Data Amplification: Instance-Optimal Property Estimation

Yi Hao and Alon Orlitsky

{yih179, alon}@**ucsd**.edu

Outline

Definitions

Estimators

Prior results

Data amplification

Example: Shannon entropy

Ideas to take away: Instance-optimal algorithm

Data amplification

Definitions

Discrete support set $\mathcal{X}$

$\{\text{heads, tails}\} = \{\text{h, t}\}$ $\quad\quad \{\ldots, -1, 0, 1, \ldots\} = \mathbb{Z}$

Distribution p over $\mathcal{X}$, probability p_x for $x \in \mathcal{X}$

$p_x \geq 0$ $\quad\quad \sum_{x \in \mathcal{X}} p_x = 1$

$p = (p_{\mathsf{h}}, p_{\mathsf{t}})$ $\quad\quad p_h = .6,\ p_t = .4$

$\mathcal{P}$ collection of distributions

$\mathcal{P}_{\mathcal{X}}$ all distributions over $\mathcal{X}$

$\mathcal{P}_{\{\mathsf{h,\ t}\}} = \{(p_{\mathsf{h}}, p_{\mathsf{t}})\} = \{(.6, .4), (.4, .6), (.5, .5), (0, 1), \ldots\}$

$f : \mathcal{P} \to \mathbb{R}$

Maps distribution to real value

Shannon entropy	$H(p)$	$\sum_x p_x \log \frac{1}{p_x}$
Rényi entropy	$H_\alpha(p)$	$\frac{1}{1-\alpha} \log \left(\sum_x p_x^\alpha\right)$
Support size	$S(p)$	$\sum_x \mathbb{1}_{p_x > 0}$
Support coverage	$S_m(p)$	$\sum_x (1 - (1 - p_x)^m)$
Expected # distinct symbols in m samples		
Distance to fixed $\boldsymbol{q}$	$L_q(p)$	$\sum_x \lvert p_x - q_x \rvert$
Highest probability	$\max(p)$	$\max \{p_x : x \in \mathcal{X}\}$
...		

Many applications

Unknown: $p \in \mathcal{P}$

Given: property f and samples $X^n \sim p$

Estimate: $f(p)$

Entropy of English words

Given: $\mathcal{X} = \{\text{English words}\}$, unknown: p, estimate: $H(p)$

species in habitat

Given: $\mathcal{X} = \{\text{bird species}\}$, unknown: p, estimate: $S(p)$

How to estimate $f(p)$ when p is unknown?

Estimators

Observe n independent samples $X^n = X_1, \ldots, X_n \sim p$

Reveal information about p

Estimate $f(p)$

Estimator: $f^{\text{est}} : \mathcal{X}^n \to \mathbb{R}$

Estimate for $f(p)$: $f^{\text{est}}(X^n)$

Simplest estimators?

N_x # times x appears in $X^n \sim p$

$p_x^{\mathsf{emp}} := \frac{N_x}{n}$

$f^{\mathsf{emp}}(X^n) = f(p^{\mathsf{emp}}(X^n))$ a.k.a. MLE estimator in literature

Advantages

- plug-and-play: simple two steps
- universal: applies to all properties
- intuitive and stable

Best-known, most-used {distribution, property} estimator

Performance?

Classical Alternative to PAC Formulation

Absolute error $|f^{\text{est}}(X^n) - f(p)|$

$L_{f^{\text{est}}}(p,n) := \mathbb{E}_{X^n \sim p} \, |f^{\text{est}}(X^n) - f(p)|$ mean absolute error

$L_{f^{\text{est}}}(\mathcal{P},n) := \max_{p \in \mathcal{P}} L_{f^{\text{est}}}(p,n)$ worst-case MAE over $\mathcal{P}$

$L(\mathcal{P},n) := \min_{f^{\text{est}}} L_{f^{\text{est}}}(\mathcal{P},n)$ min-max MAE over $\mathcal{P}$

MSE – similar definitions, similar results, but slightly more complex expressions

Prior Results

if $|\mathcal{X}|$ is finite, write

$|\mathcal{X}| = k$

$\mathcal{P}_{\mathcal{X}} = \Delta_k$, the k-dimensional standard simplex

$\Delta_{\geq 1/k} := \{p : \ p_x \geq \frac{1}{k}$ or $p_x = 0, \ \forall x\}$ for support size

References: P03, VV11a/b, WY14/19, JVHW14, AOST14, OSW16, JHW16, ADOS17

Property	Base function	$L_{f^{\text{emp}}}(\Delta_k, n)$	$L(\Delta_k, n)$
Entropy [1]	$p_x \log \frac{1}{p_x}$	$\frac{k}{n} + \frac{\log k}{\sqrt{n}}$	$\frac{k}{n\log n} + \frac{\log k}{\sqrt{n}}$
Supp. coverage[2]	$(1-(1-p_x)^m)$	$m \exp\left(-\Theta\left(\frac{n}{m}\right)\right)$	$m \exp\left(-\Theta\left(\frac{n\log n}{m}\right)\right)$
Power sum [3] [4]	$p(x)^\alpha,\ \alpha \in (0, \frac{1}{2}]$	$\frac{k}{n^\alpha}$	$\frac{k}{(n\log n)^\alpha}$
	$p(x)^\alpha,\ \alpha \in (\frac{1}{2}, 1)$	$\frac{k}{n^\alpha} + \frac{k^{1-\alpha}}{\sqrt{n}}$	$\frac{k}{(n\log n)^\alpha} + \frac{k^{1-\alpha}}{\sqrt{n}}$
Dist. to fixed $\boldsymbol{q}$ [5]	$\lvert p_x - q_x \rvert$	$\sum_x q_x \wedge \sqrt{\frac{q_x}{n}}$	$\sum_x q_x \wedge \sqrt{\frac{q_x}{n\log n}}$
Support size [6]	$\mathbb{1}_{p(x)>0}$	$k \exp\left(-\Theta\left(\frac{n}{k}\right)\right)$	$k \exp\left(-\Theta\left(\sqrt{\frac{n\log n}{k}}\right)\right)$

$\star\, \boldsymbol{n}$ to $\boldsymbol{n \log n}$ when comparing the worst-case performances

[1] $n \gtrsim k$ for empirical; $n \gtrsim k/\log k$ for minimax

[2] $k = \infty$; $n \gtrsim m$ for empirical; $n \gtrsim m/\log m$ for minimax

[3] $\alpha \in (0, \frac{1}{2}]$: $n \gtrsim k^{1/\alpha}$ for empirical; $n \gtrsim \frac{k^{1/\alpha}}{\log k}$ and $\log k \gtrsim \log n$ for minimax

[4] $\alpha \in (\frac{1}{2}, 1)$: $n \gtrsim k^{1/\alpha}$ for empirical; $n \gtrsim \frac{k^{1/\alpha}}{\log k}$ for minimax

[5] additional assumptions required, see JHW18

[6] consider $\Delta_{\geq 1/k}$ instead of Δ_k; $k \log k \gtrsim n \gtrsim k/\log k$ for minimax

Data Amplification

Min-max approach is overly pessimistic: practical distributions often possess nice structures and are rarely the worst possible

⋆ Derive "competitive" estimators

– needs no knowledge on distribution structures, yet adaptive to the simplicity of underlying distributions

⋆ Achieve $\boldsymbol{n}$ to $\boldsymbol{n \log n}$ "amplification"

– distribution by distribution, the performance of our estimator with n samples is as good as that of the empirical with $n \log n$

For a broad class of properties, we derive an "instance-optimal" estimator which does as well with n samples as the empirical estimator would do with $n \log n$, for every distribution.

Example: Shannon Entropy

Theorem 1 Estimator f^{new} such that for any $\varepsilon \le 1$, n, and p,

$$L_{f^{\mathsf{new}}}(p,n) - L_{f^{\mathsf{emp}}}(p,\varepsilon n \log n) \lesssim \varepsilon$$

Comments

f^{new} requires only X^n and ε, and runs in near-linear time

$\log n$ amplification factor is optimal

$\log n \ge 10$ for $n \ge 22{,}027$ – "order-of-magnitude improvement"

ε can be a vanishing function of n

finite support S_p, then ε improves to $\varepsilon \wedge \left(\frac{S_p}{n} + \frac{1}{n^{0.49}}\right)$

Empirical entropy estimator

– has been studied for a long time

G. A. Miller, "Note on the bias of information estimates", **1955**.

– much easier to analyze compared to minimax estimators

* Our result holds on a *distribution level*, hence strengthens many results derived in the past half-century, in a *unified manner*

– large-alphabet regime $n = o(k/\log k)$

$$L(\Delta_k, n) \leq (1 + o(1)) \log\left(1 + \frac{k-1}{n \log n}\right)$$

Proof of $L_{f^{\text{emp}}}(\Delta_k, n) \le (1+o(1))\log\left(1+\frac{k-1}{n}\right)$ for $n = o(k)$

– absolute bias [P03]

$$0 \le H(p) - \mathbb{E}H(p^{\text{emp}}) = \mathbb{E}\,\mathrm{D}_{\mathsf{KL}}(p^{\text{emp}} \,\|\, p) \le \mathbb{E}\log(1+\chi^2(p^{\text{emp}} \,\|\, p))$$
$$\le \log(1+\mathbb{E}\,\chi^2(p^{\text{emp}} \,\|\, p)) = \log(1+\tfrac{k-1}{n})$$

– mean deviation

changing a sample modifies f^{emp} by $\le \frac{\log n}{n}$

apply the Efron-Stein inequality → mean deviation $\le \frac{\log n}{\sqrt{n}}$

⋆ The proof is very simple compared to that of min-max estimators

Large-Alphabet Entropy Estimation (Cont')

Theorem 1 strengthens the result and yields, for $n = o(k/\log k)$,

$$L(\Delta_k, n) \le \log\left(1 + \frac{k-1}{n\log n}\right) + o(1)$$

* Right expression for entropy estimation?
  - meaningful since $H(p)$ can be as large as $\log k$
  - for $n = \Omega(k/\log k)$, by [VV11a/b, WY14/19, JVHW14]
    $L(\Delta_k, n) \asymp \frac{k}{n\log n} + \frac{\log n}{\sqrt{k}} \asymp \underline{\log\left(1 + \frac{k-1}{n\log n}\right) + o(1)}$
  - should write $L(\Delta_k, n)$ in the latter form

Ideas to Take Away

Instance-optimal algorithm

worst-case algorithm analysis is pessimistic

modern data science calls for instance-optimal algorithms

better performance on easier instances – data is intrinsically simpler

Data amplification

designing optimal learning algorithms directly might be hard

instead, find a simple algorithm that works

emulate its performance by an algorithm that uses fewer samples

Thank you!