Learning-based Property Estimation with Polynomials

[Technical Report]

Anonymous Author(s)

A ALGORITHM SUPPLEMENT

A.1 Polynomial and Property estimation

We study the relationship between polynomials and the additive property. Given n elements sampled from a population of size N, we focus on the property:

$$\Psi = \sum_{j=1} \psi\left(\frac{j}{N}\right) F_j. \tag{1}$$

We divide the elements into sampled elements and unsampled elements. For the sampled elements, the frequency ratio of samples, $\frac{n_i}{n}$ is a good estimate of the true frequency ratio, $\frac{N_i}{N}$. We use $\sum_{i=1} \psi(\frac{i}{n}) f_i$ to estimate the property of the elements sampled, which is known as a plug-in estimator. Because $\frac{n_i}{n}$ is an unbiased estimation for $\frac{N_i}{N}$, $\sum_{i=1} \psi(\frac{i}{n}) f_i$ is an unbiased estimation for the sum of sampled elements' properties. The unbiasedness of plug-in estimators for discrete distributions is analyzed in [1], especially entropy.

Then we need to consider the parts that are not being sampled Considering the unsampled elements, we use a linear estimator, $\hat{\Psi}_0 = \sum_{t=1}^{L} b_t f_t$ to estimate Ψ_0. The expectation of the *frequency of frequency* of sample, f_t is:

$$\mathbf{E}\left[f_t\right] = \sum_{i=1}^{D} \binom{n}{t} \left(\frac{N_i}{N}\right)^t \left(1 - \frac{N_i}{N}\right)^{n-t}. \tag{2}$$

By putting Equation (2) into $\hat{\Psi}_0 = \sum_{t=1}^{L} b_t f_t$, we have that:

$$\begin{aligned} \hat{\psi}_0 &= \sum_{t=1}^{L} b_t f_t = \sum_{t=1}^{L} b_t \sum_{i=1}^{D} \binom{n}{t} \left(\frac{N_i}{N}\right)^t \left(1 - \frac{N_i}{N}\right)^{n-t} \\ &= \sum_{t=1}^{L} b_t \sum_{j=1} \binom{n}{t} \left(\frac{j}{N}\right)^t \left(1 - \frac{j}{N}\right)^{n-t} F_j. \end{aligned} \tag{3}$$

As with the estimate of f_0, we consider the expectation of the property of the unsampled elements,

$$\mathbf{E}[\Psi_0] = \sum_{j=1} \left(1 - \frac{j}{N}\right)^n \psi\left(\frac{j}{N}\right) F_j. \tag{4}$$

Because $\sum_{i=1} \psi(\frac{i}{n}) f_i$ is the unbiased estimation of the sampled elements, the bias of $\hat{\psi}_0 + \sum_{i=1} \psi(\frac{i}{n}) f_i$ comes from $\hat{\psi}_0$. Therefore, the error of Ψ, $\mathcal{E}_\Psi$ can be determined as follows:

$$\mathcal{E}_\Psi = \sum_{j=1} \left(\sum_{t=1}^{L} \binom{n}{t} \left(\frac{j}{N}\right)^t \left(1 - \frac{j}{N}\right)^{n-t} b_t \right) F_j - \sum_{j=1} \left(1 - \frac{j}{N}\right)^n \psi\left(\frac{i}{n}\right) F_j.$$

Merging elements with the same frequency, we have:

$$\mathcal{E}_\Psi = \sum_{j=1} \left(\sum_{t=1}^{L} \binom{n}{t} \left(\frac{j}{N}\right)^t \left(1 - \frac{j}{N}\right)^{n-t} b_t - \left(1 - \frac{j}{N}\right)^n \psi\left(\frac{i}{n}\right) \right) F_j.$$

We can also obtain an L-order polynomial approximation formula by simplifying the equation above by extracting $\left(1 - \frac{j}{N}\right)^n$, and we have the following formula:

$$\mathcal{E}_{\Psi_0} = \sum_{j=1} \left[\left(\sum_{t=1}^{L} Poly(N, n, j, t) b_t - \psi\left(\frac{j}{N}\right) \right) F_j \left(1 - \frac{j}{N}\right)^n \right], \tag{5}$$

where $Poly(N, n, j, t) = \binom{n}{t} \left(\frac{j}{N-j}\right)^t$. When N, n is fixed, the value of $Poly(N, n, j, t)$ can be directly calculated. For the arbitrary property estimation, the bias can be computed by Equation (5). For the special case, such as entropy estimation, notice that $\psi(x) = -x \log x$ in entropy estimation, so the bias of the error is:

$$\mathcal{E}_{entropy} = \sum_{j=1} \left[\left(\sum_{t=1}^{L} Poly(N, n, j, t) b_t - \frac{j}{N} \log \frac{N}{j} \right) F_j \left(1 - \frac{j}{N}\right)^n \right].$$

For the α-order power sum, the corresponding function is defined as $\psi(x) = x^\alpha$. Using the definition of ψ of power sum in Equation (5), the bias of $\alpha-$ order power sum is:

$$\mathcal{E}_{PS} = \sum_{j=1} \left[\left(\sum_{t=1}^{L} Poly(N, n, j, t) b_t - \left(\frac{j}{N}\right)^\alpha \right) F_j \left(1 - \frac{j}{N}\right)^n \right].$$

REFERENCES

[1] András Antos and Ioannis Kontoyiannis. 2001. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms* 19, 3-4 (2001), 163–193.