsymbol{z}_n\boldsymbol{z}_n^\top$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \Theta^{old})}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]$
    - .. and so on..

The Expected CLL via an example

- Let's consider a latent factor model for dimensionality reduction

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}) = \mathcal{N}(\mathbf{W}\boldsymbol{z}_n, \sigma^2\mathbf{I}_K) \qquad p(\boldsymbol{z}_n) = \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$$

- A linear Gaussian model: Low-dim $\boldsymbol{z}_n \in \mathbb{R}^K$ mapped to high-dim $\boldsymbol{x}_n \in \mathbb{R}^D$ via $\mathbf{W} \in \mathbb{W}^{D\times K}$
- The complete data log-likelihood for this model will be

$$\log p(\mathbf{X}, \mathbf{Z}|\mathbf{W}, \sigma^2) = \log \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\mathbf{W}, \sigma^2) = \log \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}, \sigma^2)p(\boldsymbol{z}_n) = \sum_{n=1}^{N}\{\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}, \sigma^2) + \log p(\boldsymbol{z}_n)\}$$

- Plugging in the expressions for $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}, \sigma^2)$ and $p(\boldsymbol{z}_n)$ and simplifying

$$CLL = -\sum_{n=1}^{N}\left\{\frac{D}{2}\log\sigma^2 + \frac{1}{2\sigma^2}||\boldsymbol{x}_n||^2 - \frac{1}{\sigma^2}\boldsymbol{z}_n^\top\mathbf{W}^\top\boldsymbol{x}_n + \frac{1}{2\sigma^2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^\top\mathbf{W}^\top\mathbf{W}) + \frac{1}{2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^\top)\right\}$$

- Expected CLL will require replacing $\boldsymbol{z}_n$ by $\mathbb{E}[\boldsymbol{z}_n]$ and $\boldsymbol{z}_n\boldsymbol{z}_n^\top$ by $\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]$
    - These expectations can be easily obtained from the posterior $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ (computed in E step)
- The M step maximizes the expected CLL w.r.t. the parameters ($\mathbf{W}, \sigma^2$ in this case)

Online or Incremental EM

- Needn't compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ for every $\boldsymbol{x}_n$ in each EM iteration (computational/storage efficiency)
  - Recall that the expected CLL is often a sum over all data points

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{n=1}^{N} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]$$

  - Can compute this quantity recursively using small minibatches of data

$$\mathcal{Q}_t = (1-\gamma_t)\mathcal{Q}_{t-1} + \gamma_t \left[\sum_{n=1}^{N_t} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]\right]$$

  .. where $\gamma_t = (1+t)^{-\kappa}, 0.5 < \kappa \leq 1$ is a decaying learning rate

  - Requires computing $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ only for data in current mini-batch (computational/storage efficiency)
  - MLE on above $\mathcal{Q}_t$ can be shown to be equivalent to a simple recursive updates for Θ

$$\Theta^{(t)} = (1-\gamma_t) \times \Theta^{(t-1)} + \gamma_t \times \arg\max_{\Theta} \underbrace{\mathcal{Q}(\Theta, \Theta^{t-1})}_{\substack{\text{computed using only} \\ \text{the } N_t \text{ examples} \\ \text{from this minibatch}}}$$

Some Applications of EM

- Mixture of (multivariate) Gaussians/Bernoullis,multinoullis, Mixture of experts models
- Problems with missing labels/features (treat these as latent variables)
- Note that EM not only gives estimates of the parameters Θ but also infers latent variables $\mathbf{Z}$
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression
  $$p(\boldsymbol{y}|\mathbf{X}, \lambda, \beta) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$
  - As an alternative, can treat $\boldsymbol{w}$ as a latent variable and β, λ as parameters and use EM to learn these
  - Note: In this case, the latent variable $\boldsymbol{w}$ is not "local" (but EM still applies)
  - E step computes posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ assuming β, λ fixed from the previous M step
  - M step maximizes $\mathbb{E}[\log p(\boldsymbol{y}, \boldsymbol{w}|\mathbf{X}, \beta, \lambda)] = \mathbb{E}[\log p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta) + \log p(\boldsymbol{w}|\lambda)]$ w.r.t. λ, β
    - This requires using expectations of quantities like $\boldsymbol{w}$ and $\boldsymbol{w}\boldsymbol{w}^\top$ which can be obtained easily from the posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ which we compute in the E step

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 $p(\mathbf{Z}|\mathbf{X},\Theta)$ may not be easy to find
    - Would need to approximate $p(\mathbf{Z}|\mathbf{X},\Theta)$ in such a case
  - Even if $p(\mathbf{Z}|\mathbf{X},\Theta)$ is easy, the expected CLL, i.e., $\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)]$ may still not be tractabe

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

    .. 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