Lecture 13

Testing distributions:

the case of uniformity (cont)

Announcements:

- new pset out
- move instructions on project on website

  last 2 lectures are in-class presentations

A new model:

Probability distributions: get samples

Known

Discrete Domain D s.t. $|D| = n$

$\boxed{p}$

$\Downarrow$

this is all we see $\{$ iid samples

$p_i = \Pr[p \text{ outputs } i]$ ← unknown

Examples: lottery data
shopping choices
experimental outcomes
⋮

What do we need to know? is it uniform?
high entropy?
large support? (many distinct elts with > 0 probability)
monotone increasing, k-modal?
k-histogram?

Methods ?

learn distribution

χ^2- test

plug-in estimate

Max likelihood estimate

Goal : sample complexity sublinear in n

domain size

Testing Uniformity

uniform dist on domain D

goal: if $p = U_D$ then output PASS

with prob $\geq 3/4$

if $\underbrace{\text{dist}(p, U_D)}_{\text{which measure of distance?}} > \varepsilon$ then output FAIL

ℓ_1, ℓ_2, KL-divergence, Earthmover, Jensen-Shannon ....

today's focus

Distances

ℓ_1-distance: $\|p-q\|_1 = \sum_{i\in D} |p_i - q_i|$

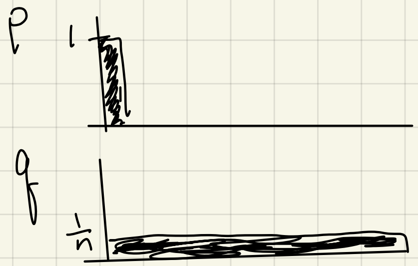
ℓ_2-distance: $\|p-q\|_2 = \sqrt{\sum_{i\in D} (p_i - q_i)^2}$

$$\|p-q\|_2 \le \|p-q\|_1 \le \sqrt{n}\cdot\|p-q\|_2$$

examples:

① $p = (1,0,0,0,\dots,0)$

$q = (\frac{1}{n},\frac{1}{n},\frac{1}{n},\dots,\frac{1}{n})$

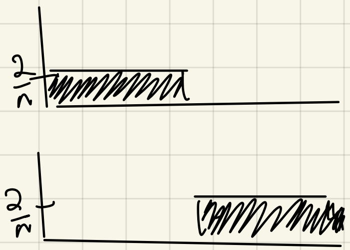


$\|p-q\|_1 = (1-\frac{1}{n}) + (n-1)(\frac{1}{n}) \approx 2$

$\|p-q\|_2 = (1-\frac{1}{n})^2 + (n-1)(\frac{1}{n^2}) \approx 1$

② $p = (\frac{2}{n},\frac{2}{n},\frac{2}{n},\dots\frac{2}{n},0,0,\dots 0)$

$q = (0,0,\dots 0,\frac{2}{n},\frac{2}{n},\dots\frac{2}{n})$



$\|p-q\|_1 = n\cdot\frac{2}{n} = 2$

$\|p-q\|_2^2 = n\cdot\left(\frac{2}{n}\right)^2 = \frac{4}{n}$ so

$\|p-q\|_2 = \frac{2}{\sqrt{n}}$ tiny even though L_1 is big

Via "Plug-in" Estimate:

- take m samples from p
- estimate $p(x)$ $\forall x$ via $\hat{p}(x) = \frac{\#\text{ times } x \text{ occurs in sample}}{m}$
- if $\sum_x |\hat{p}(x) - \frac{1}{n}| > \varepsilon$ reject

  else accept

How many samples?

can "learn" (approximately) any distribution w.r.t. L_1 distance in $\Theta(\frac{n}{\varepsilon^2})$ samples

Let's consider L_2-distance (squared):

$$\|p - \underbrace{U_{[n]}}_{\text{uniform on } 1..n}\|_2^2 = \sum_{i\in[n]} \left(p_i - \frac{1}{n}\right)^2 = \sum \left(p_i^2 - \frac{2p_i}{n} + \frac{1}{n^2}\right)$$

$$= \sum p_i^2 - \frac{2}{n} \underbrace{\sum p_i}_{=1} + \underbrace{\sum_{i=1}^{n} \frac{1}{n^2}}_{\frac{1}{n}}$$

for $p = U$: $\|p\|_2^2 = \frac{1}{n}$

for $p \neq U$: $\|p\|_2^2 > \frac{1}{n}$

$$= \underbrace{\sum p_i^2}_{\text{collision prob of } p:\ \|p\|_2^2 = \Pr_{s,t\in p}[s=t] = \sum p_i^2} - \frac{1}{n}$$

$\frac{1}{n}$ — collision prob of uniform distribution $= \|U_{[n]}\|_2^2$

we know this since we know n

$$= \|p\|_2^2 - \|U_{[n]}\|_2^2$$

Algorithm to estimate:

- take s samples of p
- let $\hat{c} \leftarrow$ estimate of $\|p\|_2^2$ from sample
- if $\hat{c} < \frac{1}{n} + \delta$ pass

  else fail

① how big is s?

② how to estimate?

③ what should δ be

How well do we need to estimate $\|p\|_2^2$?

i.e. what should δ be ?

Assumption ✭ : $|\hat{c} - \|p\|_2^2| < \Delta$

will take enough samples s.t. this holds with prob $\geq 3/4$

this is our parameter that determines whether our approximation is good.

recall: $\|p - U_{[n]}\|_2^2 = \|p\|_2^2 - \|U_{[n]}\|_2^2$

What if ✭ holds with $\Delta = \frac{\varepsilon^2}{2}$?

- if $p = U_{[n]}$ then $\hat{c} \leq \|U_{[n]}\|_2^2 + \frac{\varepsilon^2}{2} \leq \frac{1}{n} + \frac{\varepsilon^2}{2}$ so if use $\delta = \frac{\varepsilon^2}{2}$ test should PASS

- if $\|p - U_{[n]}\|_2 > \varepsilon$ then $\|p - U_{[n]}\|_2^2 > \varepsilon^2$

but $\|p\|_2^2 = \|p - U_{[n]}\|_2^2 + \frac{1}{n} > \varepsilon^2 + \frac{1}{n}$

& ✭ $\Rightarrow$ $\hat{c} > \left(\varepsilon^2 + \frac{1}{n}\right) - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{2} + \frac{1}{n}$ so if we use $\delta = \frac{\varepsilon^2}{2}$ test should FAIL

How to estimate $\|p\|_2^2$?

Naive idea: • repeat several times:

• take two samples & set $X_i \leftarrow \begin{cases} 1 & \text{if two samples equal} \\ 0 & \text{o.w.} \end{cases}$

• output average of X_i's

Better idea: "recycle" use all pairs in sample

gives $\Theta(k^2)$ samples of collision prob from k samples of p

• Take s samples from p: $x_1, \dots x_s$

• For each $1 \le i < j \le s$

$$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{o.w.} \end{cases}$$

σ_{ij}'s are not independent

$\Rightarrow$ can't use Chernoff

• Output $\hat{C} \leftarrow \dfrac{\sum_{i<j} \sigma_{ij}}{\binom{s}{2}}$

Analysis:

$$E[\hat{c}] = \frac{1}{\binom{s}{2}} \cdot E\Big[\sum_{i<j} \sigma_{ij}\Big] = \frac{1}{\binom{s}{2}} \sum_{i<j} E[\sigma_{ij}] = \frac{\binom{s}{2}}{\binom{s}{2}} E[\sigma_{ij}] = \Pr[\sigma_{ij}=1] = \|p\|_2^2$$

$$\Pr\big[\,|\hat{c} - \|p\|_2^2| > \rho\,\big] \le \frac{Var[\hat{c}]}{\rho^2}$$

Chebyshev's $\neq$

recall $Var[X] = E[(X-E[X])^2]$

$$Var[\hat{c}] = \frac{1}{\binom{s}{2}^2} \underbrace{Var\Big[\sum_{i<j} \sigma_{ij}\Big]}_{\text{need to bound}}$$

by fact: $Var[aX] = a^2 Var[X]$

difficulty: σ_{ij}'s not independent

Lemma $$Var\Big[\sum_{i<j} \sigma_{ij}\Big] \le \binom{s}{2}\|p\|_2^2 + 4\left(\binom{s}{2}\|p\|_2^2\right)^{3/2}$$

so $Var[\hat{c}]$ is $O\left(\frac{\|p\|_2^2}{s^2} + \frac{\|p\|_2^3}{s}\right)$

Lemma $\mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] \leq \binom{s}{2} \|p\|_2^2 + 4\left(\binom{s}{2}\|p\|_2^2\right)^{3/2}$

Proof

<u>def</u> $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$

so $E[\bar{\sigma}_{ij}] = 0$

$\leftarrow$ trick: rewrite variance as $E[\sum \bar{\sigma}_{ij}]$

why? $\mathrm{Var}[\sum \bar{\sigma}_{ij}] = E[(\sum \bar{\sigma}_{ij} - \underbrace{E[\sum \bar{\sigma}_{ij}]}_{=0})^2]$

$= E[(\sum \bar{\sigma}_{ij})^2]$

$= E[(\sum \sigma_{ij} - E(\sigma_{ij}))^2]$

$= E[(\sum \sigma_{i\cdot}$

$= \mathrm{Var}(\sum \sigma_{ij})$

Lemma $\mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] \le \binom{s}{2}\|p\|_2^2 + 4\cdot\left[\binom{s}{2}\|p\|_2^2\right]^{3/2}$

Proof

<u>def</u> $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$ $\Longleftarrow$ trick: rewrite variance as $E[\sum \bar{\sigma}_{ij}^2]$

why? $\mathrm{Var}[\sum \bar{\sigma}_{ij}] = E[(\sum \bar{\sigma}_{ij} - \underbrace{E[\sum \bar{\sigma}_{ij}]}_{=0})^2]$

$= E[(\sum \sigma_{ij} - E[\sigma_{ij}])^2]$

$= \mathrm{Var}[\sum \sigma_{ij}]$

So can equivalently bound $\mathrm{Var}[\sum \bar{\sigma}_{ij}]$

so $E[\bar{\sigma}_{ij}] = 0$

Facts:

- $E[\bar{\sigma}_{ij}\bar{\sigma}_{kl}] \le E[\sigma_{ij}\sigma_{kl}]$
- $\left(\sum_x p(x)^3\right)^{1/3} \le \left(\sum_x p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

(Verify @home)

Lemma $\mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] \le \binom{s}{2}\|p\|_2^2 + 4\cdot\left(\binom{s}{2}\|p\|_2^2\right)^{3/2}$

Proof

$$\mathrm{Var}\left[\sum_{i<j}\sigma_{ij}\right] = \mathrm{Var}\left[\sum_{i<j}\bar\sigma_{ij}\right] = E\left[\left(\sum_{i<j}\bar\sigma_{ij}\right)^2\right]$$

$$= E\Bigg[\underbrace{\sum_{i<j}\bar\sigma_{ij}^2}_{①} + \underbrace{\sum_{\substack{i<j\\k<l\\ i,j,k,l \text{ are all distinct}}}\bar\sigma_{ij}\bar\sigma_{kl}}_{②} + \underbrace{\sum_{\substack{i<j\\i<l\\ i,j,l \text{ distinct}}}\bar\sigma_{ij}\bar\sigma_{il}}_{③} + \underbrace{\sum_{\substack{i<j\\k<j\\ i,j,k \text{ distinct}}}\bar\sigma_{ij}\bar\sigma_{kj}}_{④}$$

$$+ \underbrace{\sum_{i<j<l}\bar\sigma_{ij}\bar\sigma_{jk}}_{⑤}\Bigg]$$

$\sigma_{ij} \leftarrow \begin{cases}1 & \text{if } x_i = x_j\\ 0 & \text{o.w.}\end{cases}$

def $\bar\sigma_{ij} = \sigma_{ij} - E[\sigma_{ij}]$

so $E[\bar\sigma_{ij}] = 0$

Facts:

- $E[\bar\sigma_{ij}\bar\sigma_{kl}] \le E[\sigma_{ij}\sigma_{kl}]$
- $\left(\sum_x p(x)^3\right)^{1/3} \le \left(\sum_x p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

Lets bound each term:

$\sigma_{ij}^2 = \sigma_{ij}$ since σ_{ij} is an indicator var

$$① \quad E\left[\sum_{i<j}\bar\sigma_{ij}^2\right] \le E\left[\sum_{i<j}\sigma_{ij}^2\right] = \binom{s}{2}\cdot \underbrace{\Pr[\sigma_{ij}=1]}_{\text{prob of collision}} = \binom{s}{2}\|p\|_2^2$$

② $$E\Big[\sum_{\substack{i<j\\k<l\\\text{all distinct}}} \bar{\sigma}_{ij}\cdot\bar{\sigma}_{kl}\Big] \le \sum_{\substack{i<j\\k<l\\\text{all distinct}}} E[\bar{\sigma}_{ij}\cdot\bar{\sigma}_{kl}] \overset{\text{independence}}{=} \sum_{\substack{i<j\\k<l\\\text{all distinct}}} E[\bar{\sigma}_{ij}]\cdot E[\bar{\sigma}_{kl}]$$

$E[\bar{\sigma}_{ij}] = 0$, $E[\bar{\sigma}_{kl}] = 0$

$$= 0$$

③ and ④ + ⑤

$$E\Big[\sum_{\substack{i<j\\i<l\\i,j,l\\\text{distinct}}} \bar{\sigma}_{ij}\bar{\sigma}_{il}\Big] \le E\Big[\sum \sigma_{ij}\sigma_{il}\Big] = \sum E[\sigma_{ij}\sigma_{il}]$$

$\sigma_{ij}\sigma_{il}$: 1 iff saw same element in ith, jth + lth sample, "3-way collision"

$$= \sum_{\substack{i,j,l\\\text{distinct}}} \Pr[X_i = X_j = X_l]$$

$$= \binom{s}{3}\sum_x p(x)^3$$

$$\le \frac{s^3}{6}\cdot\Big(\sum_x p(x)^2\Big)^{3/2}$$

$$\le \frac{\sqrt{3}}{2}\binom{s}{2}^{3/2}\left(\|p\|_2^2\right)^{3/2}$$

by facts

$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } X_i = X_j \\ 0 & \text{o.w.} \end{cases}$

def $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$

so $E[\bar{\sigma}_{ij}] = 0$

Facts:

- $E[\bar{\sigma}_{ij}\bar{\sigma}_{kl}] \le E[\sigma_{ij}\sigma_{kl}]$
- $\left(\sum_x p(x)^3\right)^{1/3} \le \left(\sum_x p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

So $\mathrm{Var}\left[\sum_{i<j} b_{ij}\right] = \mathrm{Var}\left[\sum_{i<j} \bar{b}_{ij}\right]$

$$\leq \binom{s}{2}\|p\|_2^2 + 0 + 3 \cdot \frac{\sqrt{3}}{2} \binom{s}{2}^{3/2} \left(\|p\|_2^2\right)^{3/2}$$

$$\leq \binom{s}{2} \|p\|_2^2 + 4\left(\binom{s}{2}\|p\|_2^2\right)^{3/2} \quad \blacksquare$$

We have:

$$\mathrm{Var}[\hat{C}] = O\left(\frac{\|p\|_2^2}{s^2} + \frac{\|p\|_2^3}{s}\right)$$

$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{o.w.} \end{cases}$

$\hat{C} \leftarrow \frac{\sum_{i<j} \sigma_{ij}}{\binom{s}{2}}$

where $s =$ # samples

Put into Chebyshev with $\rho = \frac{\varepsilon^2}{2}$:

$$\Pr\left[\left|\hat{C} - \|p\|_2^2\right| > \frac{\varepsilon^2}{2}\right] \le \frac{\mathrm{Var}[\hat{C}]}{\varepsilon^4} \cdot 4$$

$$\le \underbrace{\frac{\text{const}}{\varepsilon^4 \cdot s^2}}_{\text{want this to be} \le 1} \cdot \underbrace{\|p\|_2^2}_{\le 1} + \text{Const} \cdot \underbrace{\frac{1}{\varepsilon^4} \cdot \frac{1}{s}}_{\text{want to be} \ll 1} \cdot \underbrace{\|p\|_2^3}_{\le 1}$$

need $s = \Omega\left(\frac{1}{\varepsilon^2}\right)$

need $s = \Omega\left(\frac{1}{\varepsilon^4}\right)$

samples s to be $O\left(\frac{1}{\varepsilon^4}\right)$

<u>note</u> can get better bounds $s = O(1/\varepsilon^2)$

s is independent of n !!!!

recall:

$\|p-u_{[n]}\|_2^2 = \|p\|_2^2 - \|u_{[n]}\|_2^2$

$L_2 \le L_1 \le \sqrt{n} \cdot L_2$

How to estimate $\|p-u\|_1$?

1) $\|p-u\|_1 = 0 \iff \|p-u\|_2 = 0 \iff \|p\|_2^2 = \frac{1}{n}$

2) if $\|p-u\|_1 > \varepsilon \Rightarrow \|p-u\|_2 > \frac{\varepsilon}{\sqrt{n}}$

$\Rightarrow \|p-u\|_2^2 > \frac{\varepsilon^2}{n}$

$\Rightarrow \|p\|_2^2 > \frac{\varepsilon^2}{n} + \frac{1}{n}$

So either additive estimate of $\|p\|_2^2$ to within $\frac{\varepsilon^2}{2n}$

or mult " " " to within $(1 \pm \frac{\varepsilon^2}{3})$

suffices

turns out that picking # samples $s \gg \frac{\sqrt{n}}{\varepsilon^4}$ suffices

(+ actually $s = \frac{\sqrt{n}}{\varepsilon^2}$ sufficient)

$s = \sqrt{n}$ suffices

Generalizations: Given another distribution q,

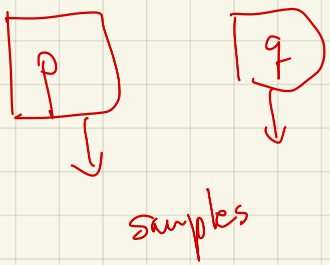
is $p = q$ or is p "far" from q?

1. "Identity Testing"

q is Known to algorithm, no samples of q needed } focus on sample complexity but runtime can be made similar

in "DNA"

2. "Closeness Testing"



q is given via samples

What is complexity in terms of n, ε?

Will see more on these ...

(e.g. pset, lecture ...)

A difficulty in analyzing distribution testers:

typical algorithm:

take m samples $\{s_1 \ldots s_m\} = S$

let $X_i :=$ # times i occured in sample

$\vdots$

$Poi(\lambda)$: $Pr[x=k] = \frac{e^{-\lambda} \lambda^k}{k!}$

$E[x] = Var[x] = \lambda$

Can we make the X_i's independent? Poissonization

<u>new algorithm 1</u>

$\hat{m} \leftarrow Poi(m)$

Take $\hat{m}$ samples to get $\hat{S}$

let $X_i :=$ # times i occured in $\hat{S}$

$\vdots$

<u>new algorithm 2</u>

For each $i \in [n]$

$X_i \leftarrow Poi(m \cdot p_i)$

add X_i coples of i to sample

Randomly permute the sample

$\vdots$