CSci 8980: Advanced Topics in Graphical Models

Mixture Models, EM

Instructor: Arindam Banerjee

September 6, 2007

Convex Functions

- Let $f : S \mapsto \mathbb{R}$, where $S \subseteq \mathbb{R}$

Convex Functions

- Let $f : S \mapsto \mathbb{R}$, where $S \subseteq \mathbb{R}$
- f is said to be convex on S if

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall x_1, x_2 \in S, \lambda \in [0,1]$$

Convex Functions

- Let $f : S \mapsto \mathbb{R}$, where $S \subseteq \mathbb{R}$
- f is said to be convex on S if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall x_1, x_2 \in S, \lambda \in [0,1]$$

- If f is twice differentiable, and $f''(x) \geq 0, \forall x \in S$, then f is convex on S

Convex Functions

- Let $f : S \mapsto \mathbb{R}$, where $S \subseteq \mathbb{R}$
- f is said to be convex on S if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall x_1, x_2 \in S, \lambda \in [0,1]$$

- If f is twice differentiable, and $f''(x) \geq 0, \forall x \in S$, then f is convex on S
- f is concave if $-f$ is convex

Convex Functions

- Let $f : S \mapsto \mathbb{R}$, where $S \subseteq \mathbb{R}$
- f is said to be convex on S if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \quad \forall x_1, x_2 \in S, \lambda \in [0,1]$$

- If f is twice differentiable, and $f''(x) \geq 0, \forall x \in S$, then f is convex on S
- f is concave if $-f$ is convex
- $log\ x$ is a concave function

Jensen's Inequality

- Let f be a convex function on S

Jensen's Inequality

- Let f be a convex function on S
- Let X be a random variable on S

Jensen's Inequality

- Let f be a convex function on S
- Let X be a random variable on S
- Jensen's inequality states

$$f(E[X]) \leq E[f(X)]$$

Jensen's Inequality

- Let f be a convex function on S
- Let X be a random variable on S
- Jensen's inequality states

$$f(E[X]) \leq E[f(X)]$$

- For the discrete case, can be proved by induction

Jensen's Inequality

- Let f be a convex function on S
- Let X be a random variable on S
- Jensen's inequality states

$$f(E[X]) \leq E[f(X)]$$

- For the discrete case, can be proved by induction
- For the general case, proof is even simpler

Proof of Jensen's Inequality

- f is convex if $\forall x_0 \; \exists$ a linear map $\ell(x) = ax + b$ s.t.

Proof of Jensen's Inequality

- f is convex if $\forall x_0 \ \exists$ a linear map $\ell(x) = ax + b$ s.t.
  - $\ell(x_0) = f(x_0)$

Proof of Jensen's Inequality

- f is convex if $\forall x_0 \; \exists$ a linear map $\ell(x) = ax + b$ s.t.
  - $\ell(x_0) = f(x_0)$
  - $\forall x \in \mathbb{R}, \; \ell(x) \leq f(x)$

Proof of Jensen's Inequality

- f is convex if $\forall x_0 \; \exists$ a linear map $\ell(x) = ax + b$ s.t.
  - $\ell(x_0) = f(x_0)$
  - $\forall x \in \mathbb{R}, \; \ell(x) \leq f(x)$
- Let ℓ be the linear map at $x_0 = E[X]$

Proof of Jensen's Inequality

- f is convex if $\forall x_0 \ \exists$ a linear map $\ell(x) = ax + b$ s.t.
  - $\ell(x_0) = f(x_0)$
  - $\forall x \in \mathbb{R}$, $\ell(x) \leq f(x)$
- Let ℓ be the linear map at $x_0 = E[X]$
- Then

$$f(E[X]) = \ell(E[X]) = E[\ell(X)] \leq E[f(X)]$$

Proof of Jensen's Inequality

- f is convex if $\forall x_0 \; \exists$ a linear map $\ell(x) = ax + b$ s.t.
  - $\ell(x_0) = f(x_0)$
  - $\forall x \in \mathbb{R}, \; \ell(x) \leq f(x)$
- Let ℓ be the linear map at $x_0 = E[X]$
- Then

$$f(E[X]) = \ell(E[X]) = E[\ell(X)] \leq E[f(X)]$$

- Uses linearity and monotonicity of expectation

Example

- Let $\lambda_i, [i]_1^n$ be a discrete distribution over $x_i, [i]_1^n$

Example

- Let $\lambda_i, [i]_1^n$ be a discrete distribution over $x_i, [i]_1^n$
- From Jensen's inequality

$$\log\left(\sum_{i=1}^{n}\lambda_i x_i\right) \geq \sum_{i=1}^{n}\lambda_i \log x_i$$

Example

- Let $\lambda_i, [i]_1^n$ be a discrete distribution over $x_i, [i]_1^n$
- From Jensen's inequality

$$\log\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq \sum_{i=1}^{n} \lambda_i \log x_i$$

- Can be applied to prove the AM-GM inequality

$$\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \geq \sum_{i=1}^{n} \frac{1}{n} \log x_i = \frac{1}{n} \log\left(\prod_{i=1}^{n} x_i\right)$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

ML Parameter Estimation

- The goal is to maximize the log-likelihood of the observations

$$L(\theta) = \log p(x|\theta)$$

ML Parameter Estimation

- The goal is to maximize the log-likelihood of the observations

$$L(\theta) = \log p(x|\theta)$$

- EM is an iterative procedure for maximizing $L(\theta)$

ML Parameter Estimation

- The goal is to maximize the log-likelihood of the observations

$$L(\theta) = \log p(x|\theta)$$

- EM is an iterative procedure for maximizing $L(\theta)$
- Applicable to a variety of settings (with missing variables)

ML Parameter Estimation

- The goal is to maximize the log-likelihood of the observations

$$L(\theta) = \log p(x|\theta)$$

- EM is an iterative procedure for maximizing $L(\theta)$
- Applicable to a variety of settings (with missing variables)
- Our focus will be primarily on mixture models

ML Parameter Estimation

- The goal is to maximize the log-likelihood of the observations

$$L(\theta) = \log p(x|\theta)$$

- EM is an iterative procedure for maximizing $L(\theta)$
- Applicable to a variety of settings (with missing variables)
- Our focus will be primarily on mixture models
- If θ_n is n^{th} iterate, want to maximize

$$L(\theta) - L(\theta_n) = \log p(x|\theta) - \log p(x|\theta_n)$$

ML Parameter Estimation

- The goal is to maximize the log-likelihood of the observations

$$L(\theta) = \log p(x|\theta)$$

- EM is an iterative procedure for maximizing $L(\theta)$
- Applicable to a variety of settings (with missing variables)
- Our focus will be primarily on mixture models
- If θ_n is n^{th} iterate, want to maximize

$$L(\theta) - L(\theta_n) = \log p(x|\theta) - \log p(x|\theta_n)$$

- If z denotes the latent variable, then $p(X|\theta) = \sum_z p(x, z|\theta)$

A Lower Bound

- Now

$$
\begin{aligned}
L(\theta) - L(\theta_n) &= \log\left(\sum_z p(x, z|\theta)\right) - \log p(x|\theta_n) \\
&= \log\left(\sum_z p(z|x, \theta_n)\frac{p(x, z|\theta)}{p(z|x, \theta_n)}\right) - \log p(x|\theta_n) \\
&\geq \sum_z p(z|x, \theta_n)\log\left(\frac{p(x, z|\theta)}{p(z|x, \theta_n)}\right) - \log p(x|\theta_n) \\
&= \sum_z p(z|x, \theta_n)\log\left(\frac{p(x, z|\theta)}{p(x, z|\theta_n)}\right) \\
&= \Delta(\theta, \theta_n)
\end{aligned}
$$

A Lower Bound (Contd.)

- Hence, we have a lower bound

$$L(\theta) \geq Q(\theta, \theta_n) = L(\theta_n) + \Delta(\theta, \theta_n)$$

A Lower Bound (Contd.)

- Hence, we have a lower bound

$$L(\theta) \geq Q(\theta, \theta_n) = L(\theta_n) + \Delta(\theta, \theta_n)$$

- Further, at $\theta = \theta_n$,

$$L(\theta) = Q(\theta, \theta_n)$$

A Lower Bound (Contd.)

- Hence, we have a lower bound

$$L(\theta) \geq Q(\theta, \theta_n) = L(\theta_n) + \Delta(\theta, \theta_n)$$

- Further, at $\theta = \theta_n$,

$$L(\theta) = Q(\theta, \theta_n)$$

- $Q(\theta, \theta_n)$ is an *auxliary function*

A Lower Bound (Contd.)

- Hence, we have a lower bound

$$L(\theta) \geq Q(\theta, \theta_n) = L(\theta_n) + \Delta(\theta, \theta_n)$$

- Further, at $\theta = \theta_n$,

$$L(\theta) = Q(\theta, \theta_n)$$

- $Q(\theta, \theta_n)$ is an *auxliary function*
- Goal: Find θ such that $Q(\theta, \theta_n)$ is maximized

Maximizing the lower bound

- Note that

$$\begin{aligned}\theta_{n+1} &= \text{argmax}_\theta \, Q(\theta, \theta_n) \\ &= \text{argmax}_\theta \left\{ \sum_z p(z|x, \theta_n) \log p(x, z|\theta) \right\} \\ &= \text{argmax}_\theta \, E_{z|x,\theta_n}[\log p(x, z|\theta)]\end{aligned}$$

Maximizing the lower bound

- Note that

$$\begin{aligned}\theta_{n+1} &= \text{argmax}_\theta\, Q(\theta, \theta_n) \\ &= \text{argmax}_\theta \left\{ \sum_z p(z|x, \theta_n) \log p(x, z|\theta) \right\} \\ &= \text{argmax}_\theta\, E_{z|x,\theta_n}[\log p(x, z|\theta)]\end{aligned}$$

- Same as maximizing the expected complete log-likelihood

Maximizing the lower bound

- Note that

$$\begin{aligned}\theta_{n+1} &= \text{argmax}_\theta\, Q(\theta, \theta_n) \\ &= \text{argmax}_\theta \left\{ \sum_z p(z|x, \theta_n) \log p(x, z|\theta) \right\} \\ &= \text{argmax}_\theta\, E_{z|x,\theta_n}[\log p(x, z|\theta)]\end{aligned}$$

- Same as maximizing the expected complete log-likelihood
- Exact update will depend on the distribution/family

Optimizing the lower bound (Contd)

- There are two steps in the EM update

Optimizing the lower bound (Contd)

- There are two steps in the EM update
- E-step: Determines the expectation $E_{z|x,\theta_n}[\log p(x, z|\theta)]$

Optimizing the lower bound (Contd)

- There are two steps in the EM update
- E-step: Determines the expectation $E_{z|x,\theta_n}[\log p(x, z|\theta)]$
- M-step: Maximize the expectation w.r.t. θ

Optimizing the lower bound (Contd)

- There are two steps in the EM update
- E-step: Determines the expectation $E_{z|x,\theta_n}[\log p(x, z|\theta)]$
- M-step: Maximize the expectation w.r.t. θ
- Determining $p(z|x, \theta_n)$ often forms the core of the E-step

Optimizing the lower bound (Contd)

- There are two steps in the EM update
- E-step: Determines the expectation $E_{z|x,\theta_n}[\log p(x,z|\theta)]$
- M-step: Maximize the expectation w.r.t. θ
- Determining $p(z|x,\theta_n)$ often forms the core of the E-step
- For FMMs, it can be computed using Bayes rule

$$p(z|x,\theta_n) = \frac{p(z|\theta_n)p(x|z,\theta_n)}{\sum_{z'} p(z'|\theta_n)p(x|z',\theta_n)}$$

Lower Bounding Function

- Both E- and M-steps solve a maximization problem

Lower Bounding Function

- Both E- and M-steps solve a maximization problem
- Consider the function

$$F(\tilde{p}, \theta) = E_{\tilde{p}}[\log p(x, z|\theta)] + H(\tilde{p})$$

Lower Bounding Function

- Both E- and M-steps solve a maximization problem
- Consider the function

$$F(\tilde{p}, \theta) = E_{\tilde{p}}[\log p(x, z|\theta)] + H(\tilde{p})$$

- $H(\tilde{p}) = E_{\tilde{p}}[-\log \tilde{p}(z)]$ is the Shannon entropy

Lower Bounding Function

- Both E- and M-steps solve a maximization problem
- Consider the function

$$F(\tilde{p}, \theta) = E_{\tilde{p}}[\log p(x, z|\theta)] + H(\tilde{p})$$

- $H(\tilde{p}) = E_{\tilde{p}}[-\log \tilde{p}(z)]$ is the Shannon entropy
- Both steps can be seen as alternately maximizing $F(\tilde{p}, \theta)$

Lower Bounding Function

- Both E- and M-steps solve a maximization problem
- Consider the function

$$F(\tilde{p}, \theta) = E_{\tilde{p}}[\log p(x, z|\theta)] + H(\tilde{p})$$

- $H(\tilde{p}) = E_{\tilde{p}}[-\log \tilde{p}(z)]$ is the Shannon entropy
- Both steps can be seen as alternately maximizing $F(\tilde{p}, \theta)$
- Can be viewed in terms of KL-divergence between $p_\theta = p(z|x, \theta)$ and $\tilde{p}$

$$F(\tilde{p}, \theta) = L(\theta) - KL(p_\theta || \tilde{p})$$

Optimizing w.r.t. $\tilde{p}$

- For a fixed θ, there is a unique distribution, p_θ, that maximizes $F(\tilde{p}, \theta)$

Optimizing w.r.t. $\tilde{p}$

- For a fixed θ, there is a unique distribution, p_θ, that maximizes $F(\tilde{p}, \theta)$
- The maximizer $p_\theta(z) = p(z|x, \theta)$

Optimizing w.r.t. $\tilde{p}$

- For a fixed θ, there is a unique distribution, p_θ, that maximizes $F(\tilde{p}, \theta)$
- The maximizer $p_\theta(z) = p(z|x, \theta)$
- Follows from the KL-divergence based expression for $F(\tilde{p}, \theta)$

Optimizing w.r.t. $\tilde{p}$

- For a fixed θ, there is a unique distribution, p_θ, that maximizes $F(\tilde{p}, \theta)$
- The maximizer $p_\theta(z) = p(z|x, \theta)$
- Follows from the KL-divergence based expression for $F(\tilde{p}, \theta)$
- Alternatively, can be derived using direct optimization

Optimizing w.r.t. $\tilde{p}$

- For a fixed θ, there is a unique distribution, p_θ, that maximizes $F(\tilde{p}, \theta)$
- The maximizer $p_\theta(z) = p(z|x, \theta)$
- Follows from the KL-divergence based expression for $F(\tilde{p}, \theta)$
- Alternatively, can be derived using direct optimization
- p_θ varies continuously with θ

Optimizing w.r.t. $\tilde{p}$ (Contd.)

- If $\tilde{p}(z) = p(z|x,\theta)$, then $F(\tilde{p},\theta) = \log p(x|\theta) = L(\theta)$

Optimizing w.r.t. $\tilde{p}$ (Contd.)

- If $\tilde{p}(z) = p(z|x,\theta)$, then $F(\tilde{p},\theta) = \log p(x|\theta) = L(\theta)$
- For $\tilde{p}(z) = p(z|x,\theta)$,

$$\begin{aligned} F(\tilde{p},\theta) &= E_{\tilde{p}}[\log p(x,z|\theta)] + H(\tilde{p}) \\ &= E_{\tilde{p}}[\log p(x,z|\theta)] - E_{\tilde{p}}[\log p(z|x,\theta)] \\ &= E_{\tilde{p}}[\log p(x,z|\theta) - \log p(z|x,\theta)] \\ &= E_{\tilde{p}}[\log p(x|\theta)] \\ &= \log p(x|\theta) \end{aligned}$$

EM as Alternate Maximization

- EM can be viewed as an alternate maximization algorithm

EM as Alternate Maximization

- EM can be viewed as an alternate maximization algorithm
- E-step: Set $\tilde{p}_{n+1}$ to the maximizer of $F(\tilde{p}, \theta_n)$

EM as Alternate Maximization

- EM can be viewed as an alternate maximization algorithm
- E-step: Set $\tilde{p}_{n+1}$ to the maximizer of $F(\tilde{p}, \theta_n)$
- M-step: Set θ_{n+1} to the maximizer of $F(\tilde{p}_{n+1}, \theta)$

EM as Alternate Maximization

- EM can be viewed as an alternate maximization algorithm
- E-step: Set $\tilde{p}_{n+1}$ to the maximizer of $F(\tilde{p}, \theta_n)$
- M-step: Set θ_{n+1} to the maximizer of $F(\tilde{p}_{n+1}, \theta)$
- The iterations are equivalent to the ones discussed earlier