FoSML: Lecture 4

- So far, we have consider exclusively the case when the label y_i is <u>deterministic</u> as a function of x_i. That is, we have assumed there is some latent function $f: X \to Y$ such that $f(x_i) = y_i$, for all i.

- In practice, this is not always reasonable. Indeed, there may be ~~many~~ cases in which $\exists\, y_i, y_i'$ distinct labels, and in which $f(x_i) = y_i$ and $f(x_i) = y_i'$ are both plausible. In this case, it is valuable to consider the $\{y_i\}_{i=1}^m$ label set as randomly dependent on $\{x_i\}_{i=1}^n$, rather than deterministically dependent.

- In this case, we model data generation as follows. Let D be a probability distribution over $X \times Y$, and let $\{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} D$ be a random sample of both $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$.

- What does generalization error mean? The same: for $h \in H$ a hypothesis, we seek to make small

$$R(h) = \mathbb{P}_{(x,y)\sim D}\left(h(x) \neq y\right)$$

$$= \mathbb{E}_{(x,y)\sim D}\left(\mathbb{1}_{h(x)\neq y}\right).$$

- We can consider PAC learning in this <u>stochastic setting</u>, via the notion of

agnostic PAC learning. Let n, size(c) be as in the definition of PAC learning.

Defn (agnostic PAC learning): Let $\mathcal{H}$ be a hypothesis set. An algorithm $\mathcal{A}$ is agnostic PAC-learnable if $\exists$ a polynomial $poly(\cdot,\cdot,\cdot,\cdot)$ such that $\forall \varepsilon, \delta > 0$, and for all distributions D over $\mathcal{X} \times \mathcal{Y}$, $m \geq poly\left(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, size(c)\right)$

$$\Rightarrow \mathbb{P}_{S \sim D}\left(R(h_S) - \min_{h \in \mathcal{H}} R(h) \leq \varepsilon \right) \geq 1 - \delta.$$

- Before, in the deterministic case, there was some function (not necessarily in $\mathcal{H}$) such that $f(x_i) = y_i, \forall i$. In particular, $R(f) = 0$; 0 generalization error was in principal achievable.

- Now, in the stochastic setting, it may be that $R(h) > 0 \ \forall h$. This motivates the notion of Bayes error, which is the best generalization error that can be reasonably hoped for. More precisely

Defn (Bayes error): Given a distribution D over $\mathcal{X} \times \mathcal{Y}$, the Bayes error R^* is defined as $R^* = \inf_{\substack{h \text{ s.t.} \\ h \text{ is measurable}}} R(h)$.

A hypothesis h with $R(h) = h^*$ is a Bayes hypothesis/classifier.

- So, agnostic PAC learnable algorithms can, with polynomially many samples in the relevant parameters, learn a nearly Bayes classifier w.h.p.

Remark: When the problem is deterministic and $R(h) = 0$ is possible, we recover the notion of PAC learning from earlier. We will generally focus on the deterministic case, for ease of exposition.

- We can similarly consider the notion of Bayes classifier h_{Bayes}, where $R(h_{Bayes}) = R^*$.
- It is not hard to see that $h_{Bayes}(x) = \underset{y \in \{0,1\}}{\arg\max}\ \mathbb{P}(y|x)$ (Exercise).

- So, the average error made by h_{Bayes} on $x \in X$ is $\min\{\mathbb{P}(0|x), \mathbb{P}(1|x)\}$.
In particular, if $\mathbb{P}(0|x) = \mathbb{P}(1|x) = 1/2$, then even a Bayes classifier is just flipping an unbiased coin. But, in this case, there is no predictive value from x at all!
- We can indeed quantify a notion of noise based on the Bayes error:

Defn (noise): Given D a distribution on $X \times Y$, the noise at $x \in X$ is

$$\text{noise}(x) = \min\{\mathbb{P}(1|x), \mathbb{P}(0|x)\}.$$

The expected noise is $\mathbb{E}_x(\text{noise}(x))$.
- In this sense, the deterministic case ($\exists f$ s.t. $y = f(x)$) may be called the noiseless case. The closer $\text{noise}(x)$ gets to $1/2$, the noisier it is.

Remark: $R^* = \mathbb{E}(\text{noise}(x))$.

Q: What to do when $|\mathcal{H}| = \infty$? Many of our earlier theorems fail to hold in that case, as the sample bound m involves $\log|\mathcal{H}|$.

- We shall now introduce two notions to allow us to handle $|\mathcal{H}| = \infty$: Rademacher complexity and VC dimension.

- As before, let $\mathcal{H}$ be a hypothesis class of functions $h: \mathcal{X} \to \mathcal{Y}$.
- Let $L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function. Let G be a family of loss functions parametrized by $\mathcal{H}$, for a fixed L: $G = \{g: (x,y) \mapsto L(h(x), y) \mid h \in \mathcal{H}\}$.
- We will phrase our definitions in terms of a more general set $G = \{g: \mathcal{Z} \to [a,b]\}$, for some arbitrary input space $\mathcal{Z}$.
- The notion of Rademacher complexity quantifies the complexity / richness of a collection of functions G by measuring how well it can fit random noise.

Defn (Empirical Rademacher complexity): Let G be a family of functions $G = \{g: \mathcal{Z} \to [a,b]\}$. Let $S = \{z_i\}_{i=1}^m$ be a fixed sample from $\mathcal{Z}$. The empirical Rademacher complexity of G with respect to S is

$$\hat{R}_S(G) = \mathop{\mathbb{E}}_{\sigma}\left(\sup_{g \in G} \sum_{i=1}^{m} \sigma_i g(z_i) \right),$$

where $\sigma = (\sigma_1, \dots, \sigma_m)$ with each σ_i a uniform r.v. on $\{-1, 1\}$.

Remark: Each σ_i is a Rademacher r.v., i.e. $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

- Let $g_S = (g(z_1), \dots, g(z_m))$. We may also write

$$\hat{R}_S(G) = \mathop{\mathbb{E}}_{\sigma}\left(\sup_{g \in G} \frac{\sigma \cdot g_S}{m} \right).$$

- Suppose there is g for σ fixed, some g s.t. $g_S = \sigma$. Then

$$\sup_{g \in G} \frac{\sigma \cdot g_S}{m} = \frac{\|\sigma\|_2^2}{m} = \frac{m}{m} = 1.$$

- So, if for all σ, $\exists\, g^\sigma$ s.t. $g^\sigma_S = \sigma$, then $E_\sigma\left(\sup_{g \in G} \frac{\sigma \cdot g_S}{m}\right) = 1$,

i.e. $\hat{R}_S(G) = 1$. In this sense, empirical R.C. captures how easy it is to fit random noise (i.e. σ) with the elements of G (i.e. g_S).

- Taking an expected value over samples S yields:

Defn (Rademacher complexity): Let D be a distribution over X. For any $m \geq 1$, the Rademacher complexity of G is

$$R_m(G) = \mathop{\mathbb{E}}_{S \sim D}\left(\hat{R}_S(G)\right).$$

- Our goal is to use R.C. to prove generalization bounds, even for infinite H.