Expectation Maximization

Piyush Rai
IIT Kanpur

Probabilistic Machine Learning (CS772A)

Jan 27, 2016

Parameter Estimation with Latent Variables

- Model $p(\mathbf{X}, \mathbf{Z}|\theta)$, observed data $\mathbf{X}$, latent variables $\mathbf{Z}$, model parameters θ
- Recall GMM, $\mathbf{Z}$: cluster assignments, θ: GMM parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$
- Goal: Estimate the model parameters θ via MLE

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

- Doing MLE in such models can be difficult because of the log-sum
- *If* we "knew" $\mathbf{Z}$, sum over all possible $\mathbf{Z}$ not needed. Just define "complete data" $\{\mathbf{X}, \mathbf{Z}\}$, and do MLE on the complete data log-lik. $\log p(\mathbf{X}, \mathbf{Z}|\theta)$
- Assumption: MLE on $\log p(\mathbf{X}, \mathbf{Z}|\theta)$ is easy
  - It often indeed is, especially when $p(\mathbf{X}, \mathbf{Z}|\theta)$ is exponential family distribution (or product of exponential family distributions)

Parameter Estimation with Latent Variables

- If MLE on $\log p(\mathbf{X}, \mathbf{Z}|\theta)$ is easy then let's do it!
- Problem: Well, we don't actually know $\mathbf{Z}$, so we are still stuck. ☹
- Solution: Use the posterior $p(\mathbf{Z}|\mathbf{X}, \theta)$ over latent variables $\mathbf{Z}$ to compute the expected complete data log-likelihood and do MLE on *that* objective

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\theta)] \\ &= \arg\max_{\theta} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \log p(\mathbf{X}, \mathbf{Z}|\theta) \end{aligned}$$

- But now we have a chicken-and-egg problem: the posterior $p(\mathbf{Z}|\mathbf{X}, \theta)$ over $\mathbf{Z}$ itself depends on the parameters θ

Solution: An Iterative Scheme (EM Algorithm)

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

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

$$\mathcal{Q}(\theta, \theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \log p(\mathbf{X}, \mathbf{Z}|\theta)$$

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

$$\begin{aligned} \theta^{new} &= \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}) \quad \text{(if doing MLE)} \\ \theta^{new} &= \arg\max_{\theta} \{\mathcal{Q}(\theta, \theta^{old}) + \log p(\theta)\} \quad \text{(if doing MAP)} \end{aligned}$$

- If the log-likelihood or the parameter values not converged then set $\theta^{old} = \theta^{new}$ and go to the E step.

Why is this doing the right thing? We'll see details in the next class. Informally, it's because we are maximizing $\mathcal{Q}(\theta, \theta^{old})$, a tight lower-bound on $\log p(\mathbf{X}|\theta)$.

Illustration: EM for GMM

- Recall that the GMM parameters $\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\} = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$
- The complete data likelihood

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} p(\boldsymbol{z}_n)p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)^{z_{nk}}$$

- Taking the log, we get:

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)\}$$

- E-step computes the expected complete data log-likelihood:

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta)}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[z_{nk}]\{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)\}$$

where $\mathbb{E}[z_{nk}]$ is the expected value of z_{nk} under the posterior

Illustration: EM for GMM (Contd.)

- The only expectation we need to compute $\mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta)}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})]$ is

$$\mathbb{E}[z_{nk}] = \sum_{z_{nk}=\{0,1\}} z_{nk}\, p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\mu_j, \Sigma_j)} = \gamma_{nk}$$

- Thus the expected complete data log-likelihood

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta)}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}\{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)\}$$

- M-step maximizes the the exp. complete data log-likelihood w.r.t. π_k, μ_k, Σ_k
- The update equations for these will be (shown on the board)

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}\boldsymbol{x}_n, \quad \Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}(\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^\top, \quad \pi_k = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N} \gamma_{nk}$ is "effective" num. of examples assigned to k^{th} Gaussian

Next class:
Why does EM work?