EN.601.776 Modern Topics in Generalization Spring 2026

SQs, PAC learning finite hypothesis classes

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

1 Statistical Queries

Definition 1.1 (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 1.2. *Fix any domain $\mathcal{X}$, any distribution D, any statistical query ϕ on $\mathcal{X}$. Then with probability at least $1-\delta$ over $S \sim_{i.i.d.} D^m$,*

$$|\mathop{\mathbb{E}}_{D}[\phi] - \mathop{\mathbb{E}}_{S}[\phi]| \leq \sqrt{\frac{\log(2/\delta)}{2m}}$$

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!

1.1 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

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

Claim 1.3. *With probability at least* $1-\delta$ *over* $S \sim_{i.i.d.} D^m$,

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

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

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

□

2 PAC Learning

Definition 2.1 (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 $h \leftarrow \mathcal{L}(S)$,

$$\Pr_{S}[\Pr_{(x,y)\sim D}[h(x) \neq y] \geq OPT + \varepsilon] \leq \delta$$

where $OPT = \min_{h\in\mathcal{H}} \Pr_{(x,y)\sim D}[h(x) \neq y]$.

Claim 2.2. *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**
  $R_S(h) = \frac{1}{m}\sum_{j=1}^{m} \ell(h_i(x_j), y_j)$
end for
return $\operatorname{argmin}_{h\in\mathcal{H}} R_S(h)$

Let $\ell(h(x), y) = \mathbb{1}[h(x) \neq y]$ be the 0-1 loss, let $h^* = \operatorname{argmin}_{h\in\mathcal{H}} R_D(h)$, and let $h = \operatorname{argmin}_{h\in\mathcal{H}} R_S(h)$. To show that this algorithm is a PAC learner for finite hypothesis classes, we will need to bound the probability that it returns a bad hypothesis, i.e., a hypothesis h such that $\Pr_{(x,y)\sim D}[h(x) \neq y] \geq OPT + \varepsilon$. Our algorithm returns a hypothesis h such that

$R_S(h) \leq R_S(h')$ for all $h' \in \mathcal{H}$, so it suffices to bound the probability that there exists some $h \in \mathcal{H}$ such that $R_S(h) \leq R_S(h^*)$, but $R_D(h) \geq R_D(h^*) + \varepsilon$.

What is the probability that a single hypothesis with population risk greater than $OPT+\varepsilon$ has empirical risk less than $R_S(h^*)$?

$$\begin{aligned}
\Pr_S[R_S(h^*) - R_S(h) \geq 0] &= \Pr_S[(R_S(h^*) - R_D(h^*)) + (R_D(h^*) - R_D(h)) + (R_D(h) - R_S(h)) \geq 0] \\
&\leq \Pr_S[(R_S(h^*) - R_D(h^*)) + (R_D(h) - R_S(h)) \geq \varepsilon] \\
&\leq \Pr_S[R_S(h^*) - R_D(h^*) \geq \varepsilon/2] + \Pr_S[R_D(h) - R_S(h)] \geq \varepsilon/2]
\end{aligned}$$

So it suffices to bound the probability that *any* hypothesis $h \in \mathcal{H}$ has empirical risk that deviates from its expectation by more than $\varepsilon/2$. Taking $m = \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2}$, we have that for any given $h \in \mathcal{H}$,

$$\Pr_S[|R_D(h) - R_S(h)| \geq \varepsilon/2] \leq 2e^{-2\varepsilon^2 m/4} = \delta/|\mathcal{H}|.$$

Union bounding over all $h \in \mathcal{H}$ then gives

$$\Pr_S[\max_{h \in \mathcal{H}} |R_D(h) - R_S(h)| \geq \varepsilon/2] \leq \delta.$$

Therefore

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

for the hypothesis h returned by ERM, and so ERM is a PAC learner. □

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.