Lecture 4 Jan 28, 25

Distribution testing

- uniformity testing

distribution testing

An (ε, δ)-tester for property P

we have an unknown distribution d

We aim to design an algorithm $\mathcal{A}$ that distinguishes the following w.p. $\geq 1-\delta$:

- if $d \in P$, $\mathcal{A}$ outputs accept
- if d is ε-far from P, $\mathcal{A}$ outputs reject

what is a property?

P = a set of distributions

$P = \{U_n\}$ $\rightarrow$ a uniform dist. on $[n]$

$P = \{$ a set of unimodal distributions $\}$

d is ε-far iff $\text{dist}(d, P) > \varepsilon$

$$\text{dist}(d, P) = \min_{d' \in P} \text{dist}(d, d')$$

Example distances:

ℓ_1-distance: $\|d - d'\|_1 = \sum_{x \in \Omega} |d(x) - d'(x)|$

ℓ_2 - distance : $\|d - d'\|_2 = \sqrt{\sum_{x \in \Omega} (d(x) - d(x'))^2}$

Total variation distance : $\|d - d'\|_{TV} = \max_{E \subseteq \Omega} |d(E) - d(E')|$

(statistical distance) ↳ every event

Turns out $\|d - d'\|_{TV} = \frac{1}{2} \|d - d'\|_1$

Today's question : uniformity testing

Design algorithm A that receives n, ε, δ, and samples from d and outputs

- accept w.p. $\geq 1 - \delta$ if $d = U_n$
- reject w.p. $\geq 1 - \delta$ if $\|d - U_n\|_1 > \varepsilon$

Q_1: which one look like a real dice ?

2 3 1 4 6 1

4 6 4 3 4 5

Q_2 what did give it away ?

A_2 repetitions! $\rightsquigarrow$ samples from a uniform distribution looks "less" repeated.

Let's formalize this intuition...

collisions : two samples that are equal to each other

\# collisions in the sample set, tells us if a distribution is uniform or not.

Algorithm:

Draw m samples from $\mathcal{D}$: $X_1, \ldots, X_m$

$\forall\ i < j \in [m]$: $\alpha_{ij} = \begin{cases} 1 & \text{if } X_i = X_j \\ 0 & \text{o.w.} \end{cases}$

$$Y \leftarrow \sum_{i=1}^{m} \sum_{j>i}^{m} \alpha_{ij} \Big/ \binom{m}{2}$$

if $Y < t$

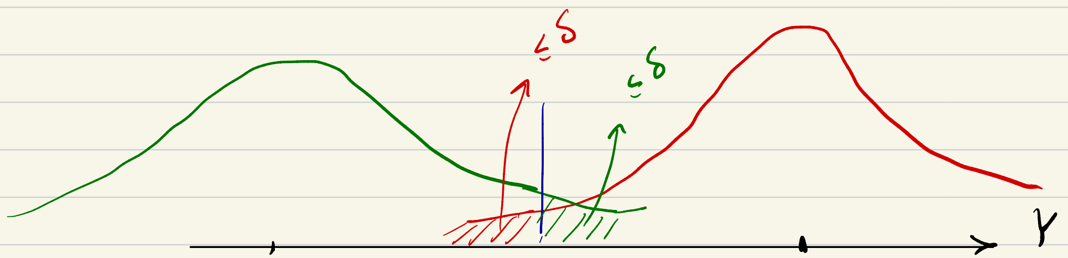
output accept

else

output reject

Our goal here: what should m & t be?

Visual description



$E[Y \mid d = U_n]$ $\qquad$ $E\left[Y \mid d \text{ is } \varepsilon\text{-far from } U_n\right]$

accept $\qquad$ reject

slack $= \varepsilon^2/n$

First step: slack exists

$$E[\sigma_{ij}] = \sum_{a=1}^{n} \Pr[X_i = a] \cdot \Pr[X_j = a]$$

$$= \sum_{a=1}^{n} d_a^2 = \|d\|_2^2$$

$$E[Y] = \frac{1}{\binom{m}{2}} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \sigma_{ij} = \|d\|_2^2$$

Case 1: d is uniform

if $d = U_n$: $\|d\|_2^2 = \sum_{a=1}^{n} d_a^2 = n \times \frac{1}{n^2} = \frac{1}{n}$

Case 2: d is ε-far from uniform

if $\|d - U_n\|_1 > \varepsilon$:

$$\|d\|_2^2 = \sum_{a=1}^{n} d_a^2 = \sum_{a=1}^{n} \left(\frac{1}{n} + \left(d_a - \frac{1}{n}\right)\right)^2$$

$$= \sum_{a=1}^{n} \frac{1}{n}^2 + \frac{2}{n}\left(d_a - \frac{1}{n}\right) + \left(d_a - \frac{1}{n}\right)^2$$

$$= \frac{1}{n} + \frac{2}{n}\underbrace{\left(\sum_{a=1}^{n} d_a - \frac{1}{n}\right)}_{=0} + \sum_{a=1}^{n}\left(d_a - \frac{1}{n}\right)^2$$

$$= \frac{1}{n} + \underbrace{\|d - U_n\|_2^2}_{\text{our slack}}$$

- Our conjecture is correct Y "tends" to be larger when d is ε-far from uniform.

How far?

we know $\| d - U_n \|_1 > \varepsilon$ $\Big\} \Rightarrow$

Cauchy-schwarz: $(\sum x_i^2) . (\sum y_i^2) \geq (\sum x_i y_i)^2$

$$\left(\sum_a \left(d_a - \frac{1}{n}\right)^2 \right) . \left(\sum_{a=1}^{n} 1^2 \right) \geq \left(\sum \left|d_a - \frac{1}{n}\right| \right)^2$$

$\Rightarrow$

$$\| d - U_n \|_2^2 = \sum_{a=1}^{n} \left(d_a - \frac{1}{n}\right)^2 \geq \frac{\left(\sum \left|d_a - \frac{1}{n}\right|\right)^2}{n}$$

$$= \frac{\| d - U_n \|_1^2}{n} > \frac{\varepsilon^2}{n}$$

$E[Y \mid d = U_n]$ $\qquad\qquad$ $E[Y \mid d \text{ is } \varepsilon\text{-far}]$

$\frac{1}{n}$ $\qquad$ $t := \frac{1 + \varepsilon^2/2}{n}$ $\qquad$ $\geq \frac{1+\varepsilon^2}{n}$

Next step : Concentration

Let set t to be in the middle : $t \leftarrow \frac{1+\varepsilon^2/2}{n}$

If we show the following, we get an (ε, δ) _tester

$\delta = 0.1$

1. $\Pr\left[Y \geq \frac{1 + \varepsilon^2/2}{n} \;\middle|\; d = U_n \right] \leq \cancel{\delta}$

$\delta = 0.1$

2. $\Pr\left[Y \leq \frac{1 + \varepsilon^2/2}{n} \;\middle|\; d \text{ is } \varepsilon\text{-far from } U_n \right] \leq \cancel{\delta}$

$Y = \frac{1}{\binom{m}{2}} \sum_{i<j} \sigma_{ij}$ not a great candidate for Chernoff. bound

(why?)

Our plan: Using chebyshev's

Lets compute the variance of Y

Lemma 1 $\mathrm{Var}(Y) = \frac{1}{\binom{m}{2}^2} \cdot \left(\binom{m}{2} \|d\|_2^2 + 6\binom{m}{3} \|d\|_3^3 \right)$

proof is deferred for now.

Case 1 : $d = U_n$

$$\Pr\left[\,\left|Y - E[Y]\right| \geq \frac{\varepsilon^2}{2n}\right] \leq \frac{\operatorname{Var}(Y)}{\left(\varepsilon^2/2n\right)^2}$$

$$\leq \frac{1}{\binom{m}{2}^2} \cdot \left(\binom{m}{2} \|d\|_2^2 + 6\binom{m}{3} \|d\|_3^3 \right) \cdot \frac{4n^2}{\varepsilon^2}$$

$$= \Theta\left(\frac{n^2}{m^4 \varepsilon^4} \cdot \left(m^2 \cdot \frac{1}{n} + \frac{m^3}{n^2} \right) \right)$$

$$= \Theta\left(\frac{n}{m^2 \varepsilon^4} + \frac{1}{m \varepsilon^4} \right) \leq 0.1$$

if $m = c \cdot \left(\frac{1}{\varepsilon^4} + \frac{\sqrt{n}}{\varepsilon^2} \right)$

for sufficiently large c

Case 2: $\|d - U_n\|_1 > \varepsilon$

The bound on the variance can be large.

$\binom{m}{2}\|d\|_2^2 + 6\binom{m}{3}\|d\|_3^3$

Could be problematic if we require $|Y - E[Y]| \leq \frac{\varepsilon}{n}$

↳ adjust the length accordingly

$$\Pr\left[Y - \mathbb{E}[Y] \geq \frac{\varepsilon^2}{2} \mathbb{E}[Y] \right] \leq \frac{4\,\mathrm{Var}[Y]}{\varepsilon^4\, \mathbb{E}[Y]^2}$$

$$\leq \frac{1}{\binom{m}{2}^2} \cdot \frac{\binom{m}{2}\|d\|_2^2 + 6\binom{m}{3}\|d\|_3^3}{\varepsilon^4 \|d\|_2^4} =$$

$$= \Theta\left(\frac{1}{m^2 \cdot \varepsilon^4 \|d\|_2^2} + \frac{\|d\|_3^3}{m\, \varepsilon^4 \|d\|_2^4} \right) \leq 0.1$$

$$= \Theta\left(\frac{n}{m^2 \varepsilon^4} + \frac{\sqrt{n}}{m\, \varepsilon^4} \right) \qquad m = c \cdot \frac{\sqrt{n}}{\varepsilon^4}$$

using $\|d\|_3^3 \leq \|d\|_2^3$ & $\|P\|_2^2 \geq \frac{1}{n}$

$\Uparrow$

ℓ_p-norm inequality $\|d\|_3 \leq \|d\|_2$

Lemma 1 $\mathrm{Var}(Y) = \frac{1}{\binom{m}{2}^2} \cdot \left(\binom{m}{2} \|d\|_2^2 + 6\binom{m}{3} \|d\|_3^3 \right)$

proof:

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{1}{\binom{m}{2}} \sum_{i<j} \sigma_{ij} \right)$$

$$= \frac{1}{\binom{m}{2}^2} \mathrm{Var}\left(\sum_{i<j} \sigma_{ij} \right)$$

$$= \frac{1}{\binom{m}{2}^2} \left(E\left[\left(\sum_{i<j} \sigma_{ij} \right)^2 \right] - \left(\sum_{i<j} \underbrace{E[\sigma_{ij}]}_{\|d\|_2^2} \right)^2 \right)$$

$$= \frac{1}{\binom{m}{2}^2} E\left[\sum_{i<j} \sum_{l<k} \sigma_{ij} \sigma_{lk} \right]$$

$$- \|d\|_2^4$$

$$\mathbb{E}[\sigma_{ij}^2] = \|d\|_2^2$$ ① $|\{i,j,\ell,k\}| = 2 \Rightarrow i=\ell,\ j=k$

$$\mathbb{E}[\sigma_{ij}\ \sigma_{\ell k}] = \|d\|_3^3$$ ② $|\{i,j,\ell,k\}| = 3$

↳ Pr[three samples are equal]

$$\mathbb{E}[\sigma_{ij}\ \sigma_{\ell k}] = \mathbb{E}[\sigma_{ij}] \cdot \mathbb{E}[\sigma_{\ell k}]$$ ③ $|\{i,j,\ell,k\}| = 4$

$$= \|d\|_2^4$$

$$\Rightarrow \mathrm{Var}[Y] = \frac{1}{\binom{m}{2}^2}\Big[\binom{m}{2}\cdot\|d\|_2^2 + \overbrace{6}^{\binom{3}{2}\cdot\left(\frac{3}{2}-1\right)}\binom{m}{3}\|d\|_3^3$$

$$+ \binom{m}{2}\binom{m-2}{2}\|d\|_2^4 - \binom{m}{2}^2\|d\|_2^4\Big]$$

$$\leq \frac{1}{\binom{m}{2}^2}\left[\binom{m}{2}\|d\|_2^2 + 6\binom{m}{3}\|d\|_3^3\right] \quad \square$$

Exercise: verify that

$$\binom{m}{2} + 6\binom{m}{3} + \binom{m}{2}\binom{m-2}{2} = \binom{m}{2}^2$$

We need independence

Poissonization method

Binomial $(n,p) \approx$ Poisson (np)

$$\Pr_{Bin}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

small k

$$\approx \frac{\cancel{n(n-1)\cdots(n-k+1)}}{k!} \frac{\lambda^k}{\cancel{n^k}} \left(1-\frac{\lambda}{n}\right)^n$$

large n

$$\approx \frac{\lambda^k e^{-\lambda}}{k!}$$