Lecture 3

Instructor: *Jess Sorrell* Scribe: *Jess Sorrell*

Last time we considered the problem of "replicating" an estimate of the 0-1 loss of a given model h on distribution D. We proved Hoeffding's Inequality:

Theorem 0.1 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \dots, X_m$ be independent, bounded random variables with $X_i \in [a_i, b_i]$. Let $S_m = \sum_{i=1}^m X_i$. Then*

$$\Pr_{X_1,X_2,\dots,X_m}[S_m \geq \mathbb{E}[S_m] + t] \leq e^{-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}}.$$

and applied it to our loss estimation task. We concluded that to ensure

$$\Pr_{S \sim D^m}[|\ell_S(h) - \ell_D(h)| \geq \varepsilon] \leq \delta$$

it suffices to take $m \geq \frac{\log(2/\delta)}{2\varepsilon^2}$ independent samples from D.

The only thing we used about our replication setting here was that we were computing the average 0-1 loss, and so the random variables $X_i \in [0, 1]$ with probability 1. These kinds of queries are broadly useful in learning and data analysis beyond just estimating losses. So much so, that they have their own name.

Definition 0.2 (Statistical Queries [Kearns, 1998])**.** Let $\mathcal{X}$ denote a domain. A *statistical query* is a function of the form $\phi : \mathcal{X} \to [0, 1]$. Let D be a distribution on $\mathcal{X}$. The *value* of a statistical query ϕ on D is defined $\mathbb{E}_{x\sim D}[\phi(x)]$ (abbreviated $\mathbb{E}_D[\phi]$). We will similarly use the abbreviation $\mathbb{E}_S[\phi] = \frac{1}{m}\sum_{x\in S} \phi(x)$.

The inequality we proved last time applies just as well to any statistical query, and so we have the following theorem.

Theorem 0.3. *Fix any domain $\mathcal{X}$, any distribution D, any statistical query ϕ on $\mathcal{X}$. Then with probability at least $1 - \delta$ over $S \sim_{i.i.d.} D^m$,*

$$|\mathbb{E}_D[\phi] - \mathbb{E}_S[\phi]| \leq \sqrt{\frac{\log(2/\delta)}{2m}}$$

It turns out that this statistical query framework captures a lot of our favorite algorithms:

- gradient descent
- Markov chain monte carlo
- PCA
- K-means clustering

can all be expressed as a sequence of statistical queries. Any algorithm that interacts with its sample exclusively through statistical queries is called a *statistical query algorithm*. Understanding the limits of statistical query algorithms is an active area of research in learning theory!

Non-Adaptive Statistical Queries

What happens now if we want to make not just one, but multiple statistical queries? Say I don't just have one model, but a set $\mathcal{H} = \{h_i\}_{i=1}^t$, and I want to estimate the loss for all of them. Letting $\phi_h(x, y) = \ell(h(x), y)$, we want

$$\Pr[\exists h \in \mathcal{H} \text{ s.t. } |\mathbb{E}_S[\phi_h] - \mathbb{E}_D[\phi_h]| \geq \varepsilon] \leq \delta.$$

Claim 0.4. *With probability at least $1 - \delta$ over $S \sim_{i.i.d.} D^m$,*

$$\max_{h \in \mathcal{H}} |\mathbb{E}_S[\phi_h] - \mathbb{E}_D[\phi_h]| \leq \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2m}}$$

Proof. Let $\varepsilon = \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2m}}$

$$\begin{aligned}
\Pr[\exists h \in \mathcal{H} \text{ s.t. } |\mathbb{E}_S[\phi_h] - \mathbb{E}_D[\phi_h]| \geq \varepsilon] &= \Pr[\cup_{i=1}^t |\mathbb{E}_S[\phi_{h_i}] - \mathbb{E}_D[\phi_{h_i}]| \geq \varepsilon] \\
&\leq \sum_{i=1}^t \Pr[|\mathbb{E}_S[\phi_{h_i}] - \mathbb{E}_D[\phi_{h_i}]| \geq \varepsilon] \qquad \text{union bound} \\
&\leq 2|\mathcal{H}|e^{-2\varepsilon^2 m} \qquad \text{Hoeffding} \\
&= \delta
\end{aligned}$$

□

Definition 0.5 (Probably Approximately Correct (PAC) Learning [Valiant, 1984])**.** Fix a data domain $\mathcal{X}$ and let $\mathcal{Y} = \{0, 1\}$. A model class $\mathcal{H}$ is PAC learnable if there exists an algorithm $\mathcal{L}$ and a function $m_0 : (0, 1)^2 \to \mathbb{N}$ such that for all distributions D over $\mathcal{X} \times \mathcal{Y}$, any $\varepsilon, \delta \in (0, 1)$, and any $m \geq m_0(\varepsilon, \delta)$, letting $S \sim_{i.i.d.} D^m$ and $h \leftarrow \mathcal{L}(S)$,

$$\Pr_S[\ell_D(h) \geq \min_{h^* \in \mathcal{H}} \ell_D(h^*) + \varepsilon] \leq \delta$$

Corollary 0.6. *Finite hypothesis classes $\mathcal{H}$ are PAC-learnable for $m_0(\varepsilon, \delta) \in O(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2})$.*

Proof. Consider the following candidate PAC-learning algorithm for a class $\mathcal{H}$.

Algorithm 1 ERM Learner $\mathcal{L}(S)$

for $h \in \mathcal{H}$ **do**
  $\ell_S(h) = \frac{1}{m} \sum_{j=1}^m \ell(h_i(x_j), y_j)$ (one SQ per h)
end for
return $\text{argmin}_{h \in \mathcal{H}} \ell_S(h)$

Let $h^* = \text{argmin}_{h \in \mathcal{H}} \ell_D(h)$ and let $h \, \text{argmin}_{h \in \mathcal{H}} \ell_S(h)$.

$$\begin{aligned}
\Pr_S[\ell_D(h) \geq \ell_D(h^*) + \varepsilon] &= \Pr_S[\ell_D(h) - \ell_D(h^*) \geq \varepsilon] \\
&= \Pr_S[(\ell_D(h) - \ell_S(h)) + (\ell_S(h) - \ell_S(h^*)) + (\ell_S(h^*) - \ell_D(h^*)) \geq \varepsilon] \\
&\leq \Pr_S[(\ell_D(h) - \ell_S(h)) + (\ell_S(h^*) - \ell_D(h^*)) \geq \varepsilon] \\
&\leq \Pr_S[\ell_D(h) - \ell_S(h) \geq \varepsilon/2] + \Pr_S[\ell_S(h^*) - \ell_D(h^*)] \geq \varepsilon/2]
\end{aligned}$$

From Claim 0.4, taking $m = \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2}$, we have that except with probability δ,

$$\max_{h \in \mathcal{H}} |\ell_S(h) - \ell_D(h)| \leq \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2m}} \leq \varepsilon/2.$$

Therefore

$$\Pr_S[\ell_D(h) \geq \ell_D(h^*) + \varepsilon] \leq \delta.$$

□

References

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.