Lecture 23

Adaptive Data Analysis.

Method

↓

Sample

↓

Conclusion

Revise
Method / hypothesis

Logistics:

Project Presentation (May 3 & 5)

Schedule announced this week

20 mins.

No more HW.

Model for Adaptive Data Analysis

Statistical / Linear Queries

$\phi : \mathcal{X} \longrightarrow [0,1]$ "predicate"

$q_\phi(P) = \mathbb{E}_{x\sim P}[\phi(x)]$ "Population value"

P: population

$q_\phi(D) = \hat{\mathbb{E}}_{d_i \sim D}[\phi(d_i)] = \frac{1}{n}\sum_{i=1}^{n}[\phi(d_i)]$ "Empirical Average"

$D = (d_1, \dots, d_n) \in \mathcal{X}^n$

Example: mean, correlation, variance, error/risk, gradient

Extension:

① Low-Sensitive Queries

$\forall$ neighbors D & D', $|q(D) - q(D')| \leq \Delta$

② Minimization Queries

Query is given by some loss function $L : \mathcal{X}^n \times \Theta \longrightarrow [0,1]$

($\mathcal{X}^n$: Dataset; Θ: Parameter Space.)

Answer $\theta \in \Theta$

$\forall$ neighbors D & D' $\quad |L(D,\theta) - L(D',\theta)| \leq \Delta$.

$\forall \theta \in \Theta$

Interaction of Adaptive Data Analysis.

$P \xrightarrow{i.i.d.} D \longrightarrow$ Mechanism $\mathcal{M}$ $\rightleftarrows$ Analyst A

P: probability distribution "Population"

D: $d_1, \ldots, d_n$ Empirical Sample

hope: generalize to P

$q_1, a_1, q_2, a_2, q_3, \ldots, q_k, a_k$ — $I(M, A, D)$

q_j can depend on $(q_1, a_1, q_2, a_2, \ldots, q_{j-1}, a_{j-1})$.

Transcript $\Pi = ((q_1, a_1), \ldots, (q_k, a_k)) \longleftarrow \underset{\text{Interaction}}{I}(M, A, D)$

"Goal" : $\forall j$,

$$|a_j - q_j(P)| \le \text{small}.$$

$q_j(P)$: Population Value. Not just empirical averages

Avoid : queries q s.t.

$$|q(P) - q(D)| \ge \text{Large}$$

D is not representative.

DP $\Longrightarrow$ Generalization in ADA

(α, β) - sample accuracy

$$\mathop{\mathbb{P}}_{D \sim P^n, \pi} \left[\max_j | q_j(D) - a_j | \geq \alpha \right] \leq \beta$$

(α, β) - distributional accuracy

$$\mathop{\mathbb{P}}_{D \sim P^n, \pi} \left[\max_j | q_j(P) - a_j | \geq \alpha \right] \leq \beta$$

Sample Complexity / Accuracy.

Non-adaptive Queries.

- Take empirical averages : $a_j = q_j(D)$

$$\max_j |a_j - q_j(P)| \lesssim \sqrt{\frac{\log(k)}{n}}$$

Adaptive Queries

- Sample Splitting Method: $D_1, \ldots, D_k$, $a_j = q(D_j)$

$$\max_j |a_j - q_j(P)| \leq \sqrt{\frac{k}{n}}$$

- Differential Privacy. Gaussian Mechanism

$$a_j = q_j(D) + N(0, \sigma^2).$$

$$\max_j |a_j - q_j(P)| \lesssim \underbrace{\frac{k^{\frac{1}{4}}}{\sqrt{n}}}_{\frac{\sqrt{k}}{n\sigma} + \sigma} \qquad O(\alpha + \varepsilon).$$

Adding Noise Reduces Error.

ε privacy / Gen Bound

α sample Accuracy Bound.

Transfer Theorem. (ε,δ)-version [JLNRSS20]

Suppose $I(M,A,D)$ is (α,β)-sample accurate $\leftarrow$

& (ε,δ)-DP. $\leftarrow$

Then for every $c,d>0$, $I(M,A,D)$ is (α',β') $\Leftarrow$ distributionally accurate, for

$$\alpha' = \alpha + (\underset{\uparrow \atop \approx\varepsilon}{e^{\varepsilon}-1}) + \underset{\uparrow \atop \alpha}{c} + 2\underset{\uparrow \atop \varepsilon}{d}, \quad \beta' = \frac{\beta}{c} + \frac{\delta}{d}$$

$$\alpha' = O(\alpha+\varepsilon) \qquad \beta' = \left(\frac{\beta}{\alpha} + \frac{\delta}{\varepsilon}\right)$$

Simpler version $(\varepsilon,0)$-DP

(α,β) - sample accuracy

$(\varepsilon,0)$ - DP.

$\Rightarrow$ (α',β') - distributionally accurate.

$$\alpha' = \alpha + (e^{\varepsilon}-1) + \sqrt{\frac{2\ln(1/\eta)}{n}}, \quad \beta' = \beta + \eta, \text{ for all } \eta > 0.$$

$$\simeq \alpha + \varepsilon + \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$$

Sampling error bound

Proof Sketch.

$$D \xleftarrow{\text{i.i.d.}} P^n$$

Transcript: $\pi = (q_1, a_1, \dots, q_k, a_k) \longleftarrow I(M, A, D)$

$$Q_\pi = P^n \,|\, \pi$$

" posterior distribution over D conditioned on π"

- Suppose M is (α, β) - sample accurate.

$$\mathbb{P}\Big[\underbrace{\max_j |a_j - q_j(D)| \geq \alpha}_{\text{Event } E \text{ about } (D, \pi)} \Big] \leq \beta$$

Lemma (Bayesian Resampling)

$$\mathbb{P}_{\substack{D \sim P^n \\ \pi \leftarrow I(M,A,D)}} \big[(D, \pi) \in E \big] = \mathbb{P}_{\substack{D \sim P^n \\ \pi \sim I(M,A,D) \\ D' \sim Q_\pi}} \big[(D', \pi) \in E \big]$$

- Generic : Nothing to do w/ DP.
- Sample accuracy w.r.t. D

  $\Longrightarrow$ Sample accuracy w.r.t. $D' \sim Q_\pi$

- $q(D) \approx q(Q_\pi) = \mathbb{E}_{D' \sim Q_\pi} [q(D')]$

- Goal: $q(D) \approx q(P)$

  $a \approx q(P)$

Missing Step: $q(Q_\pi) \approx q(P)$

Do you want to see the proof?

Bayesian Resampling Lemma

Proof. $\mathbb{P}_{\substack{D\sim p^n \\ \Pi\leftarrow I(M,A,D) \\ D'\sim Q_\pi}}\left[(D',\pi)\in E\right]$

$$= \sum_{D=x}\sum_{\Pi=\pi}\sum_{D'=x'} \mathbb{1}\left[(x',\pi)\in E\right]\cdot \mathbb{P}[D=x]\cdot\mathbb{P}[\pi|x]\cdot\mathbb{P}[x'|\pi]$$

$$= \sum_{D=x}\sum_{\Pi=\pi}\sum_{D'=x'} \mathbb{1}\left[(x',\pi)\in E\right]\cdot \mathbb{P}\left[D=x,\ \Pi=\pi\right]\ \mathbb{P}_{\substack{D'\sim p^n \\ \Pi\leftarrow I(M,A,D)}}\left[D'=x' \,\middle|\, \Pi=\pi\right]$$

$$= \sum_{\Pi=\pi}\sum_{D'=x'} \mathbb{1}\left[(x',\pi)\in E\right]\cdot \mathbb{P}\left[\Pi=\pi\right]\cdot \mathbb{P}_{\substack{D'\sim p^n \\ \Pi\leftarrow I}}\left[D'=x' \,\middle|\, \Pi=\pi\right]$$

$$= \sum_{\Pi=\pi}\sum_{D'=x'} \mathbb{1}\left[(x',\pi)\in E\right]\cdot \mathbb{P}\left[D'=x'\right]\cdot \mathbb{P}_{\substack{D'\sim p^n \\ \Pi\leftarrow I}}\left[\Pi=\pi \,\middle|\, D'=x'\right]$$

$$= \mathbb{P}_{\substack{D\sim p^n \\ \Pi\leftarrow I(M,A,D)}}\left[(D,\Pi)\in E\right]$$

Differential Privacy $\Rightarrow$ $q(Q_\pi) \approx q(P)$

Proof (Sketch).

$$q(Q_\pi) = \mathbb{E}_{D' \sim Q_\pi} [q(D')]$$

$$= \mathbb{E}_{\substack{D' \sim Q_\pi \\ i \leftarrow unif\{1,\dots,n\}}} [q(d_i')] = \mathbb{E}_{\substack{D \sim P^n \\ i \leftarrow unif\{1..n\}}} [q(d_i)]$$

Conditioning

$$= \int_{x \in \mathcal{X}} q(x) \cdot \mathbb{P}_{\substack{D \sim P^n \\ i \sim unif\{1,\dots,n\}}} [d_i = x \mid \pi]$$

Bayes Rule

$$= \int_{x \in \mathcal{X}} q(x) \cdot \frac{\mathbb{P}_{\substack{D \sim P^n \\ i \leftarrow unif\{1,\dots,n\}}} [\pi \mid d_i = x] \cdot \mathbb{P}_{\substack{D \sim P^n \\ i \leftarrow \{1..n\}}} [d_i = x]}{\mathbb{P}_{D \sim P^n} [\pi]}$$

Approximate Cancellation?

$$\leq \int_{x \in \mathcal{X}} q(x) \frac{e^\varepsilon \mathbb{P}[\pi] \cdot \mathbb{P}[d_i = x]}{\mathbb{P}[\pi]}$$

$\pi \leftarrow I(M, A, D)$

$\uparrow$

$(\varepsilon, 0)$-DP

$$= e^\varepsilon \cdot \int_{x \in \mathcal{X}} q(x) \cdot \mathbb{P}_{\substack{D \sim P^n \\ i \sim \{1 \dots n\}}} [d_i = x] = q(P).$$

$$q(Q_\pi) \leq e^\varepsilon \; q(P)$$

$$q(Q_\pi) \geq e^{-\varepsilon} \; q(P).$$

Putting Together

- $q(Q_\pi) \in [e^{-\varepsilon} q(P), \; e^{\varepsilon} q(P)]$

  Due to $(\varepsilon, 0)$-DP.

- " $q(D) \approx q(Q_\pi)$ "

$$\mathbb{P}_{\substack{D \leftarrow P^n \\ \pi \leftarrow I}} \left[|q(Q_\pi) - q(D)| \geq C_\eta \right] = \mathbb{P}_{\substack{D \leftarrow P^n \\ \pi \leftarrow I \\ D' \leftarrow Q_\pi}} \left[|q(Q_\pi) - q(D')| \geq C_\eta \right] \leq \eta$$

$$C_\eta = \sqrt{\frac{2 \ln(1/\eta)}{n}}$$

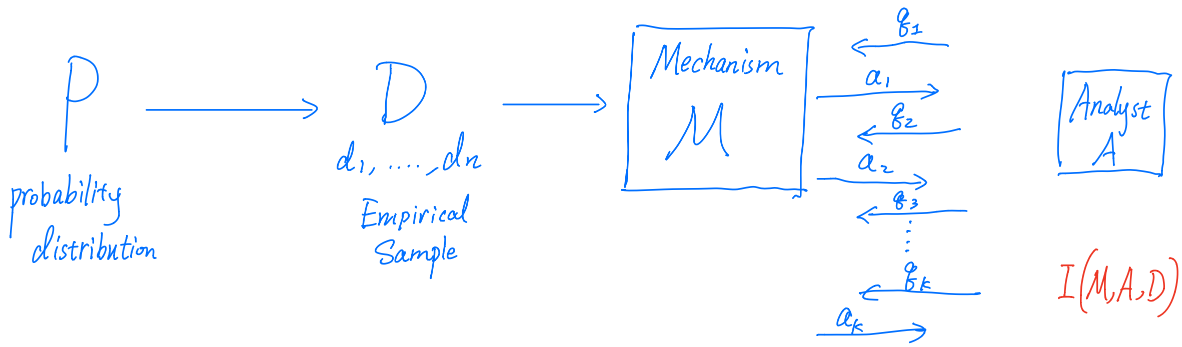
Azuma Ineq.

- Sample accuracy, w.p. $1-\beta$.

  $|a - q(D)| \leq \alpha$ for all queries. $q = q_j$.

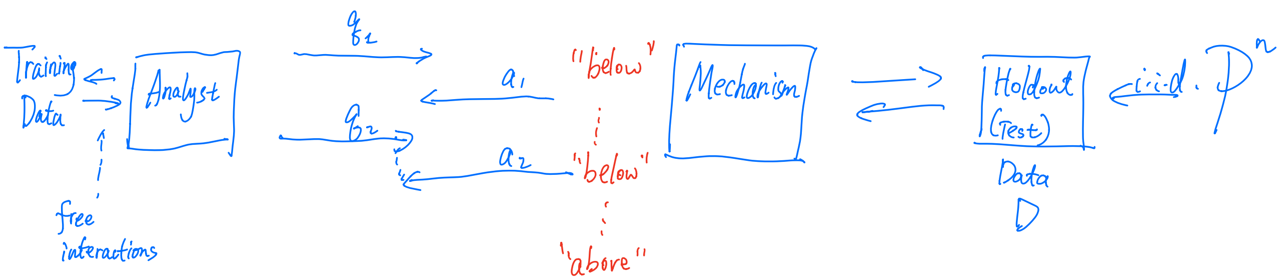
Add 3 sources of error $\square$

Mental Model of ADA.



q_j can depend on

$(q_1, a_1, q_2, a_2, \ldots, q_{j-1}, a_{j-1})$.



Example : $q_1(D) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{1}[h(x_i) = y_i] \right]$

~~error~~

accuracy

Reusable Holdout.

Sparse Vector or Thresholdout

Dwork et al. Science '2016

Algorithm 2: SparseVector$(\mathbf{s}, T, \Delta, \epsilon, q_1, q_2, \ldots)$:

Input: $q_1, q_2 \ldots$ is a stream of Δ-sensitive queries

AllDone $\leftarrow$ **FALSE**; T: threshold Stop when $q(s)$ is above T. "$q(D)$"

$\tilde{T} = T + Z_0$ where $Z_0 \sim \text{Lap}(2\Delta/\epsilon))$;

while *not* **AllDone do**

  Accept the next query q_i;

  $a_i \leftarrow q_i(\mathbf{s})$;

  $\tilde{a}_i \leftarrow a_i + Z_i$ where $Z_i \sim \text{Lap}(4\Delta/\epsilon)$;

  if $\tilde{a}_i < \tilde{T}$ **then**

    return $b_j = \perp$;

  else

    return $b_j = \top$;

    AllDone $\leftarrow$ **TRUE** ;

- Only Noise T once.
- Could release "$\perp$" many times.