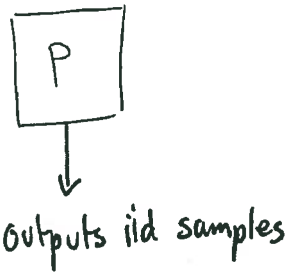
Lecture 12:

Testing Distributions

- Uniformity

Turning to a new model: prob dists

Probability distributions - get samples of distribution



Known ↙

Domain D, $|D| = n$

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

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:

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

Uniform dist on D

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

today's focus

③
P.D.

Distances

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

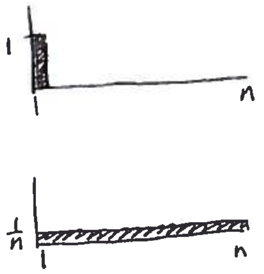
l_2-distance : $\|p-q\|_2 = \sqrt{\sum_{i\in D} (p_i - q_i)^2}$

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

examples:

① $p = (1, 0, 0, \dots 0)$

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



l_1 distance:

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

≈ 2

l_2-distance:

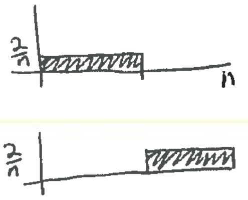
$\|p-q\|_2^2 = (1-\frac{1}{n})^2 + (n-1)(\frac{1}{n})^2$

≈ 1

②

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

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



l_1 distance:

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

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

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

"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 $\sum_x \left|\hat{p}(x) - \frac{1}{n}\right| > \varepsilon$ reject

  else accept.

Analysis: (better analyses exist)

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

So, if $p = U_n$ then p passes

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

then $\|p - U\|_1 < 2\varepsilon$.

So, if $\|p - U_n\| \overset{?}{>} 2\varepsilon$ this test is likely to Fail

how many samples? $\Omega\left(\frac{n}{\varepsilon}\right)$ 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

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

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

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

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

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

Jensen's ≠

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

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

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

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

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

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

■

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

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

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

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

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

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

Algorithm

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

⑦

P.D

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 <u>all</u> pairs in sample

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

Estimate by recycling:

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

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

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

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

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

$= \|p\|_2^2$

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

Assumption $\star$: $|\hat{c} - \|p\|_2^2| < \Delta$ ← this is our parameter that determines whether our approximation is good. Spoiler: will set $\Delta = \frac{\varepsilon^2}{2}$

will take enough samples so that this holds with prob ≥ 3/4

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

Correct behavior!

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

  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

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

  \+ $\hat{c} > \|p\|_2^2 - \Delta$ ← $\star$

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

  so test will FAIL

Remaining Question:

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

Analysis

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

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

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

Chebyshev ‡

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

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

Fact $\Rightarrow$
$Var[\hat{c}]$
$\le 4 \cdot \frac{\left(\binom{s}{2} \|p\|_2^2\right)^{3}}{\binom{s}{2}^2}$
$\le \theta(\|p\|_2^3 / s)$

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

Why? (proof...)

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

← trick - will rewrite variance as $E[\bar{\sigma}_{ij}^2]$.

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

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

verify at home? (or trust...)

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

So

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

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

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

$$\overset{⑤}{+\sum\bar\sigma_{ij}\bar\sigma_{jl}} \qquad \overset{⑥}{+\sum\bar\sigma_{ij}\bar\sigma_{ki}}$$

① $E\Big[\sum_{i<j}\bar\sigma_{ij}^{\,2}\Big] \le E\Big[\sum\sigma_{ij}^2\Big] = \binom{s}{2}\|p\|_2^2$

$\boxed{E[\sigma_{ij}] = E[\sigma_{ij}^2] \text{ since } \sigma_{ij} \text{ is indicator var}}$

② $E\Big[\sum_{\substack{i<j\\k<l\\\text{all 4 distinct}}}\bar\sigma_{ij}\,\bar\sigma_{kl}\Big] \overset{\text{independent}}{\le} \sum E[\bar\sigma_{ij}]\,E[\bar\sigma_{kl}] = 0$

③ $E\Big[\sum\bar\sigma_{ij}\,\bar\sigma_{il}\Big] \le E\Big[\sum_{\substack{i,j,l\\\text{distinct}}}\sigma_{ij}\cdot\sigma_{il}\Big] = \sum_{\substack{i,j,l\\\text{distinct}}} pr[X_i = X_j = X_l]$

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

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

$\le \frac{\sqrt3}{2}\binom{s}{2}^{3/2}\Big(\|p\|_2^2\Big)^{3/2}$ by the facts

$$\boxed{\frac{1}{6}\big(s^2\big)^{3/2} < \frac{\big(3\binom{s}{2}\big)^{3/2}}{6} = \frac{\sqrt3}{2}\binom{s}{2}^{3/2}}$$

④ same as 3

⑤

⑥

In total:

$$\mathrm{Var}\Big[\sum_{i<j} \sigma_{ij}\Big] \leq \mathrm{Var}\Big[\sum_{i<j} \bar{\sigma}_{ij}\Big]$$

$$\leq \binom{5}{2}\|p\|_2^2 + 0 + 4\cdot\frac{\sqrt{3}}{2}\left(\binom{5}{2}\|p\|_2^2\right)^{3/2}$$

$$\leq 4\left[\binom{5}{2}\|p\|_2^2\right]^{3/2}$$

∎

Putting lemma into Chebyshev:

use $\rho = \frac{\varepsilon^2}{2}$

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

note $\frac{1}{\left(\frac{s}{2}\right)^{3/2}} \leq \frac{1}{\sqrt{\frac{s^2}{2}}} \leq \frac{2}{s}$

recall this comes from const in Proof of $\hat{c}$

$$\leq \frac{4\left[\left(\frac{s}{2}\right)\|p\|_2^2\right]^{3/2}}{\left(\frac{s}{2}\right)^2 \quad \varepsilon^4} \cdot 4 \leq \underbrace{\frac{32}{\varepsilon^4} \cdot \frac{1}{s}}_{\text{also want this to be } \leq 1} \cdot \underbrace{\|p\|_2^3}_{\leq 1}$$

So Pick $s \geq \Omega\left(\frac{1}{\varepsilon^4}\right)$

Note: Can get better bnd

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

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

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

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

or mult error $\leq \underbrace{\left(1 \pm \frac{\varepsilon^2}{3}\right)}$

suffices

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

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

suffices to have

$$s \geq \frac{\text{const} \cdot \sqrt{n}}{\varepsilon^2} \quad \text{samples}$$

since $\Pr\left[\left|\hat{c} - \|p\|_2^2\right| \geq \gamma \|p\|_2^2\right] \leq \frac{\overset{\text{const}}{K} \cdot \|p\|_2^3}{s \cdot \gamma^2 \left(\|p\|_2^2\right)^2} \leq \frac{K}{s \cdot \gamma^2 \cdot \|p\|_2}$

[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 $\gamma \approx \frac{\varepsilon^2}{3}$]

so picking $s >> \frac{\sqrt{n}}{\varepsilon^4}$ will give small probability of error $\Longrightarrow$ $\approx \frac{K \cdot \sqrt{n}}{s} \cdot \frac{1}{\varepsilon^4}$