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

\[
\max_{\theta} \log P(D | \theta) = \max_{\theta} \sum_{l} \log P(x^l | \theta)
\]

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

\[
\log P(x | \theta) = \log \sum_{z} P(x, z | \theta)
\]

• 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

$$\log P(x \mid \theta) = \log \sum_z P(x \mid z, \theta)P(z \mid \theta)$$

$$= \log \sum_z \prod_i P(x_i \mid z_i, \theta)P(z_i \mid \theta)$$

$$= \log \prod_i \sum_{z_i} P(x_i \mid z_i, \theta)P(z_i \mid \theta)$$

• 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 parameters $Q(Z^l=z^l)$ for each sample $(x^l,z^l)$

  \[
  l(\theta) = \log P(x \mid \theta) = \log \sum_z Q(z) \frac{P(x,z \mid \theta)}{Q(z)} \geq \sum_z Q(z) \log \frac{P(x,z \mid \theta)}{Q(z)} = l^{EM}(\theta, Q)
  \]

  – Jensen’s inequality: \( \log \sum_z P(z) f(z) \geq \sum_z P(z) \log f(z) \)

• Sometimes, we constrain the distribution $Q$ to have factorized form

  \[
  Q(z) = \prod_i Q(z_i)
  \]

  – 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 $Q(z')$, therefore the gap can be zero

\[
\begin{align*}
  l(\theta) - l^{EM}(\theta, Q) &= \log P(x | \theta) - \sum_z Q(z) \log \frac{P(x, z | \theta)}{Q(z)} \\
  &= \sum_z Q(z) \log P(x | \theta) - \sum_z Q(z) \log \frac{P(x, z | \theta)}{Q(z)} \\
  &= \sum_z Q(z) \log \frac{P(z | x, \theta)}{Q(z)} \\
  &= KL(Q(z) \| P(z | x, \theta))
\end{align*}
\]

- KLD
  - measures the difference of two distributions
  - is never negative
  - is zero iff the two distributions are identical
E-step

• Actually still a maximization step

\[ Q^{\text{new}} = \arg \max_{Q} l^{EM}(\theta, Q) = \arg \min_{Q} KL(Q(z) \| P(z | x, \theta)) \]

• For GMM, just set \( Q(z^l) = P(z^l | x^l, \theta) \)
  – here we got the name “E-step”
M-step

- Another maximization step

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

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

$$\pi_i^{(t+1)} = \frac{\sum_j P(y = i|x_j, \lambda_i)}{m} \quad \mu_j^{(t+1)} = \frac{\sum_j P(y = i|x_j, \lambda_i) x_j}{\sum_j P(y = i|x_j, \lambda_i)} \quad \Sigma_i^{(t+1)} = \frac{\sum_j P(y = i|x_j, \lambda_i) (x_j - \mu_i^{(t+1)}) (x_j - \mu_i^{(t+1)})^T}{\sum_j P(y = i|x_j, \lambda_i)}$$
  - You’ve done it in HW2~~~

- For other applications (e.g. undirected graphic model)
  - this step itself may be an optimization procedure like 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 $\theta$
• The End
• Thanks