The Algorithmic Foundations of Adaptive Data Analysis August, 2017

Lecture 1

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Adaptive Data Analysis

This class will focus on a fundamental topic in statistics and machine learning: how to learn about a data distribution $\mathcal{D}$, when we only have access to a *data set* consisting of samples from $\mathcal{D}$. This is almost always how learning takes place: with access only to samples from the population we care about, and not the entire population itself[1]. Samples of course contain useful information about the distribution from which they were drawn, but the concern is that without care, we might fool ourselves into "discoveries" that are actually just sampling aberrations, and not actually reflective of the underlying distribution. This is the phenomenon that is referred to as "overfitting" in machine learning, and "false discovery" in empirical sciences. This is a well-understood problem in the non-adaptive setting, in which the "questions" that we ask of the data set are fixed up-front, before the data is gathered. The problem is greatly exacerbated (and much less well understood) in the "adaptive" setting, in which the questions that we ask of the data are themselves a function of the data. This will be the main focus of this course. But first, let us understand the basics of "non-adaptive" data analysis.

1 Non-Adaptive Data Analysis

Let's consider a simple inference task: we have a coin of unknown bias p, and a "data-set" consisting of n coin-flips (represented as a vector x, in which $x_i = 1$ corresponds to the i'th flip being heads, and $x_i = 0$ corresponds to the i'th flip being tails. We assume $\Pr[x_i = 1] = p$.) We wish to estimate p. A reasonable guess is the empirical frequency of heads: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. But clearly, we cannot take $\hat{p}$ as truth, especially when n is small. For example, if $p = 1/2$ and $n = 3$, with probability 1/4, we have $|\hat{p} - p| = 1/2$, the largest error possible. Instead, we should attempt to quantify the uncertainty inherent in our estimate $\hat{p}$ using a confidence interval: i.e. a high probability bound on the error $|\hat{p} - p|$. We can do this with some standard techniques, which will be useful throughout this course.

Proposition 1 (Markov's Inequality) *For any non-negative random variable X, and any $a > 0$:*

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

Proof Let f denote the probability density function of X. Since X is non-negative:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{v=0}^{\infty} v \cdot f(v)dv \\
&\geq \int_{v=a}^{\infty} v \cdot f(v)dv \\
&\geq a\int_{v=a}^{\infty} f(v)d(v) \\
&= a \cdot \Pr[X \geq a]
\end{aligned}$$

Rearranging gives Markov's inequality. ■

We know that $\mathbb{E}[\hat{p}] = p$, and so Markov's inequality already gives us *some* weak form of information about $\hat{p}$, but not enough to give a interesting confidence intervals in our example. That is because $\mathbb{E}(\hat{p})$ is the same as the expectation of a single coin flip. So knowing $\mathbb{E}(\hat{p})$ alone won't allow us to prove that $\hat{p}$ is close to p. Markov's inequality will nevertheless be a useful tool. For example, it allows us to derive Chebyshev's inequality:

[1] There are of course certain settings in which we have all of the data, rather than just samples. The US Census might be an example.

Theorem 2 (Chebyshev's Inequality) *For any random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[(X-\mu)^2]$ we have:*

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

Proof

$$\begin{aligned} \Pr[|X-\mu| \geq k\sigma] &= \Pr[(X-\mu)^2 \geq k^2\sigma^2] \\ &\leq \frac{\mathbb{E}[(X-\mu)^2]}{k^2\sigma^2} \\ &= \frac{1}{k^2} \end{aligned}$$

where the inequality follows from Markov's inequality. ■

This already is enough to give us non-trivial confidence intervals around $\hat{p}$. Recall the properties of variance: For any random variable X and $a > 0$, $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, and for a pair of independent random variables X_1, X_2, $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$. Therefore we have: $\mathrm{Var}(\hat{p}) = \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{Var}(X_i) \leq \frac{1}{4n}$.

Therefore, Chebyshev's inequality tells us that:

$$\Pr[|\hat{p} - p| \geq \frac{k}{2\sqrt{n}}] \leq \frac{1}{k^2}.$$

Setting $1/k^2 \leq \delta$, we find that with probability $1-\delta$, $|\hat{p} - p| \leq \sqrt{\frac{1/\delta}{4n}}$. In other words, we can derive a *confidence interval* of *width* $\sqrt{\frac{1/\delta}{4n}}$ and *coverage probability* $1-\delta$ around our estimator $\hat{p}$. We can also look at such an interval another way: suppose we desire a confidence interval of width ϵ: how much data do we need? To find out, we just solve for n: we find that if $n \geq \frac{1/\delta}{4\epsilon^2}$, then with probability $1-\delta$, $|\hat{p} - p| \leq \epsilon$.

Ok, so this is pretty good: we have a confidence interval with width that we can drive to 0 by taking larger sample sizes. But the number of samples we need scales linearly with $1/\delta$, which is a bummer if we want an extremely high confidence bound (i.e. a very tiny value of δ). It turns out we can do better: for this, we want a Chernoff bound which takes advantage of the fact that each coinflip x_i is independent of the others.

We will prove a more general bound, which applies to the summation of any family of independent random variables each taking values in $[0, 1]$.

Theorem 3 *Let $X := \sum_{i=1}^{n} X_i$ where X_i are independently distributed random variables taking values in $[0, 1]$. Then:*

$$\Pr[X \geq (p+t)n] \leq \left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right)^{n}$$

where $p = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]$ and $q = 1-p$.

A more convenient (but weaker) form of this bound is:

$$\Pr[X \geq \mathbb{E}[X] + t] \leq e^{-2t^2/n}$$

Remark By simply considering the random variable $-X$, we also get the symmetric bounds:

$$\Pr[X \leq (p-t)n] \leq \left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right)^{n}$$

$$\Pr[X \leq \mathbb{E}[X] - t] \leq e^{-2t^2/n}$$

Proof We will prove the stronger, but less convenient form of the bound. Manipulating this bound into the more convenient form is left as an exercise.

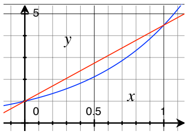
The first step is to observe that by algebraic manipulation and Markov's inequality, we have that for any $\lambda > 0$:

$$\begin{aligned}\Pr[X > m] &= \Pr[e^{\lambda X} > e^{\lambda m}] \\ &\leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda m}}\end{aligned}$$

Next, we analyze the quantity in the numerator. Because the X_i are independent of one another:

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &= \mathbb{E}[e^{\lambda \sum_i X_i}] \\ &= \mathbb{E}[\prod_i e^{\lambda X_i}] \\ &= \prod_i \mathbb{E}[e^{\lambda X_i}] \\ &\leq \prod_i (p_i e^{\lambda} + q_i)\end{aligned}$$

where we recall that $p_i = \mathbb{E}[X_i]$ and $q_i = 1 - p_i$. The inequality follows because $e^{\lambda x}$ is convex, and takes value 1 at $x = 0$ and value e^{λ} at $x = 1$. Hence, the linear interpolation between these two points (the line $y = (e^{\lambda} - 1)x + 1$) lies above the curve for any $0 \leq x \leq 1$:



Hence,

$$\mathbb{E}[e^{\lambda X_i}] \leq \mathbb{E}[(e^{\lambda} - 1)X_i + 1] = e^{\lambda} p_i + (1 - p_i) = e^{\lambda} p_i + q_i$$

Continuing, we have:

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &\leq \prod_i (p_i e^{\lambda} + q_i) \\ &\leq \left(\frac{\sum_i p_i e^{\lambda} + q_i}{n}\right)^n \\ &= (pe^{\lambda} + q)^n\end{aligned}$$

Here, the inequality follows from the "AM-GM" inequality: for non-negative a_i, $\frac{1}{n}\sum_{i=1}^n a_i \geq (\prod_{i=1}^n a_i)^{1/n}$.

Plugging this in, and setting $m = (p + t)n$, we get that for any $\lambda \geq 0$:

$$\Pr[X \geq (p + t)n] \leq \left(\frac{pe^{\lambda} + q}{e^{\lambda(p+t)}}\right)^n$$

All that remains is to pick $\lambda > 0$ to minimize the above expression. Doing so ($e^{\lambda} = \frac{-q(p+t)}{p(p+t-1)}$) yields the theorem. ■

Let's compare the strength of the bound we get from a Chernoff bound to the one we get from Chebyshev's. Applying the Chernoff bound to our estimator $\hat{p}$, we find that with probability $1 - \delta$, $|p - \hat{p}| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}$. If we want the width of this confidence interval to be ϵ, we can solve for n and see that we only need $n \geq \frac{\ln 2/\delta}{2\epsilon^2}$. Note that our dependence on the coverage probability δ is now only logarithmic — so at relatively little cost in samples, we can drive δ to be tiny!

So now we know how to estimate the bias of a coin. But this is a surprisingly useful primitive. Suppose we have a distribution $\mathcal{D}$ over arbitrary labeled datapoints $(x, y) \in \mathbb{R}^d \times \{0, 1\}$, and we have some classifier

$f : \mathbb{R}^d \to \{0, 1\}$, and we want to evaluate the error rate of our classifier: $err(f) = \Pr_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$. This can be viewed exactly as estimating the bias of a coin that has bias $p = \Pr_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$. And there is nothing special about the error rate of a classifier: the same method can be used to provide confidence intervals around the estimate of the average value of *any* predicate evaluated on individual data points that are drawn independently from some distribution. We will call these *statistical queries*:

Definition 4 *Given some data domain $\mathcal{X}$, a* statistical query *ϕ is a function $\phi : \mathcal{X} \to [0, 1]$. The* value *of the statistical query on a distribution $\mathcal{D}$ supported on $\mathcal{X}$ is: $\mathbb{E}_{x\sim\mathcal{D}}[\phi(x)]$, which we will write as $\mathbb{E}_{\mathcal{D}}[\phi]$ for brevity when $\mathcal{D}$ is understood.*

Given a dataset $S \sim \mathcal{D}^n$ consisting of n i.i.d. samples from $\mathcal{D}$, we will sometimes write $\mathbb{E}_S[\phi] = \frac{1}{n}\sum_{x\in S}\phi(x)$ to denote the empirical average of ϕ on S.

So, to summarize what we have proven in this lecture:

Theorem 5 *Fix any distribution $\mathcal{D}$, and any statistical query ϕ. Let $S \sim \mathcal{D}^n$ consist of a set of n points sampled i.i.d. from $\mathcal{D}$., with probability $1 - \delta$ over the sample:*

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

We will see going forward simple ways to extend this guarantee to *many* statistical queries when they are chosen independently of the data S, and how this fails when they might be chosen as a function of the data.

Bibliographic Information The proof of the Chernoff bound presented here follows the proof given in [DP09]. The abstraction of *statistical queries* was introduced by Kearns in [Kea98] in the context of the "statistical query" model of learning, which has been shown to capture (almost) the full power of learning in the presence of random classification noise.

References

[DP09] Devdatt Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.