Lecture 16 : EM Theory.

Quiz 2 → Up to & incl. boosting (proof)

Recall:

GMM → easy ex. of EM alg.

↳ there are many application

⇓

$$p(x;\theta) = \sum_{j=1}^{n} \underbrace{P(j)}_{\sum P(j)=1} N(x; \mu_j, \underbrace{\boxed{\Sigma_j}}_{P\text{-semi }D})$$

Optimization Method → EM Alg.

① E → posterior prob. ⇒ $p^{(l)}(j|t) = P(j \mid x_t, \theta^{(l)})$

Note: $\underbrace{p^{(l)}(j|t) = \mathbb{E}[\delta(j|t) \mid X, \theta^{(l)}]}$

illustrate why we can substitue $\delta(j|t) \rightarrow p^{(l)}(j|t)$

② M : $\hat{n}(j) = \sum_{t=1}^{n} p^{(l)}(j|t)$ ⇒ effective # of samples in j-th component.

$$\sum_{j=1}^{m} \hat{n}(j) = n$$

Mixture proportions: $p^{(l+1)}(j) = \dfrac{\hat{n}(j)}{n}$

Component means: $\mu_j^{(l+1)} = \dfrac{1}{\hat{n}(j)} \sum_{t=1}^{n} p^{(l)}(j|t)\, x_t$

Component covariance matrix :

$$\Sigma_j^{(l+1)} = \frac{1}{\hat{n}(j)} \sum_{t=1}^{n} p^{(l)}(j|t)(x_t - \mu_j^{(l)})(x_t - \mu_j^{(l)})^T$$

Posterior prob. of x_t belonging to component j:

→ since it is Bernoulli r.v.

$p^{(l)}(j|t) = \boxed{\mathbb{E}[\delta(j|t) \mid X, \theta^{(l)}]}$ → Question

$= \Pr(\text{ sample } x_t \text{ belongs to component } j \mid \underline{\theta}^{(\ell)})$

$$= \frac{P^{(\ell)}(j)\, P(x_t \mid j, \underline{\theta}^{(\ell)})}{\sum_{j'=1}^{m} P^{(\ell)}(j')\, P(x_t \mid j', \underline{\theta}^{(\ell)})} \rightsquigarrow \text{Bayes Rule}$$

Theorem to guarantee.

Let $f(\underline{\theta}) = \sum_t \log P(x_t ; \underline{\theta})$ be the log-likelihood on $X = \{x_1, \dots, x_n\}$.

Specially, $f(\underline{\theta}^{(\ell)})$: log-likelihood of EM alg. ($\ell \in \mathbb{N}$)

Theorem : [EM Thm]

Unless $\underline{\theta}^{(\ell)}$ is a stationary point of $\max_{\underline{\theta}} f(\underline{\theta})$,

we have $f(\underline{\theta}^{(\ell+1)}) > f(\underline{\theta}^{(\ell)})$ $\forall \ell \in \mathbb{N}$ in any case !

very beautiful

Idea: Majorization - Minimization (MM) Alg.

Consider the opt. (minimum) prob. $\min_{x \in \mathbb{R}^d} f(x) \Rightarrow$ Bad function

We assume that this is difficult. (directly)

Method

Iterative

Successively minimize an auxiliary function $u(x, x^{(\ell)})$

We hope $\{x^{(\ell)}\}_{\ell=1}^{\infty}$ converges to a stationary point !

to strong

we want to guarantee $f(x^{(\ell+1)}) \leq f(x^{(\ell)})$

Defn: An auxiliary function of $f: \mathbb{R}^d \to \mathbb{R}$ is a bivariate func.

$u: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ s.t.

① $u(x,x) = f(x)$

② $u(x, x^{(\ell)}) \geq f(x) \quad \forall (x, x^{(\ell)}) \in \mathbb{R}^d \times \mathbb{R}^d$

$\downarrow$

says that $u(\cdot, x^{(\ell)})$ is a Majorizor of f^{n}!

Say we have an auxiliary f^{n} $u(x, x^{(\ell)})$ of $f(x)$

MM procedure is

$$x^{(\ell+1)} = \boxed{\underset{x}{\operatorname{argmin}}\ u(x, x^{(\ell)})}$$

$\Downarrow$

assume this is easier than prime prob. (substaintically)

[Props]: $f(x^{(\ell+1)}) \leq f(x^{(\ell)}) \Rightarrow$ monotonically decrease!

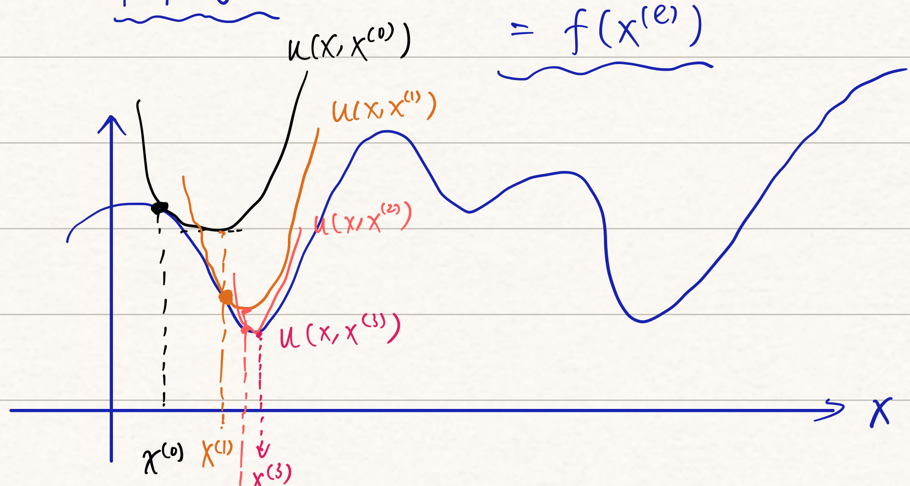
Pf: $f(x^{(\ell+1)}) \leq u(x^{(\ell+1)}, x^{(\ell)}) = \min u(x, x^{(\ell)})$

$\uparrow$

property 2

$\leq u(x^{(\ell)}, x^{(\ell)})$

$= f(x^{(\ell)})$



$x^{(2)}$

Rmk: Equality holds iff $f(x^{(\ell+1)}) = u(x^{(\ell+1)}, x^{(\ell)})$ &

$u(x^{(\ell+1)}, x^{(\ell)}) = u(x^{(\ell)}, x^{(\ell)})$

Rmk: would like $x^{(\ell)} \longrightarrow$ stationary point of f $\boxed{— (*)}$

Under additional condition → continuity & differentiability of u, we can show $(*)$

Rmk: How to get $u(\cdot,\cdot)$? → To be more precise, we should find a GOOD u !

① first-order Taylor

② Jenson inequality

Application of MM framework to EM alg.

$X = \{X_t\}_{t=1}^{n}$ → observation

Latent / Hidden variable $Z = \{Z_t\}_{t=1}^{n}$

Complete Dataset (X, Z)

X ↑ incomplete

Goal: Estimate $\underline{\theta}$ from incomplete X

MLE $\rightarrow$ $\underline{\hat{\theta}}_{MLE} = \underset{\underline{\theta}}{\operatorname{argmin}} \; -\log p(X \mid \underline{\theta})$

Intractable since $P(X \mid \theta) = \sum_{z} P(X, Z \mid \theta)$

E - step: Evaluate $p(Z \mid X, \underline{\theta}^{(\ell)}) = \{ p^{(\ell)}(j \mid t) \}$

$\downarrow$

Soft guess of latent variables Z given dataset X and current $\underline{\theta}^{(\ell)}$

M - step: Update $\underline{\theta}$ as:

complete dataset with soft Z (guess)

$(*) \leftarrow \underline{\theta} = \operatorname{argmin} - \mathbb{E}[\log P(X, Z \mid \theta) \mid X, \underline{\theta}^{(\ell)}]$

$Z \sim Z \mid X, \theta^{(\ell)}$

We like to show that $-\log P(X \mid \underline{\theta}^{(\ell+1)}) \leq -\log P(X \mid \underline{\theta}^{(\ell)})$

Pf: Notice $(*)$ is the same as,

$$\underline{\theta}^{(\ell+1)} = \underset{\underline{\theta}}{\operatorname{argmin}}\ u(\underline{\theta}, \underline{\theta}^{(\ell)})$$

(not related to $\underline{\theta}$)

$\boxed{\text{const}}$

$\Uparrow$

$$\text{and } u(\underline{\theta}, \underline{\theta}^{(\ell)}) = -\mathbb{E}[\log P(X, Z \mid \theta) \mid X, \underline{\theta}^{(\ell)}] - \boxed{H(p(\cdot \mid X, \underline{\theta}^{(\ell)}))}$$

$$H(p(\cdot)) = -\sum_{z} p(z) \log(p(z))$$

If we can check $u(\theta, \underline{\theta}^{(\ell)})$ is an auxilary f^{n} of $f(\underline{\theta}) = -\log P(X \mid \underline{\theta})$, then our conclusion comes from MM props.

Then, check:

① $f(\theta) = u(\underline{\theta}, \underline{\theta})$

$$u(\underline{\theta}, \underline{\theta}) = -\mathbb{E}[\log P(X, Z \mid \theta) \mid X, \underline{\theta}] - H(p(\cdot \mid X, \theta))$$

$$= -\sum_z \left[\log P(X, Z \mid \theta)\right] P(Z \mid X, \theta) + \sum_z P(Z \mid X, \theta) \log P(Z \mid X, \theta)$$

$$= -\sum_z P(Z \mid X, \theta) \log P(X \mid \theta)$$

$$= -\log P(X \mid \theta)$$

$$= f(\underline{\theta})$$

② To check $u(\underline{\theta}, \underline{\theta}^{(t+1)}) \geq f(\underline{\theta})$

[Jenson Inequality] Let $g \to$ convex $\Rightarrow g(\mathbb{E}\,Y) \leq \mathbb{E}[g(Y)]$

$\downarrow$

Verify Part ②

$$P(X \mid \theta) \begin{cases} \dfrac{P(X, Z \mid \theta)}{P(Z \mid X, \theta)} \\ \sum P(X \mid Z, \theta)\, P(Z \mid \theta) \end{cases}$$

$$\boxed{\begin{aligned} & -\log \sum P(X, Z \mid \theta) \\ & = -\log \sum p_1 \frac{P(X, Z \mid \theta)}{p_1} \\ & \leq -\sum p_1 \log \frac{P(X, Z \mid \theta)}{p_1} \end{aligned}}$$

$$\text{RHS} = f(\theta) = -\log P(X \mid \theta)$$

$$= -\log \sum_z P(X \mid Z, \theta)\, P(Z \mid \theta)$$

$$= -\log \sum_z P(Z \mid \theta) \left(\frac{P(Z \mid X, \theta^{(1)})\, P(X \mid Z, \theta)}{P(Z \mid X, \theta^{(1)})} \right)$$

$$= -\log \sum_z P(Z \mid X, \theta^{(1)}) \frac{P(Z \mid \theta)\, P(X \mid Z, \theta)}{P(Z \mid X, \theta^{(1)})}$$

Jenson Inequality

$$\leq -\sum_z P(Z \mid X, \theta^{(1)}) \log \frac{P(Z \mid \theta)\, P(X \mid Z, \theta)}{P(Z \mid X, \theta^{(1)})}$$

$$= -\sum_z P(Z \mid X, \theta^{(1)}) \log P(X, Z \mid \theta) - H(p(\cdot \mid X, \theta^{(1)}))$$

$$= -\underset{z\sim P(z|X,\theta^{(l)})}{\mathbb{E}}\left[\log P(X,z|\theta) \mid X,\theta^{(l)}\right] - H(P(\cdot|X,\theta^{(l)}))$$

$$= u(\theta,\theta^{(l)})$$

Hence, by the properties of Auxiliary function $u(\theta,\theta^{(l)})$ of function f.

we have $-\log P(X|\theta^{(l+1)}) \leq -\log P(X|\theta^{(l)})$

$$\Leftrightarrow \quad f(\theta^{(l+1)}) \leq f(\theta^{(l)})$$

Application of Jensen's Ineq. explained.

$$= -\log \sum_z P(z|X,\underline{\theta}^{(l)}) \frac{P(X|z,\theta)\,P(z|\theta)}{P(z|X,\theta^{(l)})}$$

$$= -\log \underset{z\sim P(z|X,\theta^{(l)})}{\mathbb{E}}\left[\frac{P(X|z,\theta)\,P(z|\theta)}{P(z|X,\underline{\theta}^{(l)})}\right]$$

$$= g\,\mathbb{E}(Y)$$

$$\leq \mathbb{E}[g(Y)]$$

$$= \sum_z P(z|X,\underline{\theta}^{(l)})\, -\log \frac{P(X|z,\theta)\,P(z|\theta)}{P(z|X,\theta^{(l)})}$$

$$= -\sum_z P(z|X,\theta^{(l)}) \log \frac{P(X|z,\theta)\,P(z|\theta)}{P(z|X,\theta^{(l)})}$$

② ①

$$= f(\underline{\theta})$$

Other Application of MM : Ranking

$m = 20$ teams

Each team has a skill level $\theta_i \in [0,1]$

BTL Model

$$Pr(i \text{ beats } j) = \frac{\theta_i}{\theta_i + \theta_j}$$

$\mathcal{D} = \{ b_{ij} : i,j \in [m] \}$ $\qquad$ $\underline{\theta} = \{\theta_1, \dots, \theta_m\}$

b_{ij} = # of times i beat j in one season

$$\mathcal{L}(\mathcal{D}; \underline{\theta}) = \prod_{i,j} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{b_{ij}} \Longrightarrow \text{Here, } i \neq j \,!$$

$$\text{MLE} \rightarrow \ell(\mathcal{D}; \underline{\theta}) = \sum_{i,j} b_{ij} \left[\log \theta_i - \log(\theta_i + \theta_j) \right]$$

$$\text{Minimize } -\ell(\mathcal{D}; \underline{\theta}) = f(\underline{\theta}) = -\sum_{\boxed{i \neq j}} b_{ij} \left[\log \theta_i - \log(\theta_i + \theta_j) \right]$$

① Gradient $\Rightarrow$ Ⓠ ! $\Rightarrow \theta_i \leftrightarrow \theta_j$ $\boxed{\text{related}}$ in second term $\underline{\log(\theta_i + \theta_j)}$

$\boxed{\text{Couple}}$

$$\frac{\partial f}{\partial \theta_k} = -\sum_j b_{kj} \left(\frac{1}{\theta_k} - \frac{1}{\theta_k + \theta_j} \right) + \sum_{i \neq k} b_{ik} \cdot \frac{1}{\theta_i + \theta_k}$$

$\begin{cases} i = k \\ i \neq k, j = k \end{cases}$

② $\underline{\text{MM framework}} \rightarrow$ $\boxed{\text{find Auxiliary } u(\cdot,\cdot) \text{ of } f^{\frown}}$

For a concave $f^{\frown}$ $h(\cdot)$,

$$h(y) \leq h(x) + h'(x)(y - x)$$

$$\Downarrow \quad u(x,y)$$

upper Bound → property $\boxed{2}$

we want to apply this to $h(y) = \log(y)$

$$\Downarrow$$

$$h(y) \leq \log x + \frac{y - x}{x}$$

A majorizer of f is:

$$u(\underline{\theta}, \underline{\theta}^{(l)}) = -\sum_{ij} b_{ij} \left[\log \theta_i - \log(\theta_i^{(l)} + \theta_j^{(l)}) - \frac{\theta_i + \theta_j}{\theta_i^{(l)} + \theta_j^{(l)}} + 1 \right]$$

$$= -\sum_{ij} b_{ij} \log \theta_i + \sum_{ij} b_{ij} \frac{\theta_i + \theta_j}{\theta_i^{(l)} + \theta_j^{(l)}}$$

$$\frac{\partial}{\partial \theta_i} = -\sum_j b_{ij} \frac{1}{\theta_i} + \sum_j \frac{(b_{ij} + b_{ji})}{\theta_i^{(l)} + \theta_j^{(l)}}$$

$$\Downarrow$$

Update Way: $\underline{\theta}_i^{(l+1)} = \underset{\underline{\theta}}{\operatorname{argmin}}\, u(\underline{\theta}, \underline{\theta}^{(l)})$

= closed form

$$= \frac{\sum_{j \neq i} b_{ij}}{\sum_{j \neq i} (b_{ij} + b_{ji}) / (\theta_i^{(l)} + \theta_j^{(l)})}$$