Problem Set 3: Learning Theory and Unsupervised Learning

Eitan Joseph Caroline Wang

July 5, 2020

Problem 1

Uniform convergence and Model Selection
In this problem, we will prove a bound on the error of a simple model selection procedure. Let there be a binary classification problem with labels $y \in 0, 1$, and let $H_1 \subseteq H_2 \subseteq \cdots \subseteq H_k$ be k different finite hypothesis classes ($|H_i| < \infty$) — ¡). Given a dataset S of m iid training examples, we will divide it into a training set Strain consisting of the first $(1-\beta)m$ examples, and a hold-out cross validation set S_cv consisting of the remaining βm examples. Here, $\beta \in (0, 1)$. Let $\mathcal{H}^i = \arg\min_{h \in H_i} \hat{\varepsilon}_{Strain}(h)$ be the hypothesis in $\mathcal{H}$ with the lowest training error (on S_train). Thus, $\hat{h}_i$ would be the hypothesis returned by training (with empirical risk minimization) using hypothesis class $\mathcal{H}_i$ and dataset Strain. Also let $h_i^* = \arg\min_{h \in \mathcal{H}_\rangle} \varepsilon(h)$ be the hypothesis in $\mathcal{H}_i$ with the lowest generalization error. Suppose that our algorithm first finds all the h_i^*'s using empirical risk minimization then uses the hold-out cross validation set to select a hypothesis from this the $\hat{h_1} \ldots \hat{h_k}$ with minimum training error. That is, the algorithm will output

$$\hat{h} = \arg\min_{h \in \{\hat{h_1}, \ldots, \hat{h_k}\}} \hat{\varepsilon}_{S_{cv}}(h)$$

For this question you will prove the following bound. Let any $\delta > 0$ be fixed. Then with probability at least 1δ, we have that

$$\varepsilon(\hat{h}) \leq \min_{i=1,\ldots,k} \left(\varepsilon(h_i^*) + \sqrt{\frac{2}{(1-\beta)m} \log \frac{4|\mathcal{H}_i|}{\delta}} \right) + \sqrt{\frac{1}{2\beta m} \log \frac{4k}{\delta}}$$

(a) Prove that with probability at least $1 - \frac{\delta}{2}$, for all $\hat{h}_i$,

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

answer:

We can apply the Hoeffding inequality on h_i. We total have βm elements in the cross-validation set. And the distribution of each h_i will have the mean as $\hat{\varepsilon}(\hat{h}_i)$, with other $\varepsilon(\hat{h}_i)$ randomly distributed. Therefore,

$$P(|\varepsilon(\hat{h}_i) - \hat{\varepsilon}(\hat{h}_i)| \geq \gamma) \leq 2\exp(-2\gamma^2 \beta m)$$

Similar in class, for k elements, the probability will hold as follows:

$$P(\exists i \text{ such that } |\varepsilon(\hat{h}_i) - \hat{\varepsilon}(\hat{h}_i)| \geq \gamma) \leq 2k\exp(-2\gamma^2 \beta m)$$

With probability 1-$\frac{\delta}{2}$, we want to solve:

$$\frac{\delta}{2} = 2k \exp(-2\gamma^2 \beta m)$$

We obtain:

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

(b) Use part (a) to show that with probability $1 - \frac{\delta}{2}$,

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

answer: From part (a), we can derive

$$|\varepsilon(\hat{h}) - \hat{\varepsilon}_{Scv}(\hat{h})| \leq \sqrt{\frac{2}{\beta m} \log \frac{4k}{\delta}}$$

Therefore, we can rewrite as:

$$\varepsilon(\hat{h}) \leq \hat{\varepsilon}_{Scv}(\hat{h}) + \sqrt{\frac{2}{\beta m} \log \frac{4k}{\delta}}$$

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

(c) Let $j = \arg\min_i \varepsilon(\hat{h}_i)$. We know from class that for $\mathcal{H}_j$, with probability $1 - \frac{\delta}{2}$

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

Use this to prove the final bound given at the beginning of this problem.

answer:

Given by the problem we have with probability $1 - \frac{\delta}{2}$ that

$$|\varepsilon(\hat{h}_j) - \hat{\varepsilon}_{S_{train}}(h_j^*)| \leq \sqrt{\frac{2}{(1-\beta)m} \log \frac{4|\mathcal{H}_j|}{\delta}} \tag{1}$$

and from part (a) we know have with probability $1 - \frac{\delta}{2}$ also that

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

Therefore, both (1) and (2) hold with probability $(1 - \frac{\delta}{2})^2 = 1 - \delta + \frac{\delta^2}{4} > 1 - \delta$, and noticing that $\varepsilon(\hat{h}_j) = \min_{i=1,\ldots,k} \varepsilon(\hat{h}_i)$ allows us to write

$$\varepsilon(\hat{h}) \leq \varepsilon(\hat{h}_j) + \sqrt{\frac{1}{2\beta m} \log \frac{4k}{\delta}}$$

So now by using equation (1) we can write the inequality

$$\varepsilon(\hat{h}) \leq \hat{\varepsilon}_{S_{train}}(\hat{h}_j^*) + \sqrt{\frac{2}{(1-\beta)m}\log\frac{4|\mathcal{H}_j|}{\delta}} + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}}$$

Clearly $\hat{\varepsilon}_{S_{train}}(\hat{h}_j^*) < \hat{\varepsilon}(\hat{h}_j^*)$ so

$$\begin{aligned}\varepsilon(\hat{h}) \leq& \hat{\varepsilon}(\hat{h}_j^*) + \sqrt{\frac{2}{(1-\beta)m}\log\frac{4|\mathcal{H}_j|}{\delta}} + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}} \\ \leq& \varepsilon(\hat{h}_j^*) + \sqrt{\frac{2}{(1-\beta)m}\log\frac{4|\mathcal{H}_j|}{\delta}} + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}} + \gamma\end{aligned}$$

Our γ can be thought of as $\gamma_{S_{train}} + \gamma_{S_{cv}}$ which is

$$\gamma = \sqrt{\frac{2}{(1-\beta)m}\log\frac{4|\mathcal{H}_j|}{\delta}} + \sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}}$$

Meaning our final equation becomes

$$\begin{aligned}\varepsilon(\hat{h}) &\leq \varepsilon(\hat{h}_j^*) + 2\sqrt{\frac{2}{(1-\beta)m}\log\frac{4|\mathcal{H}_j|}{\delta}} + 2\sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}} \\ &= \min_{i=1,\ldots,k}\left(\varepsilon(h_i^*) + 2\sqrt{\frac{2}{(1-\beta)m}\log\frac{4|\mathcal{H}_i|}{\delta}}\right) + 2\sqrt{\frac{2}{\beta m}\log\frac{4k}{\delta}}\end{aligned}$$

Problem 2

VC dimension

Let the input domain of a learning problem be $\mathcal{X} = \mathbb{R}$. Give the VC dimension for each of the following classes of hypotheses. In each case, if you claim that the VC dimension is d, then you need to show that the hypothesis class can shatter d points, and explain why there are no $d+1$ points it can shatter.

- $h(x) = \mathbf{1}\{a < x\}$, with parameter $a \in \mathbb{R}$.

  answer: $\mathbf{d = 1}$
  When there's only one point $x \in \mathbb{R}$, we can take a to be smaller than x to satisfy the true label on the point, and take a to be larger than x to satisfy false label at the point. It cannot shatter two points because when $x_1 < x_2$, there does not exists an a such that $a < x_1$ and $a > x_2$ to correctly label x_1 true and x_2 false.

- $h(x) = \mathbf{1}\{a < x < b\}$, with parameter $a, b \in \mathbb{R}$.

  answer: $\mathbf{d = 2}$
  Assume there are two points $x_1, x_2 \in \mathbb{R}$ such that $x_1 < x_2$. There are four different permutation we need to consider. When both x_1, x_2 labels are true, then we can take $a < x_1 < x_2$ and $b > x_2 > x_1$. When both x_1, x_2 labels false, then we can take $a > x_1 > x_2$ and $b > x_2 > x_1$ with $b > a$. If we have x_1 true and x_2 false, then we simply take $a < x_1 < b$ and $b < x_2$. Lastly if x_1 is false and x_2 is true, then we take $a > x_1$ and $a < x_2 < b$.
  The three points case doesn't work because, let's assume $x_1 < x_2 < x_3$. We are unable to find an $a, b \in \mathbb{R}$ that $a < x_1 < b$ and $a < x_3 < b$ but that $a < x_2 < b$ is not true.

- $h(x) = \mathbf{1}\{a \sin x > 0\}$, with parameter $a \in \mathbb{R}$.

  answer: $\mathbf{d = 1}$
  When there is only one point, if $\sin x > 0$, we can find an $a > 0$ to satisfy the true label and $a < 0$ to satisfy the false label. If $\sin x < 0$, then we can just do the opposite. In the two point case, first take $\sin x_1$ and $\sin x_2$ to have the same sign (both positive or negative). We then have to choose a (either the same sign or opposite sign) to make both of the labels true or false respectively. However, if we take $\sin x_1$ and $\sin x_2$ to have the same sign then there is no $a \in \mathbb{R}$ which can flip the sign of one (making it true) and retain the sign of the other (making it false). On the other hand, if we take x_1 and x_2 to ensure that $\sin x_1$ and $\sin x_2$ have different signs, then we just need to find a to be either positive or negative to obtain one true label and one false label. However, there does not exists an $a \in \mathbb{R}$ that make both points positive or both points negative at the same time.

- $h(x) = \mathbf{1}\{\sin(a + x) > 0\}$, with parameter $a \in \mathbb{R}$.

  answer: $\mathbf{d = 2}$
  First we assume that $\sin x_1 > 0$ and $\sin x_2 > 0$, meaning both points are in the first two quadrant on the unit circle. If we want to have both points labeled true, then we make $a = 0$. If we want both to be false, we can just find an a that shifts the two points on the unit circle to the last two quadrants. If the first point is true and the second point is false, then we want to find an a that shifts point 2 to one of the last two quadrants but keeps point 1 in one of the first two. If the first point is false and the second point is true, then we do the opposite.

  This hypothesis class, however, cannot shatters three points. If the three points span one or two consecutive quadrants of the unit circle, then it's impossible to shift them such to make only the middle point true and other two points false (or vice versa). If the three points span three consecutive quadrants on the unit circle, then it's impossible to get all three of the points to have the same label because we cannot shift them all into the first two quadrants or the last two quadrants at the same time if their distance already spans three quadrants.

Problem 5

The Generalized EM algorithm
When attempting to run the EM algorithm, it may sometimes be difficult to perform the M step exactly — recall that we often need to implement numerical optimization to perform the maximization, which can be costly. Therefore, instead of finding the global maximum of our lower bound on the log-likelihood, and alternative is to just increase this lower bound a little bit, by taking one step of gradient ascent, for example. This is commonly known as the Generalized EM (GEM) algorithm. Put slightly more formally, recall that the M-step of the standard EM algorithm performs the maximization

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} \mathcal{Q}_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{\mathcal{Q}_i(z^{(i)})}$$

The GEM algorithm, in contrast, performs the following update in the M-step:

$$\theta := \theta + \alpha \nabla_\theta \sum_i \sum_{z^{(i)}} \mathcal{Q}_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{\mathcal{Q}_i(z^{(i)})}$$

where α is a learning rate which we assume is chosen small enough such that we do not decrease the objective function when taking this gradient step.

(a) Prove that the GEM algorithm described above converges. To do this, you should show that the the likelihood is monotonically improving, as it does for the EM algorithm — i.e., show that $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$.

answer:

To prove that $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$ after each iteration of the GEM algorithm we first observe the assumption that the learning rate (α) is chosen to be small enough such that the algorithm will never take a step that eclipses the global maximum.

What is left to prove is that $\ell(\theta^{(t+1)}) > \ell(\theta^{(t)})$. We can show this with the following steps

$$\begin{aligned}\ell(\theta^{(t+1)}) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta^{(t+1)}) \\ &= \sum_i \log \sum_{z^{(i)}} \mathcal{Q}_i{}^{(t)}(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{\mathcal{Q}_i{}^{(t)}(z^{(i)})}\end{aligned}$$

By utilizing Jensen's Inequality we can write

$$\ell(\theta^{(t+1)}) \geq \sum_i \sum_{z^{(i)}} \mathcal{Q}_i{}^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{\mathcal{Q}_i{}^{(t)}(z^{(i)})}$$

then by observing that $\theta^{(t+1)}$ is chosen to by the optimization problem $\arg\max_\theta \sum_i \sum_{z^{(i)}} \mathcal{Q}_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{\mathcal{Q}_i(z^{(i)})}$ we can write

$$\ell(\theta^{(t+1)}) \geq \sum_i \sum_{z^{(i)}} \mathcal{Q}_i{}^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{\mathcal{Q}_i{}^{(t)}(z^{(i)})}$$

Finally, in order to make Jensen's Inequality hold with equality for any random variable, it suffices to ensure that the random variable in question be constant valued. We can accomplish this by requiring

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{\mathcal{Q}_i(z^{(i)})} = c$$

for some constant c not dependent on $z^{(i)}$. We then see that

$$\begin{aligned}&\sum_i \sum_{z^{(i)}} \mathcal{Q}_i{}^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{\mathcal{Q}_i{}^{(t)}(z^{(i)})} = \ell(\theta^{(t)}) \\ \implies &\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})\end{aligned}$$

Therefore we can be assured that at every iteration of the GEM algorithm the algorithm will increase the value of the objective function at every iteration - implying that it will converge monotonically.

(b) Instead of using the EM algorithm at all, suppose we just want to apply gradient ascent to maximize the log-likelihood directly. In other words, we are trying to maximize the (non-convex) function

$$\ell(\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

so we could simply use the update

$$\theta := \theta + \alpha \nabla_\theta \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

Show that this procedure in fact gives the same update as the GEM algorithm described above.

answer:

To determine whether this procedure gives the same update as the GEM algorithm we need to compare the results of the gradients.

First lets take a look at the derivative of this procedure with respect to each θ_j

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

By using the fact that $\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) = p(x^{(i)}; \theta)$ by the nature of the join distribution we can write this expression as

$$= \sum_i \frac{1}{p(x^{(i)}; \theta)} \sum_{z^{(i)}} \frac{\partial}{\partial \theta_j} p(x^{(i)}, z^{(i)}; \theta)$$

Next lets look at the derivative of the original GEM algorithm with respect to each θ_j

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell_{GEM}(\theta) =& \frac{\partial}{\partial \theta_j} \sum_i \sum_{z^{(i)}} \mathcal{Q}_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{\mathcal{Q}_i(z^{(i)})} \\
=& \sum_i \sum_{z^{(i)}} \mathcal{Q}_i(z^{(i)}) \frac{\mathcal{Q}_i(z^{(i)})}{p(x^{(i)}, z^{(i)}; \theta)} \frac{\frac{\partial}{\partial \theta_j} p(x^{(i)}, z^{(i)}; \theta)}{\mathcal{Q}_i(z^{(i)})} \\
=& \sum_i \sum_{z^{(i)}} \frac{\mathcal{Q}_i(z^{(i)})}{p(x^{(i)}, z^{(i)}; \theta)} \frac{\partial}{\partial \theta_j} p(x^{(i)}, z^{(i)}; \theta) \\
=& \sum_i \sum_{z^{(i)}} \frac{\mathcal{Q}_i(z^{(i)})}{p(x^{(i)}, z^{(i)}; \theta)} \frac{\partial}{\partial \theta_j} p(x^{(i)}, z^{(i)}; \theta)
\end{aligned}$$

We can then observe the following

$$\sum_{z^{(i)}} \mathcal{Q}_i(z^{(i)}) = 1 \tag{3}$$

$$\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) = p(x^{(i)}; \theta) \tag{4}$$

Using both (3) and (4) we finally derive

$$\sum_i \frac{1}{p(x^{(i)}; \theta)} \sum_{z^{(i)}} \frac{\partial}{\partial \theta_j} p(x^{(i)}, z^{(i)}; \theta)$$