Lec 6: Missing Data & EM Algorithm

Yanjun Han

Oct. 8, 2024

Missing data in exponential families.

$$(x_1, y_1), \cdots, (x_n, y_n) \overset{i.i.d.}{\sim} p_\theta(x,y) = \exp(\langle \theta, T(x,y)\rangle - A(\theta))\, h(x,y)$$

with $(x_1, \cdots, x_n)$: unobserved variables $(y_1, \cdots, y_n)$: observed variables

Goal. Find the MLE for θ.

Incomplete log-likelihood:

$$\begin{aligned} \ell_\theta(y_1, \cdots, y_n) &= \sum_{i=1}^n \log p_\theta(y_i) \\ &= \sum_{i=1}^n \log \int_x p_\theta(x, y_i)\, dx \\ &= \sum_{i=1}^n \log \int_x \exp(\langle \theta, T(x,y_i)\rangle - A(\theta))\, h(x, y_i)\, dx \\ &= \sum_{i=1}^n \Big(\underbrace{\log \int_x \exp(\langle \theta, T(x,y_i)\rangle)\, h(x,y_i)\, dx}_{=:\, A_{y_i}(\theta)} - A(\theta) \Big) \end{aligned}$$

The conditional distribution $p_\theta(x \mid y_i)$ also belongs to an exponential family, with log-partition function $A_{y_i}(\theta) \Rightarrow A_{y_i}(\theta)$ is convex in θ

$\ell_\theta(y_1, \cdots, y_n) = \sum_{i=1}^n (A_{y_i}(\theta) - A(\theta))$ is the difference of two convex functions, which may not be concave in θ !!

Detour: a short introduction of convex duality

Def (convex conjugate) The convex conjugate of a function f on $\mathbb{R}^d$ is

$$f^*(t) = \max_{x \in \mathbb{R}^d} \langle t, x\rangle - f(x)$$

Properties. 1) The maximizer $x^* = (\nabla f)^{-1}(t) = \nabla f^*(t)$, for convex f.

Pf. differentiation gives $t = \nabla f(x^*) \Longrightarrow x^* = (\nabla f)^{-1}(t)$.

$$\begin{aligned} \nabla f^*(t) &= \nabla \Big(\max_{x \in \mathbb{R}^d} \langle t, x\rangle - f(x) \Big) \\ &= \{ \nabla_t (\langle t, x^*\rangle - f(x^*)) : x^* \in \arg\max_x \langle t, x\rangle - f(x) \} = x^* \end{aligned}$$

2) $f^*(t)$ is convex in t

Pf. Because $f^*(t)$ is a maximum over linear functions of t.

3) For convex f, $f(x) = \max_{t \in \mathbb{R}^d} \langle x, t \rangle - f^*(t)$ (in other words, $f^{**} = f$)

Pf. Definition of f^* $\Rightarrow$ $f^*(t) + f(x) \geq \langle x, t \rangle \quad \forall x, t$

$\Rightarrow f(x) \geq \max_t \langle x, t \rangle - f^*(t)$.

Property 1) $\Rightarrow f^*(t) = \langle t, (\nabla f)^{-1}(t) \rangle - f((\nabla f)^{-1}(t))$

$\Rightarrow f^*(\nabla f(x)) = \langle \nabla f(x), x \rangle - f(x)$

$\Rightarrow f(x) = \langle \nabla f(x), x \rangle - f^*(\nabla f(x)) \leq \max_t \langle t, x \rangle - f^*(t)$

EM algorithm. Using convex duality,

$$
\begin{aligned}
\max_\theta \ell_\theta(y_1, \cdots, y_n) &= \max_\theta \sum_{i=1}^n (A_{y_i}(\theta) - A(\theta)) \\
&= \max_\theta \sum_{i=1}^n \left(\max_{\mu_i} \langle \mu_i, \theta \rangle - A^*_{y_i}(\mu_i) - A(\theta) \right) \\
&= \max_\theta \max_{\mu_1, \cdots, \mu_n} \underbrace{\sum_{i=1}^n \left(\langle \mu_i, \theta \rangle - A^*_{y_i}(\mu_i) - A(\theta) \right)}_{=: f(\theta, \mu_1, \cdots, \mu_n)}
\end{aligned}
$$

Key intuitions: 1. for fixed θ, $f(\theta, \mu_1, \cdots, \mu_n)$ is concave in $(\mu_1, \cdots, \mu_n)$

2. for fixed $(\mu_1, \cdots, \mu_n)$, $f(\theta, \mu_1, \cdots, \mu_n)$ is concave in θ

(Warning: $f(\theta, \mu_1, \cdots, \mu_n)$ is NOT "jointly" concave in θ and $(\mu_1, \cdots, \mu_n)$!)

Idea. Successive maximization starting from some $\theta^{(0)}$: $\theta^{(0)} \to \mu^{(1)} \to \theta^{(1)} \to \mu^{(2)} \to \cdots$

1) E-step: fix $\theta^{(t)}$, find the maximizer $\mu^{(t+1)}$

$$\mu_i^{(t+1)} = \nabla A_{y_i}(\theta^{(t)}) = \mathbb{E}_{x \sim p_{\theta^{(t)}}(x \mid y_i)}[T(x, y_i)] \quad (\text{"expectation" step})$$

2) M-step: fix $\mu^{(t+1)}$, find the maximizer $\theta^{(t+1)}$

$$\nabla A(\theta^{(t+1)}) = \frac{1}{n} \sum_{i=1}^n \mu_i^{(t+1)} \quad (\text{"maximization" step})$$

EM Intuition:

1. E-step: for each sample i with missing data x_i, think of a fake $\tilde{x}_i \sim p_\theta(x_i \mid y_i)$ and compute sufficient statistic $\mu_i = \mathbb{E}[T(\tilde{x}_i, y_i)]$.
2. M-step: aggregate the sufficient statistics "as if" there were no missing data problem: $\frac{1}{n}\sum_{i=1}^{n} \mu_i = \nabla A(\theta)$
3. Iterate the above process.

Example: Gaussian mixture model

$p_\theta(x, y)$: $\quad \mathbb{P}(x = j) = \pi_j, \quad j = 1, 2, \cdots, k$

$\qquad y \mid x = j \sim N(\mu_j, 1), \quad j = 1, 2, \cdots, k.$

Unknown parameter: $\theta = (\pi_1, \cdots, \pi_k, \mu_1, \cdots, \mu_k)$

Unobserved variable: $x_1, \cdots, x_n$

Observed variable: $y_1, \cdots, y_n$.

φ: pdf of $N(0,1)$

E-step: Given $\theta = \theta^{(t)}$, understand $p_\theta(x \mid y)$:

$$p_\theta(x = j \mid y) = \frac{p_\theta(x = j, y)}{p_\theta(y)} = \frac{p_\theta(y \mid x = j)\, p_\theta(x = j)}{\sum_{i=1}^{k} p_\theta(y \mid x = i)\, p_\theta(x = i)} = \frac{\pi_j\, \varphi(y - \mu_j)}{\sum_i \pi_i\, \varphi(y - \mu_i)}$$

$$\Rightarrow \quad p_{i,j}^{(t+1)} := p_{\theta^{(t)}}(x_i = j \mid y_i) = \frac{\pi_j^{(t)}\, \varphi(y_i - \mu_j^{(t)})}{\sum_{\ell=1}^{k} \pi_\ell^{(t)}\, \varphi(y_i - \mu_\ell^{(t)})}, \quad \begin{array}{l} i = 1, \cdots, n \\ j = 1, \cdots, k \end{array}$$

M-step: Pretend that $x_i \sim p_{\theta^{(t)}}(\cdot \mid y_i)$, maximize the log-likelihood

$$\sum_{i=1}^{n} \mathbb{E}_{x_i \sim p_{\theta^{(t)}}(\cdot \mid y_i)}[\log p_\theta(x_i, y_i)] = \sum_{i=1}^{n}\sum_{j=1}^{k} p_{i,j}^{(t+1)} \log p_\theta(x_i = j, y_i)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} p_{i,j}^{(t+1)} \left(\log \pi_j - \frac{(y_i - \mu_j)^2}{2} - \log\sqrt{2\pi}\right)$$

$$= \sum_{j=1}^{k} \left[\left(\sum_{i=1}^{n} p_{i,j}^{(t+1)}\right) \log \pi_j - \frac{1}{2}\sum_{i=1}^{n} (y_i - \mu_j)^2 p_{i,j}^{(t+1)} + \text{Const.} \right]$$

Carrying out the maximization over $\theta = (\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K)$ gives

$$\begin{cases} \pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{i,j}^{(t+1)} . \\ \mu_j^{(t+1)} = \dfrac{\sum_{i=1}^{n} p_{i,j}^{(t+1)} y_i}{\sum_{i=1}^{n} p_{i,j}^{(t+1)}} . \end{cases}$$

General EM via evidence lower bound

Def. For probability distributions P, Q over $\mathcal{X}$, the Kullback-Leibler (KL) divergence is

$$D_{KL}(P \| Q) = \begin{cases} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} , & \text{for pmfs} \\ \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx , & \text{for pdfs} \end{cases}$$

Thm. $D_{KL}(P \| Q) \geq 0$.

Pf. Since $\log t \geq 1 - \frac{1}{t}$ for all $t \geq 0$,

$$D_{KL}(P \| Q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq \sum_x p(x) \left(1 - \frac{q(x)}{p(x)} \right)$$

$$= \sum_x p(x) - \sum_x q(x) = 1 - 1 = 0$$

Evidence lower bound (ELBO).

$$\log p_\theta(y^n) = \max_{q(\cdot)} \underbrace{\mathbb{E}_{x^n \sim q(\cdot)} \left[\log \frac{p_\theta(x^n, y^n)}{q(x^n)} \right]}_{\text{ELBO}}$$

Pf. $\log p_\theta(y^n) - \text{ELBO} = \mathbb{E}_{x^n \sim q(\cdot)} \left[\log \frac{p_\theta(y^n) q(x^n)}{p_\theta(x^n, y^n)} \right]$

$$= \mathbb{E}_{x^n \sim q(\cdot)} \left[\log \frac{q(x^n)}{p_\theta(x^n | y^n)} \right] = D_{KL}(q(x^n) \| p_\theta(x^n | y^n)).$$

So $\log p_\theta(y^n) \geq \text{ELBO}$, with equality iff $q(x^n) = p_\theta(x^n | y^n)$.

General EM: $\max_\theta \log p_\theta(y^n) = \max_\theta \max_{q(x^n)} \mathbb{E}_{x^n \sim q(\cdot)}\left[\log \frac{p_\theta(x^n, y^n)}{q(x^n)}\right]$

- Fix θ: the maximizer is $q^*(x^n) = p_\theta(x^n | y^n)$
- Fix q: solve the maximization $\theta \mapsto Q(\theta | q) := \mathbb{E}_{x^n \sim q(\cdot)}[\log p_\theta(x^n, y^n)]$, which is often tractable

Example 1: exponential family $p_\theta(x, y) = \exp(\langle \theta, T(x, y)\rangle - A(\theta))h(x, y)$.
If $q(x^n) = p_{\theta^{(t)}}(x^n | y^n)$, then

$$Q(\theta | q) := Q(\theta | \theta^{(t)}) = \mathbb{E}_{x^n \sim p_{\theta^{(t)}}(\cdot | y^n)}[\log p_\theta(x^n, y^n)]$$

$$= \sum_{i=1}^n \langle \theta, \mathbb{E}_{x_i \sim p_{\theta^{(t)}}(\cdot | y_i)}[T(x_i, y_i)] \rangle - nA(\theta) + \text{const}$$

So maximizing $\theta \mapsto Q(\theta | \theta^{(t)})$ requires

- evaluation of $\mathbb{E}_{x_i \sim p_{\theta^{(t)}}(\cdot | y_i)}[T(x_i, y_i)]$ (E-step)
- solving the equation $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x_i \sim p_{\theta^{(t)}}(\cdot | y_i)}[T(x_i, y_i)] = \nabla A(\theta^{(t+1)})$ (M-step)

Example 2: gradient descent

$$Q(\theta | \theta^{(t)}) = \mathbb{E}_{x^n \sim p_{\theta^{(t)}}(\cdot | y^n)}[\log p_\theta(x^n, y^n)]$$

GD update for maximizing $\theta \mapsto Q(\theta | \theta^{(t)})$:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha\, \mathbb{E}_{x^n \sim p_{\theta^{(t)}}(\cdot | y^n)}\left[\nabla_\theta \log p_\theta(x^n, y^n)\big|_{\theta = \theta^{(t)}}\right]$$

Case study: variational autoencoders (VAE)

Target: given $y_1, \dots, y_n$ (e.g. images), find θ (e.g. params of a deep net) s.t.

$$y \mid x \sim N(\mu_\theta(x), \sigma_\theta^2(x) I), \quad \text{with} \quad x \sim N(0, I)$$

(once we learn θ, we can generate new images by first sampling $x \sim N(0, I_d)$ and then drawing $y \mid x \sim N(\mu_\theta(x), \sigma_\theta^2(x) I)$)

MLE:
$$\max_\theta p_\theta(y^n) \approx \max_\theta \max_\phi \mathbb{E}_{x^n \sim \underbrace{q_\phi(\cdot \mid y^n)}_{x_i \mid y_i \sim N(\mu_\phi(y_i), \sigma_\phi^2(y_i) I) \text{ parametrized by another neural network } \phi}} \left[\log \frac{p_\theta(x^n, y^n)}{q_\phi(x^n \mid y^n)} \right]$$

Aim to perform SGD jointly over (θ, ϕ)

Idea of VAE:
$$\mathbb{E}_{x^n \sim q_\phi(\cdot \mid y^n)} \left[\log \frac{p_\theta(x^n, y^n)}{q_\phi(x^n \mid y^n)} \right]$$
$$= -D_{KL}(q_\phi(x^n \mid y^n) \,\|\, p_\theta(x^n)) + \mathbb{E}_{x^n \sim q_\phi(\cdot \mid y^n)} [\log p_\theta(y^n \mid x^n)]$$

- First term: as $q_\phi(x_i \mid y_i) = N(\mu_\phi(y_i), \sigma_\phi^2(y_i) I)$ and $p_\theta(x_i) = N(0, I)$, the KL divergence has an explicit form in (θ, ϕ), so easy to compute the gradient.
- Second term: ∇_θ: easy as $\nabla_\theta \log p_\theta(y \mid x)$ is quite simple, and

$$\nabla_\theta \mathbb{E}_{x \sim q_\phi(\cdot \mid y)} [\log p_\theta(y \mid x)]$$
$$\approx \nabla_\theta \left(\frac{1}{L} \sum_{\ell=1}^{L} \log p_\theta(y \mid x_\ell) \right)$$
$$= \frac{1}{L} \sum_{\ell=1}^{L} \nabla_\theta \log p_\theta(y \mid x_\ell) \quad \text{for } x_1, \dots, x_L \sim q_\phi(\cdot \mid y)$$

∇_ϕ: 1) Approach I (REINFORCE):

$$\nabla_\phi \mathbb{E}_{x \sim q_\phi(\cdot \mid y)} [f(x)] = \mathbb{E}_{x \sim q_\phi(\cdot \mid y)} [f(x) \underbrace{\nabla_\phi \log q_\phi(x \mid y)}_{\text{simple expression due to Gaussian}}]$$
$$\approx \frac{1}{L} \sum_{\ell=1}^{L} f(x_\ell) \nabla_\phi \log q_\phi(x_\ell \mid y)$$

2) Approach II (reparametrization):

$$\nabla_\phi \mathbb{E}_{x \sim N(\mu_\phi(y), \sigma_\phi^2(y) I)}[f(x)] = \nabla_\phi \mathbb{E}_{\varepsilon \sim N(0, I)}[f(\mu_\phi(y) + \sigma_\phi(y)\varepsilon)]$$

$$= \mathbb{E}_{\varepsilon \sim N(0, I)}[\nabla_\phi f(\mu_\phi(y) + \sigma_\phi(y)\varepsilon)]$$

$$\approx \frac{1}{L}\sum_{\ell=1}^{L} \nabla_\phi f(\mu_\phi(y) + \sigma_\phi(y)\varepsilon_\ell)$$

for $\varepsilon_1, \cdots, \varepsilon_L \sim N(0, I)$