How Much is Unseen Depends Chiefly On Information About the Seen

MAX PLANCK INSTITUTE FOR SECURITY AND PRIVACY

SPOTLIGHT ICLR

Seongmin Lee and Marcel Böhme

Try it out! https://github.com/niMgnoeSeeL/UnseenGA

Abstract & Contribution

***Unknown* Multinomial Distribution over S:** $\mathscr{D}_S(\mathbf{p})$; $\mathbf{p} = \langle p_1, p_2, \cdots, p_S \rangle$ ⊃ X^n Samples

Then, what is $\Pr(X_{n+1} \mid X^n)$? (Unseen class)

Missing Mass Problem.

Given the sample X^n from the unknown multinomial distribution $\mathscr{D}$, what is the probability M_0 that the next sample X_{n+1} has never been seen before?

- Missing mass represents the ***representativeness of the training data***; if the missing mass regarding the training data is ***high*** 😨, the model trained by the data will ***often face unseen labels*** in the test data 😮.

In this work,

- ***We show exactly to what extent $\mathbb{E}[M_0]$ can be estimated from the sample X^n and how much remains ⇒ minimal bias estimator $\hat{M}_0^B$ 🎯.***
- ***We define a class of nearly unbiased estimators of M_0 👜 from different representations of $\mathbb{E}[M_0]$.***
- ***We cast the distribution-free estimation of M_0 as a search problem from 👜 whose goal is to find a distribution-specific estimator with a minimized MSE ⇒ minimal MSE estimator M_0^{Evo} 🤩.***

Problem Definition

- Frequency: $N_x(X^n) = \sum_{i=1}^{n} \mathbf{1}(X_i = x)$
- Frequency of frequencies: $\Phi_k(X^n) = \sum_{x=1}^{S} \mathbf{1}(N_x(X^n) = k)$
- Expected Φ: $f_k(n) = \mathbb{E}_{X^n \sim \mathscr{D}}\left[\Phi_k(X^n)\right]$

$$= \sum_{x=1}^{S} \mathbb{E}_{X^n \sim \mathscr{D}}\left[\mathbf{1}(N_x(X^n) = k)\right]$$

$$= \binom{n}{k} \sum_{x=1}^{S} p_x^k (1-p_x)^{n-k} \rightarrow g_k(n)$$

Example. Sample X^n of size 10

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	1	2	3	4	2	5	6	3	1

- $N_1 = 3, N_2 = 2, N_3 = 2, N_4 = 1, N_5 = 1, N_6 = 1$
- $\Phi_1 = 3, \Phi_2 = 2, \Phi_3 = 1$

Missing Mass: $\Pr(X_{11} \notin [1,6])$?

- Missing Mass: $M_0 = \sum_{x=1}^{S} p_x \mathbf{1}(N_x(X^n) = 0)$
- Good-Turing estimator [1]: $\hat{M}_0^G = \frac{\Phi_1}{n}$

Two key equations

(1) Expected Missing Mass

$$\mathbb{E}[M_0] = \sum_{x=1}^{S} p_x(1-p_x)^n = g_1(n+1)$$

(2) Relation between $g_k(n)$

$$g_k(n+1) = \sum_x p_x^k (1-p_x)^{n+1}$$
$$= \sum_x p_x^k (1-p_x)^n - \sum_x p_x^{k+1}(1-p_x)^{n+1}$$
$$= g_k(n) - g_{k+1}(n+1)$$

Minimal Bias Estimator $\hat{M}_0^B$

- By recursively applying (2) on (1), we get the following:

$$\mathbb{E}(M_0) =^{(1)} g_1(n+1)$$
$$=^{(2)} g_1(n) - g_2(n+1)$$
$$=^{(2)} g_1(n) - g_2(n) + g_3(n+1)$$
$$=^{(2)} \cdots$$
$$=^{(2)} g_1(n) - g_2(n) - \cdots + (-1)^n g_{n+1}(n+1)$$
$$= \frac{\mathbb{E}(\Phi_1)}{n} - \frac{\mathbb{E}(\Phi_2)}{\binom{n}{2}} + \frac{\mathbb{E}(\Phi_3)}{\binom{n}{3}} - \cdots + R_{n+1}$$

$\mathbb{E}[\hat{M}_0^B]$ ——— $\mathrm{Bias}(\hat{M}_0^B)$ ·····

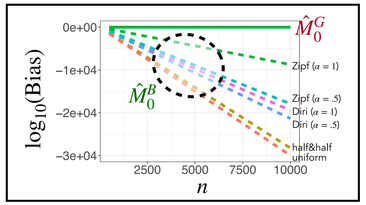
$\mathbb{E}[\hat{M}_0^G]$ ——— $\mathrm{Bias}(\hat{M}_0^G)$ ·····

Minimal Bias Estimator $\hat{M}_0^B$ 🎯

$$\hat{M}_0^B = -\sum_{i=1}^{n} (-1)^i \frac{\Phi_i}{\binom{n}{i}}$$

- $\mathrm{Bias}(\hat{M}_0^B) = |R_{n+1}| = g_{n+1}(n+1) = \sum p_i^{n+1}$
  = chance of the same class picked n + 1 times
  ⇒ ***Decays exponentially by n.***
- $\mathrm{Bias}(\hat{M}_0^B)$ is ***smaller by exponential*** factor than $\mathrm{Bias}(\hat{M}_0^G)$.

[Experiment 1] $n \sim$ Bias



*"The average **missing mass** $\mathbb{E}[M_0]$ depends chiefly on the average **frequencies of frequencies** $\mathbb{E}[\Phi_k]$!"*

[Experiment 2] $n \sim$ MSE

n	$\hat{M}_0$	Bias	Var	MSE
100	$\hat{M}_0^G$	3.6973E-03	**2.3372E-03**	**2.3508E-03**
	$\hat{M}_0^B$	**1.0000E-200**	2.3515E-03	2.3515E-03
500	$\hat{M}_0^G$	6.6369E-05	**1.1430E-05**	**1.1434E-05**
	$\hat{M}_0^B$	**1.0000E-200**	1.1445E-05	1.1445E-05
1,000	$\hat{M}_0^G$	4.3607E-07	**4.3439E-08**	**4.3439E-08**
	$\hat{M}_0^B$	**1.0000E-200**	4.3441E-08	4.3441E-08

"The MSE of $\hat{M}_0^B$ is larger than $\hat{M}_0^G$ due to the variance terms, $\mathrm{Var}(\Phi_k)$, $\mathrm{Cov}(\Phi_k, \Phi_l)$."

How to find a minimal MSE estimator?

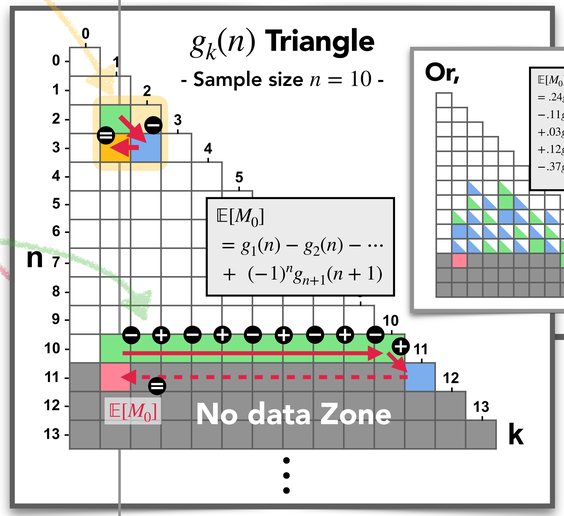
① Define a class of nearly unbiased estimators of M_0

- In the $g_k(n)$ triangle, $\mathbb{E}[M_0]$ remains invariant under any sequence of the **three mutations**.
  ⇒ ***Numerous representations of*** $\mathbb{E}[M_0]$

Three mutations

1. $g_k(n) = g_{k-1}(n-1) - g_{k-1}(n)$
2. $g_k(n) = g_k(n-1) - g_{k+1}(n)$
3. $g_k(n) = g_k(n+1) + g_{k+1}(n+1)$

+ 0. $g_k(n) = c g_k(n) + (1-c) g_k(n)$



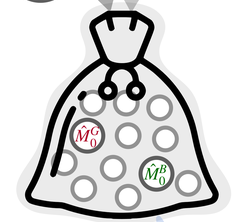
Or,

$\mathbb{E}[M_0]$
$= .24g_1(n-1) - .03g_2(n-3)$
$- .11g_2(n-4) + .5g_3(n-3)$
$+ .03g_3(n-1) + .01g_3(n-6)$
$+ .12g_3(n-2) - .01g_6(n-4)$
$- .37g_2(n-4) - .04g_4(n-7)$
⋮

Estimator form: $\hat{M}_0^c = \sum_{1 \le i \le j \le n} c_{i,j} \Phi_i(j)$

- Each representation of $\mathbb{E}[M_0]$ gives the estimator $\hat{M}_0$

$\hat{M}_0 = .12\Phi_1(X^{n+1}) - .01\Phi_3(X^{n-3}) - .01\Phi_2(X^{n-4}) + .15\Phi_3(X^{n-3}) + .01\Phi_1(X^{n-1}) + .01\Phi_6(X^{n-6}) + .02\Phi_5(X^{n-2}) - .01\Phi_4(X^{n-4}) - .07\Phi_2(X^{n-4}) - .01\Phi_4(X^{n-7})$ ⋮

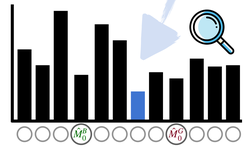


A sack 👜 of nearly unbiased estimators

② Search for the minimal MSE estimator from 👜

- Computes the MSE of each estimator in 👜 using empirically estimated $\mathbf{p}_{emp}$ [2].

$$\widehat{\mathrm{MSE}}_{emp}(\hat{M}_0^c) = \widehat{\mathrm{Var}}_{emp}(\hat{M}_0^c) + \widehat{\mathrm{Bias}}_{emp}(\hat{M}_0^c, M_0)$$
$$= \sum_{1 \le i \le j \le n} c_{i,j}^2 \widehat{\mathrm{Var}}_{emp}(\Phi_i(j)) + \sum_{\substack{1 \le i_1 \le j_1 \le n \\ 1 \le i_2 \le j_2 \le n \\ (i_1,j_1) \neq (i_2,j_2)}} c_{i_1,j_1} c_{i_2,j_2} \widehat{\mathrm{Cov}}_{emp}(\Phi_{i_1}(j_1), \Phi_{i_2}(j_2)) + \left(\hat{\mathbb{E}}_{emp}[\hat{M}_0^c] - \sum_{i=1}^{S} \hat{p}_{emp}^{n+1}\right)^2$$



[Experiment 3] Minimal MSE estimator

Dist.	$n = S/2$ $MSE(\hat{M}_0^G)$	$MSE(M_0^{Evo})$	$\hat{A}_{12}$	Ratio	$n = S$ $MSE(\hat{M}_0^G)$	$MSE(M_0^{Evo})$	$\hat{A}_{12}$	Ratio	$n = 2S$ $MSE(\hat{M}_0^G)$	$MSE(M_0^{Evo})$	$\hat{A}_{12}$	Ratio
uniform	1.09e-02	7.94e-03	0.88	72%	6.05e-03	4.29e-03	0.97	70%	1.93e-03	1.73e-03	0.96	89%
half&half	1.14e-02	7.16e-03	0.90	63%	5.46e-03	4.07e-03	0.98	74%	1.57e-03	1.42e-03	0.93	90%
zipf-1	8.09e-03	7.37e-03	0.87	91%	3.42e-03	3.04e-03	0.89	88%	1.26e-03	1.08e-03	0.94	85%
zipf-0.5	1.08e-02	8.13e-03	0.91	75%	5.23e-03	4.16e-03	0.96	79%	1.73e-03	1.54e-03	0.97	88%
diri-1	1.10e-02	7.97e-03	0.92	72%	4.36e-03	3.47e-03	0.92	79%	1.23e-03	1.05e-03	0.91	85%
diri-0.5	9.90e-03	8.02e-03	0.87	81%	3.47e-03	2.86e-03	0.88	82%	9.41e-04	8.08e-04	0.86	85%
Avg.			0.89	76%			0.93	79%			0.93	87%

*"The found estimator has **~80% of MSE** compared to the **Good-Turing estimator** $\hat{M}_0^G$."*

[1] Irving J Good. The population frequencies of species and the estimation of population parameters. Biometrika, 40(3-4):237–264, 1953.
[2] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is good-turing good. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015