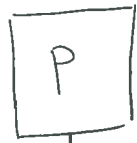


Turning to a new model:

Probability distributions - get samples of distribution



Outputs iid samples

Domain D , $|D|=n$ ← known
 $p_i = \Pr[p \text{ outputs } i]$ ← unknown

← this is all we can learn from

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 it 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 $P \equiv U_D$ then tester outputs PASS \leftarrow with prob $\geq 3/4$

• if $\underbrace{\text{dist}(P, U_D)} > \epsilon$ then tester outputs FAIL

which measure of distance?

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

$\uparrow \uparrow$
today's focus

Distances

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

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

$$\|p-q\|_2 \leq \|p-q\|_1 \leq n^{1/2} \|p-q\|_2$$

examples:

① $p = (1, 0, 0, \dots, 0)$



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



l_1 distance:

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

l_2 -distance:

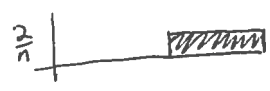
$$\|p-q\|_2^2 = \left(1 - \frac{1}{n}\right)^2 + (n-1) \left(\frac{1}{n}\right)^2 \approx 1$$

②

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



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



l_1 distance:

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

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

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

"Plug-in" Estimate:

Algorithm:

- take m samples from p
- estimate $p(x) \forall x$ via

$$\hat{p}(x) = \frac{\# \text{ times } x \text{ occurs in sample}}{m}$$

- if $\sum_x |\hat{p}(x) - \frac{1}{n}| > \epsilon$ reject
- else accept.

Analysis: (better analyses exist)!

so, if $p = U_n$ then p passes

pick m st. $\forall x, |\hat{p}(x) - p(x)| < \frac{\epsilon}{n} \Rightarrow \|\hat{p} - p\|_1 < \epsilon$

by ΔF , if $\|p - p\|_1 < \epsilon + \|\hat{p} - U_n\|_1 < \epsilon$
then $\|p - U_n\|_1 < 2\epsilon$.

so, if $\|p - U_n\|_1 > 2\epsilon$ this test is likely to Fail

how many samples? $\Omega(\frac{n}{\epsilon})$ maybe even worse...

for each x , need to see it at least once in order to give non zero estimate.

$\Theta(n)$? Can we do better?

Better analysis:

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

Pf

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

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

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

$$\leq \sum_x \sqrt{\frac{p_i}{m}}$$

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

note:

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

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

$$= \frac{m \cdot p_x}{m} = p_x$$

Jensen's \neq

$$\text{Var}(\hat{p}(x)) = \frac{1}{m^2} \sum_{i=1}^m p_i(1-p_i)$$

$$\leq \frac{p_i}{m}$$

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

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

$$E[\|\hat{p}-p\|_1] \leq \frac{\epsilon}{2}$$

by Markov's \neq : with prob $1-\frac{1}{2}$, $\|\hat{p}-p\|_1 \leq \epsilon$

Note, this says we can "learn" (approximate) any dist wrt. L_1 distance in $\Theta\left(\frac{n}{\epsilon^2}\right)$ samples

L₂ - Distance (squared):

$$\begin{aligned} \|p - \frac{U}{n}\|_2^2 &= \sum_{i \in [n]} (p_i - \frac{1}{n})^2 \\ &= \sum p_i^2 - \frac{2}{n} \sum p_i + \sum (\frac{1}{n})^2 \\ &= \sum p_i^2 - \frac{1}{n} \end{aligned}$$

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

$$= \|p\|_2^2 - \|U/n\|_2^2$$

we can estimate this

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

First:

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

Naive idea:

take two new samples:

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

" 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 "

Estimate by recycling:

• Take s samples from p : X_1, \dots, X_s

• for each $1 \leq i < j \leq s$

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

• Output $\hat{c} \leftarrow \frac{\sum_{i < j} b_{ij}}{\binom{s}{2}}$

b_{ij} 's not independent
so can't use Chernoff

Analysis: $E[\hat{c}] = \frac{1}{\binom{s}{2}} \cdot \binom{s}{2} \cdot E[b_{ij}]$
 $= \|p\|_2^2$

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

Assumption \star :

$$|\hat{C} - \|p\|_2^2| < \Delta$$

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

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

What happens if \star holds with $\Delta = \frac{\epsilon^2}{2}$?

• if $p = U_{[n]}$ then $\hat{C} \leq \|U_{[n]}\|_2^2 + \Delta = \frac{1}{n} + \frac{\epsilon^2}{2}$

so test will PASS

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

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

+ $\hat{C} > \|p\|_2^2 - \Delta$
 $\geq \epsilon^2 + \frac{1}{n} - \Delta = \epsilon^2 + \frac{1}{n} - \frac{\epsilon^2}{2} = \frac{\epsilon^2}{2} + \frac{1}{n}$

so test will FAIL

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

Analysis

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

$$E[\hat{c}] = \frac{1}{\binom{s}{2}} \binom{s}{2} E[\delta_{ij}] = \|p\|_2^2$$

$$\Pr[|\hat{c} - \|p\|_2^2| > \rho] \leq \frac{\text{Var}[\hat{c}]}{\rho^2}$$

Chebyshev \neq

Fact $\text{Var}[aX] = a^2 \text{Var}[X]$

So
$$\text{Var}[\hat{c}] = \text{Var}\left[\frac{1}{\binom{s}{2}} \sum_{i < j} \delta_{ij}\right] \\ = \frac{1}{\binom{s}{2}^2} \text{Var}\left[\sum_{i < j} \delta_{ij}\right]$$

Lemma $\text{Var}\left[\sum \delta_{ij}\right] \leq 4 \left(\binom{s}{2} \|p\|_2^2\right)^{3/2}$

Why? (proof...)

def. $\bar{\delta}_{ij} = \delta_{ij} - E[\delta_{ij}]$

← trick - will rewrite variance as $E\left[\left(\sum \delta_{ij}\right)^2\right]$

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

Also $\because E[\bar{\delta}_{ij} \bar{\delta}_{kl}] \leq E[\delta_{ij} \delta_{kl}]$

Verify at home? (or trust...)

- $(\sum p(x)^3)^{1/3} \leq (\sum p(x)^2)^{1/2}$
- $s^2 \leq 3 \binom{s}{2}$
- $\binom{s}{3} \leq s^3/6$

e.g. $(a^3 + b^3)^2 \leq (a^2 + b^2)^3$
 $a^6 + 2a^3b^3 + b^6 \leq a^6 + b^6 + 3a^4b^2 + 3a^2b^4$

So

$$\text{Var} \left[\sum_{i < j} \bar{b}_{ij} \right] = E \left[\left(\sum_{i < j} \bar{b}_{ij} - E \left[\sum_{i < j} \bar{b}_{ij} \right] \right)^2 \right]$$

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

(5) $+ \sum \bar{b}_{ij} \bar{b}_{il}$
 (6) $+ \sum \bar{b}_{ij} \bar{b}_{ki}$

$$= E \left[\sum_{i < j} \bar{b}_{ij}^2 + \sum_{\substack{i < j \\ k < l \\ i, j, k, l \text{ distinct}}} \bar{b}_{ij} \bar{b}_{kl} + \sum_{\substack{i < j \\ k < l \\ i, j, l \text{ distinct}}} \bar{b}_{ij} \bar{b}_{il} + \sum_{\substack{i < j \\ k < l \\ i, k, j \text{ distinct}}} \bar{b}_{ij} \bar{b}_{kj} \right]$$

(1) (2) (3) (4)

(1) $E \left[\sum_{i < j} \bar{b}_{ij}^2 \right] \leq E \left[\sum \bar{b}_{ij}^2 \right] = \binom{s}{2} \|p\|_2^2$

$E[\bar{b}_{ij}] = E[\bar{b}_{ij}^2]$ since \bar{b}_{ij} is indicator var

(2) $E \left[\sum_{\substack{i < j \\ k < l \\ \text{all 4 distinct}}} \bar{b}_{ij} \bar{b}_{kl} \right] \leq \sum E[\bar{b}_{ij}] E[\bar{b}_{kl}] = 0$

independent

(3) $E \left[\sum \bar{b}_{ij} \bar{b}_{il} \right] \leq E \left[\sum_{\substack{i, j, l \\ \text{distinct}}} \bar{b}_{ij} \cdot \bar{b}_{il} \right] = \sum_{\substack{i, j, l \\ \text{distinct}}} \text{pr}[X_i = X_j = X_l]$

$\leq \binom{s}{3} \sum_x p(x)^3$ expected # 3-way collisions

$\frac{1}{6} \binom{s}{2}^{3/2} < \frac{(3 \binom{s}{2})^{3/2}}{6}$
 $= \frac{\sqrt{3}}{2} \binom{s}{2}^{3/2}$

$\leq \frac{s^3}{6} \left(\sum_x p(x)^2 \right)^{3/2}$
 $\leq \frac{\sqrt{3}}{2} \binom{s}{2}^{3/2} (\|p\|_2^2)^{3/2}$ by the facts

- ④ same as 3
- ⑤
- ⑥

In total:

$$\begin{aligned} \text{Var} \left[\sum_{i < j} \delta_{ij} \right] &\leq \text{Var} \left[\sum_{i < j} \bar{\delta}_{ij} \right] \\ &\leq \binom{s}{2} \|p\|_2^2 + 0 + 2 \cdot \frac{\sqrt{3}}{2} \left(\binom{s}{2} \|p\|_2^2 \right)^{3/2} \\ &\leq 2 \left[\binom{s}{2} \|p\|_2^2 \right]^{3/2} \end{aligned}$$



Putting lemma into Chebyshev

1) use $p = \frac{\epsilon^2}{2}$

$$\Pr[|\hat{c} - \|p\|_2^2| > \frac{\epsilon^2}{2}] \leq \frac{\text{Var}[\hat{c}]}{\epsilon^4} \cdot 4$$

Recall this comes from const. in front of \hat{c}

$$\leq \frac{\binom{s}{2} \|p\|_2^2}{\binom{s}{2}^2 \epsilon^4} \cdot 4 = \frac{8}{\epsilon^4} \cdot \frac{1}{s} \cdot \|p\|_2^3$$

also want this to be ≤ 1

So pick $s \geq \left(\frac{1}{\epsilon^4}\right)$

Note: Can get better bound

1) Testing closeness to any known distribution — reduce to uniform case!

2) lower bound

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

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

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

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

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

either additive estimate with error $\leq \frac{\epsilon^2}{2n}$

or mult error $\leq (1 \pm \frac{\epsilon^2}{3})$

suffices

would have this if have additive error $\leq \frac{\epsilon^2}{3n} \cdot \|p\|_2^2$

to get additive error $\leq \frac{\epsilon^2}{3n} \|p\|_2^2$

suffices to have

$s \geq \frac{\text{const} \cdot \sqrt{n}}{\epsilon^2}$

samples

since $\Pr[|\hat{C} - \|p\|_2^2| \geq \gamma \|p\|_2^2] \leq \frac{k \cdot \|p\|_2^3}{s \cdot \gamma^2 (\|p\|_2^2)^2} \leq \frac{k}{s \cdot \gamma^2 \cdot \|p\|_2}$

[note $\|p\|_2^2 > \frac{1}{n}$ so $\|p\|_2 > \frac{1}{\sqrt{n}}$ so $\frac{1}{\|p\|_2} < \sqrt{n}$]

$\leq \frac{k \cdot \sqrt{n}}{s \cdot \gamma^2}$

[note: we need $\gamma^2 \approx \frac{\epsilon^2}{3}$]

so picking $s \gg \frac{\sqrt{n}}{\epsilon^4}$ will give small probability of error \Rightarrow

$\approx \frac{k \cdot \sqrt{n}}{s} \cdot \frac{1}{\epsilon^4}$

Lower bound of $\Omega(\sqrt{n})$ samples for testing uniformity :

$$p = U_{[n]}$$

$$q = U_S \quad \text{where } S \text{ is a set of } n/2 \text{ elements chosen randomly from } [n]$$

Note: If you can test uniformity, you can distinguish p from q in $t(n)$ samples in $t(n)$ samples

But how can you tell if samples come from p or q ?

- Notice that "names" of elements tell you nothing
- only collision statistics give you information
- first collisions happen only after $\Omega(\sqrt{n})$ samples in both p, q