Expectation-Maximization

10-701/15-781, Recitation

Feb 18, 2010

Ni Lao

What's EM

- Used for finding maximum likelihood estimates of parameters in probabilistic models

- Useful when there are latent variables (incomplete data)
  - No closed form solution to the objective/gradient due to the summation over hidden variables
  - Or when we don't want the standard optimization procedures

- It alternates between two steps
  - Expectation (E) step
    - computes an expectation of the latent variables

  - Maximization (M) step
    - computes the parameters which maximize the expected log likelihood given the expectations from E-step

MLE with Hidden Variables

- We have a MLE problem

$$\max_{\theta} \log P(\mathrm{D} \mid \theta) = \max_{\theta} \sum_{l} \log P(\mathrm{x}^{l} \mid \theta)$$

- For most applications, the existence of latent variables z makes it nasty to compute expectations (here we omit the superscript *l*)

$$\log P(\mathrm{x} \mid \theta) = \log \sum_{\mathrm{z}} P(\mathrm{x}, \mathrm{z} \mid \theta)$$

- e.g.
  - z is a binary vector of length *n*, z_i are not independent
  - then there are 2^n terms in the summation
  - not affordable if dynamic programming is not applicable

MLE with GMM

- For GMM, z_i x_i are indeed independent to each other, and we can calculate the objective function efficiently

$$\log P(\mathrm{x} \mid \theta) = \log \sum_{\mathrm{z}} P(\mathrm{x} \mid \mathrm{z}, \theta) P(\mathrm{z} \mid \theta)$$

$$= \log \sum_{\mathrm{z}} \prod_{i} P(\mathrm{x}_i \mid \mathrm{z}_i, \theta) P(\mathrm{z}_i \mid \theta)$$

$$= \log \prod_{i} \sum_{\mathrm{z}_\mathrm{i}} P(\mathrm{x}_i \mid \mathrm{z}_i, \theta) P(\mathrm{z}_i \mid \theta)$$

- But we still cannot get close form solution to the parameters
  - after introducing hidden variables, the objective function is not convex anymore

- And we hate gradient ascent
  - especially with constrained optimization $\pi' 1 = 1$

Variational Method

- The variational method
  - approximates the original objective function by adding extra parameters
  - Here we introduce a set of parameter $Q(z^l)$ for each sample (x^l, z^l)

$$l(\theta) = \log P(\mathrm{x} \mid \theta) = \log \sum_{\mathrm{z}} Q(\mathrm{z}) \frac{P(\mathrm{x}, \mathrm{z} \mid \theta)}{Q(\mathrm{z})} \geq \sum_{\mathrm{z}} Q(\mathrm{z}) \log \frac{P(\mathrm{x}, \mathrm{z} \mid \theta)}{Q(\mathrm{z})} = l^{EM}(\theta, Q)$$

  - Jensen's inequality: $\log \sum_{\mathbf{z}} P(\mathbf{z})\, f(\mathbf{z}) \geq \sum_{\mathbf{z}} P(\mathbf{z}) \log f(\mathbf{z})$

- Sometimes, we constrain the distribution Q to have factorized form

$$Q(\mathrm{z}) = \prod_i Q(z_i)$$

  - therefore, we can enumerate each z_i independently instead of jointly in the summation

KL Divergence

- l^{EM}(x) is an lower bound of l(x), and the gap is a KL divergence.
  - for GMM, there is no constraint on Q(z^l), therefore the gap can be zero

$$
\begin{aligned}
l(\theta) - l^{EM}(\theta, Q) &= \log P(\mathrm{x} \mid \theta) - \sum_{\mathrm{z}} Q(\mathrm{z}) \log \frac{P(\mathrm{x}, \mathrm{z} \mid \theta)}{Q(\mathrm{z})} \\
&= \sum_{\mathrm{z}} Q(\mathrm{z}) \log P(\mathrm{x} \mid \theta) - \sum_{\mathrm{z}} Q(\mathrm{z}) \log \frac{P(\mathrm{x}, \mathrm{z} \mid \theta)}{Q(\mathrm{z})} \\
&= \sum_{\mathrm{z}} Q(\mathrm{z}) \log \frac{P(\mathrm{z} \mid \mathrm{x}, \theta)}{Q(\mathrm{z})} \\
&= KL(Q(\mathrm{z}) \,\|\, P(\mathrm{z} \mid \mathrm{x}, \theta))
\end{aligned}
$$

- KLD
  - measures the difference of two distributions
  - is never negative
  - Is zero iff the two distribution are identical

E-step

- Actually still a maximization step

$$Q^{new} = \arg\max_Q l^{EM}(\theta, Q) = \arg\min_Q KL(Q(\mathrm{z}) \| P(\mathrm{z} \mid \mathrm{x}, \theta))$$

- For GMM, just set $Q(\mathrm{z}^l) = P(\mathrm{z}^l \mid \mathrm{x}^l, \theta)$
  - here we got the name “E-step”

M-step

- Another maximization step

$$\theta^{new} = \arg\max_{\theta} l^{EM}(\theta, Q) = \arg\max_{\theta} \sum_{\mathrm{z}} Q(\mathrm{z}) \log P(\mathrm{x}, \mathrm{z} \mid \theta)$$

- For GMM (and many other directed graphic models)
  - there are closed form solutions

$$\pi_i^{(t+1)} = \frac{\sum_j \mathrm{P}\left(y = i \middle| x_j, \lambda_t\right)}{m} \qquad \mu_i^{(t+1)} = \frac{\sum_j \mathrm{P}\left(y = i \middle| x_j, \lambda_t\right) x_j}{\sum_j \mathrm{P}\left(y = i \middle| x_j, \lambda_t\right)} \qquad \Sigma_{\mathrm{i}}^{(\mathrm{t+1})} = \frac{\sum_{\mathrm{j}} \mathrm{P}\left(\mathrm{y} = \mathrm{i} \middle| \mathrm{x}_{\mathrm{j}}, \lambda_{\mathrm{t}}\right)\left(\mathrm{x}_{\mathrm{j}} - \mu_{\mathrm{i}}^{(\mathrm{t+1})}\right)\left(\mathrm{x}_{\mathrm{j}} - \mu_{\mathrm{i}}^{(\mathrm{t+1})}\right)^{\mathrm{T}}}{\sum_{\mathrm{j}} \mathrm{P}\left(\mathrm{y} = \mathrm{i} \middle| \mathrm{x}_{\mathrm{j}}, \lambda_{\mathrm{t}}\right)}$$

  - You’ve done it in HW2~~~

- For other applications (e.g. undirected graphic model)
  - this step itself may be an optimization procedure (gradient ascent, or Newton’s method)

Summery

- EM is useful when there are latent variables (incomplete data)
  - No closed form solution to the parameters
  - Hard to estimate objective/gradient due to the summation over hidden variables
  - Or when we don’t like the standard optimization procedures

- It alternates between two steps
  - Maximizing the variational parameter Q(z)
  - Maximizing the model parameter θ

- The End
- Thanks