Property Testing on Distributions

Lekshmi Ramesh

Indian Institute of Science
Bangalore

September 29, 2018

Property testing

- Property testing broadly refers to testing structural properties of data
- Some examples
  - Testing whether a graph can be clustered
  - Testing whether a boolean function is monotone
  - Testing whether samples are generated from the uniform distribution
- Decide whether an object has a property or is far from having it

Testing uniformity

- Given i.i.d. samples $\{X_i\}_{i=1}^n$ from an unknown discrete distribution P, determine whether P is the uniform distribution or far from it

- Test $\quad \mathcal{H}_0 : \; P = U$
  vs $\quad \mathcal{H}_1 : \; \|P - U\| \geq \epsilon$

  where U denotes the uniform distribution on $[k]$, and $\| \cdot \|$ is some notion of distance between distributions

- We will consider the ℓ_1 distance between discrete distributions

$$\|P - Q\|_1 := \sum_{i=1}^{k} |P_i - Q_i|$$

Testing uniformity

- We first consider testing in ℓ_2 and extend it to ℓ_1
- For testing in ℓ_2, need a good estimate for $\|P - U\|_2^2$

$$\begin{aligned} \|P - U\|_2^2 &= \sum_{i=1}^{k} (P_i - \frac{1}{k})^2 \\ &= \|P\|_2^2 - \frac{1}{k} \end{aligned}$$

So we need to estimate $\|P\|_2^2$

Additive vs multiplicative accuracy

- How much error can we allow in our estimate for $\|P\|_2^2$?
- First note that $\|P\|_2^2 = \frac{1}{k}$, when $P = U$
- If $\|P - U\|_2 \geq \epsilon$, then

$$\begin{aligned} \|P\|_2^2 &= \|P - U\|_2^2 + \frac{1}{k} \\ &\geq \epsilon^2 + \frac{1}{k} \end{aligned}$$

- Under ℓ_2 distance, can allow additive error of $\frac{\epsilon^2}{2}$

Additive vs multiplicative accuracy

- How much error can we allow in our estimate for $\|P\|_2^2$?
- If $\|P - U\|_1 \geq \epsilon$, then

$$\begin{aligned}\|P\|_2^2 &= \|P - U\|_2^2 + \frac{1}{k} \\ &\geq \frac{1}{k}\|P - U\|_1^2 + \frac{1}{k} \\ &\geq \frac{1}{k}\epsilon^2 + \frac{1}{k} = \frac{1}{k}(1 + \epsilon^2)\end{aligned}$$

- Under ℓ_1 distance, can allow multiplicative error of $\frac{\epsilon^2}{3}\|P\|_2^2$

A test

- How to estimate $\|P\|_2^2$ using the samples?
- Use collision probabilities

$$Y_{ij} := \begin{cases} 1, & \text{if } X_i = X_j \\ 0, & \text{otherwise} \end{cases}$$

$$T := \frac{1}{\binom{n}{2}} \sum_{i<j} Y_{ij}$$

Analysis of the collision-based tester

- Note that

$$\begin{aligned}\mathbb{E}T &= \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbb{E}Y_{ij} \\ &= \frac{1}{\binom{n}{2}} \sum_{i<j} (P_1^2 + \ldots + P_k^2) \\ &= \|P\|_2^2\end{aligned}$$

T is an unbiased estimate of $\|P\|_2^2$

- Also

$$\mathbb{E}\Big(\sum_{i<j} Y_{ij}\Big)^2 = \sum_{i<j} Y_{ij}^2 + \sum_{\substack{i<j,k<l \\ \text{3 distinct indices}}} Y_{ij}Y_{kl} + \sum_{\substack{i<j,k<l \\ \text{all indices distinct}}} Y_{ij}Y_{kl}$$

$$= \binom{n}{2}\|P\|_2^2 + 6\binom{n}{3}\|P\|_3^3 + \binom{n}{2}\binom{n-2}{2}\|P\|_2^4$$

Analysis of the collision-based tester

- Therefore,

$$\begin{aligned}\mathrm{Var}(T) &= \frac{1}{\binom{n}{2}^2}\mathrm{Var}(\sum_{i<j} Y_{ij}) \\ &\le \frac{2}{n(n-1)}\|P\|_2^2 + \frac{4}{n}\|P\|_3^3\end{aligned}$$

Analysis of the collision-based tester

- For testing in ℓ_2,

$$\mathbb{P}\left(T - \|P\|_2^2 \geq \frac{\epsilon^2}{2}\right) \leq \frac{1}{(\epsilon^2/2)^2}\mathrm{Var}(T)$$

which gives $n \geq \frac{60}{\epsilon^4}$ for error probability $\leq 1/3$

Analysis of the collision-based tester

- For testing in ℓ_1,

$$\mathbb{P}\left(T - \|P\|_2^2 \leq \frac{\epsilon^2}{3}\|P\|_2^2\right) \leq \frac{9}{\epsilon^4\|P\|_2^4}\mathrm{Var}(T)$$

Simplifying and using $\|P\|_2^2 \geq \frac{1}{k}$ and $\|P\|_3 \leq \|P\|_2$, we get $n \geq \frac{216}{\epsilon^4}\sqrt{k}$ for error probability $\leq 1/3$

Conclusion

- The tester we saw doest not give optimal dependence on ϵ
- A $O(\frac{\sqrt{k}}{\epsilon^2})$ dependence was shown to be optimal [1]
- Some other problems in distribution testing
  - Independence testing
  - Identity testing
  - Testing closeness of distributions
  - Testing unimodality

[1] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. IEEE Trans. on Information Theory, 2008

Reference

- *D. Ron.* Property Testing. In Handbook on Randomized Computing (Vol. II), Kluwer Academic Publishers, 2001.
- *P. Beame.* Sublinear and Streaming Algorithms. Lecture notes (2014).