The algorithmic foundations of adaptive data analysis: outline

Kelvin Li

October 14, 2019

Abstract

This is the outline of the course *The algorithmic foundations of adaptive data analysis* given by Aaron Roth at the University of Pennsylvania.

The fundamental problem of machine learning is to study the underlying (unknown) distribution $\mathcal{D}$ through samples from $\mathcal{D}$. *Overfitting* (or *false discovery*) is a problem that people try but not always avoid during learning (or experiment). In machine learning, models are trained and tested using training and testing data sets respectively to avoid overfitting. In experimental science, statistical hypotheses are stated up-front before collecting data (a.k.a., pre-registration) to avoid *p-hacking (p-fishing)*.

The way how research is done today is in an *adaptive* fashion, where the current trained model or hypothesis is selected based on some (if not all) of the results achieved in previous studies. In other words, the current model or hypothesis is a function of the data (indirectly). If the learning process is not treated properly, there is a high risk that the model will overfit or the discovery is a false finding.

This course studies some of the fundamental topics in *adaptive data analysis*, including some of the basic results from *statistical learning theory* and *differential privacy*. The entire course consists of 20 lectures. It can be separated into three sections.

talk about generalization, stability, etc

1 Preliminary

1.1 Lecture 1

Throughout this course, we use a (SQL) *query* to replace the concept of learning or experiment. In particular, we focus on *statistical queries*. Let $\mathcal{X}$ denote an input domain, X a random variable, $\mathcal{D}$ denote a probability distribution of $\mathcal{X}$ and $S \sim \mathcal{D}$ denote a data set whose records are sampled *i.i.d.* from $\mathcal{D}$.

Definition 1.1. *A **statistical query** is a function $\phi : \mathcal{X} \to [0, 1]$ such that the value of ϕ on a distribution $\mathcal{D}$ is*

$$\phi(\mathcal{D}) = E_{X \sim \mathcal{D}}(\phi(X)) \tag{1}$$

and the value of ϕ on a data set S is

$$\phi(S) = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i). \tag{2}$$

The former is also known as *population parameter* and the latter is known as *sample parameter*. mention statistical query is also low sensitivity query? define Δ-sensitivity queries?

The rest of this lecture introduces some of the basic concentration inequalities that provide bounds on how a random variable deviates from its expected value (or other values that we are interested in). The reason to bring in these inequalities is because we want to develop *mechanisms* that ensure a (adaptive or non-adaptive) query is accurate w.r.t. the data or underlying distribution with high probability.

Theorem 1.1. *(**Markov inequality**) For any non-negative random variable X and any $a > 0$,*

$$P(X \geq a) \leq \frac{E(x)}{a}.$$

This result is not so useful for our purpose except it derives the Chebyshev's inequality.

Theorem 1.2. *(**Chebyshev's inequality**) For any random variable X with mean $\mu = E[X]$ and variance $\sigma^2 = E[(X - \mu)^2]$, we have*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Here is an application of Chebyshev's inequality. Let $\sigma^2 = \frac{1}{4n}$ and $k^2 = \frac{1}{\delta}$, we get

$$P\left(|X - \mu| < \sqrt{\frac{1/\delta}{4n}}\right) \geq 1 - \delta.$$

To obtain a small error $|X - \mu| < \epsilon$, it suffices to have $\sqrt{\frac{1/\delta}{4n}} \leq \epsilon$. Solve for n, we get the *sample complexity* $n \geq \frac{1/\delta}{4\epsilon^2}$. Both ϵ and δ are arbitrarily small. The sample size increases linearly with $1/\delta$.

A better bound is the *Chernoff bound* that uses the assumption that each trial is independent. There are several variations of Chernoff bound, based on slightly different assumptions.

Theorem 1.3. *(**Chernoff bound**) Let $X := \sum_{i=1}^{n} X_i$ where X_i are i.i.d . random variables taking values in $[0,1]$ and $\mu = E[X]$. Then*

$$P(|X - \mu| \geq t) \leq 2e^{-2t^2/n}.$$

To obtain $P(|X - \mu| \leq \epsilon) \geq 1 - \delta$, it suffices if the sample complexity $n \geq \frac{\ln 2/\delta}{2\epsilon^2}$. This is an improvement over the previous result, since the sample size increases logarithmically with $1/\delta$.

1.2 Lecture 2

Define a query answering *mechanism* M that takes a query ϕ_i and a data set $S \sim \mathcal{D}$ as inputs and outputs an answer a_i. There are different types of mechanisms, such as the *empirical mechanism* that answers a query using the sample mean $a_i = \phi_i(S)$, the *noise addition mechanism* that adds some noise to the sample mean $a_i = \phi_i(S) + Lap(\cdot)$, etc.

The performance of a mechanism is measured by the worst absolute error w.r.t. the distribution $\mathcal{D}$ (or sample $S \sim \mathcal{D}$).

Definition 1.2. *A query answering mechanism M is (α, β)-**accurate** w.r.t. the underlying distribution $\mathcal{D}$ for k adaptively chosen statistical queries $K = \{q_1, \ldots, q_k\} \subseteq Q$ from a query class if for its answers $\{a_1, \ldots, a_k\}$ the following holds*

$$P\left(\sup_{q_i \in K} |q_i(\mathcal{D}) - a_i| \leq \alpha\right) \geq 1 - \beta.$$

The definition of (α, β)-**accurate** w.r.t. the data set $S \sim \mathcal{D}$ is defined analogously by replacing $q_i(\mathcal{D})$ with $q_i(S)$.

Theorem 1.4. *(**Union bound**) For every set of k events $E_1, \ldots, E_k$ in the same probability space, the probability of their union is at most the sum of their probabilities*

$$P\left(\bigcup_{j=1}^{k} E_j\right) \leq \sum_{j=1}^{k} P(E_j).$$

By Chernoff bound and Union bound, we can obtain the following bound for k nonadaptive statistical queries.

Theorem 1.5. *(**Nonadaptive queries**) Let $\mathcal{D}$ be a distribution on the set $\mathcal{X}$ and $\phi_1, \ldots, \phi_k$ be statistical queries with expectations $\mu_j = E_{\mathcal{D}}[\phi_j]$. If $S \sim D^n$ is a data set of size n drawn i.i.d. from $\mathcal{D}$, then*

$$P\left(\max_{j=1,\ldots,k} |E_S[\phi_j] - \mu_j| \leq \sqrt{\frac{\ln(2k/\delta)}{2n}}\right) \geq 1 - \delta.$$

The sample complexity is $n \geq \frac{\ln(2k/\delta)}{2\epsilon^2}$, which increases logarithmically with the number of queries k, given ϵ and δ are fixed. For example, if $n = 100000$ and $\epsilon = \delta = 0.01$, we can ask $k = 2425826$ nonadaptive queries without violating the probability bound.

1.3 Lecture 3

Adaptive queries can easily overfit. For example, if a data set $S = \{1, 2, 3, 4\}$, then construct the first query $q_1(S) = binary(\sum_{x \in S} 1/2^x)$ to output the binary representation of the sum of each $1/2^x$. This is a histogram of the data set S. Then construct the second query to overfit.

Remark 1.1. *Overfitting (or adversarial attack) appears when learning the identity of each record in a data set S, then carefully tailoring the subsequent queries to perfectly fit (or attack) S. So any method of answering statitsical queries which promises protection from overfitting must in some sense prevent the analyst (or adversary) from learning too much about too many individual data points in S.*

To get a probability bound on the worst error of adaptive queries, we cannot just use the worst error among the k queries that have been asked, because the queries are not stated up-front like in the nonadaptive case. To use Chernoff bound and union bound, the union should be taken over all possible queries that could have been asked, had the answers of the previous queries come up differently. Assuming each of the k queries that was asked is a binary function. The space of all queries can be thought as a *Garden of Forking Paths* (or decision tree). The k queries that were asked is a particular path on the decision tree. But there are in total 2^k possible combinations of queries that could have been asked. Therefore, by the same argument as in Theorem 1.5 (i.e., Chernoff bound and Union bound), the sample complexity of adaptive queries is $n \geq \frac{k + \ln 2/\delta}{2\epsilon^2}$, which scales linearly with the number of queries k. The question to ask is: Can we improve the sample complexity?

2 Description length bounds

Develop *transcript compressible* statistical estimators that allow us to control how much the analyst overfits the data, while allowing the analyst to still perform useful analyses.

Here, description length is the two part description length (according to **Occam's Razor**), where the first part corresponds to the model itself (including the number of parameters and their precision) and the second part corresponds to the compression of the data set given the model. The idea is that shorter description length estimators have lower generalization error. This can be thought from the model complexity perspective where an overly complicated model has longer description length and also higher risk of capturing random noise in the data. It can also be understood from the privacy perspective, where an overly complicated model fits every data point almost perfectly, hence reveals too much information about the data set, which then leads to a higher risk of being attacked.

2.1 Lecture 4

Let $\mathcal{A}, \mathcal{O}, \mathcal{Q}$ and S denote an analyst, a statistical estimator, a space of statistical queries and a data set respectively. For $i \in [1, k]$, the analyst $\mathcal{A}$ chooses a query $q_i \in \mathcal{Q}$ and gives it to the estimator $\mathcal{O}$. The estimator $\mathcal{O}$ generates an answer a_i to q_i based on the data set S and gives a_i to the analyst $\mathcal{A}$. Denote this **interaction** between $\mathcal{A}$ and $\mathcal{O}$ by $GT(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})$.

Definition 2.1. *The output $T = (q_1, a_1, \ldots, q_k, a_k)$ of the interaction $GT(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})$ is called the **transcript**.*

Definition 2.2. *A statistical estimator $\mathcal{O}$ is (α, β)-**accurate** w.r.t. the underlying distribution $\mathcal{D}$ for k queries from $\mathcal{Q}$ if for every algorithm $\mathcal{A}$ and for every distribution $\mathcal{D}$, the transcript T generated by $GT(\mathcal{A}, \mathcal{D}, \mathcal{O}, \mathcal{Q})$ satisfies*

$$P\left(\max_{i=1}^{k} |q_i(\mathcal{D}) - a_i| \leq \alpha\right) \geq 1 - \beta.$$

(α, β)-**accurate** w.r.t. the data set $S \sim \mathcal{D}$ can be defined similarly by replacing $q_i(\mathcal{D})$ by $q_i(S)$. Although we want to design random estimators, but it suffices to look at deterministic estimators and prove they are accurate.

Definition 2.3. *A statistical estimator $\mathcal{O}$ for k queries from $\mathcal{Q}$ is **transcript compressible** to $b(n, k)$ bits if for every analyst $\mathcal{A}$ there is a set of transcripts H_A of size $|H_A| \leq 2^{b(n,k)}$ such that for every data set S*

$$P(GT_{n,k}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q}) \in H_A) = 1.$$

Theorem 2.1. *Any $b(n, k)$-compressible estimator $\mathcal{O}$ for statistical queries will have the property that for every data analyst $\mathcal{A}$ and every distribution $\mathcal{D}$, the following is satisfied*

$$P\left(\max_{i=1}^{k} |q_i(S) - q_i(\mathcal{D})| \leq \sqrt{\frac{(b(n,k)+1)\ln 2 + \ln(k/\beta)}{2n}}\right) \geq 1 - \beta.$$

This theorem tells us that the queries do not substantially overfit the data set S, but we do not know if the answers to these queries are accurate w.r.t the distribution. For example, we can easily construct a estimator that always return a constant value. This estimator has description length 0, but is not accurate at all.

The following transfer theorem is the main result for this section. A similar transfer theorem can also be proved for differential private mechanism (later in this course).

Theorem 2.2. ***(Transcript compressibility transfer theorem)*** *For any $\beta'' > 0$, a statistical estimator $\mathcal{O}$ for statistical queries that is*

1. *$b(n, k)$-compressible and*
2. *(α', β')-accurate w.r.t. a data set $S \sim \mathcal{D}$*

is (α, β)*-accurate w.r.t. the distribution* $\mathcal{D}$*, where* $\beta = \beta' + \beta''$ *and*

$$\alpha = \alpha' + \sqrt{\frac{(b(n,k)+1)\ln 2 + \ln(k/\beta'')}{2n}}$$

Note that the theorem is also applicable to *low sensitivity queries*.

2.2 Lecture 5

Two important characteristics that compressible estimators must have are the *robustness to post-processing* and *composability*.

Robustness to post-processing guarantees that an analyst, without additional knowledge about the private data set, cannot compute a function of the output of a statistical estimator and make its second part description length longer. That is, compress/fits the data set better.

Theorem 2.3. *(**Post-processing for transcript compressibility**) Suppose* $\mathcal{O} : \mathcal{Q} \to \mathcal{R}$ *is b-transcript compressible. Let* $f : \mathcal{Q} \cup \mathcal{R} \to \mathcal{Q} \cup \mathcal{R}$ *be an arbitrary stateful algorithm. Then* $f \circ \mathcal{O}$ *is also b-transcript compressible.*

Composability ensures that if all estimators are transcript compressible, so will their combination. Hence, we can build more sophisticated estimators using a combination of several compressible estimtors. Later in the course we will see that given the privacy budget, we can work out the privacy budget for each query.

Theorem 2.4. *(**Composition for transcript compressibility**) Suppose* $\mathcal{O}_1 : \mathcal{Q} \to \mathcal{R}$ *is transcript compressible to* $b_1(n, k_1)$ *bits, and* $\mathcal{O}_2 : \mathcal{Q} \to \mathcal{R}$ *is transcript compressible to* $b_2(n, k_2)$ *bits. Then the composition* $(\mathcal{O}_1, \mathcal{O}_2)$ *is transcript compressible to* $b(n, k_1 + k_2) = b_1(n, k_1) + b_2(n, k_2)$ *bits.*

An example is the statistcal estimator that returns the experical mean of statistical queries, but to a truncated number of digits of precision. It can be proved that this estimator is

2.3 Lecture 6

More examples of accurate transcript compressible statistical estimators.

3 Stability and generalization

3.1 Lecture 7