ECE 6980
An Algorithmic and Information-Theoretic Toolbox for Massive Data

Instructor: Jayadev Acharya — Lecture #8
Scribe: Lifan Wu — 20th September, 2016

In today's lecture, we talked about:

- Poisson sampling
- Good-Turing probability estimation and missing mass

1 Poisson Sampling

Let P be a discrete distribution, and X_1^n be the independent samples drawn from P. Let N_x be the number of times that x appears in X_1^n, then we know that $N_x \sim \text{Bin}(n, p_x)$, and hence

$$\Pr(N_x = n_x) = \binom{n}{n_x} p_x^{n_x} (1-p_x)^{n-n_x}.$$

Notice that in this case, $N_x's$ are dependent, and $\sum_x N_x = n$.

Now considering the following sampling process:

- Generate a random number $N \sim \text{Poi}(n)$.
- Generate $X_1, \ldots, X_N \sim P$.

Theorem 1. *Let $X_1, \ldots, X_N$ be the samples generated using above procedure, and N_x is the number of time that x appears in X_1^N, then*

1. *$N_x \sim \text{Poi}(np_x)$;*
2. *N_x's are independent, that is,*

$$\Pr(N_{x_1} = n_{x_1}, N_{x_2} = n_{x_2}, \ldots) = \prod_{x_i} \Pr(N_{x_i} = n_{x_i});$$

3. *Conditioned on $N = n_0$,*

$$P^{\text{Poi}(n_0)} = P^{n_0}$$

Proof. 1. Recall that the pmf of a $\text{Poi}(n)$ is:

$$\Pr(N = N^*) = e^{-n} \frac{n^{N^*}}{N^*!}.$$

Using the concept of conditional probability, we have

$$\begin{aligned}
\Pr(N_x = n_x) &= \sum_{N^* \geq n_x} \Pr(N = N^*) \Pr(N_x = n_x | N = N^*) \\
&= \sum_{N^* \geq n_x} e^{-n} \frac{n^{N^*}}{N^*!} \binom{N^*}{n_x} p_x^{n_x} (1-p_x)^{N^*-n_x} \\
&= e^{-n} \sum_{N^* \geq n_x} \frac{n^{n_x+(N^*-n_x)}}{N^*!} \frac{N^*!}{n_x!(N^*-n_x)!} p_x^{n_x} (1-p_x)^{N^*-n_x} \\
&= \frac{e^{-n}}{n_x!} (np_x)^{n_x} \sum_{N^* \geq n_x} \frac{n^{N^*-n_x}}{(N^*-n_x)!} (1-p_x)^{N^*-n_x} \\
&= \frac{e^{-n}}{n_x!} (np_x)^{n_x} \sum_{N^* \geq n_x} \frac{(n-np_x)^{N^*-n_x}}{(N^*-n_x)!} \\
&= \frac{e^{-n}}{n_x!} (np_x)^{n_x} e^{n-np_x} = e^{-np_x} \frac{(np_x)^{n_x}}{n_x!},
\end{aligned}$$

where in the last line, we use the fact that $e^x = \sum_{i \geq 0} \frac{x^i}{i!}$. We recognize that the resulting probability is the pmf of $\text{Poi}(np_x)$.

2. To show the independence, it is enough for us to show that

$$\Pr(N_x = n_x, N_y = n_y) = \Pr(N_x = n_x) \Pr(N_y = n_y).$$

Since

$$\begin{aligned}
&\Pr(N_x = n_x, N_y = n_y) \\
=& \sum_{N^* \geq n_x+n_y} e^{-n} \frac{n^{N^*}}{N^*!} \binom{N^*}{n_x, n_y, N^*-n_x-n_y} p_x^{n_x} p_y^{n_y} (1-p_x-p_y)^{N^*-n_x-n_y} \\
=& e^{-n} \sum_{N^* \geq n_x+n_y} \frac{n^{n_x+n_y+(N^*-n_x-n_y)}}{N^*!} \frac{N^*!}{n_x! n_y! (N^*-n_x-n_y)!} p_x^{n_x} p_y^{n_y} (1-p_x-p_y)^{N^*-n_x-n_y} \\
=& \frac{e^{-n}}{n_x! n_y!} (np_x)^{n_x} (np_y)^{n_y} \sum_{N^* \geq n_x+n_y} \frac{n^{N^*-n_x-n_y}}{(N^*-n_x-n_y)!} (1-p_x-p_y)^{N^*-n_x-n_y} \\
=& \frac{e^{-n}}{n_x! n_y!} (np_x)^{n_x} (np_y)^{n_y} \sum_{N^* \geq n_x+n_y} \frac{(n-np_x-np_y)^{N^*-n_x-n_y}}{(N^*-n_x-n_y)!} \\
=& \frac{e^{-np_x-np_y}}{n_x! n_y!} (np_x)^{n_x} (np_y)^{n_y} = e^{-np_x} \frac{(np_x)^{n_x}}{n_x!} e^{-np_y} \frac{(np_y)^{n_y}}{n_y!} = \Pr(N_x = n_x) \Pr(N_y = n_y),
\end{aligned}$$

we know that N_x and N_y are independent for $x \neq y$, and hence N_x's are independent.

□

Theorem 2. *If there exists an algorithm on a problem P such that $\Pr(\text{error}) < 1/4$ when using n samples, then there exists an algorithm on the same problem P such that $\Pr(\text{error}) < 1/4 + 1/16$ when using $\text{Poi}(n + 4\sqrt{n})$ samples.*

2 Good-Turing Probability Estimation

2.1 Missing mass problem

Let P be a discrete distribution, and we observe samples $X_1^n \sim P$. Let M_j be the total probabilities of symbols that appear exactly j times, i.e.,

$$M_j = \sum_x p_x \mathbb{I}(N_x = j),$$

where

$$\mathbb{I}(N_x = j) \begin{cases} 1, & \text{if } N_x = j \\ 0, & \text{otherwise} \end{cases}.$$

M_0 is the missing mass.

Example 3. *Let P be a discrete distribution on the sample space $\{a, \ldots, z\}$. Suppose we observe samples $X_1^n = abracadabra$, then*

$$\begin{aligned} M_0 &= (p(e) + p(f) + \cdots + p(z)) - p(r) \\ M_1 &= p(d) + p(c) \\ M_2 &= p(b) + p(r) \\ M_3 &= 0 \\ M_4 &= 0 \\ M_5 &= p(a) \end{aligned}$$

Theorem 4. *Let φ_j be the number of symbols in X_1^n appearing exactly j times, we know that $\varphi_j = \sum_x \mathbb{I}(N_x = j)$. Then*

$$\mathbb{E}[M_j] = \frac{j+1}{n} \mathbb{E}[\varphi_{j+1}].$$

Proof. Since

$$\begin{aligned} \mathbb{E}[M_j] = \mathbb{E}\left[\sum_x p_x \mathbb{I}(N_x = j)\right] &= \sum_x p_x \mathbb{E}\left[\mathbb{I}(N_x = j)\right] = \sum_x p_x \Pr\left(N_x = j\right) \\ &= \sum_x p_x e^{-np_x} \frac{(np_x)^j}{j!} = \frac{j+1}{n} \sum_x e^{-np_x} \frac{(np_x)^{j+1}}{(j+1)!} \end{aligned}$$

and

$$\mathbb{E}[\varphi_j] = \mathbb{E}\left[\sum_x \mathbb{I}(N_x = j)\right] = \sum_x \mathbb{E}\left[\mathbb{I}(N_x = j)\right] = \sum_x \Pr\left(N_x = j\right) = \sum_x e^{-np_x} \frac{(np_x)^j}{j!},$$

we have

$$\mathbb{E}[M_j] = \frac{j+1}{n} \sum_x e^{-np_x} \frac{(np_x)^{j+1}}{(j+1)!} = \frac{j+1}{n} \mathbb{E}[\varphi_{j+1}]$$

□

The Good-Turing probability estimator is:

$$\hat{M}_j = \frac{j+1}{n}\varphi_{j+1},$$
$$\hat{M}_0 = \frac{1}{n}\varphi_1.$$

Theorem 5. *The mean squared error of* M_0

$$\mathbb{E}\left[\left(M_0 - \frac{\varphi_1}{n}\right)^2\right] \leq \frac{1}{n}$$

Proof. Consider the MSE of M_j.

$$M_j - \frac{j+1}{n}\varphi_{j+1} = \sum_x \left(p_x \mathbb{I}(N_x = j) - \mathbb{I}(N_x = j+1)\frac{j+1}{n}\right)$$

Since $\mathbb{E}[M_j] = \frac{j+1}{n}\mathbb{E}[\varphi_{j+1}]$ and N_x's are independent under Poisson sampling,

$$\begin{aligned}
&\mathbb{E}\left[\left(M_j - \frac{j+1}{n}\varphi_{j+1}\right)^2\right] \\
=&\sum_x \mathbb{E}\left[\left(p_x \mathbb{I}(N_x = j) - \mathbb{I}(N_x = j+1)\frac{j+1}{n}\right)^2\right] \\
=&\sum_x \mathbb{E}\left[p_x^2 \mathbb{I}(N_x = j) + \frac{(j+1)^2}{n^2}\mathbb{I}(N_x = j+1)\right] \\
=&\sum_x \left(p_x^2 e^{-np_x}\frac{(np_x)^j}{j!} + \frac{(j+1)^2}{n^2}e^{-np_x}\frac{(np_x)^{j+1}}{(j+1)!}\right) \\
=&\frac{1}{n^2}\sum_x \left(e^{-np_x}\frac{(np_x)^{j+2}}{(j+2)!}(j+1)(j+2) + (j+1)^2 e^{-np_x}\frac{(np_x)^{j+1}}{(j+1)!}\right) \\
=&\frac{1}{n^2}\left((j+1)(j+2)\mathbb{E}\left[\varphi_{j+2}\right] + (j+1)^2\mathbb{E}\left[\varphi_{j+1}\right]\right).
\end{aligned}$$

Now, take $j = 0$. Then

$$\mathrm{MSE}(M_0) = \frac{1}{n^2}\left(2\mathbb{E}\left[\varphi_2\right] + \mathbb{E}\left[\varphi_1\right]\right).$$

Recall that $n = \sum_j j\mathbb{E}[\varphi_j]$, and hence $2\mathbb{E}\left[\varphi_2\right] + \mathbb{E}\left[\varphi_1\right] \leq n$. Therefore, $\mathrm{MSE}(M_0) \leq 1/n$. □