6.S896 - Algorithmic Statistics

Lecture 1: Introduction

Uniformity Testing

Recent News Titles

① "Biden's Job Approval 42%" - Gallup
8/25/23

② "Obesity spreads through Social Networks"
- Harvard Medical School
7/25/07

③ "Better Cybersecurity w/a New Quantum Random Number Generator"
- SciTechDaily
9/4/23

④ "EHR-Safe: generating high-fidelity & privacy preserving synthetic electronic health records"
- Nature Digital Medicine
8/11/23

<u>Q</u>: How much stock to put on such claims?

<u>Challenge</u>: all these claims involve high-dimensional distributions which are, in general, hard to understand

This Class: Develop $\left\{\begin{array}{l}\text{models}\\ \text{mathematical tools}\\ \text{algorithmic tools}\\ \text{complexity tools}\end{array}\right\}$ that

are useful to make sense of high-dimensional data

Application 1: Polling

underlying assumption: n people are asked whether they approve Biden & their answers, $X_1, X_2, \dots, X_n$, are independently & identically sampled from Bernoulli(p)

← estimand

natural estimator: $\hat{p} = \frac{1}{n} \sum_i X_i$

$$\mathbb{E}[\hat{p}] = p$$

$$\mathrm{Var}[\hat{p}] = \frac{1}{n} p(1-p) \leq \frac{1}{4n}$$

$$\hookrightarrow \mathrm{std}(\hat{p}) \leq \frac{1}{2\sqrt{n}}$$

so I expect $\hat{p} \approx p \pm \frac{1}{\sqrt{n}}$ w pr $\gtrsim 95\%$

$\hookrightarrow 2 \times \mathrm{std}(\hat{p})$

a.k.a. to approximate unknown p to within $\pm\varepsilon$ it suffices to ask $\approx \frac{1}{\varepsilon^2}$ people

Question: is $\frac{1}{\varepsilon^2}$ necessary?

Le Cam's two-point method

setting: two distributions P, Q over $\mathcal{X}$ (say finite)

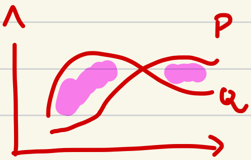
goal: find a test $T: \mathcal{X} \to \{P, Q\}$ such that error probability of T is small

$$\max\left\{ \Pr_{x\sim P}[T(x)=Q],\ \Pr_{x\sim Q}[T(x)=P] \right\}$$

Lemma: error prob. $\geq \frac{1}{2}\left(1 - TV(P,Q)\right)$

$$TV(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$$

$$= \max_{A \subseteq \mathcal{X}} |P(A) - Q(A)|$$



$TV = \frac{1}{2} \cdot$

let's use Le Cam's method to show

Lemma: $\Omega\left(\frac{1}{\varepsilon^2}\right)$ iid samples from a Bernoulli are necessary to estimate its mean to within $\pm\varepsilon$.

Stronger Lemma: $\Omega\left(\frac{1}{\varepsilon^2}\right)$ iid samples are necessary to distinguish Bernoulli$\left(\frac{1}{2}\right)$ vs Bernoulli$\left(\frac{1}{2}+\varepsilon\right)$

Proof: Suppose $P = \mathrm{Ber}(\frac{1}{2})^{\otimes n}$

$$Q = \mathrm{Ber}(\tfrac{1}{2}+\varepsilon)^{\otimes n}$$

n iid samples from $\mathrm{Ber}(\frac{1}{2}) \Leftrightarrow 1$ sample from P

— // — $\mathrm{Ber}(\frac{1}{2}+\varepsilon) \Leftrightarrow$ — // — Q

distinguisher between $\mathrm{Ber}(\frac{1}{2})$ & $\mathrm{Ber}(\frac{1}{2}+\varepsilon)$ is a function:

$$T : \{0,1\}^n \to \{P, Q\}$$

Le Cam: Probability of error $\geq \frac{1}{2}\left(1 - TV(P,Q)\right)$ ↳ ??

Claim: $TV(P,Q) \leq O(\sqrt{n}\cdot\varepsilon)$

$$\geq \frac{1}{2}\left(1 - O(\sqrt{n}\cdot\varepsilon)\right)$$

$$\geq \frac{1}{4} \text{ unless } n = \Omega\left(\frac{1}{\varepsilon^2}\right)$$

$\boxtimes$

Proof of Claim: Tensorization of Kullback-Leibler Divergence

Def: If P, Q distn's over (finite) $\mathcal{X}$

$$KL(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Fact: $TV(P,Q) \le \sqrt{\frac{1}{2} KL(P\|Q)}$

Now suppose $P = Ber(p)^{\otimes n}$ & $Q = Ber(q)^{\otimes n}$

then

$$KL(P\|Q) = \sum_{x \in \{0,1\}^n} \prod_i p(x_i) \log \frac{\prod_i p(x_i)}{\prod_i q(x_i)}$$

$$= \sum_x \prod_i p(x_i) \sum_i \log \frac{p(x_i)}{q(x_i)}$$

$$= \sum_i \underbrace{\sum_{x_{-i}} \prod_{j \neq i} p(x_j)}_{= 1} \cdot \underbrace{\sum_{x_i} p(x_i) \log \frac{p(x_i)}{q(x_i)}}_{= KL(Ber(p)\|Ber(q))}$$

$$= n \cdot KL(Ber(p)\|Ber(q))$$

now if $p = \frac{1}{2}$ $q = \frac{1}{2} + \varepsilon$

Made with Goodnotes

$$KL\left(Ber\left(\tfrac{1}{2}\right) \| Ber\left(\tfrac{1}{2}+\varepsilon\right)\right) = \frac{1}{2} \log \frac{1/2}{1/2+\varepsilon} + \frac{1}{2} \log \frac{1/2}{1/2-\varepsilon}$$

$$= \frac{1}{2} \log \frac{1}{1+2\varepsilon} + \frac{1}{2} \log \frac{1}{1-2\varepsilon}$$

$$= -\frac{1}{2} \log \left(1 - 4\varepsilon^2\right)$$

$$= -\frac{1}{2}\left(-4\varepsilon^2 + O(\varepsilon^4)\right)$$

$$= 2\varepsilon^2 + O(\varepsilon^4)$$

$$\Rightarrow KL(P \| Q) = 2n \cdot \varepsilon^2 + O(n \cdot \varepsilon^4)$$

$$= O(n \cdot \varepsilon^2)$$

$$\Rightarrow TV(P \| Q) = O(\sqrt{n} \cdot \varepsilon)$$

$$\boxtimes$$

Proof of Le Cam:

$$\max\left\{ \Pr_{x \sim P}(T(x) = Q), \Pr_{x \sim Q}(T(x) = P) \right\}$$

$$\geq \frac{1}{2}\left(\Pr_{x \sim P}(T(x) = Q) + \Pr_{x \sim Q}(T(x) = P) \right\}$$

$$= \frac{1}{2}\left(\sum_x P(x) \cdot \mathbb{1}_{\{T(x) = Q\}} + \sum_x Q(x) \cdot \mathbb{1}_{\{T(x) = P\}} \right)$$

Made with Goodnotes

$$= \frac{1}{2} \left(\sum_x P(x) \cdot \left(1 - \mathbb{1}_{\{T(x)=P\}}\right) + \sum_x Q(x) \cdot \mathbb{1}_{\{T(x)=P\}} \right)$$

$$= \frac{1}{2} \cdot \left(1 - \sum_x \left(P(x) - Q(x)\right) \cdot \mathbb{1}_{\{T(x)=P\}} \right)$$

$$\geq \frac{1}{2} \left(1 - \sum_{x : P(x) \geq Q(x)} \left(P(x) - Q(x)\right) \right)$$

$$= \frac{1}{2} \left(1 - TV(P, Q) \right)$$

$\boxtimes$

Application 2: Uniformity Testing

- Big domain $\mathcal{X}$, $|\mathcal{X}| = N$
- Distribution P over $\mathcal{X}$ (e.g. that Quantum random number generator)

Question: is P uniform?

More precise question:

distinguish using iid samples from P between

$$P = \mathcal{U}(\mathcal{X}) \quad \text{vs} \quad TV(P, \mathcal{U}(\mathcal{X})) \geq \varepsilon$$

How many samples do we need?

<u>Thm:</u> $\Theta\left(\frac{\sqrt{N}}{\varepsilon^2}\right)$ samples are necessary & sufficient
[Paninski]

trouble: if $\mathcal{X} = \{0,1\}^n$, #samples $\frac{2^{n/2}}{\varepsilon^2}$!!!

Today: proof of lower bound of $\Omega\left(\frac{\sqrt{N}}{\varepsilon^2}\right)$

Intuition: unless $\gtrsim \sqrt{N}$ samples are drawn how can we distinguish between P uniform over full or uniform over half of the domain?

<u>Proof (of lower bound)</u> use Le Cam!

w.l.o.g. $\mathcal{X} = [N] = \{1, \dots, N\}$

$P = \mathcal{U}([N])^{\otimes n}$ (P is dist'n of n samples from $\mathcal{U}([N])$)

Ok and what is a bad dist'n Q?

if Q is uniform over a specific half of the domain (e.g. all even numbers in arange it would be easy to distinguish from P...

Idea: take Q to be uniform on a random half of the domain!

Q: - sample $z_1, \ldots, z_{N/2}$ uar from $\{\pm 1\}$

importantly note that any q thus defined satisfies that $TV(q, U[N]) \geq \varepsilon$

- define dist'n q over $[N]$ s.t. $q_{2i-1} = \frac{1 + z_i \cdot \varepsilon}{2}$

$$q_{2i} = \frac{1 - z_i \cdot \varepsilon}{2}$$

& generate n samples from q

Le Cam: $\Pr[\text{error}] \geq \frac{1}{2}(1 - TV(P, Q))$

↳ how small is this as a function of N, n, ε?

TL;DR Lemma: $TV(P,Q) \simeq \sqrt{e^{O(n^2 \varepsilon^4/N)} - 1}$

so, unless $n \geq \Omega\left(\frac{\sqrt{N}}{\varepsilon^2}\right)$, $\Pr[\text{error}] \geq \frac{1}{4}$

Proof of TL;DR lemma: will sketch this for a slightly different setting where rather than n samples we draw $\tilde{n} \sim \text{Poisson}(n)$ samples

- So now: P : - sample $\tilde{n} \sim \text{Poisson}(n)$
  - draw $\tilde{n}$ samples from $\mathcal{U}([N])$

  Q : - sample $\tilde{n} \sim \text{Poisson}(n)$
  - sample $Z_1, \ldots, Z_{N/2}$ uar from $\{\pm 1\}$
  - define dist'n q over $[N]$ s.t. $q_{2i-1} = \frac{1 + Z_i \cdot \varepsilon}{N}$

$$q_{2i} = \frac{1 - Z_i \cdot \varepsilon}{N}$$

  - draw $\tilde{n}$ samples from q

- Suppose X_i is how many times $i \in [N]$ appeared in the sample

under P : $X_1, \ldots, X_N$ independent

& $X_i \sim \text{Poisson}\left(\frac{n}{N}\right)$

$$P(X_1 = n_1, \ldots, X_N = n_N) = e^{-n} \prod_{i=1}^{N} \frac{(n/N)^{n_i}}{n_i!}$$

under Q : (X_{2i-1}, X_{2i}) independent from (X_{2j-1}, X_{2j}) if $j \neq i$

& $\Pr[X_{2i-1}, X_{2i}]$ is a mixture

$$Q(X_1 = n_1, X_2 = n_2) = \frac{1}{2} e^{-\frac{n(1+\varepsilon)}{N}} \frac{\left(\frac{n(1+\varepsilon)}{N}\right)^{n_1}}{n_1!} \cdot e^{-\frac{n(1-\varepsilon)}{N}} \frac{\left(\frac{n(1-\varepsilon)}{N}\right)^{n_2}}{n_2!}$$

$$+ \frac{1}{2} e^{-\frac{n(1+\varepsilon)}{N}} \frac{\left(\frac{n(1+\varepsilon)}{N}\right)^{n_1}}{n_2!} \cdot e^{-\frac{n(1-\varepsilon)}{N}} \frac{\left(\frac{n(1-\varepsilon)}{N}\right)^{n_2}}{n_1!}$$

$$= \frac{1}{2} e^{-\frac{2n}{N}} \frac{\left(\frac{n}{N}\right)^{n_1}\left(\frac{n}{N}\right)^{n_2}}{n_1!\, n_2!} \left((1+\varepsilon)^{n_1}(1-\varepsilon)^{n_2} + (1-\varepsilon)^{n_1}(1+\varepsilon)^{n_1} \right)$$

$$Q(X_1 = n_1, \ldots, X_N = n_N) = e^{-n} \prod_{i=1}^{N} \frac{(n/N)^{n_i}}{n_i!} \prod_{i=1}^{N/2} \left((1+\varepsilon)^{n_{2i-1}} (1-\varepsilon)^{n_{2i}} + (1-\varepsilon)^{n_{2i-1}} (1+\varepsilon)^{n_{2i}} \right)$$

next we use that $TV(P,Q) \leq \sqrt{\frac{1}{2} KL(Q \| P)}$

$$\leq \sqrt{\frac{1}{2} \underbrace{\chi^2(Q \| P)}}$$

$$\chi^2(Q \| P) = \mathbb{E}_{x \sim Q}\left[\frac{Q(x)}{P(x)}\right] - 1$$

$$\hookrightarrow \left(\frac{e^{2n\varepsilon^2/N} + e^{-2n\varepsilon^2/N}}{2}\right)^{N/2} - 1$$