Lecture 12

Testing distributions:

the case of uniformity

A new model:

Probability distributions: get samples

$$\boxed{p}$$
$$\Downarrow$$

this is all we see $\Big\{$ iid samples

Discrete Domain D s.t. $|D| = n$ $\leftarrow$ known

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

Examples: lottery data

shopping choices

experimental outcomes

$\vdots$

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 $p \equiv U_D$ 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: $\quad \|p-q\|_1 = \sum_{i \in D} |p_i - q_i|$

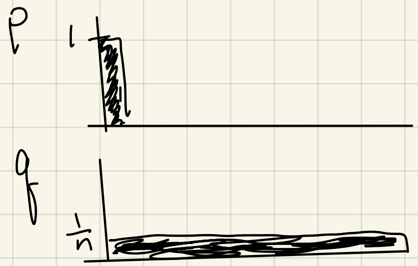
ℓ_2-distance: $\quad \|p-q\|_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}$

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

examples:

(1) $p = (1, 0, 0, 0, \ldots, 0)$

$q = \left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

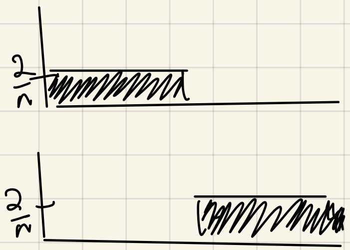


$\|p-q\|_1 = \left(1 - \frac{1}{n}\right) + (n-1)\left(\frac{1}{n}\right) \approx 2$

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

(2) $p = \left(\frac{2}{n}, \frac{2}{n}, \frac{2}{n}, \ldots \frac{2}{n}, 0, 0, \ldots 0\right)$

$q = \left(0, 0, \ldots 0, \frac{2}{n}, \frac{2}{n}, \ldots \frac{2}{n}\right)$



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

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

Naive Analysis: (better analyses exist)

so if $p=U$ likely to pass

pick m s.t. $\forall x$ $|\hat{p}(x) - p(x)| < \frac{\varepsilon}{n}$ $\Rightarrow$ $\|\hat{p} - p\|_1 < \varepsilon$

by $\Delta \neq$, if $\|p - \hat{p}\|_1 < \varepsilon$ & $\|\hat{p} - U\|_1 < \varepsilon$ then $\|p - U\|_1 < 2\varepsilon$

if $|p-U|>2\varepsilon$ likely to fail

how big should m be?

do you need to see each x at least once? $\log n$ times?

$\Omega(n)$? $\Omega(n/\varepsilon)$? $\Omega(n/\varepsilon^2)$? $\Omega(n\log/\varepsilon^2)$?

Claim $E[\|\hat{p}-p\|_1] \leq \sqrt{\frac{n}{m}}$

So pick $m = \Theta(\frac{n}{\varepsilon^2})$ gives $E[\|\hat{p}-p\|_1] \leq \frac{\varepsilon}{c}$

+ by Markov's ≠ with prob $1-\frac{1}{c}$

$\|\hat{p}-p\|_1 < \varepsilon$

Pf of claim

$$E[\|\hat{p}-p\|_1] = \sum_x E[|\hat{p}(x)-p(x)|]$$

$$E[\hat{p}(x)] = \frac{1}{m} \cdot E\left[\sum_{i=1}^{m} \mathbb{1}_{i\text{th sample is } x}\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m} E[\mathbb{1}_{i\text{th sample is } x}]$$

$$= \frac{1}{m} \cdot m \cdot p(x) = p(x)$$

$$\leq \sum_x \sqrt{E[(\hat{p}(x)-p(x))^2]}$$ Jensen's ≠

$$= \sum_x \sqrt{\mathrm{Var}(\hat{p}(x))}$$

fact $\mathrm{Var}(aX) = a^2 \cdot X$

$$= \sum_x \sqrt{\frac{p(x)}{m}}$$

$$\mathrm{Var}(\hat{p}(x)) = \frac{1}{m^2} \cdot \mathrm{Var}[\#\text{ times } x \text{ occurs in sample}]$$

$$= \frac{1}{m^2}\mathrm{Var}\left[\sum_{i=1}^{m} \mathbb{1}_{i\text{th sample is } x}\right]$$ indep for $i \neq j$

$$= \frac{1}{m^2}\sum \mathrm{Var}[\underbrace{\mathbb{1}_{i\text{th sample is } x}}_{p(x)(1-p(x))}]$$

$$= \frac{1}{m^2} \cdot m \cdot p(x)\underbrace{(1-p(x))}_{<1} \leq \frac{p(x)}{m}$$

$$\leq \frac{\sqrt{n}}{\sqrt{m}}$$ ■

$\max_p \sum \sqrt{p(x)}$ is $\sqrt{n}$

So 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):

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

$U_{[n]}$: uniform on $1 \dots n$

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

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

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

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

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

we know this since we know n

$\sum p_i^2$: collision prob of p: $\|p\|_2^2 = \Pr_{s,t \in p}[s = t] = \sum p_i^2$

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

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

& ★ $\Rightarrow$ $\hat{C} > \left(\varepsilon^2 + \frac{1}{n}\right) - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{n} + \frac{1}{n}$ 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 $X_i \leftarrow \begin{cases} 1 & \text{if two samples equal} \\ 0 & \text{o.w.} \end{cases}$
  - increment i
- output average of X_i's

$O(k)$ samples of collisions from k samples of p

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

$O(k)$ samples of collisions from k samples of p

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: $x_1, \ldots x_s$
- 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: $E[\hat{c}] = \frac{1}{\binom{s}{2}} \cdot E\left[\sum_{i<j} \sigma_{ij}\right] = \frac{1}{\binom{s}{2}} \sum_{i<j} E[\sigma_{ij}] = \frac{\binom{s}{2}}{\binom{s}{2}} E[\sigma_{ij}] = \Pr[\sigma_{ij}=1]$

$= \|p\|_2^2$

$$\Pr\left[\left|\hat{c} - \|p\|_2^2\right| > \rho\right] \le \frac{Var[\hat{c}]}{\rho^2}$$ Chebyshev's $\neq$

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

$$Var[\hat{c}] = \frac{1}{\binom{s}{2}^2} Var\left[\sum_{i<j} \sigma_{ij}\right]$$ by fact from before

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

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

Proof of lemma in next lecture