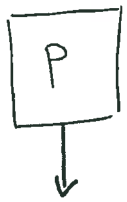
Testing Distributions:

Uniformity

Turning to a new model: prob dists

Probability distributions - get samples of distribution



outputs iid samples

↖ this is all we can learn from

Domain D, $|D| = n$ ← known

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

Examples:

Lottery data

Shopping choices

experimental outcomes

⋮

What do we want to know?

is it uniform? eg. lottery

is it high entropy?

large support? (many distinct elements have >0 probability

is p monotone increasing, k-modal, monotone hazard rate...?

how can we do it?

χ^2 test

plug in estimate

learn distribution, Maximum likelihood estimates

Goal: sample complexity SUBLINEAR in n

Testing Uniformity

The goal:

Uniform dist on D

- if $P = U_D$ then tester outputs PASS ← with prob ≥ 3/4
- if $\underbrace{dist(P, U_D)}_{\text{which measure of distance?}} > \varepsilon$ then tester outputs FAIL

ℓ_1, ℓ_2, KL-divergence, Earth mover, Jensen-Shannon

today's focus

good direction for projects!

Distances

$\|p-q\|_1 = 2 \cdot \underbrace{\text{TVD dist}(p,q)}$

$= \max_{S \subseteq D} \{ \sum_{i \in S} |p_i - q_i| \}$

"Total variation distance"

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

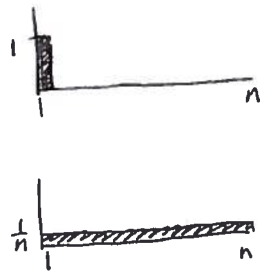
ℓ_2-distance : $\|p-q\|_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}$

Fact: $\|p-q\|_2 \le \|p-q\|_1 \le n^{1/2} \|p-q\|_2$

examples:

① $p = (1, 0, 0, \ldots 0)$

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



ℓ_1 distance:

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

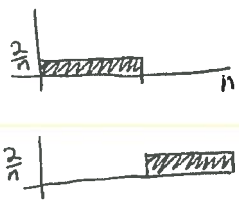
ℓ_2-distance:

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

②

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

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



as far as possible

ℓ_1 distance:

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

ℓ_2-distance: $\|p-q\|_2^2 = n \cdot \left(\frac{2}{n}\right)^2 = \frac{4}{n}$

$\|p-q\|_2 = \frac{2}{\sqrt{n}}$

pretty close

$\Rightarrow$ so ℓ_2-distance can be weird

(but above fact limits how weird)

"Plug-in Estimate"

Algorithm:

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

  else accept

Analysis: (better analyses exist, e.g. next page)

Correct behavior

pick m s.t. whp $\forall x$ $|p(x) - \hat{p}(x)| < \varepsilon/n$ (assume this holds in following)

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

so if $p = U$, test will accept

to show if $\|p - U\|_1 > 2\varepsilon$, likely to reject,

will show contrapositive:

If test accepts,

good approx test acepts

by $\Delta \neq$: if $\|p - \hat{p}\|_1 < \varepsilon$ + $\|\hat{p} - U\|_1 < \varepsilon$

then $\|p - U\|_1 < 2\varepsilon$ ∎

but how big should m be?

$\Omega(n)$? Coupon collector? . . . ☹

Above algorithm gives good approx in $O(n/\varepsilon^2)$ samples:

Thm if $m = \Theta\left(\frac{n}{\varepsilon^2}\right)$, $\Pr\left[\|\hat{p}-p\|_1 \le \varepsilon\right] \ge 3/4$

Better analysis: (not done in lecture)

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

Pf

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

$$\le \sum_x \sqrt{E[(\hat{p}(x)-p(x))^2]}$$

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

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

$$\le \frac{1}{\sqrt{m}} \cdot \sqrt{n}$$

def of $\hat{p}$ + lin of expectation

note: $E[\hat{p}(x)] = \frac{1}{m} E[\sum 1_{i^{th}\text{ sample is } x}]$

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

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

Jensen's ≠

$Var(\hat{p}(x)) = \frac{1}{m^2} m\, p(x)(1-p(x))$

$\le \frac{p(x)}{m}$

since $\max_{p \in \text{prob dist over domain of size } n} \sum \sqrt{p(x)}$ is $\sqrt{n}$

■

So picking $m = \Omega\left(\frac{n}{\varepsilon^2}\right)$ gives

$$E[\|\hat{p}-p\|_1] \le \frac{\varepsilon}{2}$$

by Markov's ≠: with prob $1-\frac{1}{2}$, $\|\hat{p}-p\|_1 \le \varepsilon$

Note, this says can "learn" (approximate) any dist w.r.t. L_1 distance in $\Theta(n/\varepsilon^2)$ samples

Let's consider an "easier" problem — L_2-distance

⑥ pd

L_2 - Distance (squared):

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

$$= \sum p_i^2 - \frac{2}{n} \underbrace{\sum p_i}_{=1} + \underbrace{\sum_i \left(\frac{1}{n}\right)^2}_{=\frac{1}{n}}$$

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

Collision probability of p:

$$\|p\|_2^2 \equiv \Pr_{s,t \in p}[s = t] = \sum p_i^2$$

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

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

$$= \underbrace{\|p\|_2^2}_{\text{we can estimate this}} - \underbrace{\|U_{(n)}\|_2^2}_{\text{we know this since we know } n}$$

Algorithm

1. take s samples from p ← ① how many samples?
2. let $\hat{c} \leftarrow$ estimate of $\|p\|_2^2$ from sample ← ② how?
3. if $\hat{c} < \frac{1}{n} + \delta$ pass ← ③ what should δ be?

   else fail

<u>Thm</u> if $s = \Theta(\sqrt{n}/\varepsilon^4)$, $\Pr\left[\left|\hat{c} - \|p\|_2^2\right| > \frac{\varepsilon^2}{2}\right] \leq 1/4$

+ Algorithm is property tester for uniformity under L_2-dist.

Question ⑦

P.D

② How to estimate $\|p\|_2^2$?

Naive idea: (pair off samples)

take two new samples:

$$\sigma_i \leftarrow \begin{cases} 1 & \text{if samples are equal} \\ 0 & \text{o.w} \end{cases}$$

$\}$ σ_i's are independent

"gives $\theta(k)$ samples of collision probability from k samples of p"

Better idea: recycle - use all pairs in sample

"gives $\theta(k^2)$ samples of collision probability from k samples of p"

$\}$ σ_{ij}'s are not independent

$$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if sample } i \text{ \& } j \text{ are equal} \\ 0 & \text{o.w.} \end{cases}$$

Estimate by recycling:

- Take s samples from p: $X_1 \cdots X_s$
- for each $1 \le i < j \le s$

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

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

$\}$ σ_{ij}'s not independent so can't use Chernoff

Analysis: $E[\hat{c}] = \frac{1}{\binom{s}{2}} \cdot \binom{s}{2} \cdot E[\sigma_{ij}]$

$= \|p\|_2^2$

Question ③

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

(& how to pick δ?)

Pick $\delta = \Delta = \varepsilon^2/2$

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

will take enough samples so that this holds with prob $\geq 3/4$

this is our parameter that determines whether our approximation is good. Spoiler: will set $\Delta = \frac{\varepsilon^2}{2}$

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

Correct behavior!

- if $p = U_{[n]}$ then $\hat{c} \overset{\text{by ✡}}{\leq} \|U_{[n]}\|_2^2 + \Delta = \frac{1}{n} + \frac{\varepsilon^2}{2}$

  so test will PASS

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

  but $\|p\|_2^2 = \|p - U_{[n]}\|_2^2 + \frac{1}{n}$ ← see p. 6

  $> \varepsilon^2 + \frac{1}{n}$

  & $\hat{c} > \|p\|_2^2 - \Delta$ ← ✡

  $\geq \varepsilon^2 + \frac{1}{n} - \Delta = \varepsilon^2 + \frac{1}{n} - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{2} + \frac{1}{n}$

  so test will FAIL

Remaining Question: (Question 1)

How many samples do we need to estimate $\hat{c}$ to within Δ?

Question ①:

Analysis: $E[\sigma_{ij}] = Pr[\sigma_{ij} = 1] = \|p\|_2^2$

$E[\hat{C}] = \frac{1}{\binom{s}{2}} \cdot \binom{s}{2} \cdot E[\sigma_{ij}] = \|p\|_2^2$

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

Fact $Var[aX] = a^2 Var[X]$

So $Var[\hat{C}] = Var\left[\frac{1}{\binom{s}{2}} \sum_{i<j} \sigma_{ij}\right]$

$= \frac{1}{\binom{s}{2}^2} Var\left[\sum_{i<j} \sigma_{ij}\right]$

Lemma $Var\left[\sum_{i<j} \sigma_{ij}\right] \le O(s^3 \cdot \|p\|_2^3)$

Corr $Var[\hat{C}] \le O(\|p\|_2^3 / s)$

Proof of lemma def $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$ (nice trick)

note $E[\bar{\sigma}_{ij}] = 0$ & $\bar{\sigma}_{ij} < \sigma_{ij}$ since $E[\sigma_{ij}] > 0$

also $E[\bar{\sigma}_{ij}\,\bar{\sigma}_{kl}] \le E[\sigma_{ij} \cdot \sigma_{kl}]$

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

$$= E\Big[\Big(\sum_{i<j} \bar{\sigma}_{ij}\Big)^2\Big]$$

$$= E\Big[\underbrace{\sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=2}} \bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{1} + \underbrace{\sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=3}} \bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{2} + \underbrace{\sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=4}} \bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{3}\Big]$$

Bounding 1:

$$E\Big[\sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=2}} \bar{\sigma}_{ij}\bar{\sigma}_{kl}\Big] = E\Big[\sum_{i<j} \bar{\sigma}_{ij}^2\Big] \le E\Big[\sum_{i<j} \sigma_{ij}^2\Big] = \binom{s}{2}\|p\|_2^2$$

note $\sigma_{ij}^2 = \sigma_{ij}$ since indicator variable

Bounding 3:

independence + linearity of expectations

$$E\Big[\sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=4}} \bar{\sigma}_{ij}\bar{\sigma}_{kl}\Big] = \sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=4}} E[\bar{\sigma}_{ij}]\,E[\bar{\sigma}_{kl}] = 0$$

note: moving to $\bar{\sigma}_{ij}$ means all of these terms drop out!

Bounding 2:

$$E\Big[\sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=3}} \bar{\sigma}_{ij}\bar{\sigma}_{kl}\Big] \le E\Big[\sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=3}} \sigma_{ij}\sigma_{kl}\Big] = \sum_{\substack{i<j \\ k<l \\ |\{i,j,k,l\}|=3}} E[\sigma_{ij}\sigma_{kl}] = \sum_{|\{a,b,c\}|=3} \Pr[X_a = X_b = X_c]$$

e.g. $i=k, j\ne k, i\ne l, j\ne l \ldots$

\# ways to pick $i<j$, $k<l$ s.t. $|\{i,j,k,l\}|=3$: (pick 3 indices, pick one to be repeated twice leaves 2 options)

$$\le 6\cdot\binom{s}{3}\cdot\sum_x p(x)^3$$

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

note $\left(\sum p(x)^3\right)^{1/3} \le \left(\sum p(x)^2\right)^{1/2}$

$$\le O\left(s^3\cdot\left(\|p\|_2^2\right)^{3/2}\right) = O\left(s^3\|p\|_2^3\right)$$

So: $\mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] = O\left(\binom{s}{2}\|p\|_2^2 + 0 + s^3 \cdot \|p\|_2^3\right)$

$$= O\left(s^3 \|p\|_2^3\right)$$

So how many samples?

for property tester wrt L_2-distance

need to estimate $\|p\|_2^2$ to within (additive) $\Delta = \frac{\varepsilon^2}{2}$

$$\Pr\left[\,\left|\hat{c} - \|p\|_2^2\right| > \frac{\varepsilon^2}{2}\,\right] \le \frac{\mathrm{Var}[\hat{c}]}{\varepsilon^4/4} = \frac{1}{\binom{s}{2}^2} \cdot \frac{s^3 \cdot \|p\|_2^3}{\varepsilon^4} \cdot \mathrm{Const}$$

$$\le O\Big(\underbrace{\frac{1}{s} \cdot \frac{1}{\varepsilon^4}}_{\text{want this to be small}} \cdot \underbrace{\|p\|_2^3}_{\le 1}\Big)$$

Pick $s = \Omega\left(\frac{1}{\varepsilon^4}\right)$ (can do better)

what about L_1-distance?

Now: Distinguish $\|p-U\|_1 \geq \varepsilon$ from $p=U$

via L_2 - testing

Thm there is distribution testing algorithm which tests uniformity (in L_1) & outputs correct answer w/ prob $\geq 1-\delta$ using $O(\frac{1}{\varepsilon^4}\sqrt{n}\log(1/\delta))$ samples

Why?

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

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

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

need to get multiplicative estimate of $\|p\|_2^2$ to w/in $(1 \pm \underbrace{\frac{\varepsilon^2}{4}}_{\gamma})$ do this

or additive " " " " $\frac{\varepsilon^2}{2n}$

$$\Pr\left[\,|\hat{c}-\|p\|_2^2| > \gamma\cdot\|p\|_2^2\right] \leq \frac{\mathrm{Var}[\hat{c}]}{\gamma^2\|p\|_2^4} \leq \frac{\text{Const}\cdot\|p\|_2^3/s}{\gamma^2\|p\|_2^4} = \frac{\text{Const}}{s\gamma^2\cdot\underbrace{\|p\|_2}_{\geq \frac{1}{\sqrt{n}}}}$$

$$\leq O\left(\frac{\sqrt{n}}{s\gamma^2}\right)$$

so pick $s = \Omega\left(\frac{\sqrt{n}}{\varepsilon^4}\right)$ ∎