CSE 544, Spring 2017, Probability and Statistics for Data Science

Assignment 1: Probability Theory review Due: 2/15, in class

(8 questions, 70 points total)

I/We understand and agree to the following:

(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.

(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

(write down the name of all collaborating students on the line below)

---

1. Alternative expression for expectation **(Total 5 points)**

Let X be a non-negative, integer-valued RV. Prove that:

$$E[X] = \sum_{x=0}^{\infty} \Pr[X > x]$$

(Hint: One approach is to consider double summations and carefully switch the summations)

$\sum_{x=0}^{\infty} Pr[X > x]$

$\Rightarrow Pr[X>0] = P(1) + P(2) + P(3) + \cdots$

$Pr[X>1] = P(2) + P(3) + \cdots$

$Pr[X>2] = P(3) + \cdots$

$\vdots$

$Pr[X>x] = P(x) + P(x+1) + \cdots$

$P(X>0) + Pr(X>1) + Pr(X>2) + \cdots \quad Pr[X>x] = Pr(1) + 2Pr(2) + 3Pr(3) + \cdots x Pr(x) + \cdots$

$$RHS = \sum_{x=0}^{\infty} P[X>x] = \sum_{x=0}^{\infty} x \Pr(x)$$

$$= E[X] = LHS$$

2. Poisson distribution (Total 10 points)

The Poisson distribution, X ~ Poisson(λ), is a discrete distribution with p.m.f. given by:

$$p_X(i) = \frac{e^{-\lambda}\lambda^i}{i!}, i \geq 0$$

(a) Ensure that the p.m.f. adds up to 1 (2 points)
(Hint: You will need to use the infinite series expansion of an Exponential)

(b) Find E[X] (3 points)

(c) Find Var[X] (5 points)

(a) $\sum_{i=0}^{\infty} p_X(i) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$

$= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right)$

$= e^{-\lambda} e^{\lambda} = \underline{1}$

(b) $E[X] = \sum_{i=0}^{\infty} i \frac{e^{-\lambda}\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{i\lambda^i}{i!}$

$= e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!}$

$= e^{-\lambda}\left(\frac{\lambda}{0!} + \frac{\lambda^2}{1!} + \frac{\lambda^3}{2!} + \cdots\right)$

$= \lambda e^{-\lambda} e^{\lambda} = \underline{\lambda}$

(c) $Var[X] = E[X^2] - E[X]^2$

$E[X^2] = \sum_{i=0}^{\infty} i^2 \frac{e^{-\lambda}\lambda^i}{i!} = \sum_{i=0}^{\infty} \frac{i(i-1+1)e^{-\lambda}\lambda^i}{i!}$

$= \sum_{i=2}^{\infty} \frac{e^{-\lambda}\lambda^i}{(i-2)!} + \sum_{i=1}^{\infty} \frac{e^{-\lambda}\lambda^i}{(i-1)!}$

$= \lambda^2 e^{-\lambda} \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} + \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}$

$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} = \lambda^2 + \lambda$

$Var[X] = \lambda^2 + \lambda - \lambda^2 = \underline{\lambda}$

Cohesity Cohesity

3. Spark with replication **(Total 10 points)**

There are n servers organized into r racks of k servers each, such that $n = r \times k$. There are n tasks in the system, and each task has an associated data set that it must work on. Each data set is replicated f times among the n servers. Further, each server has exactly f data slots, each of which can store one data set. Thus, $(n \times f)$ data sets are distributed among the available $(n \times f)$ data slots. A task is *data-local* if it is assigned to a server that has its data set (at least one copy). A task is *rack-local* if it is assigned to a rack of servers that has its data set. The n tasks are randomly assigned to servers, such that each server has exactly one task. Further, the $n f$ data sets are also randomly assigned to the $n f$ available data slots.

(a) What is the expected number of data-local tasks? (5 points)

(b) What is the expected number of rack-local tasks? (5 points)

(a) let $X_i = \begin{cases} 1, & \text{if task on node } i \text{ is data local} \\ 0, & \text{otherwise} \end{cases}$

Thus, we have

$$E[\text{data local}] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = n\,E[X_1]$$

$$E[X_1] = Pr[\text{task on node 1 is data local}]$$

$$= Pr[\text{atleast 1 of } f \text{ copies is on } f \text{ local slots}]$$

$$= 1 - Pr[\text{none of } f \text{ copies is on } f \text{ local slots}]$$

$$= 1 - Pr[\text{all the } f \text{ copies is on rest } (nf - f) \text{ slots}]$$

$$= 1 - \frac{{}^{nf-f}C_f}{{}^{nf}C_f}$$

Thus $E[\text{data local}] = \boxed{n\left(1 - \frac{{}^{nf-f}C_f}{{}^{nf}C_f}\right)}$

(b) Similarly, $X_i = \begin{cases} 1 & \text{if task on node } i \text{ is rack local} \\ 0 & \text{otherwise} \end{cases}$

$$E[\text{rack-local}] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = n\,E[X_1]$$

$$E[X_1] = Pr[\text{task on node 1 is rack local}]$$

$$= 1 - Pr[\text{none of } f \text{ copies is on rack local}]$$

$$= 1 - Pr[\text{all of } f \text{ copies is on rest } (nf - f) \text{ slots}]$$

$$= 1 - \frac{n_f - k_f}{n_f c_f}$$

Thus, $E[\text{rack-local}] = n\left(1 - \frac{n_f - k_f c_f}{n_f c_f}\right)$

4. Pokémon Go fanatic **(Total 10 points)**

Let us assume there are only *n* distinct types of Pokémons to capture in the entire Pokémon world, though there is an infinite supply of each type. Every day, you capture exactly one Pokémon. The Pokémon that you capture could be any one of the *n* types of Pokémons with equal probability. Your goal is to capture at least one Pokémon of all *n* distinct types. Let X denote the number of days needed to complete your goal.

(a) What is E[X]? (5 points)

(b) What is Var[X]? (5 points)

We do not need closed-forms here for parts (a) and (b).

We model this problem as sum of geometric random variables which are independent

$$X = t_1 + t_2 + t_3 + \dots + t_d$$

Each t_i represents the no of days to get the i^{th} new pok

(a) So, $E[t_1] = 1$ ∵ any pokemon on 1st day = new pokemo

If p_i = Pr of success (we get new pokemon on any da

⟶ if we already have found $(i-1)$ distinct poke

$$p_{i+1} = \frac{d-i}{d}$$

∵ on 2nd day; no of new pokemons $= (d-1)$

Each of these t_i, are geometric random variables with probability of success p_i

$$E[t_i] = \frac{1}{p_i}$$

$$E[X] = \sum_{i=1}^{d} E[t_i]$$

$$E[X] = 1 + \frac{d}{d-1} + \frac{d}{d-2} + \dots \frac{d}{1}$$

$$E[X] = d\left(1 + \frac{1}{2} + \frac{1}{3} + \dots \frac{1}{d-2} + \frac{1}{d-1} + \frac{1}{d}\right)$$

(b) $\text{Var}(t_i) = \frac{1-p_i}{p_i^2} = \frac{d(i-1)}{(d+1-i)^2}$

All of the t_i are independent of each other.
They depend on 'i' → (no of pokemons that are already found)

But, no of days to get new pokemon is independent

$$\text{Var}(X) = \sum_{i=1}^{d} \text{Var}[X_i]$$

$$= d \sum_{i=1}^{d} \frac{i-1}{(d+1-i)^2}$$

$$= d\left[\frac{1}{(d-1)^2} + \frac{2}{(d-2)^2} + \cdots + \frac{d-1}{1}\right]$$

4. 5. **Seating capacity planning** (Total 5 points)

An instructor for CSE 999 knows that on average every student who enrolls for the class has a 10% chance of dropping the class before the 1st day of the lecture. For the Spring 2017 edition of the class, 60 students enroll initially.

(a) How many students will end up staying in the class, on average? (2 points)

(b) If the instructor picks a class with only 50 seats, what is the probability that there will be a seat available for every student who ends up staying in the class? (3 points)

(i) 10% chance of dropping the class.

$\Rightarrow Pr(\text{drop}) = \frac{10}{100} = \frac{1}{10}$

$\Rightarrow Pr(\text{enroll}) = \frac{9}{10}$

Number of students = 60

No of students that end up staying in class on avg
= Expected value of number of students taking the class
$= E[X]$
$= \sum_{x=0}^{60} x\, Pr(x)$

$Pr(x)$ →
- $Pr(\text{all drop}) = \left(\frac{1}{10}\right)^{60}$
- $Pr(1 \text{ enroll}) = {}^{60}C_1 \left(\frac{9}{10}\right)^1 \left(\frac{1}{10}\right)^{59}$
- $Pr(2 \text{ enroll}) = {}^{60}C_2 \left(\frac{9}{10}\right)^2 \left(\frac{1}{10}\right)^{58}$
- ⋮
- $Pr(\text{all enroll}) = \left(\frac{9}{10}\right)^{60}$

$\Rightarrow Pr(x) = {}^{60}C_x \left(\frac{9}{10}\right)^x \left(\frac{1}{10}\right)^{60-x}$

$\Rightarrow E[X] = \sum_{x=0}^{60} x\ {}^{60}C_x \left(\frac{9}{10}\right)^x \left(\frac{1}{10}\right)^{60-x}$

We know, Binomial (n,p) with $Pr(i) = {}^{n}C_i\, p^i (1-p)^{n-i}$ has $E[X] = np$.

$= 60 \times \frac{9}{10} = \underline{54}$

(b) If class has 50 seats.

then probability that there is a seat for every student

= Probability that atleast 10 people drop

= 1 - Probability that atleast 51 people stay

$$= 1 - \sum_{x=51}^{60} {}^{60}C_x \left(\frac{9}{10}\right)^x \left(\frac{1}{10}\right)^{60-x}$$

=

6. **Nerdy NBA** (Total 10 points)

In the 2016 NBA Western Conference finals, Golden State Warriors (GSW) played the Oklahoma City Thunder (OKC) in a best-of-7 series where the first team to win 4 games clinches the Western Conference. Assume that the outcome of each game is independent.

(a) At the end of the regular season, GSW had an 89% win percentage and OKC had a 67% win percentage. Assuming that these are the probabilities of each team winning each game (that is, GSW wins any game w.p. 0.89 and OKC w.p. 0.67; ignore the fact that these don't add up to 1), what is the probability that after the first 4 games, OKC would be up 3-1? Clearly show all your steps. (2 points)

(b) OKC was, in fact, up 3-1 at the end of 4 games. Now assume that the probability of either team winning a game is 0.5. Starting from 3-1, for subsequent games, draw the decision tree; note that if a team ends up winning 4 games total, subsequent games will not be held. (5 points)

(c) Using the decision tree, compute the probability of OKC clinching the Western Conference finals and that of GSW clinching the Western Conference finals. (3 points)

(d) Optional: If you are an NBA fan, who did you support in this series? (0 points)

a) $P_{win}(GSW) = 0.89$

$P_{win}(OKC) = 0.67$

At the end of 4 games we have OKC - 3
GSW - 1

This implies that of 4 games OKC should win any three and GSW any one.

The way / possible ways to $= {}^4C_1 = {}^4C_3 = 4$

There can be total of 4 ways in which we can achieve 3-1 score.

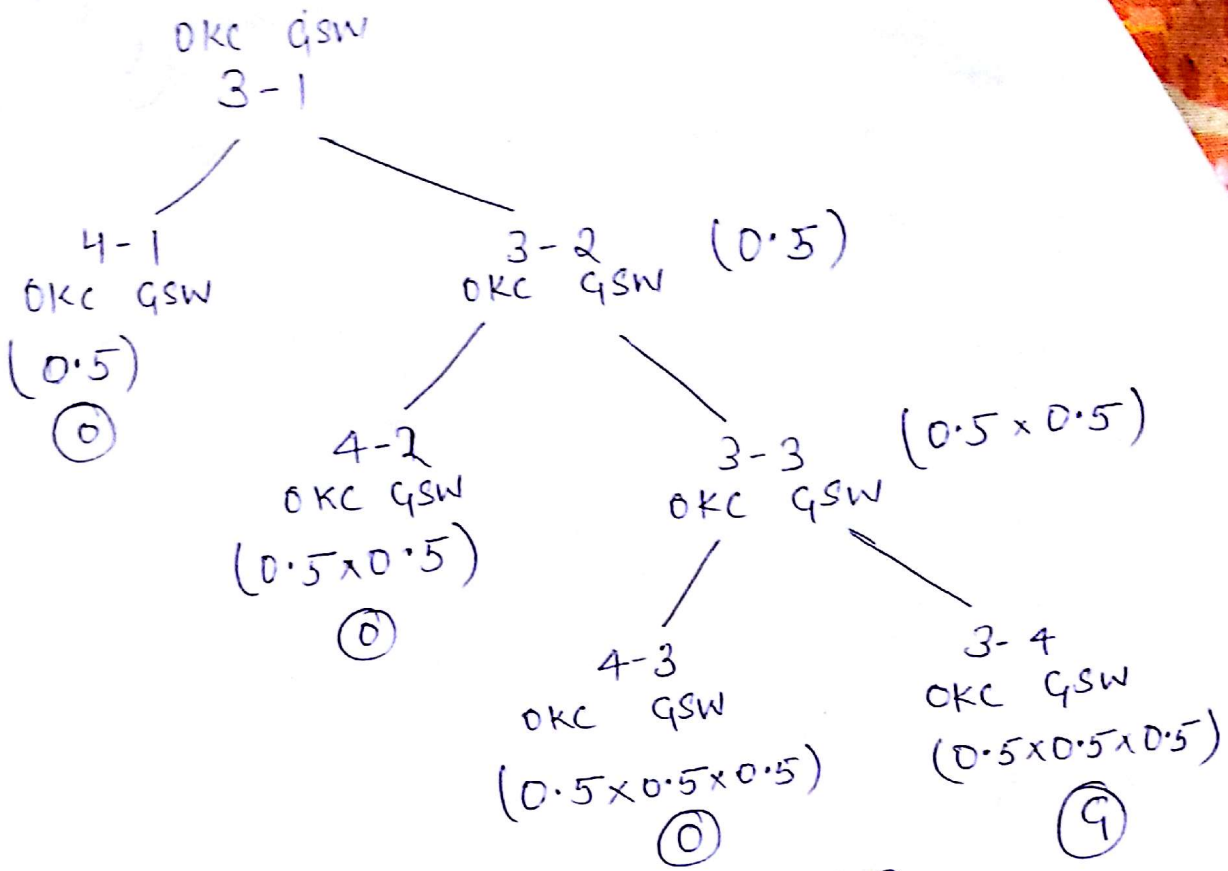
Now The probability would be:-

$${}^4C_1 \cdot (P_{win}(GSW))^1 (P_{win}(OKC))^3$$

$$= {}^4C_1 (0.89)(0.67)^3$$

$$= 1.0707 \; //$$

(b)



(c)

$P_{WIN}(GSW)$ = sum of all paths that leads to (9).

$= 0.5 \times 0.5 \times 0.5$

$P_{WIN}(GSW) = 0.125$

$P_{WIN}(OKC)$ = Sum of all paths that leads to (0)

$= 0.5 + (0.5 \times 0.5) + (0.5 \times 0.5 \times 0.5)$

$= 0.5 + .25 + .125$

$P_{WIN}(OKC) = .875$

(d) Cleveland Cavaliers.

(-1)

7. *Cubs* win World Series after 108 years! **(Total 10 points)**

In the finals of the 2016 Major League Baseball (MLB) World Series, Cleveland Indians (CLE) faced off against the eventual champions, Chicago Cubs (CHC), in a best-of-7 series. CLE had a 94-67 (58%) win-loss record and CHC had a 103-58 (64%) record in the games before the finals.

(a) Assume that the probability of CLE winning any game is 58/(58+64), and that of CHC is 64/(58+64). What is the probability that the winners of the 7 games, in sequence, are CLE, CHC, CLE, CLE, CHC, CHC, CHC? (2 points)

(b) Now assume that the probabilities are updated after every game based on the above formula. That is, if CLE wins game 1, its win record goes up to 95-67 (58.6%) and that of CHC goes down to 103-59 (63.6%); this then affects the probabilities for game 2. What is the answer to (a) in this case? (5 points)

(c) Assume it is now 1908, and Cubs just won the World Series. Assuming that each of the 30 teams has a 1/30 chance of winning each year independently, what is the probability that Cubs won't win again till 2016? (1 points)

(d) Assume it is the end of 1908, and the probability of CHC winning the World Series every subsequent year goes down by 2 if they don't win that year. Start with a win probability of 0.5. Thus, the probability of CHC winning in 1909 is 0.5, but will become 0.25 for 1910 if they lose in 1909. What is the probability that Cubs will win their next title only after 108 years in 2016? (2 points)

(e) Optional: Which team has won the most World Series titles thus far? (0 points)

(a) $P(CLE) = \frac{58}{58+64}$

$P(CHC) = \frac{64}{58+64}$

Prob of winning 7 games in sequence

$= \prod$ Prob of winning each game

$= Pr(CLE)\,Pr(CHC)\,Pr(CLE)\,Pr(CLE)\,Pr(CHC)\,Pr(CHC)\,Pr(CHC)$

$= \left(\frac{58}{58+64}\right)\left(\frac{64}{58+64}\right)\left(\frac{58}{58+64}\right)\left(\frac{58}{58+64}\right)\left(\frac{64}{58+64}\right)\left(\frac{64}{58+64}\right)\left(\frac{64}{58+64}\right)$

$= 0.008137$

(b) Probabilities are updated after every game.

Initially,

	Win loss record	
	CLE	CHC
① CLE	94-67 (58%)	103-58 (64%)
② CHC	95-67 (58.6%)	103-59 (63.6%)
③ CLE	95-68 (58.2%)	104-59 (63.8%)
④ CLE	96-68 (58.5%)	104-60 (63.4%)
⑤ CHC	97-68 (58.8%)	104-61 (63%)
⑥ CHC	97-69 (58.4%)	105-61 (63.3%)
⑦ CHC	97-70 (58.1%)	106-61 (63.5%)

Prob of sequence

$= \left(\frac{58}{58+64}\right)\left(\frac{63.6}{58.6+63.6}\right)\left(\frac{58.2}{58.2+63.8}\right)\left(\frac{58.5}{58.5+63.4}\right)\left(\frac{63}{63+58.8}\right)\left(\frac{63.3}{63.3+58.4}\right)\left(\frac{63.5}{63.5+58.1}\right)$

$= 0.00079$

(c) $Pr(win) = \left(\frac{1}{30}\right) \Rightarrow Pr(loss) = \left(\frac{29}{30}\right)$

$$P\begin{pmatrix}\text{Cubs won't}\\ \text{win till 2016}\end{pmatrix} = \prod_{i=1}^{108} Pr_i(loss)$$

$$= \left(\frac{29}{30}\right)^{108}$$

$$= \left(\frac{29}{30}\right)^{107} \cdot \frac{1}{30}$$

$\because$ Win each year is
$\Rightarrow$ Loss each year is
$\Rightarrow Pr(A \cap B) = Pr(A) \cdot$

(d) Initial winning probability is 0.5 in year 1909

What we need is a losing streak from 1909–2015 and win in

If we construct a decision tree of all the possibilities then we will get the answer at the end of the path which look like:-

1909	1910	1911	.....	2015	2016
lose	lose	lose		lose	win

Now as given in question if we try and calculate probability at leaf node it would be:-

1909	1910	1910	.....	2015	2016
$(1-0.5)$	$(1-(0.5)^2)$	$(1-(0.5)^3)$		$(1-(0.5)^{107})$	$(0.5)^{108}$
$= (1-P(win)$	$(1-P(win)$	$(1-P(win))$		$(1-P(win))$	$P(win)$

(given P(win) reduces by half every subsequent year if we lose)

$\therefore$ P(winning title only after 108 years)

$$= (1-0.5)(1-0.5^2)(1-0.5^3)(1-0.5^4) \cdots (1-0.5^{107})(0.5)^{108}$$

e) New York Yankees. ∴

8. The One Ring? **(Total 10 points)**

Bilbo Baggins of the Shire has a ring, similar to one of about 10,000 rings that exist in Middle Earth. Gandalf the Wizard, however, fears that it may, in fact, be the One Ring!

(a) If the ring is the One Ring, there is a 90% chance that the owner will have an above-average lifespan. If the ring is not the One Ring, there is a 90% chance that the owner will not have an above-average lifespan. What is the probability that, given Bilbo is pushing 111 years (above-average for Hobbits), his ring is, in fact, the One Ring? (3 points)

(b) Since Gandalf wants to be absolutely sure, he administers another independent test and throws the ring into a fireplace. If it is the One Ring, writing will appear on it with probability 0.9; if it is not the One Ring, writing may still appear on it with probability 0.02. Given that writing appears on it, and that Bilbo has an above-average lifespan, what is the probability that this is, in fact, the One Ring? (7 points)

let

(a) A = avg life span

R = One ring

$P(A|R) = 0.9$ $\quad P(\bar{A}|\bar{R}) = 0.9$ $\quad P(R) = 10^{-4}$

$P(\bar{A}|R) = 0.1$ $\quad P(A|\bar{R}) = 0.1$ $\quad \Rightarrow P(\bar{R}) = 0.9999$

then, $$P(R|A) = \frac{P(A|R)P(R)}{P(A|R)P(R) + P(A|\bar{R})P(\bar{R})}$$

$$= \frac{0.9 \times 10^{-4}}{0.9 \times 10^{-4} + 0.1 \times 0.9999}$$

$$= 8.99 \times 10^{-4}$$

(b) let W = writing appears on it

$P(W|R) = 0.9$ $\quad P(\bar{W}|R) = 0.98$

$P(\bar{W}|R) = 0.1$ $\quad P(W|\bar{R}) = 0.02$

$\frac{1}{10000}$

Since A & W are indep of tests,

we have $$P(R|WA) = \frac{P(WA|R)P(R)}{P(WA)}$$

$$= \frac{P(WA|R)P(R)}{P(W)P(A)}$$

→ 41 yrs

→ 111 | yrs = 09/10000

$\frac{1}{10000} \times 0.9 + 0.$

Since A & W are independent tests.
& accuracy of test A/R has nothing to do with accuracy of test W/R
A/R & W/R are conditionally independent.

$$\text{so, } P(R|WA) = \frac{P(W|R)\,P(A|R)\,P(R)}{P(W)\,P(A)}$$

Similarly we have

$$P(W) = P(W|R)P(R) + P(W|\bar{R})P(\bar{R})$$

$$= 0.9\times10^{-4} + 0.02 \times 0.9999$$

$$= 0.020088$$

$$\text{thus, } P(R|WA) = \frac{0.9\times0.9\times10^{-4}}{0.020088 \times 0.10008}$$

$$\boxed{P(R|WA) = 0.04}$$