EN.601.774 Theory of Replicable ML Spring 2025

Lecture 9

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

Acknowledgements. Much of this material (and the material for the next few weeks) is lifted wholesale from the course notes of Aaron Roth and Adam Smith, available at

https://www.adaptivedataanalysis.com

Domain $\mathcal{X} = \{0,1\}^d$, $\mathcal{Y} = \{0,1\}$.

Overfitting with "natural" adaptive SQs

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

```
P = ∅
for i ∈ [d] do
    φ_i(x, y) = { 1, x_i = y ; 0, o.w. }
    a_i ← (1/m) Σ_{(x,y)∈S} [φ(x, y)]
    if a_i ≥ 1/2 + 1/√m then
        P = P ∪ i
    end if
    return f(x) = ⌊ (1/|P|) Σ_{i∈P} x_i ⌉
end for
```

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

Claim 0.1. *When D is the uniform distribution over $\mathcal{X} \times \mathcal{Y}$, $\exists$ constant c such that with probability at least $1-\delta$, if $d \geq c\max\{m, \log(1/\delta)\}$:*

$$|acc_S(f) - acc_D(f)| \geq .49$$

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

$$|acc_S(f) - acc_D(f)| \in O\left(\sqrt{\frac{\log(d/\delta)}{m}}\right).$$

Proof. Let

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

Then

$$\Pr_{S\sim D^m}[X_i = 1] = \Pr_{S\sim D^m}[\frac{1}{m}\sum_{(x,y)\in S} \mathbb{1}[x_i = y] \geq \frac{1}{2} + \frac{1}{\sqrt{m}}]$$

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

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

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

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

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

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

applied to $|P|$, there is a $t \in \Omega(d)$ such that we have

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

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

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

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

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

Therefore $f(x) = y$ 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

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

So if $|P| > \frac{\ln(100)m}{2}$, $acc_S(f) = .99$. However, $acc_D(f) = 1/2$. 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 $acc_S(f) = .99$. Letting $c = \max\{c_1, c_2\}$, it follows that there exists a c such that with probability at least $1 - \delta$, if $d \geq c \max\{m, \log(1/\delta)\}$:

$$|acc_S(f) - acc_D(f)| \geq .49$$

□

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 $o(\frac{1}{\sqrt{n}})$ 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

  $$|acc_S(f) - acc_D(f)| \in \Omega(\sqrt{k}m).$$

  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.
- Do replicable SQs help? Since the first d queries are non-adaptive, we know that so long as we use a large enough sample, we can guarantee

  $$\Pr_{S_1, S_2, r}[f^r_{S_1} = f^r_{S_2}] > 1 - \rho$$

where $f^r_{S_i} = A(S_i; r)$. It follows that

$$\begin{aligned}
\Pr_{S_1,r}[acc_{S_1}(f^r_{S_1}) \geq \tfrac{1}{2} + \tau] &= \Pr_{S_1,S_2,r}[acc_{S_1}(f^r_{S_2}) \geq \tfrac{1}{2} + \tau \mid f^r_{S_2} = f^r_{S_1}] \cdot \Pr_{S_1,S_2,r}[f^r_{S_2} = f^r_{S_1}] \\
&\quad + \Pr_{S_1,S_2,r}[acc_{S_1}(f^r_{S_1}) \geq \tfrac{1}{2} + \tau \mid f^r_{S_2} = f^r_{S_1}] \cdot \Pr_{S_1,S_2,r}[f^r_{S_2} \neq f^r_{S_1}] \\
&\leq \Pr_{S_1,S_2,r}[acc_{S_1}(f^r_{S_2}) \geq \tfrac{1}{2} + \tau \mid f^r_{S_2} = f^r_{S_1}] \cdot \Pr_{S_1,S_2,r}[f^r_{S_2} = f^r_{S_1}] + \rho \\
&\leq \Pr_{S_1,S_2,r}[acc_{S_1}(f^r_{S_2}) \geq \tfrac{1}{2} + \tau] + \rho \\
&= \Pr_{S_1,S_2,r}[acc_{S_1}(f^r_{S_2}) - \mathbb{E}_{S_1,S_2,r}[acc_{S_1}(f^r_{S_2})] \geq \tau] + \rho \\
&\leq e^{-2\tau^2 m} + \rho \\
&\in O(\rho)
\end{aligned}$$

so long as we take $m \in \Omega(\frac{\log 1/\rho}{\tau^2})$.

However, to ensure that $\Pr_{S_1,S_2,r}[f^r_{S_2} \neq f^r_{S_1}] \leq \rho$, we need to make d non-adaptive replicable statistical queries with $\rho' = \rho/d$, so we need $O(\frac{d^2}{\tau^2 \rho^2})$ samples. Which is already worse than resampling!