Machine Learning
4771

Instructor: Tony Jebara
Topic 13

- Expectation Maximization as Bound Maximization
- EM for Maximum A Posteriori
EM as Bound Maximization

Let’s now show that EM indeed maximizes likelihood.

Bound Maximization: optimize a lower bound on $l(\theta)$.

Since log-likelihood $l(\theta)$ not concave, can’t max it directly.

Consider an auxiliary function $Q(\theta)$ which is concave.

$Q(\theta)$ kisses $l(\theta)$ at a point and is less than it elsewhere.

Monotonically increases log-likelihood.

But how to find a bound and guarantee we max it?
Jensen’s Inequality

• An important general bound from Jensen (1906)
  • For convex $f$: $f \left( E \{x\} \right) \leq E \left\{ f(x) \right\}$
  • For concave $f$: $f \left( E \{x\} \right) \geq E \left\{ f(x) \right\}$

• Expectation in discrete case is sum weight by probability
  • For convex $f$: $f \left( \sum_{i=1}^{M} p_i x_i \right) \leq \sum_{i=1}^{M} p_i f(x_i)$ when $\sum_{i=1}^{M} p_i = 1$, $p_i \geq 0$
  • For concave $f$: $f \left( \sum_{i=1}^{M} p_i x_i \right) \geq \sum_{i=1}^{M} p_i f(x_i)$ when $\sum_{i=1}^{M} p_i = 1$, $p_i \geq 0$

• Example: $f(x) = \log(x) = \text{concave and } M=2$

$$\log \left( px_1 + (1-p)x_2 \right) \geq p \log x_1 + (1-p) \log x_2$$

• Bound $\log(\text{sum})$ with $\text{sum}(\log)$

• How to apply this to mixture models?
Expectation-Maximization

\[
\begin{align*}
l(\theta) &= \sum_{n=1}^{N} \log p(x_n | \theta) \\
        &= \sum_{n=1}^{N} \log \sum_z p(x_n, z | \theta) \\
        &= \sum_{n=1}^{N} \log \sum_z p(x_n, z | \theta) \frac{p(z | x_n, \theta_t)}{p(z | x_n, \theta_t)} \\
        &= \sum_{n=1}^{N} \log \sum_z p(z | x_n, \theta_t) \frac{p(x_n, z | \theta)}{p(z | x_n, \theta_t)} \\
        \geq \sum_{n=1}^{N} \sum_z p(z | x_n, \theta_t) \log \frac{p(x_n, z | \theta)}{p(z | x_n, \theta_t)} \\
        &= \sum_{n=1}^{N} \sum_z p(z | x_n, \theta_t) \log p(x_n, z | \theta) \\
        &\quad - \sum_{n=1}^{N} \sum_z p(z | x_n, \theta_t) \log p(z | x_n, \theta_t) \\
        &= Q(\theta | \theta_t) - \text{const}
\end{align*}
\]
EM as Bound Maximization

• Now have the following bound and maximize it:

\[
\ell(\theta) \geq Q(\theta | \theta_t) - \sum_{n=1}^{N} \sum_z p(z | x_n, \theta_t) \log p(z | x_n, \theta_t)
\]

\[
\theta^{t+1} = \arg \max_{\theta} Q(\theta | \theta_t) = \arg \max_{\theta} \sum_{n=1}^{N} \sum_z p(z | x_n, \theta_t) \log p(x_n, z | \theta)
\]

• \(Q(\theta | \theta_t)\) is called **Auxiliary Function**... take derivatives of it
• This is easy for e-families... just weighted max likelihood!

• For example, Gaussian mixture:

\[
\frac{\partial Q(\theta)}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \sum_k \tau_{n,k} \log \pi_k N(\tilde{x}_n | \mu_k, \Sigma_k)
\]

\[
0 = \sum_{n=1}^{N} \tau_{n,k} \frac{\partial}{\partial \mu_k} \left( -\frac{1}{2} (\tilde{x}_n - \mu_k)^T \Sigma_k^{-1} (\tilde{x}_n - \mu_k) \right)
\]

\[
\mu_k = \frac{\sum_{n=1}^{N} \tau_{n,k} \tilde{x}_n}{\sum_{n=1}^{N} \tau_{n,k}}
\]

... similarly get \(\pi_k\) and \(\Sigma_k\)
EM as Expected Log-Likelihood

- **Incomplete Log-Likelihood**
  \[
  l(\theta) = \log p\left(\text{observed} \mid \theta\right) = \sum_{n=1}^{N} \log \sum_z p\left(x_n, z \mid \theta\right)
  \]

- **Complete Log-Likelihood**
  \[
  l^C(\theta) = \log p\left(\text{observed, hidden} \mid \theta\right) = \sum_{n=1}^{N} \log p\left(x_n, z_n \mid \theta\right)
  \]

- We don’t know the hidden variables \(z\)
- EM computes expected values of hidden \(z\) under current \(\theta_t\)
- EM chooses \(Q\) to be the *Expected Complete Log-Likelihood*

  \[
  E\left\{l^C(\theta)\right\} = \sum_{\text{hidden}} p\left(\text{hidden} \mid \text{observed}, \theta_t\right) l^C(\theta)
  = \sum_{z_1} \cdots \sum_{z_N} p\left(z_1, \ldots, z_n \mid x_1, \ldots, x_n, \theta_t\right) l^C(\theta)
  = \sum_{z_1} \cdots \sum_{z_N} \prod_n p\left(z_n \mid x_n, \theta_t\right) l^C(\theta)
  = \sum_{z_1} \cdots \sum_{z_N} \prod_n p\left(z_n \mid x_n, \theta_t\right) \sum_n \log p\left(x_n, z_n \mid \theta\right)
  = \sum_n \sum_{z_n} p\left(z_n \mid x_n, \theta_t\right) \log p\left(x_n, z_n \mid \theta\right) \sum_{z_1} \cdots \sum_{z_{i=n}} \prod_{i \neq n} p\left(z_i \mid x_i, \theta_t\right)
  = \sum_n \sum_{z_n} p\left(z_n \mid x_n, \theta_t\right) \log p\left(x_n, z_n \mid \theta\right) = Q(\theta \mid \theta_t)
  \]
EM for Max A Posteriori

- We can also do MAP instead of ML with EM (stabilizes sol’n)
  \[
  \log \text{posterior} (\theta) = \sum_{n=1}^{N} \log \sum_{z} p(x_n, z | \theta) + \log p(\theta)
  \]
- Prior doesn’t have log-sum
- The E-step remains the same: lower bound log-sum
  \[
  \log \text{posterior} (\theta) = l(\theta) + \log p(\theta) \geq E \{l^c (\theta) \} + \text{const} + \log p(\theta)
  \]
- The M-step becomes slightly different for each model
  
- For example, mixture of Gaussians with prior on covariance
  \[
  \log \text{posterior} (\theta) = \sum_{n=1}^{N} \log \sum_{k} \pi_k N(\tilde{x}_n | \tilde{\mu}_k, \Sigma_k) + \log \prod_{k} p(\Sigma_k | S, \eta)
  \]
  \[
  \log \text{posterior} (\theta) \geq \sum_{n=1}^{N} \sum_{k} \tau_{n,k} \log \pi_k N(\tilde{x}_n | \tilde{\mu}_k, \Sigma_k) + \sum_{k} \log p(\Sigma_k | S, \eta) + \text{const}
  \]
- Updates on \( \pi \) and \( \mu \) stay the same, only \( \Sigma \) is:
  \[
  \Sigma_k \leftarrow \frac{1}{\sum_{n=1}^{N} \tau_{n,k} + \eta} \left( \sum_{n=1}^{N} \tau_{n,k} (\tilde{x}_n - \tilde{\mu}_k)(\tilde{x}_n - \tilde{\mu}_k)^T + \eta S \right)
  \]
- Typically, we use the identity matrix I for S and a small eta.