Lecture 3

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Adaptive Data Analysis

Ok, so we now understand that with relatively little data, we can answer *a lot* of statistical queries to high confidence. To recap, one thing we know is the following:

Theorem 1 *Fix any distribution $\mathcal{D}$ and any set of k statistical queries $\phi_1, \ldots, \phi_k$. Let $S \sim \mathcal{D}^n$ consist of a set of n points sampled i.i.d. from $\mathcal{D}$. Then with probability $1-\delta$ over the sample:*

$$\max_i |\mathbb{E}_S[\phi_i] - \mathbb{E}_{\mathcal{D}}[\phi_i]| \leq \sqrt{\frac{\ln(2k/\delta)}{2n}}$$

In other words, for a fixed confidence level δ and sample size n, the maximum error over all of our queries grows only logarithmically in the total number of queries we ask. Said another way, if we fix a goal of having some constant level of error: say $\max_i |\mathbb{E}_S[\phi_i] - \mathbb{E}_{\mathcal{D}}[\phi_i]| \leq 1/100$, then we can ask as many as $k = \frac{\delta}{2}e^{2n/10000}$ queries (exponentially many in n) and still satisfy the bound.

Note that in this theorem, the identities of the queries ϕ_i are fixed *before* the dataset S is sampled. This is clearly important. Let S denote a data set of n points sampled uniformly at random from $\{0,1\}^d$. We can, after the dataset is drawn, define a query ϕ such that $\phi(x) = 1$ if $x \in S$ and $\phi(x) = 0$ otherwise. By definition, $\mathbb{E}_S[\phi] = 1$, but $\mathbb{E}_{\mathcal{D}}[\phi] \leq \frac{n}{2^d}$. So with probability 1, we have $|\mathbb{E}_S[\phi] - \mathbb{E}_{\mathcal{D}}[\phi]| \geq 1 - \frac{n}{2^d} \approx 1$ if n is only polynomial in the dimension of the data d. This is as bad as possible.

Note that just limiting the analyst's access to the dataset to the ability to compute the empirical answers to statistical queries is not enough to prevent this kind of "attack": we can design a single statistical query whose (exact) empirical answer allows us to just "read off" the elements in the sample S. For example, suppose without loss of generality that the data domain $X = \{1, 2, 3, \ldots\}$. Define the query $q(x) = 1/2^x$. Then $n \cdot \mathbb{E}_S[q] = \sum_{x \in S} 1/2^x$, and the binary representation of this value is a histogram representing the dataset. Then, such a data analyst can overfit after asking just two queries: using the first one to read off the data set, and using the 2nd one to overfit as above. Of course, this kind of attack would be foiled if we just truncated our evaluation of $\mathbb{E}_S[q]$ to a small number of bits of precision, so it isn't very robust. Perhaps that is the issue?

The above example — where an "attacker" picks such a query explicitly as a function of S — represents an adversary who is explicitly trying to overfit. As we observed, it is also very brittle. Might it be that "natural" procedures that access the data only via the answers to statistical queries, and that are robust to perturbation, don't have this problem? For such "natural" analyses, could we still derive useful confidence intervals around the answers to statistical queries?

Here we go through a case study of what might happen. Suppose our data domain consists of labeled examples $(x, y) \in \{0,1\}^d \times \{0,1\}$: i.e. each example x consists of d binary features, and is endowed with a binary label y. Our goal is to learn some classifier $f : \{0,1\}^d \to \{0,1\}$ that will classify these examples as well as possible — i.e. to maximize the accuracy $\text{acc}(f) = \Pr_{(x,y)\sim\mathcal{D}}[f(x) = y]$. Note that for a classifier f, $\text{acc}(f)$ is just a statistical query. Here is a seemingly plausible approach we might take:

1. Begin by checking how predictive each feature on its own is with the label: For each i from 1 to d, compute $c_i = \mathbb{E}_S[\mathbb{1}(x_i = y)]$.
2. Say that a feature is *predictive* if $c_i \geq \frac{1}{2} + 1/\sqrt{n}$. Let P be the set of predictive features.
3. Produce a classifier f that simply takes a majority vote over the predictive features:

$$f(x) = \begin{cases} 1, & \sum_{i \in P} x_i \geq |P|/2; \\ 0, & \text{otherwise.} \end{cases}$$

4. Check the performance of our classifier: Compute $\mathrm{acc}_S(f) = \mathbb{E}_S[\mathbb{1}(f(x) = y)]$

This sensible seeming procedure only asks $d+1$ statistical queries. So, if the theorem we proved for non-adaptively chosen statistical queries carries over to "reasonable" adaptive procedures, we might expect that our estimate of the error of our final classifier is fairly accurate: $|\mathrm{acc}_S(f) - \mathrm{acc}(f)| \leq O\left(\sqrt{\frac{\log d/\delta}{n}}\right)$. Is it?

It is not. We will show that when the features are entirely independent of one another and all uncorrelated with the true label, this procedure will significantly over fit.

Theorem 2 *When $\mathcal{D}$ denotes the uniform distribution over $\{0,1\}^d \times \{0,1\}$, there is a constant c such that with probability $1-\delta$, if $d \geq c\max(n, \log(1/\delta))$:*

$$|\mathrm{acc}_S(f) - \mathrm{acc}(f)| \geq 0.49$$

Note that this is as bad as possible, and shows that we cannot in general expect the empirical answers to statistical queries to be non-trivially correct for more than *linearly* many queries, when the queries themselves can be chosen as a function of past answers.

In fact, as we will see in the proof, the manner in which this procedure overfits is not so different from our earlier example of an over-fitting query defined in terms of the data set S itself — the query f above will essentially reconstruct the dataset S.

Proof To analyze how much f overfits, we will deploy a Chernoff bound a couple of times. Lets recall a special case of the Chernoff bound: If we have m independent random variables X_i, taking values in $\{0,1\}$ such that $\Pr[X_i = 1] = p$, and we write $X = \sum_{i=1}^m X_i$, then:

$$\Pr[X < pm - t] \leq \exp\left(-\frac{2t^2}{m}\right)$$

For each i, the quantity $c_i = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}(x_i = y)$ is an independent random variable with a (rescaled) binomial distribution with expected value $1/2$ and standard deviation $\frac{1}{2\sqrt{n}}$. Since a binomial random variable will deviate from its mean by two standard deviations with constant probability, we have that for each i, $\Pr[i \in P] = \Omega(1)$. Thus, $\mathbb{E}[|P|] = \Omega(d)$, and we can apply a Chernoff bound to conclude that so long as $d \geq c\log\frac{1}{\delta}$, we have that except with probability δ, $|P| = \Omega(d)$. Let us continue assuming this is the case.

Now consider a uniformly randomly selected $(x, y) \in S$. By definition of f, we have that $f(x) = y$ if and only if $\sum_{i\in P} \mathbb{1}(x_i = y) > |P|/2$. But by definition of P, we have that for each $i \in P$,

$$\Pr[\mathbb{1}(x_i = y)] \geq \frac{1}{2} + \frac{1}{\sqrt{n}}.$$

Hence:

$$\mathbb{E}[\sum_{i\in P} \mathbb{1}(x_i = y)] \geq \frac{|P|}{2} + \frac{|P|}{\sqrt{n}}$$

Therefore, we have that $f(x) = y$ unless $\sum_{i\in P} \mathbb{1}(x_i = y)$ differs from its expectation by at least $\frac{|P|}{\sqrt{n}}$. But this quantity is the sum of $|P|$ independent 0/1 valued random variables, and so we can apply a Chernoff bound. Thus, for a randomly selected point $(x, y) \in S$ we have:

$$\begin{aligned}
\Pr_{(x,y)\sim S}[f(x) \neq y] &= \Pr[\sum_{i\in P} \mathbb{1}(x_i = y) \leq \mathbb{E}[\sum_{i\in P} \mathbb{1}(x_i = y)] - \frac{|P|}{\sqrt{n}}] \\
&= \leq \exp\left(-\frac{2|P|^2}{n\cdot|P|}\right) \\
&= \exp\left(-\frac{2|P|}{n}\right)
\end{aligned}$$

This is less than $1/100$ when $|P| \geq \frac{n\ln(100)}{2}$.

To recap what we have shown that with probability $1-\delta$:

$$\begin{aligned}\mathrm{acc}_S(f) &= \frac{1}{n}\sum_{(x,y)\in S}[\mathbb{1}(f(x)=y)] \\ &= \Pr_{(x,y)\sim S}[f(x)=y] \\ &= 1-\Pr_{(x,y)\sim S}[f(x)\neq y] \\ &\geq 0.99\end{aligned}$$

On the other hand, we have by construction of $\mathcal{D}$:

$$\mathrm{acc}(f) = \Pr_{(x,y)\sim\mathcal{D}}[f(x)=y] = 0.5$$

since $\Pr[y=1/2]$, and y is independent of x (and therefore of $f(x)$). Therefore, with probability $1-\delta$:

$$|\mathrm{acc}_S(f)-\mathrm{acc}(f)| \geq 0.99-0.5 = 0.49$$

as claimed. ■

A couple of things to note:

1. Unlike the first "attack" we saw, the above analysis is robust to small amounts of noise. Everything continues to go through if instead of observing the exact numerical values of $\mathbb{E}_S[q_i]$ for each query q_i asked, the analyst gets back an arbitrary estimate a_i such that $|a_i - \mathbb{E}_S[q_i]| \leq o(1/\sqrt{n})$.

2. We analyzed the above procedure when the number of queries asked was $k = d+1$. But it is also possible to analyze it when $k \leq d$. This more general analysis asserts that after k queries (the correlation of the first $k-1$ features with the label, followed by the accuracy of the classifier f), we have that:

   $$|\mathrm{acc}_S(f)-\mathrm{acc}(f)| = \Omega\left(\sqrt{\frac{k}{n}}\right).$$

   This shows that in general, the best confidence interval width that the *empirical averages* of statistical queries can be endowed with, in the adaptive setting, is $O(\sqrt{\frac{k}{n}})$. Contrast this with the non-adaptive setting, in which the bound scales like $O(\sqrt{\frac{\log k}{n}})$, which is better by an exponential factor in k.

3. A seemingly conservative approach to deriving upper bounds would be to derive confidence intervals in the non-adaptive setting, uniformly over *all queries that might ever have been asked*, over all possible realizations of the data. This is the "uniform convergence" approach, and is addressed by quantities like VC-dimension. In the above case, we could do this just by counting: there is one possible linear-threshold function f for every possible subset of the d variables, so the total number of queries k' that might ever be asked in the above procedure is $k' = O(2^d)$. We can apply the standard non-adaptive confidence interval calculation to each of these k' potential queries, to derive that with high probability:

   $$|\mathrm{acc}_S(f)-\mathrm{acc}(f)| = O\left(\sqrt{\frac{\log k'}{n}}\right) = O\left(\sqrt{\frac{d}{n}}\right).$$

   In fact, this tightly matches the bound we actually obtain, so in this case, no better analysis (than simply uniformly bounding the error of every query that might ever be asked) is possible.

4. Note also that the two ways we saw of overfitting in this lecture in some sense relied on "learning" the identity of the data points in S, and then carefully tailoring a query to these points. The first (trivial) method did this explicitely. The more subtle method we just analyzed did it implicitly. But this is important: it means that any method of answering statistical queries which promises protection from overfitting must in some sense prevent the analyst from learning too much about too many individual data points in S.

Bibliographic Information A variant of the "linear threshold classifier attack" appeared in the Appendix of [DFH+15].

References

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *STOC*, pages 117–126. ACM, 2015.