Problem Set #3: Learning Theory and Unsupervised Learning

Trung H. Nguyen

1. **Uniform convergence and Model Selection**

(a) Let $Z_{ij} = 1\{\hat{h}_i(x^{(j)}) \neq y^{(j)}\}$. Thus, we have that $\varepsilon(\hat{h}_i) = E(Z_{ij})$ and $\hat{\varepsilon}(\hat{h}_i) = \dfrac{1}{\beta m}\sum_{j=1}^{\beta m} Z_{ij}$. Applying Hoeffding inequality, we have for any fixed $\gamma > 0$:

$$p(|\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{CV}}(\hat{h}_i)| > \gamma) \leq 2\exp\left(-2\gamma^2 \beta m\right)$$

Using the union bound, we have that:

$$\begin{aligned} p(\exists i \in [1,k]. \quad |\varepsilon(\hat{h} - i) - \hat{\varepsilon}_{S_{CV}}(\hat{h}_i)| > \gamma) &\leq 2k\exp\left(-2\gamma^2\beta m\right) \\ p(\forall i \in [1,k]. \quad |\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{CV}}(\hat{h}_i)| \leq \gamma) &\geq 1 - 2\exp\left(-2\gamma^2\beta m\right) \end{aligned}$$

Let $\delta = 4k\exp(-2\gamma^2\beta m)$, which gives us:

$$\gamma = \sqrt{\frac{1}{2\beta m}\log\frac{4k}{\delta}}$$

Then, with probability at least $1 - \dfrac{\delta}{2}$, for all $\hat{h_i}$,

$$\left|\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{CV}}(\hat{h}_i)\right| \leq \sqrt{\frac{1}{2\beta m}\log\frac{4k}{\delta}}$$

(b) Using the uniform convergence result obtained from part a, we have with the probability at least $1 - \dfrac{\delta}{2}$:

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}_{S_{CV}}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}_{S_{CV}}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma \\ &= \min_{i=1,\ldots,k} \varepsilon(\hat{h}_i) + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}} \end{aligned}$$

(c) From part (a) and (c), we have that with the probability at least $(1-\dfrac{\delta}{2})(1-\dfrac{\delta}{2}) = 1-\delta+\dfrac{\delta^2}{4} \geq 1-\delta$:

$$\varepsilon(\hat{h}) \leq \min_{i=1,\ldots,k} \varepsilon(\hat{h}_i) + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}}$$

$$\left|\varepsilon(\hat{h}_j) - \hat{\varepsilon}_{S_{\text{train}}}({h_j}^*)\right| \leq \sqrt{\frac{2}{(1-\beta)m}\log\frac{4\,|\mathcal{H}_j|}{\delta}}, \quad \forall h_j \in \mathcal{H}_j$$

When equality holds for both above inequality, we have:

$$\begin{aligned}
\varepsilon(\hat{h}) &\leq \min_{i=1,\ldots,k} \varepsilon(\hat{h}_i) + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}} \\
&= \varepsilon(\hat{h}_j) + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}} \\
&\leq \hat{\varepsilon}_{S_{\text{train}}}({h_j}^*) + \sqrt{\frac{2}{(1-\beta)m}\log\frac{4\,|\mathcal{H}_j|}{\delta}} + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}}, \quad \forall h_j \in \mathcal{H}_j \\
&\leq \varepsilon({h_j}^*) + 2\sqrt{\frac{2}{(1-\beta)m}\log\frac{4\,|\mathcal{H}_j|}{\delta}} + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}}, \quad \forall h_j \in \mathcal{H}_j
\end{aligned}$$

2. **VC Dimension**
   - $h(x) = \mathbf{1}\{a < x\} \Rightarrow$ VC-dim $= 1$
   - $h(x) = \mathbf{1}\{a < x < b\} \Rightarrow$ VC-dim $= 2$
   - $h(x) = \mathbf{1}\{a \sin x > 0\} \Rightarrow$ VC-dim $= 1$
   - $h(x) = \mathbf{1}\{\sin(x + a) > 0\} \Rightarrow$ VC-dim $= 2$
3. **ℓ_1 regularization for least squares**

   (a) For $s_i = 1$, we have:

$$\begin{aligned}
J(\theta) &= \frac{1}{2}\left\|X\bar{\theta} + X_i\theta_i - \vec{y}\right\|_2^2 + \lambda\left\|\bar{\theta}\right\|_1 + \lambda\theta_i \\
&= \frac{1}{2}\left(X\bar{\theta} + X_i\theta_i - \vec{y}\right)^T\left(X\bar{\theta} + X_i\theta_i - \vec{y}\right) + \lambda\left\|\bar{\theta}\right\|_1 + \lambda\theta_i \\
&= \frac{1}{2}\left((X\bar{\theta} - \vec{y})^T(X\bar{\theta} - \vec{y}) + 2X_i^T(X\bar{\theta} - \vec{y})\theta_i + X_i^TX_i\theta_i^2\right) + \lambda\left\|\bar{\theta}\right\|_1 + \lambda\theta_i
\end{aligned}$$

   Taking the derivative w.r.t θ_i, we obtain:

$$\frac{\partial J(\theta)}{\partial\theta_i} = X_i^TX_i\theta_i + X_i^TX\bar{\theta} - X_i^T\vec{y} + \lambda$$

   Setting the derivative equal to zero, we have:

$$\theta_i = \frac{-X_i^TX\bar{\theta} + X_i^T\vec{y} - \lambda}{X_i^TX_i}$$

   Thus, the optimal value of θ_i is:

$$\theta_i = \max\left\{\frac{-X_i^TX\bar{\theta} + X_i^T\vec{y} - \lambda}{X_i^TX_i}, 0\right\}$$

   Similarly, for $s_i = -1$, we have the optimal value of θ_i is:

$$\theta_i = \min\left\{\frac{-X_i^TX\bar{\theta} + X_i^T\vec{y} + \lambda}{X_i^TX_i}, 0\right\}$$

   (b)

   (c)
4. **K-Means Clustering**
5. **The Generalized EM algorithm**

(a) We have:

$$
\begin{aligned}
\ell(\theta^{(k+1)}) &\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta^{(t+1)}\right)}{Q_i^{(t)}(z^{(i)})} \\
&\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta^{(t)}\right)}{Q_i^{(t)}(z^{(i)})} \\
&= \ell(\theta^{(k)})
\end{aligned}
$$

where the first inequality comes from the fact that

$$
\ell(\theta) \geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta^{(t+1)}\right)}{Q_i^{(t)}(z^{(i)})}
$$

holds for any values of Q_i and θ due to Jensen's inequality. The second one holds by the chosen update rule that taking small step of θ without decreasing the objective function.

(b) For the case of applying the gradient ascent to maximize the log-likelihood directly, we have:

$$
\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_j} &= \frac{\partial \sum_i \log \sum_{z^{(i)}} p\left(x^{(i)}, z^{(i)}; \theta\right)}{\partial \theta_j} \\
&= \sum_i \frac{1}{\sum_{z^{(i)}} p\left(x^{(i)}, z^{(i)}; \theta\right)} \sum_{z^{(i)}} \frac{\partial p(x^{(i)}, z^{(i)}; \theta)}{\partial \theta_j} \\
&= \sum_i \frac{1}{p(x^{(i)}; \theta)} \sum_{z^{(i)}} \frac{\partial p(x^{(i)}, z^{(i)}; \theta)}{\partial \theta_j}
\end{aligned}
$$

And for GEM algorithm, we have that:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} &= \sum_i \sum_{z^{(i)}} \frac{Q_i(z^{(i)})}{p(x^{(i)}, z^{(i)}; \theta)} \frac{\partial p(x^{(i)}, z^{(i)}; \theta)}{\partial \theta_j} \\
&= \sum_i \sum_{z^{(i)}} \frac{p(z^{(i)} | x^{(i)}; \theta)}{p(x^{(i)}, z^{(i)}; \theta)} \frac{\partial p(x^{(i)}, z^{(i)}; \theta)}{\partial \theta_j} \\
&= \sum_i \sum_{z^{(i)}} \frac{1}{p(x^{(i)}; \theta)} \frac{\partial p(x^{(i)}, z^{(i)}; \theta)}{\partial \theta_j}
\end{aligned}
$$

which is equal to the above update rule.