Bi/BE/CS 183 2021-2022
Instructor: Lior Pachter
TAs: Tara Chari, Meichen Fang, Zitong (Jerry) Wang

Problem Set 5 - Midterm

You are not allowed to collaborate with others for this Midterm, though you may ask clarification questions on Piazza or at Office Hours. This Midterm is open note, which includes Lecture slides and your personal notes.

Submit your solutions as a single PDF file via Canvas by **10am Friday February 11th**.

- If writing up problems by hand, please use a pen and not a pencil, as it is difficult to read scanned submission of pencil work. Typed solutions are preferred.
- For problems that require coding, Colab notebooks will be provided. Please copy and save the shared notebook and edit your own copy, which you should then submit by including a clickable link in your submitted homework. Prior to submission make sure that you code runs from beginning to end without any error reports.

Problem 1

Given a Poisson random variable X, with probability mass function $P(X = x) = e^{-\lambda}\lambda^x/x!$, show that $E(X) = \lambda$. (Hint: $e^x = \sum_{n=0}^{\infty} x^n/n!$)

A:

$$\begin{aligned} E(X) &= \sum_{0}^{\infty} xP(X = x) \\ &= \sum_{0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\ &= \sum_{0}^{\infty} \lambda\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \lambda \end{aligned}$$

Problem 2

Suppose $x = (x_1, x_2, \ldots, x_n)$ are i.i.d. observations from a Poisson random variable with unknown parameter λ.

(a) Write down the log-likelihood function $L(\lambda \mid x)$ for this set of observations.

A:

$$L(\lambda \mid x) = \sum_{i=1}^{n} (-\lambda + x_i \log(\lambda) - \log(x_i!))$$

(b) Find the maximum-likelihood estimator of λ by taking the appropriate derivative of $L(\lambda \mid x)$. (Hint: the estimator you obtain should be a function of the set of observations x)

A:

$$\frac{\partial L(\lambda \mid x)}{\partial \lambda} = -n + \sum_{i=1}^{n} x_i \frac{1}{\lambda}$$

Set the derivative to zero:

$$n = \sum_{i=1}^{n} x_i \frac{1}{\lambda} \rightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

Check the second derivative:

$$\frac{\partial^2 L(\lambda \mid x)}{\partial \lambda^2} = -\sum_{i=1}^{n} x_i \frac{1}{\lambda^2} < 0$$

Therefore, the maximum-likelihood estimator of λ is $\frac{\sum_{i=1}^{n} x_i}{n}$.

Problem 3

Consider two independent geometric random variables, X and Y, with identical probability mass function $P(X = k) = (1-p)^{k-1}p$, where $k = 1, 2, 3, \ldots$ and p is a parameter of the distribution.

(a) Show that $P(X + Y = k) = (k-1)(1-p)^{k-2}p^2$ for $k = 2, 3, \ldots$

A:

$$\begin{aligned} P(X+Y=k) &= \sum_{l=1}^{k-1}(1-p)^{l-1}p(1-p)^{k-l-1}p \\ &= \sum_{l=1}^{k-1}(1-p)^{k-2}p^2 \\ &= (k-1)(1-p)^{k-2}p^2 \end{aligned}$$

(b) We say that Z has a negative binomial distribution with parameters r, p if,

$$P(Z = z) = \binom{z-1}{r-1}p^r(1-p)^{z-r}.$$

Show that $X + Y$ has a negative binomial distribution with parameters $r = 2$ and p.

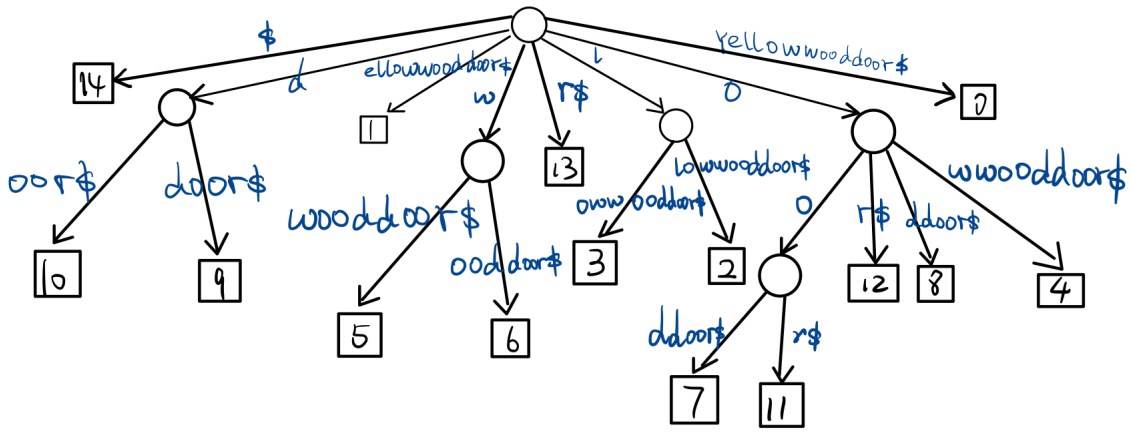
$$\begin{aligned} P(X+Y=k) &= (k-1)(1-p)^{k-2}p^2 \\ &= \binom{k-1}{1}p^r(1-p)^{k-r} \\ &= P(Z=k|r=2,p=p) \end{aligned}$$

(c) see solution notebook here

Problem 4

Construct the suffix tree for the word "Yellowwooddoor".

A:



Problem 5

For this problem you will be exploring various models which can be used to describe count data i.e. the gene-count matrices we use in single-cell genomics. The Problem 5 notebook is here.

Single-cell gene counts, which describe stochastically sampled, discrete measurements of UMI counts, are often modeled as being generated from a negative binomial (or Gamma-Poisson) distribution. However, there is a common assumption that droplet-based methods for single-cell RNA seq incur an overabundance of zeros (more zero counts) than would be predicted by random sampling. Thus it is also common to see single-cell data modeled with zero-inflated negative binomials (the ZINB distribution, with an extra parameter for the probability of zero counts). Here you will analyze these zero-inflation assumptions, following work done in [1].

The Problem 5 solution is here

References

1. Svensson, V. Droplet scRNA-seq is not zero-inflated. en. *Nat. Biotechnol.* **38,** 147–150 (Feb. 2020).