13.11.2024 ***Adaptive data analysis***

Lecture 2: The dangers of adaptivity

Source: Lecture notes by Aaron Roth and Adam Smith

Lecturer: Uri Stemmer

We analyze the exercise from the end of the last lecture (showing that even naïve analysts might overfit their data)

Reminder: The domain is $X = \{0,1\}^d \times \{0,1\}$, where for $(x, y) \in X$, we refer to x as an "element" or "example" and refer to y as the "label" of x. Given a sample S (drawn from an unknown distribution $\mathcal{D}$ over the domain X), our goal is to find a labeling function $f: \{0,1\}^d \to \{0,1\}$ that predicts, as accurately as possible, the labels of new examples from $\mathcal{D}$. In other words, we seek a function f such that the following expression is minimized:

$$\text{error}_{\mathcal{D}}(f) = \Pr_{(x,y)\sim\mathcal{D}}[f(x) \neq y] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}\{f(x) \neq y\}]$$

Notes:

- A function f like this is sometimes called a "classifier," "hypothesis," or "prediction rule."
- $\text{error}_{\mathcal{D}}(f)$ is called the "generalization error" of f.
- Note that $\text{error}_{\mathcal{D}}(f)$ is a statistical query.

How can we find such a function f? The following procedure seems reasonable at first glance:

1. For each coordinate $1 \leq i \leq d$, calculate
$$c_i = \mathbb{E}_{(x,y)\sim S}[\mathbb{1}\{x_i = y\}] = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{x_i = y\}$$
In other words, for each coordinate i, we check how "well" it predicts the label.

2. We say that a coordinate i "predicts the label well" if $c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$.
Let $P \subseteq \{1,2,\dots,d\}$ be the set of all coordinates that predict the label well.

3. Define the following prediction rule, which makes a majority decision among the coordinates that predict well:
$$f(x) = \begin{cases} 1 & , \quad \sum_{i\in P} x_i \geq \frac{|P|}{2} \\ 0 & , \quad \text{otherwise} \end{cases}$$

4. Estimate the empirical error of f on our sample, i.e., calculate
$$c_i = \mathbb{E}_{(x,y)\sim S}[\mathbb{1}\{x_i = y\}] = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{x_i = y\}$$

Note that in this procedure, we accessed the sample only by making statistical queries (and learning their empirical values). In total, we asked $d + 1$ statistical queries. Therefore, if the theorem we proved in the previous lesson also applied to analysts who appear "innocent" and access the data only through statistical queries, we could expect, with probability at least $1 - \beta$, to obtain:

$$|\text{error}_S(f) - \text{error}_{\mathcal{D}}(f)| \leq O\left(\sqrt{\frac{\log(d/\beta)}{n}}\right)$$

Is this the case?

Theorem 0: There exists a constant $c > 0$ such that the following holds. For all $\boldsymbol{d \leq c \cdot n}$, with probability at least 0.9, we have

$$|\text{error}_S(f) - \text{error}_{\mathcal{D}}(f)| \leq 0.01$$

(Why is Theorem 0 true?)

Theorem 1: There exists a constant $c > 1$ such that the following holds. For all $\boldsymbol{d \geq c \cdot n}$, if $\mathcal{D}$ is the uniform distribution over $\{0,1\}^d \times \{0,1\}$, then with probability at least 0.9, the following holds:

$$|\text{error}_S(f) - \text{error}_{\mathcal{D}}(f)| \geq 0.49$$

Note:

(1) We know that $\text{error}_{\mathcal{D}}(f) = \frac{1}{2}$, so the above theorem implies that, with high probability, we will find a prediction rule with an empirical error very close to zero.

(2) This means that, in general, we cannot expect the empirical answers to statistical queries to provide non-trivial accuracy for more than a linear number (in n) of queries when the queries are chosen adaptively based on the answers to previous queries.

Lemma: With probability at least 0.95, it holds that $|P| = \Omega(d)$.

Proof Idea for the Lemma:

a. The variables $c_1, c_2, \dots, c_d$ are independent, and for all i we have $\Pr[i \in P] = \Pr\left[c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}\right] = \Omega(1)$.

- The fact that $c_1, c_2, \dots, c_d$ are independent follows from the assumption that $\mathcal{D}$ is uniform over $\{0,1\}^d \times \{0,1\}$ and hence the different coordinates are independent.

- The fact that $\Pr\left[c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}\right] = \Omega(1)$ arises from the nature of $c_i = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{x_i = y\}$, which is a (normalized) binomial random variable with mean $\frac{1}{2}$ and standard deviation $\sigma = \frac{1}{2\sqrt{n}}$. It can be shown that such a random variable deviates from its mean by at least 2σ with a constant probability (these types of bounds are known as "anti-concentration" bounds). See further details below.

b. Thus, by Chernoff's bound, for sufficiently large $d = \Omega(1)$, with probability at least 0.95 we have

$$|P| = \mathbb{1}\{1 \in P\} + \cdots + \mathbb{1}\{d \in P\} = \Omega(d)$$

Further Details on Anti-Concentration:

Consider the random variable $W = n \cdot c_i = \sum_{(x,y)\in S} \mathbb{1}\{x_i = y\}$. This is an unnormalized binomial random variable with mean $\frac{n}{2}$. The Chernoff and Hoeffding bounds we covered in the previous lecture provide a "concentration" result, indicating that, with constant probability W is close to its mean within $\pm\sqrt{n}$. Specifically:

$$\Pr\left[\left|W - \frac{n}{2}\right| \geq \sqrt{n}\right] \leq 2e^{-2} \approx 0.27$$

Now we will show an "anti-concentration" result for this case, showing that the bound above is roughly tight. Specifically, we will show with constant probability, the random variable W does deviate from its mean by something like $\sqrt{n}$.

Anti-Concentration Theorem:

$$\Pr\left[\left|W - \frac{n}{2}\right| \geq \frac{\sqrt{n}}{2}\right] \geq 0.2$$

Proof Sketch:

The probability $\Pr[W = i]$ is maximized for $i = \frac{n}{2}$ where

$$\Pr\left[W = \frac{n}{2}\right] = \frac{number\ of\ sequences\ with \frac{n}{2} ones}{number\ of\ sequences} = \frac{\binom{n}{n/2}}{2^n} = 2^{-n} \cdot \frac{n!}{\left(\frac{n}{2}\right)! \cdot \left(\frac{n}{2}\right)!}$$

Using Stirling's approximation for factorials, which states that $n! \approx \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$, we get

$$\Pr\left[W = \frac{n}{2}\right] \approx 2^{-n} \cdot \frac{\sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n}{\sqrt{\pi n} \cdot \left(\frac{n}{2e}\right)^{\frac{n}{2}} \cdot \sqrt{\pi n} \cdot \left(\frac{n}{2e}\right)^{\frac{n}{2}}} = \sqrt{\frac{2}{\pi n}}$$

Thus,

$$\Pr\left[\left|W-\frac{n}{2}\right|<\frac{\sqrt{n}}{2}\right]=\sum_{i=\frac{n}{2}-\frac{\sqrt{n}}{2}}^{\frac{n}{2}+\frac{\sqrt{n}}{2}}\Pr[W=i]\leq 2\frac{\sqrt{n}}{2}\cdot\sqrt{\frac{2}{\pi n}}=\sqrt{\frac{2}{\pi}}\approx 0.8$$

Which means that

$$\Pr\left[\left|W-\frac{n}{2}\right|\geq\frac{\sqrt{n}}{2}\right]\geq 0.2$$

q.e.d. (This proof was simplified; the probability is actually higher than this bound)

Additionally, by symmetry around n/2n/2n/2, this anti-concentration theorem shows that:

$$\Pr\left[W\geq\frac{n}{2}+\frac{\sqrt{n}}{2}\right]\geq 0.1$$

Proof of Theorem 1:

According to the choice of the distribution $\mathcal{D}$, we have:

$$\text{error}_{\mathcal{D}}(f)=\Pr_{(x,y)\sim\mathcal{D}}[f(x)=y]=0.5$$

Now, let's analyze $\text{error}_S(f)$ and show that, with high probability, it will be much smaller. To do this, we consider an equivalent process for sampling the dataset S:

- First, we sample the column $\vec{y}=(y_1,y_2,\dots,y_n)$
- Then, we sample which coordinates will "predict well" the values in $\vec{y}$, i.e., decide which coordinates are in P
- Finally, we sample the columns $\vec{x}_i$ for coordinates $i\in P$ and for coordinates $i\notin P$ according to the appropriate conditional probabilities.

Note that given $\vec{y}, P$ the columns $\{\vec{x}_i : i\in P\}$ are independent.
Fix some $\vec{y}$ and fix $P\subseteq\{1,2,\dots,d\}$. It holds that:

$$\mathbb{E}_{\vec{x}_1,\dots,\vec{x}_d}[\text{error}_S(f)]=\mathbb{E}_{\vec{x}_1,\dots,\vec{x}_d}\left[\frac{1}{n}\sum_{r\in[n]}\mathbb{1}\{f(x^r)\neq y^r\}\right]$$

$$=\mathbb{E}_{\vec{x}_1,\dots,\vec{x}_d}\left[\mathbb{E}_{r\in[n]}[\mathbb{1}\{f(x^r)\neq y^r\}]\right]=\mathbb{E}_{\substack{\vec{x}_1,\dots,\vec{x}_d\\ r\in[n]}}[\mathbb{1}\{f(x^r)\neq y^r\}]$$

$$=\Pr_{\substack{\vec{x}_1,\dots,\vec{x}_d\\ r\in[n]}}[f(x^r)\neq y^r]=\Pr_{\substack{\vec{x}_1,\dots,\vec{x}_d\\ r\in[n]}}\left[\sum_{i\in P}\mathbb{1}\{x_i^r=y^r\}<\frac{|P|}{2}\right]=\big((1)\big)$$

where the last equality follows from the definition of f (it makes a majority decision among the coordinates in P).

Next we analyze the expectation $\mathbb{E}_{\vec{x}_1,\ldots,\vec{x}_d\ ,\ r\in[n]}[\sum_{i\in P}\mathbb{1}\{x_i^r = y^r\}]$. By definition of P, for every fixture of $\vec{x}_1,\ldots,\vec{x}_d$ (after we have fixed $P,\vec{y}$) and for every coordinate $i \in P$ we have

$$\frac{1}{n}\sum_{r\in[n]}\mathbb{1}\{x_i^r = y^r\} \geq \frac{1}{2}+\frac{1}{\sqrt{n}}$$

That is,

$$\mathbb{E}_{r\in[n]}[\mathbb{1}\{x_i^r = y^r\}] \geq \frac{1}{2}+\frac{1}{\sqrt{n}}$$

and hence

$$\mathbb{E}_{r\in[n]}\left[\sum_{i\in P}\mathbb{1}\{x_i^r = y^r\}\right] = \sum_{i\in P}\mathbb{E}_{r\in[n]}[\mathbb{1}\{x_i^r = y^r\}] \geq \frac{|P|}{2}+\frac{|P|}{\sqrt{n}}$$

and hence

$$\mathbb{E}_{\substack{\vec{x}_1,\ldots,\vec{x}_d\\ r\in[n]}}\left[\sum_{i\in P}\mathbb{1}\{x_i^r = y^r\}\right] \geq \frac{|P|}{2}+\frac{|P|}{\sqrt{n}}$$

Plugging this into $\big((1)\big)$ we get

$$\mathbb{E}_{\vec{x}_1,\ldots,\vec{x}_d}[\mathrm{error}_S(f)] \leq \Pr_{\substack{\vec{x}_1,\ldots,\vec{x}_d\\ r\in[n]}}\left[\sum_{i\in P}\mathbb{1}\{x_i^r = y^r\} \ < \ \mathbb{E}_{\substack{\vec{x}_1,\ldots,\vec{x}_d\\ r\in[n]}}\left[\sum_{i\in P}\mathbb{1}\{x_i^r = y^r\}\right]-\frac{|P|}{\sqrt{n}}\right] \leq \exp\left(-\frac{2|P|}{n}\right)$$

where the last inequality follows from Chernoff's bound. We can use Chernoff's bound here because, once $\vec{y}, P$ are fixed, then $\sum_{i\in P}\mathbb{1}\{x_i^r = y^r\}$ is the sum of $|P|$ independent 0/1 random variables.

The last expression is at most 1/2000 whenever $|P| \geq \frac{n\cdot\ln(2000)}{2}$.
In such a case, by Markov's inequality,

$$\Pr_{\vec{x}_1,\ldots,\vec{x}_d}\left[\mathrm{error}_S(f) > \frac{1}{100}\right] \leq \frac{\mathbb{E}_{\vec{x}_1,\ldots,\vec{x}_d}[\mathrm{error}_S(f)]}{1/100} \leq \frac{1}{20}$$

The above statement is true for every fixture of $\vec{y}$ and for every fixture of P satisfying $|P| \geq \frac{n\cdot\ln(2000)}{2} \triangleq \Gamma$. By the lemma we proved, such a P is obtained w.p. at least 0.95, because the lemma stated that w.p. 0.95 $|P|$ is linear in d, and so this holds for large enough $d = \Omega(n)$.

Overall,

$$\Pr_{S\sim\mathcal{D}}\left[\text{error}_S(f) > \frac{1}{100}\right] =$$

$$= \Pr[|P| < \Gamma] \cdot \Pr\left[\text{error}_S(f) > \frac{1}{100}\middle|P\right] + \Pr[|P| \geq \Gamma] \cdot \Pr\left[\text{error}_S(f) > \frac{1}{100}\middle|P\right]$$

$$\leq \frac{1}{20} \cdot \Pr\left[\text{error}_S(f) > \frac{1}{100}\middle|P\right] + \Pr[|P| \geq \Gamma] \cdot \frac{1}{20} \leq \frac{1}{20} + \frac{1}{20} = \frac{1}{10}$$

So overall, w.p. at least 0.9 the empirical error is at most 1/100.

q.e.d.

Notes:

1. Unlike the first "attack" we analyzed (using the low-order bits), the recent analysis did not require answers with excessively high precision and was robust to small perturbations in the answers (on the order of $o(1/\sqrt{n})$).

2. We observed two "attacks" that successfully overfit the sample. In some sense, both of these attacks first "learned" the sample S in some way, and then used this knowledge to craft a query that overfits (the first attack did this explicitly, and the second attack achieved it more implicitly). The conclusion here is that any method that guarantees generalization in the adaptive case must, in some way, prevent the adversary from learning too much information about the sample.

Summary of What We've Learned So Far:

1. **In the non-adaptive case:** it is possible to answer $\exp(n)$ statistical queries using a sample of size n.
2. **In the adaptive case:** we must restrict our setting somehow. Otherwise, if the analyst gains full access to the sample and can choose any query, overfitting becomes possible.
3. **We decided to limit the analyst's access to the sample.** Specifically, we attempted to restrict access by allowing the analyst to interact with the data only through statistical queries.
4. **This restriction alone is insufficient to prevent overfitting:**
    - We saw an example where an analyst can use the answer to a single (high precision) statistical query to learn the entire sample exactly. The analyst can then craft a "bad" query where the empirical value differs significantly from the true value.
    - We saw a seemingly more "innocent" analyst can arrive at a "bad" query of this kind.

Let us consider a game similar to the one we defined in the previous lecture:

Let M be a mechanism that takes a sample and answers queries. For a sample S and an analyst A, we define the following game, called $\text{AdaptiveGame}_{n,k}(A,S,M)$), or $\text{AG}_{n,k}(A,S,M)$ in short.

$\text{AdaptiveGame}_{n,k}(A,S,M)$

1. The mechanism M gets the sample S (the analyst A does not get S)
2. For $i = 1,2,\dots,k$:
   - The analyst chooses a query q_i
   - The mechanism M gets the query q_i and returns and answer a_i
   - The analyst A gets a_i
3. Return the transcript $T = (q_1, a_1, q_2, a_2, \dots, q_k, a_k)$ of the interaction between A and M

We aim to design a mechanism M such that for any distribution $\mathcal{D}$ and any analyst A, given $S \sim \mathcal{D}^n$, with high probability, the answers M provides during the entire execution are statistically accurate. That is, for all $1 \le i \le k$ we have $a_i \approx q_i(\mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}}[q_i(x)]$.

Definition 1: A mechanism M is $(\boldsymbol{\alpha},\boldsymbol{\beta})$**-statistically-accurate** for k adaptive queries given a sample of size n if for any distribution $\mathcal{D}$ and any analyst A it holds that:

$$\Pr_{\substack{S\sim\mathcal{D}^n \\ \text{AG}_{n,k}(A,S,M)}} [\exists i \text{ s.t. } |a_i - q_i(\mathcal{D})| > \alpha] \le \beta$$

Note: Sometimes we will refer to M as (α,β)-accurate instead of (α,β)-statistically accurate.

Important: For a mechanism to be (α,β)-accurate, it must ensure that its answers are accurate no matter how the analyst A behaves. In particular, the answers must be accurate even for the "worst-case" analyst who tries to cause the mechanism to fail. Therefore, we may sometimes think of the analyst as an "adversary."

Main Question of Interest:
What should n be (as a function of k,α,β) be in order to guarantee accuracy? In other words, our goal is to design a mechanism M that satisfies the above definition using a sample size $n = n(k,\alpha,\beta)$ as small as possible.

Conclusion from the previous lecture:
If the mechanism M answers each query using the exact empirical mean, then it does not satisfy the above definition — not even for $k = 2$.

Question: In the above definition, does the mechanism M know the distribution $\mathcal{D}$? Does the analyst know it?

How can we design accurate mechanisms (according to Definition 1)?

Option 1: Sample Splitting

We can design a mechanism that splits its sample into k disjoint parts and answers each query using a different part of the sample:

$\mathbf{SampleSplitting}_{n,k,\alpha,\beta}(S)$

Input: a sample S of size n

1. Partition S into k disjoint parts: $S_1, S_2, \dots, S_k$
2. For $i = 1,2, \dots, k$
   - Get the next query q_i
   - Return $q_i(S_i) = \frac{1}{|S_i|}\sum_{x \in S_i} q_i(x)$

Question: For which parameter $k = k(\alpha, \beta, n)$ can we show that the above algorithm is (α,β) - accurate for k queries given a sample of size n?