Machine Learning

4771

Instructor: Tony Jebara

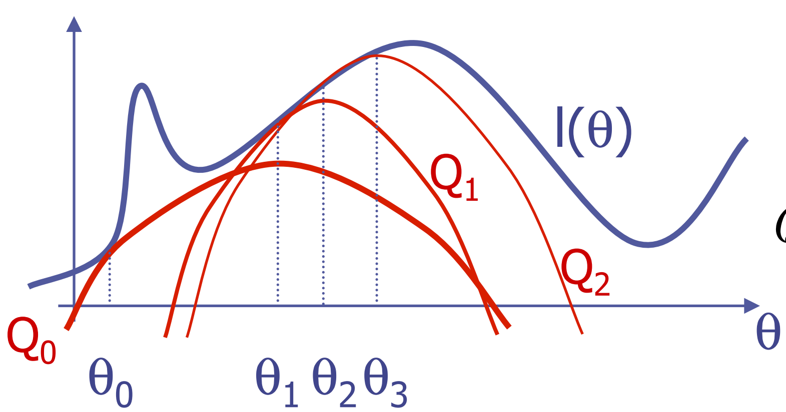
Topic 13

- 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



$$l(\theta) \geq Q_t(\theta) \quad \forall\theta \ \forall t$$

$$l(\theta_t) = Q_t(\theta_t)$$

$$\theta_{t+1} = \arg\max_\theta Q_t(\theta_t)$$

$$Q_t(\theta_{t+1}) > Q_t(\theta_t)$$

$$l(\theta_{t+1}) \geq Q_t(\theta_{t+1}) > Q_t(\theta_t)$$

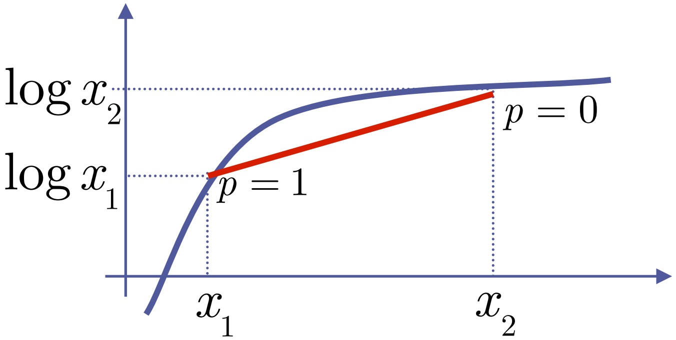
$$l(\theta_{t+1}) > l(\theta_t)$$

- 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: $f\left(E\left\{x\right\}\right) \leq E\left\{f\left(x\right)\right\}$
- For concave f: $f\left(E\left\{x\right\}\right) \geq E\left\{f\left(x\right)\right\}$
- Expectation in discrete case is sum weight by probability
- For convex f: $f\left(\sum_{i=1}^{M} p_i x_i\right) \leq \sum_{i=1}^{M} p_i f\left(x_i\right) \;\; when \; \sum_{i=1}^{M} p_i = 1, \;\; p_i \geq 0$
- For concave f: $f\left(\sum_{i=1}^{M} p_i x_i\right) \geq \sum_{i=1}^{M} p_i f\left(x_i\right) \;\; when \; \sum_{i=1}^{M} p_i = 1, \;\; p_i \geq 0$
- Example: f(x)=log(x)=concave and M=2

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



- Bound log(sum) with sum(log)
- How to apply this to mixture models?

Expectation-Maximization

$$l(\theta) = \sum_{n=1}^{N} \log p(x_n \mid \theta)$$

Original Log-Likelihood

$$= \sum_{n=1}^{N} \log \sum_z p(x_n, z \mid \theta)$$

Has Hidden Variables (messy)

$$= \sum_{n=1}^{N} \log \sum_z p(x_n, z \mid \theta) \frac{p(z \mid x_n, \theta_t)}{p(z \mid x_n, \theta_t)}$$

Multiply by 1 Ratio of hidden posterior density

$$= \sum_{n=1}^{N} \log \sum_z p(z \mid x_n, \theta_t) \frac{p(x_n, z \mid \theta)}{p(z \mid x_n, \theta_t)}$$

Rearrange

$$\geq \sum_{n=1}^{N} \sum_z p(z \mid x_n, \theta_t) \log \frac{p(x_n, z \mid \theta)}{p(z \mid x_n, \theta_t)}$$

Jensen $\log(\Sigma_i p_i x_i)$

$$= \sum_{n=1}^{N} \sum_z p(z \mid x_n, \theta_t) \log p(x_n, z \mid \theta) - \sum_{n=1}^{N} \sum_z p(z \mid x_n, \theta_t) \log p(z \mid x_n, \theta_t)$$

$$= Q(\theta \mid \theta_t) - const$$

New auxiliary function called Q (not messy)

EM as Bound Maximization

- Now have the following bound and maximize it:

$$l(\theta) \geq Q(\theta \mid \theta_t) - \sum_{n=1}^{N} \sum_z p(z \mid x_n, \theta_t) \log p(z \mid x_n, \theta_t)$$

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

$$= \arg\max_\theta \sum_{n=1}^{N} \sum_z \tau_{n,z} \log p(x_n, z \mid \theta)$$

- $Q(\theta|\theta_t)$ is called Auxiliary Function… take derivatives of it
- This is easy for e-families… just weighted max likelihood!
- For example, Gaussian mixture:

$$\frac{\partial Q(\theta)}{\partial \vec{\mu}_k} = \frac{\partial}{\partial \vec{\mu}_k} \sum_{n=1}^{N} \sum_k \tau_{n,k} \log \pi_k N(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k)$$

$$0 = \sum_{n=1}^{N} \tau_{n,k} \frac{\partial}{\partial \vec{\mu}_k} \left(-\frac{1}{2} (\vec{x}_n - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{x}_n - \vec{\mu}_k) \right)$$

$$\vec{\mu}_k = \frac{\sum_{n=1}^{N} \tau_{n,k} \vec{x}_n}{\sum_{n=1}^{N} \tau_{n,k}}$$

… similarly get π_k and Σ_k

EM as Expected Log-Likelihood

- Incomplete Log-Likelihood $$l(\theta) = \log p(observed \mid \theta) = \sum_{n=1}^{N} \log \sum_{z} p(x_n, z \mid \theta)$$
- Complete Log-Likelihood $$l^C(\theta) = \log p(observed, hidden \mid \theta) = \sum_{n=1}^{N} \log p(x_n, z_n \mid \theta)$$
- We don't know the hidden variables z
- EM computes expected values of hidden z under current θ_t
- EM chooses Q to be the Expected Complete Log-Likelihood

$$\begin{aligned}
E\left\{l^C(\theta)\right\} &= \sum_{hidden} p(hidden \mid observed, \theta_t)\, l^C(\theta) \\
&= \sum_{z_1} \cdots \sum_{z_N} p(z_1, \ldots, z_n \mid x_1, \ldots, x_n, \theta_t)\, l^C(\theta) \\
&= \sum_{z_1} \cdots \sum_{z_N} \prod_n p(z_n \mid x_n, \theta_t)\, l^C(\theta) \\
&= \sum_{z_1} \cdots \sum_{z_N} \prod_n p(z_n \mid x_n, \theta_t) \sum_n \log p(x_n, z_n \mid \theta) \\
&= \sum_n \sum_{z_n} p(z_n \mid x_n, \theta_t) \log p(x_n, z_n \mid \theta) \sum_{z_1} \cdots \sum_{z_{i \neq n}} \cdots \sum_{z_N} \prod_{i \neq n} p(z_i \mid x_i, \theta_t) \\
&= \sum_n \sum_{z_n} p(z_n \mid x_n, \theta_t) \log p(x_n, z_n \mid \theta) = Q(\theta \mid \theta_t)
\end{aligned}$$

EM for Max A Posteriori

- We can also do MAP instead of ML with EM (stabilizes sol'n)

$$\log posterior(\theta) = \sum_{n=1}^{N} \log \sum_{z} p(x_n, z \mid \theta) + \log p(\theta)$$

- Prior doesn't have log-sum
- The E-step remains the same: lower bound log-sum

$$\log posterior(\theta) = l(\theta) + \log p(\theta) \geq E\{l^C(\theta)\} + const + \log p(\theta)$$

- The M-step becomes slightly different for each model

- For example, mixture of Gaussians with prior on covariance

$$\log posterior(\theta) = \sum_{n=1}^{N} \log \sum_{k} \pi_k N(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k) + \log \prod_{k} p(\Sigma_k \mid S, \eta)$$

$$\log posterior(\theta) \geq \sum_{n=1}^{N} \sum_{k} \tau_{n,k} \log \pi_k N(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k) + \sum_{k} \log p(\Sigma_k \mid S, \eta) + const$$

- Updates on π and μ stay the same, only Σ is:

$$\Sigma_k \leftarrow \frac{1}{\sum_{n=1}^{N} \tau_{n,k} + \eta} \left(\sum_{n=1}^{N} \tau_{n,k} (\vec{x}_n - \vec{\mu}_k)(\vec{x}_n - \vec{\mu}_k)^T + \eta S \right)$$

- Typically, we use the identity matrix I for S and a small eta.