Lecture 18: Sampling and estimation

Plan/outline

Today's topic is sampling, which is one of the classic examples of a randomized algorithm. We'll develop methods to reason about sampling, obtaining error and "confidence" bounds.

Estimating the average

Suppose we have an array $A[1], A[2], \ldots, A[n]$ of numbers in the interval $[-1, 1]$, and the goal is to find the average $\frac{1}{n}\sum_i A[i]$. For convenience, let us denote the average as μ.

This can easily be done in time $O(n)$ by going over the array. But what if we just want to have a *good estimate* of μ? Suppose we are OK with an error of $\pm\epsilon$, for some parameter ϵ. Can we do this without going over the array?

The natural idea is to sample a few elements of the array and take the empirical average; this raises the questions:

- how many samples do we need to take?
- what is the *confidence* we have in our estimate?
- does the correctness depend on the entries in the array?

Today we'll formally study these questions.

Sampling basics

First off, let us formalize what we mean by sampling. The natural first suggestion is to take m of the n array elements at random. The issue with analyzing this is that the different samples are not independent -- for instance, the second element sampled is necessarily a different array element. We remedy this by sampling **with replacement**. Thus when we talk of taking k samples, we simply mean picking indices $i_1, i_2, \ldots, i_k$ independently and uniformly at random in $[1, n]$ (with replacement), and considering $A[i_1], A[i_2], \ldots, A[i_k]$. The estimate we produce is now simply $\hat{\mu} := \frac{1}{k}\sum_j A[i_j]$.

We can now analyze the procedure by defining the random variables X_j, where $1 \le j \le k$ and X_j is the value of the jth sample, i.e., $X_j = A[i_j]$.

Thus by definition, the variables $\{X_j\}$ are all independent and identically distributed (the standard abbreviation here is IID). We also have that for every j,
$\mathbb{E}[X_j] = \sum_{r=1}^{n} \Pr(X_j = A[r]) \cdot A[r] = \frac{1}{n}\sum_{r=1}^{n} A[r] = \mu$.

The first equality is by the definition of the expectation and the second one uses the fact that X_j is a uniform sample.

The empirical average $\hat{\mu} = \frac{1}{k}\sum_j X_j$ thus also has expectation equal to μ. (By the linearity of expectation.)

Our goal is to understand: how close is $\hat{\mu}$ to μ? And with what probability does it deviate?

Try 1: Markov. Markov's inequality, as we saw, gives a first cut at reasoning about how much a random variable deviates from its expected value. If we were to be able to apply it, we get that for all $t \ge 1$,
$\Pr[\hat{\mu} \ge t\mu] \le 1/t$.

But note that $\hat{\mu}$ is not a non-negative random variable! So we cannot apply Markov's inequality here! But suppose for instance that all the $A[j]$ are in $[0, 1]$ instead of $[-1, 1]$ (which we can do by shifting by 1 and dividing by 2 for example). Even so, this bound is rather weak: to see this, suppose $\mu = 1/2$. Then the bound we get on $\Pr[\hat{\mu} > 1/2 + \epsilon]$ is only roughly $1 - 2\epsilon$. The other significant problem is that the bound is *independent of the number of samples*. Intuitively, we expect that as we take more samples for averaging, we should get a better guarantee.

Variance

It turns out that a much better way to analyze this situation is using the **variance**. Formally, the variance of a random variable X is defined as $\mathbb{E}[(X - \mathbb{E}[X])^2]$. In general, this simplifies to $\mathbb{E}[X^2] - (\mathbb{E}[X])^2$; the variance is usually denoted as $\mathrm{var}(X)$.

The variance captures the notion of how much we expect a random variable to deviate from the expected value. The square root of the variance is called the *standard deviation*, for this reason.

For example, if we have an unbiased coin toss, with outcomes 0 and 1 with probability $1/2$ each, the expected value is $1/2$, but in either outcome, the value is $1/2$ away from the expectation, and the variance is $1/4$.

Let us now see how to compute the variance of the random variable $\hat{\mu}$.

Computing the variance. we have already seen that $\mathbb{E}[\hat{\mu}]$ is μ. Thus by the definition of variance, we have

$$\mathrm{var}(\hat{\mu}) := \mathbb{E}\left[\left(\frac{X_1+X_2+\cdots+X_k}{k} - \mu\right)^2\right] = \mathbb{E}\left[\left(\frac{(X_1-\mu)+(X_2-\mu)+\cdots+(X_k-\mu)}{k}\right)^2\right].$$

For convenience, let us write $Y_i = X_i - \mu$. Since the X_i are independent, so are the Y_i. The nice thing is that $\mathbb{E}[Y_i] = 0$ for all i, and by independence, we have $\mathbb{E}[Y_i Y_j] = 0$ for all $i \neq j$ (because the expectation of the product of independent variables is the product of the expectations).

Thus by expanding out the square and using the above, we have
$\mathrm{var}(\hat{\mu}) = \frac{1}{k^2}\mathbb{E}[(Y_1 + Y_2 + \cdots + Y_k)^2] = \frac{1}{k^2} \cdot \mathbb{E}\left[\sum_{i=1}^k Y_i^2\right] = \frac{1}{k^2} \cdot \sum_{i=1}^k \mathbb{E}[Y_i^2]$.

In the last step, we used the linearity of expectation. Now, since $X_i \in [-1, 1]$ by assumption the mean μ is also in this range, thus Y_i is always in the interval $[-2, 2]$. This implies that $\mathbb{E}[Y_i^2] \leq 4$. Plugging this in, because there are k terms, we get $\mathrm{var}(\hat{\mu}) \leq \frac{4}{k}$.

This implies that the standard deviation is at most $2/\sqrt{k}$. In other words, the estimate $\hat{\mu}$ deviates by "roughly" $2/\sqrt{k}$. This is nice because as the sample size k grows, the error in the estimate drops.

Chebychev's inequality

We can ask if the bound we have on the estimate holds "most of the time". Such a result can be obtained via what is known as Chebychev's inequality.

Theorem. (Chebychev's inequality) Let X be a random variable whose variance is $V = \sigma^2$. Then for any $t \geq 1$, we have
$\Pr[|X - \mathbb{E}(X)| \geq t\sigma] \leq \frac{1}{t^2}$.

Proof. The proof follows directly by applying Markov's inequality to the random variable $Z = (X - \mathbb{E}[X])^2$. In this case Z is a non-negative random variable, and its expectation is $V = \sigma^2$ by definition. Now, having $|X - \mathbb{E}| \geq t\sigma$ is equivalent to having $Z \geq t^2 V$, and thus Markov's inequality implies the theorem.

Let us now plug in $t = 2$ in our bound earlier on the variance (which gave $\sigma = 2/\sqrt{k}$). We get:
$\Pr[|\hat{\mu} - \mu| \geq \frac{4}{\sqrt{k}}] \leq \frac{1}{4}$. (**)

This is a considerably better bound than the one obtained by Markov's inequality! Note that since the proof Chebychev's inequality was only a simple application of Markov, what we *really* did was moving to the variable $(\hat{\mu} - \mu)^2$; this turns out to be a common trick: applying Markov to "higher moments" leads to much stronger bounds. The catch is that computing the higher moments is often messy -- the variance is one of the easy cases.

Samples vs accuracy. The bound (**) above tells us that if the number of samples $k = 16 \cdot 10^4$, then we get an accuracy of 0.01 in the estimate, with probability at least $3/4$.

Interestingly, the same number of samples can end up with a worse bound for the error but higher confidence. For example, setting $t = 10$ and using Chebychev's inequality, we get
$\Pr[|\hat{\mu} - \mu| \geq \frac{40}{\sqrt{k}}] \leq \frac{1}{100}$.

Thus, with $k = 16 \cdot 10^4$, we have that we get an accuracy of 0.1 with probability at least $99/100$.

This tradeoff between the error bound and confidence is quite common in sampling and in many randomized algorithms.

Is this bound tight? we can ask if doing a more sophisticated analysis can lead to better bounds. This is true for the *confidence probabilities* that we obtained. Indeed, **Chernoff bounds** typically give the right bounds for such problems.

However, for say a confidence of $/34$, the simple analysis above is fairly tight. By taking k samples, we typically *do* expect an error roughly $1/\sqrt{k}$ (this is why the quantity is called the standard deviation). You will also see this in your homework problems via experiments.