EN.601.774 Theory of Replicable ML Spring 2025

Lecture 10

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 $S \sim D^m$

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

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

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

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

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

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!

Algorithmic stability

We'll now turn to other stability notions and see how they can be used to get us the data-reuse guarantees of replicability (more cheaply).

Setup:

- $\mathcal{X}$ - data domain
- $\mathcal{Y}$ - label space
- $\mathcal{Z}$ - sample space $\mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}$ - output space

Definition 0.2. Two datasets $S, S' \in \mathcal{Z}^m$ are called *neighboring* if they differ in a single element.

Definition 0.3. A deterministic algorithm $\mathcal{A} : \mathcal{Z}^m \to \mathcal{H}$ is ε-uniform change-one (UCO) stable if for all neighboring datasets $S, S' \in \mathcal{Z}^m$, and for all inputs $x \in \mathcal{X}$,

$$|h_S(x) - h_{S'}(x)| \leq \varepsilon$$

where $h_S(x) = \mathcal{A}(S)$ and $h_{S'} = \mathcal{A}(S')$.

Example: k-NN

Algorithm 2 k-NN(S, x')

Inputs/Parameters: Sample $S \in \mathbb{Z}^m$

x', a point to be classified

1: Let $i_1, \ldots, i_k$ be the indices of the k points in S that are nearest to x' (i.e., that minimize $\|x' - x_i\|$, breaking ties arbitrarily)

2: **return** $h_S(x') = \frac{1}{k} \sum_{j=1}^{k} y_j$

Claim 0.4. *k-NN classification is $\frac{1}{k}$-UCO stable.*

Proof. For every point x' and data set S, changing a single point in S changes at most one of the k nearest neighbors, so the average label can go up or down by at most $1/k$. □

Define accuracy $acc_D(h_S) = 1 - \mathbb{E}_{(x,y)\sim D}[|h_S(x) - y|]$ and

$$acc_S(h_S) = 1 - \frac{1}{m} \sum_{(x,y)\in S} |h_S(x) - y|$$

.

Theorem 0.5 (Bousquet-Elisseef'02)**.** *Let $\mathcal{A}$ be ε-UCO stable, for a hypothesis class $\mathcal{H}$ such that $h \in \mathcal{H}$ is bounded. That is, $h : \mathcal{X} \to [0, M]$ for all $h \in \mathcal{H}$. Then for every distribution D over $\mathcal{X} \times \{0, 1\}$, we have that except with probability at most δ over $S \sim D^m$:*

$$|acc_S(h_S) - acc_D(h_S)| \leq \varepsilon + (2\varepsilon m + M)\sqrt{\frac{\ln 1/\delta}{2m}}$$

Theorem 0.6 (McDiarmid's Inequality)**.** *Let $F : \mathcal{Z}^m \to \mathbb{R}$ be a function such that for all neighboring datasets S, S',*

$$|F(S) - F(S')| \leq \varepsilon.$$

Then

$$\Pr_S[|F(S) - \mathbb{E}_S[F]| > t] \leq 2e^{\frac{-2t^2}{m\varepsilon^2}}$$

Proof idea:

1. Use stability of $\mathcal{A}$ to show that $\mathbb{E}_S[acc_S(h_S) - acc_D(h_S)] \leq \varepsilon$
2. Use stability to show that $|(acc_S(h_S) - acc_D(h_S)) - (acc_{S'}(h_{S'}) - acc_D(h_{S'}))| \leq 2\varepsilon + \frac{M}{m}$
3. Apply McDiarmid's inquality to $F(S) = acc_S(h_S) - acc_D(h_S)$ to show that with high probability, $F(S)$ must be close to its expectation

Claim 0.7. *Let $\mathcal{A}$ be ε-UCO stable. Then for every distribution D over $\mathcal{X} \times \{0, 1\}$, the expected generalization error of the classifier is at most ε, that is:*

$$|\mathbb{E}_S[acc_S(h_S) - acc_D(h_S)]| \leq \varepsilon$$

Proof.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{S}[acc_S(h_S) - acc_D(h_S)] &= \mathop{\mathbb{E}}_{S}[\mathop{\mathbb{E}}_{(x,y)\sim D}[|h_S(x) - y|] - \tfrac{1}{m}\sum_{i=1}^{m} |h_S(x_i) - y_i|] \\
&= \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_S(x) - y|] - \mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_S(x_i) - y_i|]\right) && \text{lin of exp} \\
&= \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_{S_{i\to(x,y)}}(x_i) - y_i| - |h_S(x_i) - y_i|]\right) && \text{equivalent dist} \\
&\le \tfrac{1}{m}\sum_{i=1}^{m}\left(\mathop{\mathbb{E}}_{\substack{S\\(x,y)\sim D}}[|h_{S_{i\to(x,y)}}(x_i) - h_S(x_i)|]\right) && \text{triangle ineq} \\
&\le \tfrac{1}{m}\sum_{i=1}^{m}\varepsilon && \varepsilon\text{-UCO stability} \\
&= \varepsilon
\end{aligned}
$$

□