Missing mass problem and Good-Turing estimator

Daniel Hsu

April 24, 2025

Let P be a probability distribution over a discrete domain $\mathcal{X}$. Let $\mathbf{S}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote an i.i.d. sample from P of size n. The *missing mass* $\mathbf{M}_n$ is the total P-mass of domain elements not appearing in $\mathbf{S}_n$:

$$\mathbf{M}_n := P(\mathcal{X} \setminus \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}).$$

Note that $\mathbf{M}_n$ is a random variable. What is a good estimator for its expectation $\mathsf{m}_n := \mathrm{E}(\mathbf{M}_n)$?

The Good-Turing estimator [1] for m_n is:

$$\hat{\mathsf{m}}_n := \frac{\text{number of domain elements appearing exactly once in } \mathbf{S}_n}{n}.$$

This can be interpreted as a "leave-one-out cross validation" (LOOCV) estimator, as discussed in [2]. To see this, re-interpret m_n as follows:

$$\begin{aligned}
\mathsf{m}_n = \mathrm{E}[\mathbf{M}_n] &= \mathrm{E}[P(\mathcal{X} \setminus \{\mathbf{x}_1, \ldots, \mathbf{x}_n\})] \\
&= \mathrm{E}\left[\sum_{x \in \mathcal{X}} P(x) \cdot \mathbb{1}\{x \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}\}\right] \\
&= \sum_{x \in \mathcal{X}} P(x) \cdot \Pr(x \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}) \\
&= \Pr(\mathbf{x}_{n+1} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}), \qquad (1)
\end{aligned}$$

where $(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1})$ is an i.i.d. sample from P of size $n+1$. So, the natural LOOCV estimator based on the expression in (1) is

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mathbf{x}_i \notin \{\mathbf{x}_j : j \in \{1, \ldots, n\} \setminus \{i\}\}\}.$$

The indicator function in the ith term of the summation takes value 1 if and only if $\mathbf{x}_i$ appears exactly once in $\mathbf{S}_n$. Therefore, this LOOCV estimator is identical to the Good-Turing estimator. (So, we see that $\hat{\mathsf{m}}_n$ is an unbiased estimator for m_{n-1}; as an estimator for m_n, it may be biased.)

References

[1] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[2] Ashwin Pananjady, Vidya Muthukumar, and Andrew Thangaraj. Just wing it: Near-optimal estimation of missing mass in a markovian sequence. *Journal of Machine Learning Research*, 25(312):1–43, 2024.