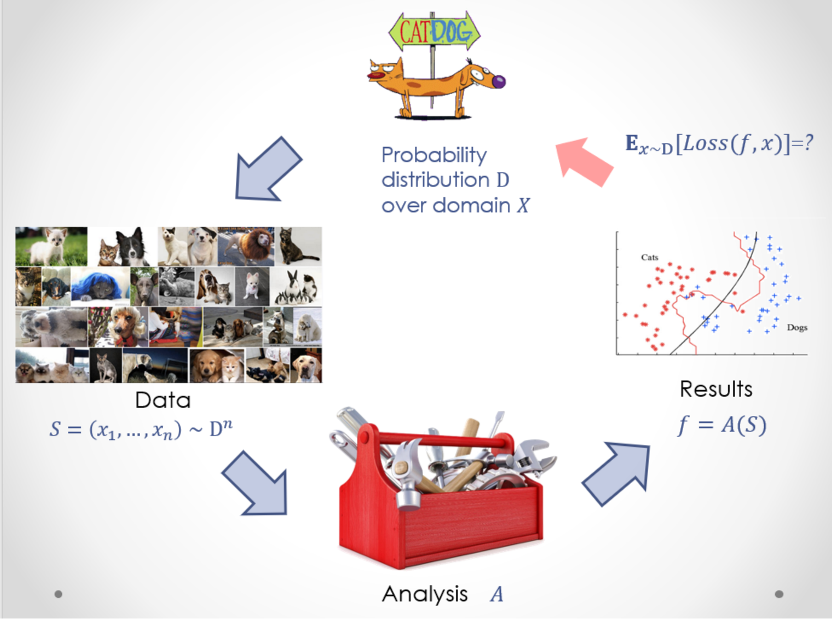
Interactive data analysis, and the dangers of adaptivity.

Shih-Ting Huang

Ruhr-Universitt Bochum

May 7, 2019

Overview : Data Analysis



Probability distribution D over domain X

$\mathbf{E}_{x\sim \mathrm{D}}[Loss(f,x)]=?$

Data

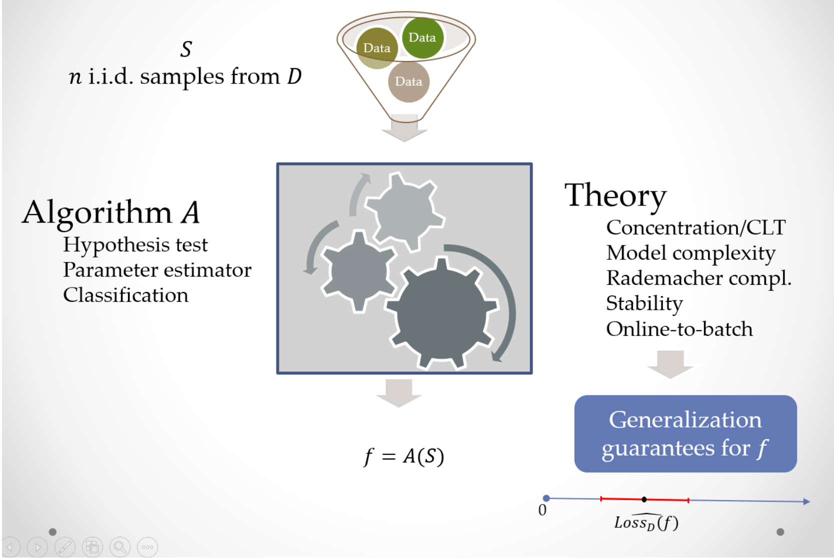
$S = (x_1, \dots, x_n) \sim \mathrm{D}^n$

Results

$f = A(S)$

Analysis A

Overview : Goal



S

n i.i.d. samples from D

Algorithm A

- Hypothesis test
- Parameter estimator
- Classification

Theory

- Concentration/CLT
- Model complexity
- Rademacher compl.
- Stability
- Online-to-batch

$f = A(S)$

Generalization guarantees for f

Review

Theorem 1

Fix any distribution D and any set of k statistical queries ϕ_1, ..., ϕ_k. Let S $\in D^n$ consist of a set of n points sampled i.i.d. from D. Then with probability $1-\delta$ over the sample:

$$max_i|E_S[\phi_i] - E_D[\phi_i]| \leq \sqrt{\frac{ln(2k/\delta)}{2n}}$$

- Fix δ and n
- Fix error
  - Example : $\sqrt{\frac{ln(2k/\delta)}{2n}} = 100 \rightarrow$ k $= \frac{\delta}{2}\ e^{2n/10000}$

Adaptive v.s. Non-adaptive Data

- Non-adaptive Data : The identities of the queries ϕ_i are fixed before the dataset S is sampled
- Adaptive Data : Steps depend on previous analyses of the same dataset S

Example

Adaptive Data

Let S denote a data set of n points sampled uniformly at random from $\{0,1\}^d$.
We can, after the dataset is drawn, defne a query ϕ such that $\phi(x) = 1$ if $x \in S$ and $\phi(x) = 0$ otherwise. By defnition, $E_S[\phi] = 1$, but $E_D[\phi] \leq \frac{n}{2^d}$.
So with probability 1, we have

$$|E_S[\phi] - E_D[\phi]| \geq 1 - \frac{n}{2^d}$$

Differential Privacy

Attack :Gain confidential information through queries

Limiting the analyst's access to the dataset to the ability to compute the empirical answers to statistical queries is not enough to prevent this kind of "attack".

We can design a single statistical query whose (exact) empirical answer allows us to just "read of" the elements in the sample S.

Example

Attack by 2 statistical queries

Suppose without loss of generality that the data domain $X = \{1, 2, 3, ...\}$. Defne the query $q(x) = 1/2^x$. Then $n \cdot E_S[q] = \sum_{x \in S} 1/2^x$, and the binary representation of this value is a histogram representing the dataset. Then, such a data analyst can overfit after asking just two queries: using the first one to read off the data set, and using the 2nd one to overfit as above.

This kind of attack would be foiled if we just truncated our evaluation of $E_S[q]$ to a small number of bits of precision

Case Study

Case

Suppose our data domain consists of labeled examples $(x, y) \in \{0,1\}^d \times \{0,1\}$: i.e. each example x consists of d binary features, and is endowed with a binary label y.

Our goal is to learn some classifier $f : \{0,1\}^d \rightarrow \{0,1\}$ that will classify these examples as well as possible, i.e. to maximize the accuracy $acc(f) = Pr_{(x,y)\sim D}[f(x) = y]$.

Note that for a classifier f, acc(f) is just a statistical query.

Algorithm

Steps

1. Begin by checking how predictive each feature on its own is with the label: For each i from 1 to d, compute $c_i = E_S[\mathbb{1}(x_i = y)]$.
2. Say that a feature is predictive if $c_i \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$. Let P be the set of predictive features.
3. Produce a classifier f that simply takes a majority vote over the predictive features:

$$f(x) = \begin{cases} 1 & \sum_{i \in P} x_i \geq \frac{|P|}{2} \\ 0 & otherwise \end{cases}$$

4. Check the performance of our classifier: Compute $acc_S(f) = E_S[\mathbb{1}(f(x) = y)]$

We might expect that our estimate of the error of our final classifier is fairly accurate:

$|acc_s(f) - acc(f)| \leq \mathcal{O}(\sqrt{\frac{log(d/\delta)}{n}})$

Theorem

Theorem 2

When D denotes the uniform distribution over $\{0,1\}^d \times \{0,1\}$, there is a constant c such that with probability $1-\delta$, if $d \geq c \cdot max(n, log(1/\delta))$:

$$|acc_S(f) - acc(f)| \geq 0.49$$

proof

Remark

If we have m independent random variables X_i, taking values in $\{0, 1\}$ such that $Pr[X_i = 1] = p$, and we write $X = \sum_{i=1}^{m} X_i$, then:

$$Pr[X < pm - t] \leq exp(\frac{-2t^2}{m})$$

Now consider a uniformly randomly selected $(x, y) \in S$. By definition of f, we have that $f(x) = y$ if and only if $\sum_{i \in P} \mathbb{1}(x_i = y) > |P|/2$. But by definition of P, we have that for each $i \in P$,

$$Pr[\mathbb{1}(x_i = y)] \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$$

Hence :

$$E[\sum_{i \in P} \mathbb{1}(x_i = y)] \geq \frac{|P|}{2} + \frac{|P|}{\sqrt{n}}$$

Proof

Therefore, we have that $f(x) = y$ unless $\sum_{i \in P} \mathbb{1}(x_i = y)$ differs from its expectation by at least $|P|/\sqrt{n}$.

Thus, for a randomly selected point $(x, y) \in S$ we have:

$Pr_{(x,y)\sim S}[f(x) \neq y] =$

$\Pr[\sum_{i \in P} \mathbb{1}(x_i = y) \leq E[\sum_{i \in P} \mathbb{1}(x_i = y)] - \frac{|P|}{\sqrt{n}}]$

$\leq exp(\frac{-2|P|^2}{n|P|}) = exp(\frac{-2|P|}{n})$

This is less than $1/100$ when $|P| \geq n \cdot ln(100)/2$

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.

Kea98Michael Kearns. Efficient noise-tolerant learning from statistical queries. Journal of the ACM (JACM), 45(6):9831006, 1998.