EN.601.776 Modern Topics in Generalization Spring 2026

Overfitting with SQs

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

1 Statistical Queries Continued

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 $\mathbb{E}_S[\phi] = \frac{1}{m}\sum_{x\in S}\phi(x)$.

1.1 Non-Adaptive Statistical Queries

What happens if we want to make not just one, but multiple statistical queries $\Phi = \{\phi_1, \ldots, \phi_k\}$?

Claim 1.2. *With probability at least* $1 - \delta$ *over* $S \sim_{i.i.d.} D^m$*,*

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

1.2 Adaptive Statistical Queries

Note that once we fix a sample $S \sim_{i.i.d.} D^m$, we can no longer meaningfully bound the deviation $|\mathbb{E}_S[\phi] - \mathbb{E}_D[\phi]|$ for arbitrary statistical queries. For instance, let D be the uniform distribution over $[N]$. Let $S \sim_{i.i.d.} D^m$. Then if we take $\phi_S(x) = \begin{cases} 1, & x \in S \\ 0, & o/w \end{cases}$ we have $\mathbb{E}_S[\phi_S] = 1$ and $\mathbb{E}_D[\phi_S] \leq \frac{m}{N}$. Therefore

$$|\mathop{\mathbb{E}}_{S}[\phi_S] - \mathop{\mathbb{E}}_{D}[\phi_S]| \geq 1 - \frac{m}{N}$$

with probability 1.

Even in our SQ model in which the learner does not get direct access to the sample, we can still run into issues with generalization when answering adaptive SQs with empirical estimates. Suppose our data domain is the integers $\mathbb{Z}_+$ and our distribution is uniform over some large subset of $\mathbb{Z}_+$. Let $\phi_1(x) = 2^{-x}$. Then $\mathbb{E}_S[\phi(x)] = \frac{1}{m}\sum_{i=1}^{m} 2^{-x}$ and the learner can determine the sample S completely by inspecting the binary representation of $m\,\mathbb{E}_S[\phi]$ (so long as there aren't duplicate elements in S, which we can assume whp as long as D is supported on sufficiently many integers). So if we answer SQs with the empirical average on

our sample, we reduce to the case where the learner has access to the sample itself, and can select its next query to overfit badly.

We'll now show that overfitting through adaptive SQs is possible, even for a comparatively benign sequence of queries. Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = \{0,1\}$. The following algorithm iterates over the d features and adds every feature with "good enough" correlation with the label to a set of predictive features P. Then the algorithm return a model which, on input x, returns the majority vote of the features in P on x.

Algorithm 1 Query learner
Inputs/Parameters: Sample $S \sim D^m$

$P = \emptyset$
for $i \in [d]$ **do**
  $\phi_i(x,y) = \begin{cases} 1, & x_i = y \\ 0, & o.w. \end{cases}$
  $a_i \leftarrow E_S[\phi]$
  if $a_i \geq \frac{1}{2} + \frac{1}{\sqrt{m}}$ **then**
    $P = P \cup i$
  end if
  return $h(x) = \lfloor \frac{1}{|P|} \sum_{i \in P} x_i \rceil$
end for

Claim 1.3. *Let $\ell(h(x), y) = \mathbb{1}[h(x) \neq y]$. When D is the uniform distribution over $\mathcal{X} \times \mathcal{Y}$, $\exists$ constant c such that with probability at least $1 - \delta$, if $d \geq c \max\{m, \log(1/\delta)\}$:*

$$|R_s(h) - R_D(h)| \geq .49$$

Compare to the accuracy guarantee we have for non-adaptive statistical queries, from which we would expect

$$|R_S(h) - R_D(h)| \in O\left(\sqrt{\frac{\log(d/\delta)}{m}}\right).$$

Proof. Let

$$X_i = \begin{cases} 1, & i \in P \\ 0, & o.w. \end{cases}$$

and let $A_i = \frac{1}{m} \sum_{(x,y) \in S} \mathbb{1}[x_i = y]$. Then

$$\Pr_{S \sim D^m}[X_i = 1] = \Pr_{S \sim D^m}[A_i \geq \tfrac{1}{2} + \tfrac{1}{\sqrt{m}}]$$

A_i is a binomial random variable with $\mathbb{E}_D[A_i] = \frac{1}{2}$ and standard deviation $\frac{1}{2\sqrt{m}}$, and therefore

$$\Pr_D[X_i = 1] = \Pr_D[A_i \geq \tfrac{1}{2} + \tfrac{1}{\sqrt{m}}] \in \Theta(1).$$

Therefore, each i gets added to P with constant probability. It follows that

$$\mathop{\mathbb{E}}_{S\sim D^m}[|P| = \sum_{i=1}^{d} X_i] = \Theta(d).$$

Recall what the Chernoff-Hoeffding inequality gives us for a sum of bounded r.v.'s $X_i \in [a_i, b_i]$, so for $|P| = \sum_{i=1}^{d} X_i$:

$$\Pr_{X_1,\ldots X_d}[|P| \leq \mathbb{E}[|P|] - t] \leq e^{\frac{-2t^2}{\sum_{i=1}^{m}(b_i-a_i)^2}}.$$

So there is a $t \in \Omega(d)$ such that we have

$$\begin{aligned}
\Pr_{S\sim D^m}[|P| \notin \Theta(d)] &\leq \Pr_{X_1,\ldots,X_d}[|P| \leq \mathbb{E}[|P|] - t] \\
&\leq e^{\frac{-2t^2}{\sum_{i=1}^{d}(b_i-a_i)^2}} \\
&= e^{\frac{-2t^2}{d}} \\
&\in e^{-\Omega(d)}
\end{aligned}$$

So there exists some constant c_1 such that so long as $d > c_1 \log(1/\delta)$,

$$\Pr_{S\sim D^m}[|P| \in \Theta(d)] > 1 - \delta$$

Now let's see how this causes us to get an unreliable empirical estimate of $R_S(h)$ if we reuse the sample S. Let $(x, y) \sim S$ be chosen uniformly at random. $h(x) = y$ iff $\sum_{i\in P} \mathbb{1}[x_i = y] \geq \frac{|P|}{2}$. We have that for each $i \in P$, $\Pr_S[x_i = y] \geq \frac{1}{2} + \frac{1}{\sqrt{m}}$, and so

$$\mathop{\mathbb{E}}_{(x,y)\sim S}[\sum_{i\in P} \mathbb{1}[x_i = y]] \geq \frac{|P|}{2} + \frac{|P|}{\sqrt{m}}$$

Therefore $h(x) = y$ unless $\sum_{i\in P} \mathbb{1}[x_i = y]$ is less than its expectation by at least $\frac{|P|}{\sqrt{m}}$. Applying Chernoff-Hoeffding to the sum of random variables $C = \sum_{i\in P} \mathbb{1}[x_i = y]$, we have

$$\begin{aligned}
R_S(h) &= \Pr_{(x,y)\sim S}[h(x) \neq y] \\
&= \Pr_{(x,y)\sim S}[C \leq \mathbb{E}[X] - \tfrac{|P|}{\sqrt{m}}] \\
&\leq e^{\frac{-2|P|^2}{m|P|}} \\
&= e^{\frac{-2|P|}{m}}
\end{aligned}$$

So if $|P| > \frac{\ln(100)m}{2}$, $R_S(h) = .01$. However, $R_D(h) = 1/2$. We already showed that there exists c_1 such that $|P| \in \Omega(d)$ except with probability δ, so long as $d > c_1 \log(1/\delta)$.

Therefore there exists c_2 such that so long as $d > c_2 m$, $|P| > \frac{\ln(100)m}{2}$, and $R_S(h) = .99$. Letting $c = \max\{c_1, c_2\}$, it follows that there exists a c such that with probability at least $1 - \delta$, if $d \geq c \max\{m, \log(1/\delta)\}$:

$$|R_S(h) - R_D(h)| \geq .49$$

□

Observations

- The argument above still goes through when we don't answer queries with an exact empirical estimate, but instead add some noise on the order $o(\frac{1}{\sqrt{m}})$ to the estimate.
- We could have done a similar analysis using only the first $k-1$ of d features. Redoing the argument using only k statistical queries instead of $d+1$ gives a bound of

  $$|R_S(h) - R_D(h)| \in \Omega(\sqrt{k}m).$$

  So the best confidence interval we can hope for with k adaptive statistical queries, answered by empirical estimate over reused data, is $O(\sqrt{\frac{k}{m}})$. Recall that for k non-adaptive statistical queries, our bound was $O(\sqrt{\frac{\log k}{m}})$.
- If we want a confidence interval of ε, reusing data in this way doesn't save us anything, since we would need $m \in \Omega(\frac{k}{\varepsilon^2})$ samples... which is k times what we would need for a single statistical query.
- Could we have made our adaptive algorithm non-adaptive? The first d queries will non-adaptive, so what if we just committed to estimating the error of every possible f we *might* have constructed in our algorithm, rather than the one that was chosen *after* looking at the results of our d non-adaptive queries. Would this give a better bound? There are 2^d many subsets of d variables that could be included in P, and therefore 2^d different functions f. So we would need to make $O(2^d)$ statistical queries, giving a bound of $O(\sqrt{\frac{\log 2^d}{m}}) = O(\sqrt{\frac{d}{m}})$, so no better than the adaptive version.

References

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