Lecture 22

Adaptive Data Analysis.

Method

$\downarrow$

Sample

$\downarrow$

Conclusion

Revise
Method / hypothesis

Issues
- Overfitting
- Empirical Sciences:
  Statistical Validity

Case Study: "Wacky Boost"

Suppose $\{x, y\} \in \{0,1\}^d \times \{0,1\}$

d-dim features; Binary labels

Learn $f : \mathcal{X} \longmapsto \mathcal{Y}$

Accuracy: $Acc(f) = \mathbb{P}_{(x,y)\sim P}[f(x) = y]$

Dataset D: $Acc_D(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[f(x_i) = y_i]$

proxy?

ALG WB:

$F = \{\ \}$ "selected feature"

For $j = 1, \dots, d$

Compute $C_j = \frac{1}{n}\sum_{i=1}^{n}[x_j^{(i)} = y^{(i)}]$

If $C_j \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$, then $F \leftarrow F \cup \{j\}$

ENDFOR

Output: $\hat{f}(x) = \begin{cases} 1 & \text{if } \sum_{j\in F} x_j \geq \frac{|F|}{2} \\ 0 & \text{o/w.} \end{cases}$ "Majority Vote"

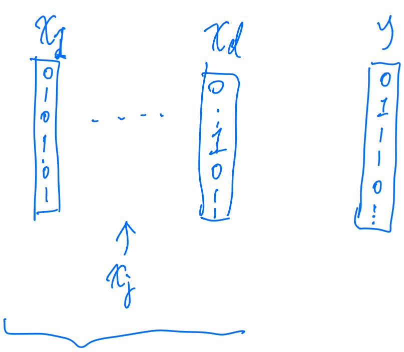
Theorem. Let P denote the uniform distribution over $\{0,1\}^d \times \{0,1\}$. There exists a constant c such that with probability $1-\delta$,

$$\left|Acc_D(\hat{f}) - Acc(\hat{f})\right| \geq 0.49$$

as long as $d \geq c \cdot \max\left(n, \log\left(\frac{1}{\delta}\right)\right)$.

- None of the features is predictive.
- $Acc(\hat{f}) \approx \frac{1}{2}$, $\quad Acc_D(\hat{f}) \longrightarrow 99\%$. gap of 49%.

Proof.

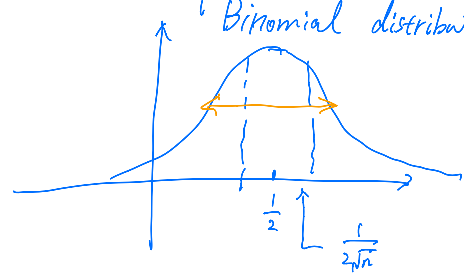


"$\frac{1}{\sqrt{n}}$ predictive features"

$\rightarrow$ 99% predictor.

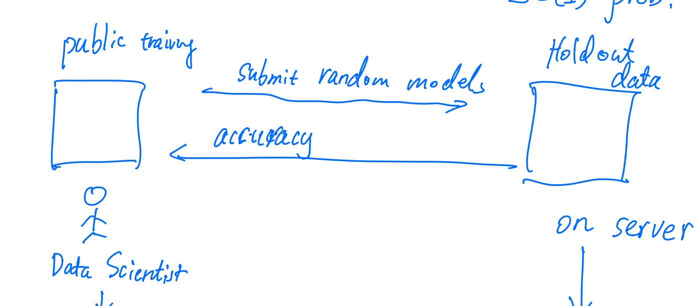
$$C_j = \frac{1}{n} \sum_i \mathbb{1}\left[x_j^{(i)} = y^{(i)}\right]$$

Binomial distribution



$C_j \geq \frac{1}{2} + \frac{1}{\sqrt{n}}$ happens with $\Omega(1)$ prob.

Background: Kaggle.



Aggregate slightly accurate random models.

Scores

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"

$\uparrow$
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"

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

Why SQ's?

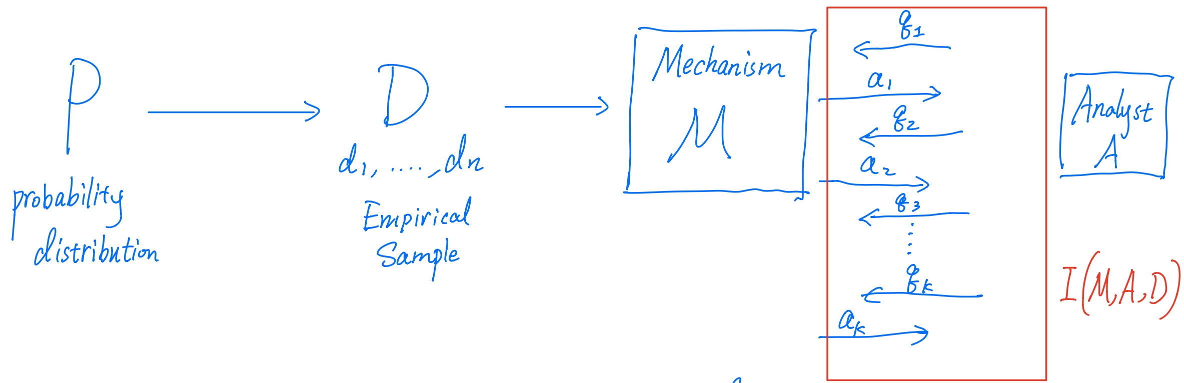
- Mean, Variances, correlations, etc.
- Error / Risk of Predictive Models

  $\mathbb{E}[\ell(f(x), y)]$
- Gradient of loss of a hypothesis

  $\mathbb{E}[\nabla \ell(f(x), y)]$
- Statistical Query Model (Kearns '94).

  "pretty much" all PAC learning problems

Interaction of Adaptive Data Analysis.



q_j can depend on

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

Transcript $\Pi = ((q_1, a_1), \ldots\ldots, (q_k, a_k)) \leftarrow I(M, A, D)$

"Goal" : $\forall j$,

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

Population Value.

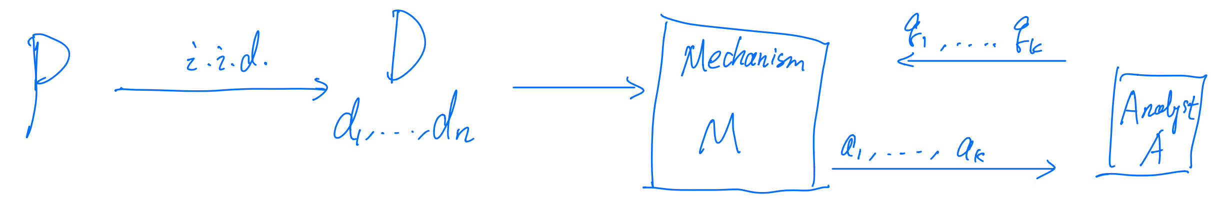
Not just empirical averages

Avoid : queries q s.t.

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

D is not representative.

Non-adaptive queries



What would M be?

Output $q(D)$

Theorem. $\max\limits_{j \in \{1,\dots,k\}} |q_j(D) - q_j(P)| \le \sqrt{\dfrac{\ln(\frac{2k}{\delta})}{2n}}$ w.p. $1-\delta$.

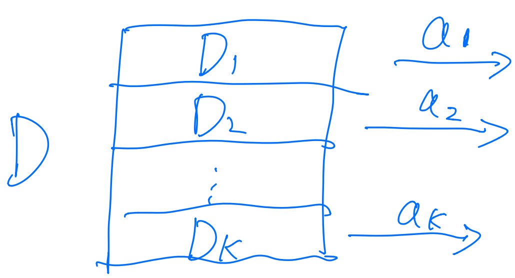
Proof Sketch. Chernoff Bound, $\forall j$

$$\Pr\left[|q_j(D) - q_j(P)| > \sqrt{\frac{\ln(2k/\delta)}{n}} \right] \le \frac{\delta}{k}.$$

Union Bound

$\Longrightarrow$ Stated Claim.

Data Splitting



Independence.

$$|D_j| = \frac{n}{K}$$

$$\max_j |q_j(D) - q_j(P)| \lesssim \sqrt{\frac{K \ln(K)}{n}}$$

DP $\Longrightarrow$ Generalization in ADA

$\boxed{(\alpha, \beta)\text{-sample accuracy}}$

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

observe

$\boxed{(\alpha, \beta)\text{-distributional accuracy}}$ $\leftarrow$ Ultimate Goal.

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

unobserve

Idea: Make sure D is representive w.r.t. q

$$q(D) \approx q(P)$$

Differential Privacy

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

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

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

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

(α, β) - sample accuracy

$(\varepsilon, 0)$ - DP.

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

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

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

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: $D_1, \dots, D_k$, $a_j = q(D_j)$

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

- Differential Privacy. Gaussian Mechanism

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

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

$O(\alpha + \varepsilon)$.

Adding Noise

Decreases Error!

ε privacy / Gen Bound

α sample Accuracy Bound.

distributional accuracy

$\beta = 5\%$.