Inconsistent Hypothesis Model and the Uniform Convergence Theorem

1 Inconsistent Hypothesis Model

There are obvious drawbacks with the PAC-learning model:

- There may not be any target concept $c \in \mathcal{C}$ which "generates" the labels for the samples.
- Due to noise, they may not be *any* function consistent with the samples. For example, we might get both $(\mathbf{x}, 1)$ and $(\mathbf{x}, 0)$ in the samples, due to noise. Perhaps out of 1000 $(\mathbf{x}_1, 1)$ there is one $(\mathbf{x}, 0)$ due to noise.
- Finding a consistent hypothesis might be an $\mathbf{NP}$-hard problem, which is quite often the case.

We want to extend PAC to deal with these drawbacks, especially to deal with noises which are very real in practice. The new learning model is called the *inconsistent hypothesis model* (IHM), which is realistic. We will stick with this model for the rest of the semester. IHM has the following assumptions:

- There is some *unknown* distribution $\mathcal{D}$ on $\Omega \times \{0, 1\}$ from which the m samples

  $$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$$

  are drawn. Here $\mathbf{x}_i \in \Omega$ and $y_i \in \{0, 1\}$ is a label for $\mathbf{x}_i$. Note that a particular $\mathbf{x}$ might appear many times in S with different labels.
- There is a hypothesis class $\mathcal{H}$ from which the learner $\mathbf{L}$ needs to pick a hypothesis h_S. Each hypothesis is a function from Ω to $\{0, 1\}$.
- The "quality" of a hypothesis h is measured by its *generalization error*

  $$\text{err}_{\mathcal{D}}(h) := \underset{(\mathbf{x},y)\sim\mathcal{D}}{\text{Prob}}[h(\mathbf{x}) \neq y]$$

  We will often drop the subscript and write err(h). Note that there is no longer a target concept! The generalization error is also called the *risk* of h. (This definition of risk comes from something called 01-*loss function*. In more general settings, the risk can be measured with other types of "loss functions," which we will come to later once we have time.)

Definition 1.1 (Ideal problem)**.** Find a hypothesis h^* which minimizes the generalization error:

$$h^* = \underset{h\in\mathcal{H}}{\text{argmin}}\{\text{err}(h)\}.$$

Suppose we do know $\mathcal{D}$, then the following hypothesis is obviously the optimal hypothesis:

$$h_{\text{OPT}}(\mathbf{x}) = \begin{cases} 1 & \text{Prob}[y = 1 \mid \mathbf{x}] \geq 1/2 \\ 0 & \text{Prob}[y = 1 \mid \mathbf{x}] < 1/2 \end{cases}$$

The function h_{OPT} is called the *Bayes optimal classifier*, and its risk err(h_{OPT}) is called the *Bayes risk*.

We cannot solve the ideal problem because we do not know $\mathcal{D}$. In particlar, we cannot evaluate err(h). A common trick for optimizing over some un-evaluatable function is to find another function which approximates the unknown function, and optimize on the approximation function instead. The natural approximation for err(h) is the *empirical error* (also called *empirical risk* or *empirical loss*):

$$\widehat{\text{err}}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(\mathbf{x}_i)\neq y_i} = \mathop{\text{E}}_{(\mathbf{x},y)\sim S}[\mathbf{1}_{h(\mathbf{x})\neq y}].$$

Here, the expectation $\text{E}_{\mathbf{x}\sim S}[\cdot]$ is on the uniform distribution on the m samples S. The expected value of the empirical error of any function is the function's generalization error. (We will make it precise below.) Hence, with enough samples we expect that the empirical error approximates the generalization error very well. We will prove a theorem to that effect. We can thus change the problem. The following is actually more like a "strategy" rather than a problem.

Definition 1.2 (Empirical Risk Minimization (ERM))**.** Find a hypothesis $\hat{h}^* \in \mathcal{H}$ which minimizes the empirical error:

$$\hat{h}^* = \mathop{\text{argmin}}_{h\in\mathcal{H}} \left\{\widehat{\text{err}}_S(h)\right\}.$$

The following relationship is obvious from definition:

$$\text{err}(\hat{h}^*) \geq \text{err}(h^*) \geq \text{err}(h_{\text{OPT}}).$$

Exercise 1. When does err$(h^*) = $ err(h_{OPT})?

Now, the quality of our solution to the ERM problem can be broken up into two parts:

$$\text{err}(\hat{h}^*) - \text{err}(h_{\text{OPT}}) = \underbrace{[\text{err}(\hat{h}^*) - \text{err}(h^*)]}_{\text{Estimation error}} + \underbrace{[\text{err}(h^*) - \text{err}(h_{\text{OPT}})]}_{\text{Approximation error}}.$$

The estimation error measures the quality of the samples S. The approximation error measures the quality of the hypothesis class $\mathcal{H}$. We can reduce the approximation error by picking a better hypothesis class $\mathcal{H}$. It would be zero if $h_{\text{OPT}} \in \mathcal{H}$, for example. Estimating the approximation error is usually hard when we do not make any assumption about the target distribution. In fact, under certain assumptions it can be shown that the rate of convergence to zero of the approximation error can be arbitrarily slow.

We will focus on bounding the estimation error, which is accomplished by showing the so-called *uniform convergence theorems* (UCT). To get a sense of how the reasoning goes, we will prove a finite version of the UCT first, and the VC-dimension analog next.

2 Uniform convergence, finite $\mathcal{H}$ case

Theorem 2.1. *Suppose $\mathcal{H}$ is finite, if we take m i.i.d. samples with*

$$m \geq \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\epsilon^2}$$

then

$$\underset{S \sim \mathcal{D}^m}{\textit{Prob}}\left[\sup_{h \in \mathcal{H}} |\mathrm{err}(h) - \widehat{\mathrm{err}}(h)| \leq \epsilon\right] \geq 1 - \delta.$$

In other words, by taking m i.i.d. samples we have

$$\underset{S \sim \mathcal{D}^m}{\textit{Prob}}\left[\sup_{h \in \mathcal{H}} |\mathrm{err}(h) - \widehat{\mathrm{err}}_S(h)| \leq \sqrt{\frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2m}}\right] \geq 1 - \delta.$$

Proof. Note that the expected empirical error (over the random samples S) is exactly the generalization error. Specifically,

$$\begin{aligned}
\underset{S \sim \mathcal{D}^m}{\mathrm{E}}[\widehat{\mathrm{err}}_S(h)] &= \underset{S \sim \mathcal{D}^m}{\mathrm{E}}\left[\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(\mathbf{x}_i) \neq y_i}\right] \\
&= \frac{1}{m}\sum_{i=1}^{m} \underset{S \sim \mathcal{D}^m}{\mathrm{E}}[h(\mathbf{x}_i) \neq y_i)] \\
&= \frac{1}{m}\sum_{i=1}^{m} \underset{(\mathbf{x}_i, y_i) \sim \mathcal{D}}{\mathrm{E}}[h(\mathbf{x}_i) \neq y_i)] \\
&= \frac{1}{m}\sum_{i=1}^{m} \underset{(\mathbf{x}_i, y_i) \sim \mathcal{D}}{\mathrm{Prob}}[h(\mathbf{x}_i) \neq y_i)] \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathrm{err}(h) \\
&= \mathrm{err}(h).
\end{aligned}$$

Hence, by Hoeffding inequality we have, for any hypothesis $h \in \mathcal{H}$, the difference between the empirical error and the generalization error is large with exponentially small probability:

$$\underset{S \sim \mathcal{D}^m}{\mathrm{Prob}}\left[|\mathrm{err}(h) - \widehat{\mathrm{err}}_S(h)| > \epsilon\right] \leq 2e^{-2\epsilon^2 m}. \tag{1}$$

The union bound gives

$$\underset{S \sim \mathcal{D}^m}{\mathrm{Prob}}\left[\sup_{h \in \mathcal{H}} |\mathrm{err}(h) - \widehat{\mathrm{err}}_S(h)| > \epsilon\right] \leq 2|\mathcal{H}|e^{-2\epsilon^2 m},$$

which yields the desired result. □

Exercise 2. We can show that the empirical error is close to the generalization error (when m large) with high probability without resorting to Hoeffding inequality. Let h be an arbitrary hypothesis. Let $(\mathbf{x}_i, y_i) \sim \mathcal{D}$ be m independent examples taken from $\mathcal{D}$.

(a) Prove that $\text{Var}\left[\mathbf{1}_{h(\mathbf{x}_i) \neq y_i}\right] \leq 1/4$ for every $i \in [m]$.

(b) Using Chebyshev's inequality, prove that

$$\operatorname*{Prob}_{S \sim \mathcal{D}^m}\left[|\widehat{\text{err}}_S(h) - \text{err}(h)| > \epsilon\right] \leq \frac{1}{4m\epsilon^2}.$$

(Of course, this bound is not as good as the Hoeffding bound asymptotically, but it is good enough in some cases.)

This theorem quantifies our intuition that the empirical error is a good approximation to the generalization error if there are enough samples. We can bound the estimation error as follows. With probability at least $1 - 2\delta$,

$$\begin{aligned}
\text{err}(\hat{h}^*) - \text{err}(h^*) &= [\text{err}(\hat{h}^*) - \widehat{\text{err}}(\hat{h}^*)] + [\widehat{\text{err}}(\hat{h}^*) - \text{err}(h^*)] \\
&\leq \epsilon + [\widehat{\text{err}}(\hat{h}^*) - \text{err}(h^*)] \\
&\leq \epsilon + [\widehat{\text{err}}(h^*) - \text{err}(h^*)] \\
&\leq 2\epsilon \\
&= 2\sqrt{\frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2m}}.
\end{aligned}$$

The first inequality comes from the uniform convergence theorem. The second inequality is the fact that $\hat{h}^*$ has the minimum empirical error among all functions in $\mathcal{H}$, including h^*. The third inequality is again the uniform convergence theorem.

A few observations are worth noticing:

- Increasing m leads to reduced estimation error. This is intuitively obvious
- The sample size is dependent on ϵ^2 instead of ϵ as in the corresponding PAC learning theorem.
- We can also try to reduce the complexity $|\mathcal{H}|$ of the hypothesis class in order to reduce the estimation error. However, reducing $\mathcal{H}$ will likely increase the approximation error!
- To reduce the approximation error (in effect, reducing $\text{err}(h^*)$) we must increase the "power" of the hypothesis class $\mathcal{H}$, which implies increasing $|\mathcal{H}|$. However, eventually the term $\sqrt{\frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2m}}$ will dominate the $\text{err}(h^*)$ term and we get to a nasty situation called *overfitting*. This is a fundamental (and difficult) problem of Machine Learning. *Structural Risk Minimization* and *Regularization* are two strategies (among others) for dealing with overfitting. We will also discuss AdaBoost and SVM which are algorithms which can cope relatively well with overfitting.

3 Uniform convergence, infinite $\mathcal{H}$ case

Theorem 3.1. *For any $\delta > 0$, if we take m i.i.d. samples then*

$$\mathop{Prob}_{S}\left[\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| > 2\sqrt{2\frac{\ln|\Pi_{\mathcal{H}}(2m)| + \ln(4/\delta)}{m}}\right] \leq \delta.$$

More concretely, if $d = \text{VCD}(\mathcal{H})$ then from Sauer lemma we can infer

$$\mathop{Prob}_{S}\left[\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| > 2\sqrt{2\frac{d\ln(2me/d) + \ln(4/\delta)}{m}}\right] \leq \delta.$$

It is possible to knock off the $\ln m$ factor inside the square root, but that's beyond the scope of this document.

Proof. The proof has three steps, of which the first step is the most important. We have kind of seen all three steps in the proof to the Vapnik-Chevonenkis theorem for PAC learning.

Step 1: Symmetrization. This is the step that brings the infinite down to the finite, using the double sampling trick which we have seen before. Let $S = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)\}$ be the first m i.i.d. samples and $S' = \{(\mathbf{x}'_1, y'_1), \cdots, (\mathbf{x}'_m, y'_m)\}$ be the second set of m i.i.d. (ghost) samples. We will prove that, for any $\epsilon > 0$ where $m\epsilon^2 \geq 2$,

$$\mathop{\text{Prob}}_{S}\left[\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| > \epsilon\right] \leq 2 \cdot \mathop{\text{Prob}}_{S,S'}\left[\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \widehat{\text{err}}_{S'}(h)| > \epsilon/2\right]. \tag{2}$$

For a given sample set S, define h_S as follows. If $\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| > \epsilon$ then let h_S be any hypothesis such that $|\widehat{\text{err}}_S(h_S) - \text{err}(h_S)| > \epsilon$. If $\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| \leq \epsilon$ then let h_S be an arbitrary hypothesis. From the definition of h_S, it follows that

$$\left\{\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| > \epsilon\right\} \text{ implies } \{|\widehat{\text{err}}_S(h_S) - \text{err}(h_S)| > \epsilon\}$$

and thus

$$\mathop{\text{Prob}}_{S}\left[\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| > \epsilon\right] \leq \mathop{\text{Prob}}_{S}\left[|\widehat{\text{err}}_S(h_S) - \text{err}(h_S)| > \epsilon\right].$$

Conditioned on S, from the triangle inequality we conclude that

$$\{|\widehat{\text{err}}_S(h_S) - \text{err}(h_S)| > \epsilon \text{ and } |\widehat{\text{err}}_{S'}(h_S) - \text{err}(h_S)| \leq \epsilon/2\} \text{ implies } \{|\widehat{\text{err}}_S(h_S) - \widehat{\text{err}}_{S'}(h_S)| > \epsilon/2.\}$$

Thus,

$$\mathbf{1}_{|\widehat{\text{err}}_S(h_S) - \text{err}(h_S)| > \epsilon}\mathbf{1}_{|\widehat{\text{err}}_{S'}(h_S) - \text{err}(h_S)| \leq \epsilon/2} \leq \mathbf{1}_{|\widehat{\text{err}}_S(h_S) - \widehat{\text{err}}_{S'}(h_S)| > \epsilon/2}. \tag{3}$$

Now, still conditioned on S, take expectation over S' on both sides of the above we obtain

$$\mathbf{1}_{|\widehat{\text{err}}_S(h_S) - \text{err}(h_S)| > \epsilon} \mathop{\text{Prob}}_{S'}\left[|\widehat{\text{err}}_{S'}(h_S) - \text{err}(h_S)| \leq \epsilon/2\right] \leq \mathop{\text{Prob}}_{S'}\left[|\widehat{\text{err}}_S(h_S) - \widehat{\text{err}}_{S'}(h_S)| > \epsilon/2\right]. \tag{4}$$

Now, let's look at the probability $\text{Prob}_{S'}\left[|\widehat{\text{err}}_{S'}(h_S) - \text{err}(h_S)| \leq \epsilon/2\right]$ on the left of (4). (Again, conditioning on S is implicit.) From Exercise 2, we have

$$\mathop{\text{Prob}}_{S'}\left[|\widehat{\text{err}}_{S'}(h_S) - \text{err}(h_S)| < \epsilon/2\right] \leq \frac{1}{4(\epsilon/2)^2 m}$$

which is at most $1/2$ if $\epsilon^2 m \geq 2$. Consequently, when $\epsilon^2 m \geq 2$ we have

$$\operatorname*{Prob}_{S'}\left[|\widehat{\text{err}}_{S'}(h_S) - \text{err}(h_S)| \geq \epsilon/2\right] \geq \frac{1}{2}. \tag{5}$$

Now, (4) implies

$$\mathbf{1}_{|\widehat{\text{err}}_S(h_S) - \text{err}(h_S)| > \epsilon} \leq 2 \cdot \operatorname*{Prob}_{S'}\left[|\widehat{\text{err}}_S(h_S) - \widehat{\text{err}}_{S'}(h_S)| > \epsilon/2\right]. \tag{6}$$

Finally, take expectation over S on both side of (6) we obtain (2). (Why?)

Step 2: Symmetrization using Rademacher variables. Let $\sigma_i \in \{-1, 1\}$, $i \in [m]$, be m independent Rademacher random variables, namely $\text{Prob}[\sigma_i = 1] = \text{Prob}[\sigma_i = -1] = 1/2$. We have

$$\begin{aligned}
& \operatorname*{Prob}_{S,S'}\left[\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \widehat{\text{err}}_{S'}(h)| > \epsilon/2\right] \\
= \; & \operatorname*{Prob}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m (\mathbf{1}_{\{h(\mathbf{x}_i)\neq y_i\}} - \mathbf{1}_{\{h(\mathbf{x}'_i)\neq y'_i\}})\right| > \epsilon/2\right] \\
= \; & \operatorname*{Prob}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\left(\mathbf{1}_{\{h(\mathbf{x}_i)\neq y_i\}} - \mathbf{1}_{\{h(\mathbf{x}'_i)\neq y'_i\}}\right)\right| > \epsilon/2\right] \\
= \; & \text{E}_{S,S'}\left[\operatorname*{Prob}_{\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\left(\mathbf{1}_{\{h(\mathbf{x}_i)\neq y_i\}} - \mathbf{1}_{\{h(\mathbf{x}'_i)\neq y'_i\}}\right)\right| > \epsilon/2 \mid S, S'\right]\right] \\
= \; & \text{E}_{S,S'}\left[\operatorname*{Prob}_{\sigma}\left[\sup_{h\in\Pi_{\mathcal{H}}(S\cup S')} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\left(\mathbf{1}_{\{h(\mathbf{x}_i)\neq y_i\}} - \mathbf{1}_{\{h(\mathbf{x}'_i)\neq y'_i\}}\right)\right| > \epsilon/2 \mid S, S'\right]\right]
\end{aligned}$$

The second equality "says" that independently swapping the ith sample from S and the ith sample from S' does not change the overall probability. The third equality comes from marginalizing over σ.

Step 2': Concentration bounding using Hoeffding inequality. Now, fix S, S', and look at the probability inside the expectation. Note that for a fixed $h \in \Pi_{\mathcal{H}}(S \cup S')$ the expressions

$$Z_i = \sigma_i\left(\mathbf{1}_{\{h(\mathbf{x}_i)\neq y_i\}} - \mathbf{1}_{\{h(\mathbf{x}'_i)\neq y'_i\}}\right)$$

are independent random variables in $[-1, 1]$ with zero expectation. Hence, Hoeffding inequality implies

$$\operatorname*{Prob}_{\sigma}\left[\frac{1}{m}\left|\sum_{i=1}^m Z_i\right| > \epsilon/2\right] < 2e^{-m\epsilon^2/8}.$$

(Here, we use the Hoeffing bound from Exercise 7 of the Tail Inequality lecture.)

Step 3: Union bound. The above inequality holds for a fixed h (and fixed S, S'). Thus, applying the union bound on $\Pi_{\mathcal{H}}(S \cup S')$ we obtain

$$\operatorname*{Prob}_{\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\left(\mathbf{1}_{\{h(\mathbf{x}_i)\neq y_i\}} - \mathbf{1}_{\{h(\mathbf{x}'_i)\neq y'_i\}}\right)\right| > \epsilon/2 \mid S, S'\right] \leq 2|\Pi_{\mathcal{H}}(2m)|e^{-m\epsilon^2/8}.$$

This means the expectation (over S, S') of the LHS is at most the RHS also. Putting everything together, we conclude that

$$\operatorname*{Prob}_{S}\left[\sup_{h\in\mathcal{H}} |\widehat{\text{err}}_S(h) - \text{err}(h)| > \epsilon\right] \leq 4|\Pi_{\mathcal{H}}(2m)|e^{-m\epsilon^2/8}.$$

Setting the last expression $\leq \delta$ and we are done. □