MAS473 Written Homework 2

20200130 Yujun Kim

#1 (Radamacher Complexity)

$\mathcal{F} \subseteq \{f : \mathcal{X} \to \{+1, -1\}\}$

$\mathcal{L} = \{(x, y) \to \mathbb{1}\{f(x) \neq y\} : f \in \mathcal{F}\}$

Note that for $y, y' \in \{\pm 1\}$, $\mathbb{1}\{y \neq y'\} = \frac{1 - yy'}{2}$. $\cdots (*)$

Let $S = \{z_1, \cdots, z_n\} \subseteq \mathcal{Z}$, for $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

$$\begin{aligned}
\hat{R}_n(\mathcal{L}) &= \mathbb{E}_{\vec{\sigma}}\left[\sup_{l \in \mathcal{L}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i l(z_i)\right] \\
&= \mathbb{E}_{\vec{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}\{f(x_i) \neq y_i\}\right] \\
&\overset{(*)}{=} \mathbb{E}_{\vec{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - y_i f(x_i)}{2}\right] \\
&= \mathbb{E}_{\vec{\sigma}}\left[\sup_{f \in \mathcal{F}} \left\{\left(\frac{1}{2n}\sum_{i=1}^{n} \sigma_i\right) - \left(\frac{1}{2n}\sum_{i=1}^{n} \sigma_i y_i f(x_i)\right)\right\}\right] \\
&= \mathbb{E}_{\vec{\sigma}}\left[\frac{1}{2n}\sum_{i=1}^{n} \sigma_i + \sup_{f \in \mathcal{F}} \frac{1}{2n}\sum_{i=1}^{n} (-y_i \sigma_i) f(x_i)\right] \\
&= \frac{1}{2n}\sum_{i=1}^{n} \cancelto{0}{\mathbb{E}_{\vec{\sigma}}[\sigma_i]} + \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} (-y_i \sigma_i) f(x_i)\right] \\
&= \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right] \quad \left(\because \sigma_i \sim \mathrm{Unif}\{\pm 1\} \iff -y_i \sigma_i \sim \mathrm{Unif}\{\pm 1\}\right) \\
&= \frac{1}{2}\hat{R}_n(\mathcal{F}).
\end{aligned}$$

By linearity of expectation

Note that expectation here is taken over $\vec{\sigma} \sim \mathrm{Unif}\{\pm 1\}^n$.

Hence, $\boxed{\hat{R}_n(\mathcal{L}) = \frac{1}{2}\hat{R}_n(\mathcal{F})}$.

/ /

#2 (a) (Joint PDF)

$\underset{\mathbb{R}^D}{x} = Bz + \mu + \varepsilon$, for $\underset{\mathbb{R}^M}{z} \sim N(0, I)$, $\varepsilon \sim N(0, \sigma^2 I)$.

Denote $\theta = (B, \sigma^2)$, $\theta^{(t)} = (B^{(t)}, (\sigma^{(t)})^2)$. Fix $\mu = \mu_{ML} = \bar{x}$.

$$
\begin{aligned}
p(x, z \mid \theta) &= p(x \mid \theta, z)\, p(z) \\
&= p(\varepsilon = x - Bz - \mu \mid \theta, z)\, p(z) \\
&= N(x - Bz - \mu \mid 0, \sigma^2 I)\, N(z \mid 0, I). \\
&= \frac{1}{(2\pi\sigma^2)^{D/2}} e^{-\frac{1}{2\sigma^2}\|x - Bz - \mu\|^2} \cdot \frac{1}{(2\pi)^{M/2}} e^{-\frac{1}{2}\|z\|^2}
\end{aligned}
$$

(b) (E-step)

$p(z \mid x, \theta, \mu) = N(z \mid m, C)$ for $m = B^T(BB^T + \sigma^2 I)^{-1}(x - \mu)$

$C = I - B^T(BB^T + \sigma^2 I)^{-1} B$

$D \underset{}{\overset{M}{\boxed{B}}}$

$$
\begin{aligned}
&\mathbb{E}_{z \mid x, \theta^{(t)}, \mu}[\log p(x, z \mid \theta, \mu)] \\
&= \int_z (\log p(x, z \mid \theta, \mu))\, p(z \mid x, \theta^{(t)}, \mu)\, dz \\
&= \int_z \left(-\frac{D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|x - Bz - \mu\|^2 - \frac{M}{2}\log(2\pi) - \frac{1}{2}\|z\|^2\right) \\
&\qquad \times N(z \mid m^{(t)}, C^{(t)})\, dz. \\
&= \alpha - \mathbb{E}_{z \sim N(m^{(t)}, C^{(t)})}\left[\frac{1}{2\sigma^2}\|x - Bz - \mu\|^2 + \frac{1}{2}\|z\|^2 - D\log\sigma\right]
\end{aligned}
$$

$\left(\alpha = -\frac{D+M}{2}\log(2\pi) \text{ is a constant}\right)$

$$
\begin{aligned}
&= \alpha - \mathbb{E}_{z \sim p^{(t)}}\Big[\frac{1}{2\sigma^2}\big\{\|x - \mu - Bm^{(t)}\|^2 + \|Bm^{(t)} - Bz\|^2 \\
&\qquad + \underbrace{2\langle x - \mu - Bm^{(t)}, Bm^{(t)} - Bz\rangle}_{0}\big\} \\
&\qquad + \frac{1}{2}\|z\|^2 - D\log\sigma\Big] \\
&= \alpha - \frac{1}{2\sigma^2}\left\{\|x - \mu - Bm^{(t)}\|^2 + \mathrm{Tr}(B C^{(t)} B^T)\right\} \\
&\qquad + \frac{1}{2}\mathrm{Tr}(C^{(t)}) - D\log\sigma. \\
&= E(B, \sigma) \quad (\text{Expectation})
\end{aligned}
$$

$$B^T(BB^T+\sigma^2 I_D)^{-1} = (B^TB+\sigma^2 I_M)^{-1}B^T$$

$$(B^TB+\sigma^2 I_M)B^T = B^T(BB^T+\sigma^2 I_D)$$

(c)
(M-Step)

① $\frac{\partial E}{\partial \sigma} = \sigma^{-3}\left(\|x-\mu-Bm^{(t)}\|^2 + Tr(B^T C^{(t)} B)\right) - \sigma^{-1} D = 0$

$\Rightarrow (\sigma^{(t+1)})^2 = \frac{1}{D}\left(\|x-\mu-\underbrace{B^{(t+1)}m^{(t)}\|^2 + Tr(B^{(t+1)T}}_{\text{See below}} C^{(t)} B)\right)$

Lemma1, $\frac{d}{dX_{ij}}(Tr(AX^T)) = A_{ij}$ $\left(\Rightarrow \frac{d}{dX}(Tr(AX^T)) = A\right)$

$\because)$ $Tr(AX^T) = \sum_i (AX^T)_{ii} = \sum_i \sum_j A_{ij} X^T_{ji} = \sum_i \sum_j A_{ij} X_{ij}$

$\Rightarrow \frac{d}{dX_{ij}}(Tr(AX^T)) = A_{ij}$.

Lemma2, $\frac{d}{dX}(Tr(XAX^T)) = X(A+A^T)$

$\because)$ $\frac{d}{dX}(Tr(XAX^T)) = \frac{d}{dX} Tr(X f(X))$, for $f(X) = AX^T$

$= Xf'(X) + f(X)^T$

$= X(A+A^T)$

Lemma3, $\frac{d}{dX}\|b-Xa\|^2 = \frac{d}{dX}(b-Xa)^T(b-Xa)$

$= -2(b-Xa)a^T$

$= 2(Xa-b)a^T$.

② Thus, $\frac{\partial E}{\partial B} = -\frac{1}{2\sigma^2}\left(\underbrace{2(Bm^{(t)}-(x-\mu))m^{(t)T}}_{\text{lemma 3}} + \underbrace{B(C^{(t)}+C^{(t)T})}_{\text{lemma 2}}\right)$

$= 0$.

$\Rightarrow B(2m^{(t)}m^{(t)T} + C^{(t)} + C^{(t)T}) = 2(x-\mu)m^{(t)T}$.

$\Rightarrow B^{(t+1)} = 2(x-\mu)m^{(t)T}(2m^{(t)}m^{(t)T} + C^{(t)} + C^{(t)T})^{-1}$

$\left(= (x-\mu)m^{(t)T}(m^{(t)}m^{(t)T} + C^{(t)})^{-1}\right)$

if $C^{(t)}$ is symmetric.

#3

(GMM EM $\Downarrow$ Loyd's)

$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \varepsilon I)$.

① Change of r_{nk} as $\varepsilon \to 0$

$$r_{nk} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \varepsilon I)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \varepsilon I)} \quad \cdots (*)$$

If $\exists!\ k_0 = \underset{k \in [K]}{\operatorname{argmin}} (\|x_n - \mu_k\|^2)$, then

$$\frac{\mathcal{N}(x_n \mid \mu_k, \varepsilon I)}{\mathcal{N}(x_n \mid \mu_{k_0}, \varepsilon I)} = e^{-\frac{1}{2\varepsilon^2}(\underbrace{\|x_n - \mu_k\|^2 - \|x_n - \mu_{k_0}\|^2}_{>0})}$$

$\longrightarrow 0$ as $\varepsilon \to 0$. $\forall\, k \neq k_0$

Using this on $(*)$, we have

$r_{nk_0} \to 1$ as $\varepsilon \to 0$, given $\pi_{k_0} \neq 0$.

$r_{nk} \to 0$ as $\varepsilon \to 0$, for $\forall k \neq k_0$, given $\pi_{k_0} \neq 0$.

② Change of μ_k^{new} as $\varepsilon \to \infty$

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}} \longrightarrow \text{Centroid of } \{x_n \mid \mu_k \text{ is closest}\}$$

Hence, we update mean as the centroid of current cluster, which is equivalent to Loyd's rule.

Here, we don't iteratively update ε.