6.11.2024 ***Adaptive data analysis***

Lecture 1: Introduction

Source: Lecture notes by Aaron Roth and Adam Smith

Lecturer: Uri Stemmer

The central question of the course:

How can we learn about an unknown distribution $\mathcal{D}$ when all we have is a set of samples from $\mathcal{D}$?

The samples naturally contain useful information about the distribution $\mathcal{D}$, but the danger is that if we are not careful, we might "discover" insights that hold for our specific sample but are not generally true for $\mathcal{D}$. This phenomenon is called **"overfitting"** or **"false discovery"**.

We have a fairly good understanding of the overfitting problem, and effective tools to handle it, in the **non-adaptive** setting. That is, when we commit in advance to the tests we will perform on the collected data, then collect the data, and finally perform exactly the tests we committed to (and nothing beyond that). The situation is much less clear in the **adaptive** setting, where the researcher decides on the tests to perform only after collecting (and seeing) the data.

Our course will focus on the adaptive case. However, today we will start with a basic review of the non-adaptive case.

Non-adaptive data analysis

Let's start with the following simple problem. Suppose we have a biased coin, meaning the probability of getting "heads" is not equal to the probability of getting "tails". We can flip the coin as many times as we want and observe the results.

How can we learn the probability of getting "heads" when flipping this coin?

Formally:
There is a distribution Bernoulli(p) with an unknown parameter p. That is, if $X \sim$ Bernoulli(p), then $\Pr[X = 1] = p$ and $\Pr[X_i = 0] = 1 - p$. We have a "database" $x = (x_1, x_2, \dots, x_n)$ containing n independent samples from this distribution (i.e., the results of n coin flips). How can we use x to learn the value of p?

One option is to use the empirical mean $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ as our estimate for p.

How accurate is this estimate? How can we quantify its "level of accuracy"?

Example:
Suppose that $p = 1/2$ and $n = 1000$. Although the expectation of $\hat{p}$ is exactly p, the probability of actually obtaining $\hat{p} = 1/2$ (exactly) is relatively small – approximately $\sqrt{2/\pi n}$. (More generally, it could be that p is irrational, in which case the probability that $\hat{p} = p$ is zero...).

This means that the following claim is false:
"As $\boldsymbol{n}$ increases, the probability that $\boldsymbol{\hat{p}}$ exactly equals $\boldsymbol{p}$ increases."

So what can we say?

- We can be "fairly confident" that $|\boldsymbol{\hat{p}} - \boldsymbol{p}| \leq \mathbf{0.05}$. Indeed, this holds except with probability at **0.02**.
- We can be "even more confident" that $|\boldsymbol{\hat{p}} - \boldsymbol{p}| \leq \mathbf{0.06}$. Indeed, this holds except with probability at **0.0015**.

Arguments like these are known as *high probability bounds* or *confidence intervals.*

Now, we want to understand how to prove such statements. That is, how can we establish claims of the form:

$$\mathbf{Pr}[|\boldsymbol{\hat{p}} - \boldsymbol{p}| > \boldsymbol{A}] \leq \boldsymbol{B}$$

where the probability is taken over the randomness of the coin flips, i.e., over the sampling of our "database" x. Our goal is to make both A and B as small as possible.

To do this, we need to recall some background from probability theory.

Theorem [Markov's Inequality]: For any non-negative random variable Y and any $a > 0$ we have

$$\Pr[Y \geq a] \leq \frac{\mathbb{E}[Y]}{a}$$

Proof (for a discrete random variable Y):

$$\mathbb{E}[Y] = \sum_{y} y \cdot \Pr[Y = y] \geq \sum_{y \geq a} y \cdot \Pr[Y = y] \geq \sum_{y \geq a} a \cdot \Pr[Y = y] = a \cdot \sum_{y \geq a} \Pr[Y = y] = a \cdot \Pr[Y \geq a]$$

In our coin-flipping example, we know that $\mathbb{E}[\hat{p}] = p$, so Markov's inequality gives us some (weak) connection between p and $\hat{p}$. However, this is still not strong enough to obtain the high probability bounds we were aiming for. Nevertheless, Markov's inequality is a very useful tool. In particular, it allows us to prove Chebyshev's inequality:

Theorem [Chebyshev's Inequality]: For any random variable Y with expectation $\mu = \mathbb{E}[Y]$ and variance $\sigma^2 = \mathrm{Var}(Y) = \mathbb{E}[(Y - \mu)^2] > 0$, and for any $a > 0$, the following holds

$$\Pr[|Y - \mu| \geq a\sigma] \leq \frac{1}{a^2}$$

Proof:

$$\Pr[|Y - \mu| \geq a\sigma] = \Pr[(Y - \mu)^2 \geq a^2\sigma^2] \leq \frac{\mathbb{E}[(Y - \mu)^2]}{a^2\sigma^2} = \frac{1}{a^2}$$

Where the inequality follows from Markov's inequality.

This is already enough to give us a non-trivial probabilistic bound on our coin-flipping problem. That is, we can use Chebyshev's inequality to obtain a probabilistic bound on $|\hat{p} - p|$. To do this, we need to analyze the variance of $\hat{p}$. For that, we need to recall basic properties of expectation and variance.

Theorem: For independent random variables, the expectation of a product equals the product of the expectations, i.e.,

$$\mathbb{E}[X_1 \cdot X_2 \cdots X_n] = \mathbb{E}[X_1] \cdot \mathbb{E}[X_2] \cdots \mathbb{E}[X_n]$$

Proof:
We prove the statement for discrete (and independent) random variables X, Y

$$\mathbb{E}[X \cdot Y] = \sum_{x,y} \Pr[X = x, Y = y] \cdot xy = \sum_{x,y} \Pr[X = x] \cdot \Pr[Y = y] \cdot xy$$
$$= \left(\sum_{x} \Pr[X = x] \cdot x \right) \cdot \left(\sum_{y} \Pr[Y = y] \cdot y \right) = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Theorem: For any random variable Y and any constant $a > 0$, the following holds

$$\mathrm{Var}(aY) = a^2 \cdot \mathrm{Var}(Y)$$

Proof: Denote $\mu = \mathbb{E}[Y]$. Then,

$$\mathrm{Var}(aY) = \mathbb{E}[(aY - a\mu)^2] = a^2 \mathbb{E}[(Y - \mu)^2] = a^2 \cdot \mathrm{Var}(Y)$$

Theorem: For a pair of independent random variables Y_1, Y_2 it holds that

$$\mathrm{Var}(Y_1 + Y_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2)$$

Proof: This follows from the definition of variance and from the fact (which we proved) that for independent RV's it holds that the expectation of a product equals the product of the expectations. Specifically,

$$\mathbb{E}\left[\left(Y_1 + Y_2 - \mathbb{E}[Y_1 + Y_2] \right)^2 \right] = \mathbb{E}\left[\left(\ (Y_1 - \mathbb{E}[Y_1]) \ + \ (Y_2 - \mathbb{E}[Y_2]) \ \right)^2 \right]$$

$$= \mathbb{E}\left[(Y_1 - \mathbb{E}[Y_1])^2 + (Y_2 - \mathbb{E}[Y_2])^2 + 2(Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2]) \right]$$

$$= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 2\,\mathbb{E}\left[(Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2]) \right]$$

$$= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 2\,\mathbb{E}\left[Y_1 Y_2 - Y_1 \mathbb{E}[Y_2] - Y_2 \mathbb{E}[Y_1] + \mathbb{E}[Y_1]\mathbb{E}[Y_2] \right]$$

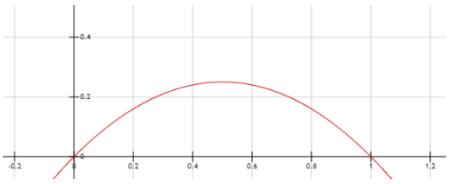
$$= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 2\left(\mathbb{E}\left[Y_1 Y_2\right] - \mathbb{E}[Y_1]\mathbb{E}[Y_2] - \mathbb{E}[Y_2]\mathbb{E}[Y_1] + \mathbb{E}[Y_1]\mathbb{E}[Y_2]\right)$$

$$= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 2\left(\mathbb{E}\left[Y_1 Y_2\right] - \mathbb{E}[Y_1]\mathbb{E}[Y_2]\right) \underbrace{=}_{\text{independent}} \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2)$$

Let's return to our coin-flipping problem. We want to analyze the variance of $\hat{p}$. Recall that we defined $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where each X_i is a Bernoulli random variable with parameter p. For each X_i we know that

$$\mathrm{Var}(X_i) = \mathbb{E}[(X_i - p)^2] = \underbrace{\mathbb{E}\left[X_i^2\right]}_{\substack{X_i \text{ is a bit} \\ \text{and so } X_i^2 = X_i \\ \text{and } \mathbb{E}[X_i^2] = \mathbb{E}[X_i] = p}} - 2p\mathbb{E}[X_i] + p^2 = p - 2p^2 + p^2 = p(1-p)$$

Fact (see plot on the right):
For every $0 \le p \le 1$ it holds that $p(1-p) \le \frac{1}{4}$.



Therefore $\mathrm{Var}(X_i) \le \frac{1}{4}$, and hence

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \le \frac{1}{4n}$$

So we can apply Chebyshev's inequality to $\hat{p}$ and obtain that for any $a > 0$, the following holds:

$$\Pr\left[|\hat{p} - p| \ge \frac{a}{2\sqrt{n}}\right] \le \frac{1}{a^2}$$

Let $\beta > 0$ be a parameter and denote $a = \sqrt{\frac{1}{\beta}}$. We get,

$$\Pr\left[|\hat{p} - p| \ge \sqrt{\frac{1/\beta}{4n}}\right] \le \beta$$

Or in other words, with probability at least $1 - \beta$ it holds that $|\hat{p} - p| \le \sqrt{\frac{1/\beta}{4n}}$.

Another way to look at this: Suppose that for certain parameters $\alpha, \beta > 0$, we want to ensure that with probability at least $(1 - \beta)$, our estimate $\hat{p}$ is close to p within an error of at most α. What sample size n do we need?

To answer this question using the above inequality, we require that $|\hat{p} - p| \le \sqrt{\frac{1/\beta}{4n}} \le \alpha$. Solving for n we see that it suffices to take

$$n \ge \frac{1/\beta}{4\alpha^2}$$

This is great! This gives us some high probability bound on $|\hat{p} - p|$. However, we don't want to stop here. In the above requirement, we see that n needs to grow proportionally to $1/\beta$, which is not ideal if we want β to be extremely small (i.e., if we want to be super confident in the validity of our estimate).

It turns out that in this case, much stronger guarantees can be obtained using what are called Chernoff/Hoeffding bounds.

Theorem [Chernoff and Hoeffding Bounds]:

- Let $X_1, X_2, \ldots, X_n$ be independent random variables where for each i we have $\Pr[X_i{=}1] = p$ and $\Pr[X_i{=}0] = 1 - p$ for some parameter $0 < p < 1$. The expected sum of the variables is $\mathbb{E}[\sum_{i=1}^{n} X_i] = p \cdot n$. Then the following holds:
  - **(a) For all $\boldsymbol{0 < \alpha < 1}$ it holds that $\boldsymbol{\Pr[\sum_{i=1}^{n} X_i \geq (1+\alpha) \cdot pn] < \exp(-\alpha^2 pn/4)}$**
  - **(b) For all $\boldsymbol{0 < \alpha < 1}$ it holds that $\boldsymbol{\Pr[\sum_{i=1}^{n} X_i \leq (1-\alpha) \cdot pn] < \exp(-\alpha^2 pn/2)}$**
- For $B > A > 0$ let $X_1, X_2, \ldots, X_n \in [A, B]$ be independent random variables, and define $\sum_{i=1}^{n} \mathbb{E}[X_i] = \mu$. Then,
  - **(c) For all $\boldsymbol{\delta > 0}$ it holds that $\boldsymbol{\Pr[|(\sum_{i=1}^{n} X_i) - \mu| \geq \delta] \leq 2\exp\left(-\frac{2\delta^2}{n \cdot (B-A)^2}\right)}$**

Before discussing the proof of this theorem, let's see what it implies for our coin example. Specifically, using inequality (c) with $\delta = n\alpha$ we obtain

$$\Pr[|\hat{p} - p| \geq \alpha] \leq 2\exp(-2\alpha^2 n)$$

To ensure that $2\exp(-2\alpha^2 n)$ is at most β (for some parameter $\beta > 0$), it suffices to require $n \geq \frac{\ln\left(\frac{2}{\beta}\right)}{2\alpha^2}$. Notice that now the dependence of n on $1/\beta$ is logarithmic. This means that for a relatively small "cost" in terms of sample size n we can make β arbitrarily small.

Proof [Chernoff and Hoeffding Bounds]:

We will prove only part (a). The proofs of (b) and (c) are similar. Let $\boldsymbol{X_1, X_2, \ldots, X_n}$ be independent RV's such that for each $\boldsymbol{i}$ we have $\boldsymbol{\Pr[X_i = 1] = p}$ and $\boldsymbol{\Pr[X_i = 0] = 1 - p}$. Let $\boldsymbol{0 < \alpha < 1}$. We need to show that

$$\boldsymbol{\Pr\left[\sum_{i=1}^{n} X_i \geq (1+\alpha) \cdot pn\right] < \exp(-\alpha^2 pn/4)}$$

Denote $\boldsymbol{t = (1+\alpha)pn}$ and denote $\boldsymbol{c = \alpha/2}$. (Note that since $\boldsymbol{0 < \alpha < 1}$ then $\boldsymbol{0 < c < 1/2}$).
We calculate:

$$\boldsymbol{\Pr[\sum X_i \geq t] = \Pr[c \cdot \sum X_i \geq c \cdot t] = \Pr\left[e^{c \cdot \sum X_i} \geq e^{c \cdot t}\right] = ((1))}$$

Now by Markov's inequality we get that

$$\big((\mathbf{1})\big) \le e^{-c\cdot t}\cdot \mathbb{E}\big[e^{c\cdot\sum X_i}\big] = e^{-c\cdot t}\cdot \mathbb{E}[e^{c\cdot X_1}\cdot e^{c\cdot X_2}\cdots e^{c\cdot X_n}] = ((\mathbf{2}))$$

As $X_1, \dots, X_n$ are independent, we get that

$$\big((\mathbf{2})\big) = e^{-c\cdot t}\cdot \mathbb{E}[e^{c\cdot X_1}]\cdot \mathbb{E}[e^{c\cdot X_2}]\cdots \mathbb{E}[e^{c\cdot X_n}] = ((\mathbf{3}))$$

And since $X_1, \dots, X_n$ are identically distributed, it holds that $\mathbb{E}[e^{c\cdot X_1}] = \mathbb{E}[e^{c\cdot X_2}] = \cdots = \mathbb{E}[e^{c\cdot X_n}]$ and therefore

$$\big((\mathbf{3})\big) = e^{-c\cdot t}\cdot \Big(\mathbb{E}[e^{c\cdot X_1}]\Big)^n$$

So we have established that

$$\mathbf{Pr}[\textstyle\sum X_i \ge t] \le e^{-c\cdot t}\cdot \Big(\mathbb{E}[e^{c\cdot X_1}]\Big)^n$$

Now note that

$$\mathbb{E}[e^{c\cdot X_1}] = p\cdot e^c + (1-p)e^0 = p\cdot e^c + 1 - p \le p\big(1 + c + c^2\big) + 1 - p = 1 + p\big(c + c^2\big) \le e^{p(c+c^2)}$$

Where the first inequality follows from the formula $e^z \le 1 + z + z^2$, which holds for all $z \le 1$, and the second inequality follows from the formula $1 + z \le e^z$, which holds for all $z \in \mathbb{R}$.

Plugging this into our previous calculation we obtain

$$\mathbf{Pr}[\textstyle\sum X_i \ge t] \le e^{-c\cdot t}\cdot \Big(\mathbb{E}[e^{c\cdot X_1}]\Big)^n \le e^{-c\cdot t}\cdot \Big(e^{p(c+c^2)}\Big)^n = e^{-c\cdot t}\cdot e^{pn(c+c^2)} = ((\mathbf{4}))$$

Recall that we chose $t = (1+\alpha)pn$ and so

$$\big((\mathbf{4})\big) = e^{-c\cdot(1+\alpha)pn}\cdot e^{pn(c+c^2)} = e^{-cpn(\alpha - c)} = ((\mathbf{5}))$$

Recall that we chose $c = \alpha/2$ and so

$$\big((\mathbf{5})\big) = e^{-\alpha^2 pn/4}$$

q.e.d.

What happened in this proof?

Markov's inequality says that a non-negative random variable cannot deviate a lot from its expectation. But in the Chernoff bound, we wanted to show that $\sum X_i$ cannot deviate even a little from its expectation. So instead of applying Markov's inequality directly to $\sum X_i$ we considered the random variable $\exp(\sum X_i)$.

Why is this helpful? Now Markov tells us that the probability that $\exp(\sum X_i)$ deviates a lot from its expectation is small — and this implies that the probability that $\sum X_i$ deviates even a little from its expectation is small (because if $\sum X_i$ deviates even a little , then $\exp(\sum X_i)$ deviates a lot...).

Additionally, thanks to the fact that $X_1, \dots, X_n$ are independent, we were able to analyze the expectation of $\exp(\sum X_i)$ since this expectation split into a product of expectations.

So we've learned how to estimate the expectation of an unknown coin. Somewhat surprisingly, this is a super useful tool. For example, suppose our data contains medical records of individuals randomly sampled from a certain population. Additionally, suppose we have a test which, given a person's medical information, indicates whether they have a certain disease or not. Using the same method, we can estimate from the data the percentage of people in the entire population who have the disease.

Definition: A statistical query over a domain $\boldsymbol{X}$ is a function $\boldsymbol{q}: \boldsymbol{X} \to [\mathbf{0}, \mathbf{1}]$. For a distribution $\boldsymbol{\mathcal{D}}$ over the domain X, the value of the query $\boldsymbol{q}$ on $\boldsymbol{\mathcal{D}}$ is

$$\boldsymbol{q}(\boldsymbol{\mathcal{D}}) \coloneqq \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\mathcal{D}}}[\boldsymbol{q}(\boldsymbol{x})]$$

For a database $\boldsymbol{S} \in \boldsymbol{X}^{\boldsymbol{n}}$, the value of $\boldsymbol{q}$ on $\boldsymbol{S}$ is

$$\boldsymbol{q}(\boldsymbol{S}) \coloneqq \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{x}\in\boldsymbol{S}} \boldsymbol{q}(\boldsymbol{x})$$

Using Chernoff/Hoeffding bounds, we can estimate $q(D)$ using $q(S)$:

Theorem: Let $\boldsymbol{\mathcal{D}}$ be a distribution over a domain $\boldsymbol{X}$, let $\boldsymbol{q}: \boldsymbol{X} \to [\mathbf{0}, \mathbf{1}]$ be any statistical query, and let $\boldsymbol{\beta} > \mathbf{0}$. For a sample $\boldsymbol{S} \sim \boldsymbol{\mathcal{D}}^{\boldsymbol{n}}$ consisting of $\boldsymbol{n}$ independent samples from $\boldsymbol{\mathcal{D}}$ it holds that

$$\mathbf{Pr}\left[|\boldsymbol{q}(\boldsymbol{S}) - \boldsymbol{q}(\boldsymbol{\mathcal{D}})| \leq \sqrt{\frac{\ln\left(\frac{\mathbf{2}}{\boldsymbol{\beta}}\right)}{\mathbf{2}\boldsymbol{n}}}\right] \geq \mathbf{1} - \boldsymbol{\beta}$$

Where the probability is over the randomness of the sample $\boldsymbol{S}$.

Question: What happens if we have k statistical queries and we want to obtain an estimate of the expectation of each of them?

Theorem: Let $\mathcal{D}$ be a distribution over a domain X, let $q_1, q_2, \dots, q_k: X \to [0,1]$ be statistical queries, and let $\beta > 0$. For a sample $S \sim \mathcal{D}^n$ containing n independent sampled from $\mathcal{D}$ we have

$$\Pr\left[\max_{i\in\{1,2,\dots,k\}} |q_i(S) - q_i(\mathcal{D})| \leq \sqrt{\frac{\ln(2k/\beta)}{2n}}\right] \geq 1 - \beta,$$

where the probability is over sampling S.

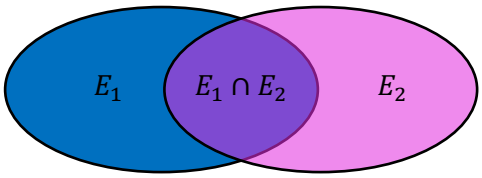
That is, for a given parameter β and sample size n, our maximum error when estimating the expectation of k queries grows only logarithmically with k. In other words, suppose there is a certain error α (for example, $\alpha = 1/100$) that we are willing to tolerate. Then we can estimate the expectation of $k = \frac{2}{\beta} \cdot e^{2\alpha^2 n}$ queries (an exponential number in n).

Proof:

Recall the union bound: For any collection of k events $E_1, E_2, \dots, E_k$ (in the same probability space), we have:

$$\Pr\left[\bigcup_{j=1,2,\dots,k} E_j\right] \leq \sum_{j=1}^{k} \Pr[E_j]$$

In a diagram:



Let $\beta > 0$. As we've seen, by the Hoeffding bound, for any (single) statistical query q_i it holds that

$$\Pr\left[|q_i(S) - q_i(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}}\right] \leq \frac{\beta}{k}$$

Therefore, by a union bound,

$$\Pr\left[\max_{i\in\{1,2,\dots,k\}} |q_i(S) - q_i(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}}\right] =$$

$$= \Pr\left[\left\{|q_1(S) - q_1(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}}\right\} \; OR \; \dots \; OR \; \left\{|q_k(S) - q_k(\mathcal{D})| > \sqrt{\frac{\ln(2k/\beta)}{2n}}\right\}\right]$$

$$\leq \Pr\left[|q_1(S) - q_1(\mathcal{D})| > \sqrt{\frac{\ln\left(\frac{2k}{\beta}\right)}{2n}}\right] + \dots + \Pr\left[|q_k(S) - q_k(\mathcal{D})| > \sqrt{\frac{\ln\left(\frac{2k}{\beta}\right)}{2n}}\right] \leq \frac{\beta}{k} + \dots + \frac{\beta}{k} = \beta$$

q.e.d.

Great. So in the non-adaptive case we can estimate the expectation of <u>many</u> queries using a relatively small sample.

Question: What do we mean by "non-adaptive" here?
Answer: The queries $q_1, q_2, \dots, q_k$ are fixed <u>before</u> the sample S is dawn.

Question: Does a similar result hold also when the queries are chosen <u>after</u> the sample S is drawn, by an analyst who sees the sample S?
Answer: In general no!

Example: suppose that our domain is $X = \{0,1\}^d$ and that the distribution $\mathcal{D}$ is uniform over this domain. After S is drawn (and we see it) we can define the following query:

$$q_S(x) = \begin{cases} 1, & \text{if } x \in S \\ 0, & \text{if } x \notin S \end{cases}$$

By definition, we have that $q_S(S) = \frac{1}{n}\sum_{x \in S} q_S(x) = 1$ but $q_S(\mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[q_S(x)] \leq \frac{n}{2^d}$.

That is, if we can select our query as a function of S, then we might select q_S, in which case with probability 1 it holds that

$$|q_S(S) - q_S(\mathcal{D})| \geq 1 - \frac{n}{2^d} \approx 1$$

(assuming that $n \ll 2^d$)

Note: In the last example, we gave the analyst unrestricted access to the sample S, which allowed him to define a statistical query that heavily overfits: the query returns 1 exactly on the elements of the sample and returns 0 outside the sample. The conclusion is that if we want to guarantee something in the adaptive case, we need to somehow restrict our setting. There are two main approaches to this:

Approach 1: Focus only on analysts who perform a pre-specified set of allowed operations. That is, we allow the analyst to receive the sample S, but we restrict his behavior afterward. For example:

- An analyst restricted to choosing queries only from a specific family, meaning not every query is "valid".
- When the analyst wants to perform a specific sequence of actions with the sample (e.g., first select variables and then run regression on them), sometimes it is possible to analyze this and show there's no danger of overfitting.

Approach 2: Restrict the analyst's access to the sample S (but otherwise do not restrict the analyst). That is, we don't give the analyst the sample S itself, but allow him only limited access to S in some way. Beyond that, we place no restrictions on the analyst and make no assumptions about how they operate.

In our course, we will focus on **Approach 2**. The advantages of Approach 2 are: (1) It is more general; (2) It is easier to enforce; and (3) It requires no assumptions about what the analyst will do with the data. The advantage of Approach 1 is that it allows for "tailored" solutions for specific cases, and so sometimes achieves better results

Question: How can one restrict, or is it reasonable to restrict, the analyst's access to the sample?

Proposal: Allow the analyst to access the data only by checking the empirical value of statistical queries of his choice.

Formally, let us consider the following game involving an analyst A and a distribution $\mathcal{D}$.

1. Draw a sample $S \sim \mathcal{D}^n$ (the analyst A does not receive the sample S)
2. For $i = 1,2,\ldots,k$:
   - The analyst A specifies a statistical query q_i
   - The analyst A receives the empirical value of the query $q_i(S) = \frac{1}{n}\sum_{x \in S} q_i(x)$

Question: Will this guarantee avoidance of overfitting with high probability? That is, does it guarantee that no matter what the analyst does, as long as they access the data in this way, then with high probability, for every such q_i, it holds that

$$q_i(S) \approx q_i(\mathcal{D})$$

Answer: No. This can fail already for $k = 2$.

Example: Suppose that our domain is $X = \{1,2,\ldots,2n\}$ and that the distribution $\mathcal{D}$ is uniform over X. We will design an analyst that asks one query q_1 and then finds a query q_2 such that $q_2(S) \gg q_2(\mathcal{D})$.

The first query q_1 is defined as follows:

$$q_1(x) = 0.000000\ldots01$$

$$\#\text{zeroes} = x \cdot \log n$$

Note: From the empirical value $q_1(S)$, the analyst learns the entire sample S, because it is "encoded" in the low-order bits of the answer. For example, if $S = (1,3,4,4,4)$, then $n \cdot q_1(S) = \sum_{x \in S} q_1(x)$ is

$$\begin{array}{r} \mathbf{0.0001} \\ \mathbf{+0.0000000001} \\ \mathbf{+0.0000000000001} \\ \mathbf{+0.0000000000001} \\ \mathbf{+0.0000000000001} \\ \hline \mathbf{0.0001000001011} \end{array}$$

After the analyst has learned S, he can define the query q_2 as we saw earlier:

$$q_2(x) = \begin{cases} 1, & \text{if } x \in S \\ 0, & \text{if } x \notin S \end{cases}$$

Again, we have that $q_S(S) = \frac{1}{n}\sum_{x \in S} q_S(x) = 1$ but $q_S(\mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[q_S(x)] \leq \frac{n}{|X|} = \frac{1}{2}$.

Discussion:

1. The last example relied on us working with very high precision and giving the analyst the exact empirical value of the query. This example breaks down if we restrict ourselves to using only a small number of digits after the decimal point. Could this solve the problem in general?
2. The analyst in the last example explicitly tried to overfit. Perhaps the problem does not exist for "benign" analysts?

Homework exercise for thought ahead of the next class — optional — not for submission:

Assume the domain is $X = \{0,1\}^d \times \{0,1\}$, where for $(x, y) \in X$ we think of x as the "example" and 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 $f: X \to \{0,1\}$ that predicts the labels of new examples from $\mathcal{D}$ as accurately as possible. That is, we seek a function f such that the following expression is as small as possible:

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

Notes:

* A function f of this kind is sometimes called a "classifier," a "hypothesis," or a "prediction rule"
* $\text{error}_{\mathcal{D}}(f)$ is called the "generalization error" of f

How can we find such a function f? In this exercise, we will examine the following algorithm

Input: $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ where $x_j \in \{0,1\}^d$ and $y_j \in \{0,1\}$.

1. For each coordinate $1 \leq i \leq d$ compute $c_i = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{x[i] = y\}$
   That is, for each coordinate i, we check how well it "predicts" the label
2. We say that coordinate i "predicts well" the label if $c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$
   Let $P \subseteq \{1,2,\ldots,d\}$ be the set of all coordinates that predict the label well.
3. We define the following prediction rule, which makes a majority vote 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, that is, compute:
$$\text{error}_S(f) = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{f(x) \neq y\}$$

Implement this algorithm and analyze its performance. Specifically, for different values of n and d:

- Create a random sample $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ where each x_j is drawn uniformly (and independently) from $\{0,1\}^d$ and each y_j is drawn uniformly (and independently) from $\{0,1\}$.
- Run the algorithm on S and examine different values of $n, d, |P|$ and any other parameter you find relevant.
- What is the relationship between the performance of the prediction rule f on the sample (i.e., $\text{error}_S(f)$) and and its performance on the distribution from which the sample was drawn (i.e., $\text{error}_{\mathcal{D}}(f)$)? Why?

Note: Once we fix the function *f*, the function $q(x, y) = \mathbb{1}\{f(x) \neq y\}$ is a statistical query. Therefore, if we had fixed *f* in advance and then drawn the sample *S* ("the non-adaptive case"), then as we saw at the beginning of the lesson, with high probability over the draw of *S*, it would hold that $\text{error}_S(f) \approx \text{error}_{\mathcal{D}}(f)$. But that's not what's happening here, because here *f* is chosen based on *S*...