CSE 544, Spring 2020, Probability and Statistics for Data Science

Assignment 1: Probability Theory review Due: 2/12, in class

(7 questions, 75 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. Nerdy NBA (Total 15 points)

In the 2019 NBA Eastern Conference semi-finals, Philadelphia 76ers (PHI) played the (eventual 2019 NBA Champion) Toronto Raptors (TOR) in a best-of-7 series where the first team to win 4 games wins the series. Assume that the outcome of each game is independent.

(a) Assuming that either team has a win probability of 0.5, what is the probability that after the first 4 games, the teams would be tied 2-2? Clearly show all your steps. (2 points)

(b) PHI-TOR were, in fact, tied 2-2 at the end of 4 games, making it effectively a best-of-3 for the remaining games. Assuming the either team has a 0.5 probability of winning each game, draw the decision tree for the subsequent games starting from 2-2; note that if a team ends up winning 4 games total, subsequent games will not be held. (3 points)

(c) Using decision tree, compute the probability of TOR winning the series 4-3. TOR in fact did win the series 4-3 at the last second, resulting in one of the most thrilling sports moments of 2019. (1 point)

(d) Repeat part (b), but now with the assumption that the home team has a 0.75 probability of winning the game. Game 5, 6, and 7 were held in TOR, PHI, and TOR, respectively. (3 points)

(e) Repeat part (c), but using the decision tree of part (d) (1 point)

(f) The frequentist interpretation of probability based on a large number N of repetitions of an experiment is $P(A) \approx \frac{N_A}{N}$, where N_A is the number of times A occurs and N is the total number of times the experiment is repeated. Similarly, the conditional probability $P(B \mid A) \approx \frac{N_{BA}}{N_A}$, where N_{BA} is the number of times $B \cap A$ occurs and N_A is the number of times A occurs. Use simulations (coded in Python) to verify the results of part (a), (c) and (e). For instance, one can let A be the event that there is a tie after 4 games and B be the event that TOR wins the game 4-3. Then we can simulate "series", a sequence of games until one of the teams wins 4 games, N times. Among those N repetitions of series, one can compute N_A, the number of times there is a tie after 4 games, N_{BA}, number of times there was a tie after 4 games and TOR won 4-3. Finally, we can approximate $P(A) \approx \frac{N_A}{N}$ and $P(B|A) \approx \frac{N_{BA}}{N_A}$. Try $N = 10^n$ for $n = 3, 4, 5, 6, 7$. What do you observe as N increases?

Hint: In Python, `numpy.random.binomial(1, p)`can simulate a Bernoulli trial with probability p.

For this programming assignment, you should **submit, via email, a Python script** named `nba.py` to the TA Supreeth (email on class website); in the email, list all your group member names and SBU

IDs. Combine the email submission with that required for Q2(b) and have the subject as “CSE 544 A1 code submission”. The script should have a variable named N, the number of times the experiment is repeated, at the very beginning of the program so that TAs can try out different values for N. The program should print the results of part (a), (c) and (e) as follows:

```
For N = ..., the simulated value for part (a) is ...
For N = ..., the simulated value for part (c) is ...
For N = ..., the simulated value for part (e) is ...
```

You should **also report the answers in your hard-copy assignment submission**. (5 points)

2. Free yourself (Total 15 points)

In the near future, you realize that you have spent far too much money on buying and hoarding phones and decide to rid yourself of all your hoarded iPhones. Turns out you have n iPhones, with each iPhone belonging to a unique generation from iPhone 1 to iPhone n. So, to cleanse your digital life, you play a risky game. In step 1, you randomly pick an iPhone from this pile; if the selected iPhone is iPhone 1, you keep it, else you discard it. In step 2, you again randomly pick an iPhone from the remaining (n-1) iPhones and if the selected iPhone is iPhone 2, you keep it, else you discard it. You repeat this immensely satisfying exercise n times. We would like to find out, at the end of this exercise, what is the probability that you have at least one undiscarded iPhone?

(a) Solve this problem using the principle of inclusion-exclusion (PIE). For n events E_1, E_2, …, E_n, PIE says

$$\Pr(\bigcup_{i=1}^{n} E_i) = \sum_i \Pr(E_i) - \sum_{i<j} \Pr(E_i \cap E_j) + \sum_{i<j<k} \Pr(E_i \cap E_j \cap E_k) - \ldots + (-1)^{n+1} \Pr(E_i \cap \ldots \cap E_n).$$

Choose the events E_1, E_2, …, E_n carefully so that you can obtain the required probability. (8 points)

(b) Now solve this problem via simulation (using Python). We first need to simulate a pile of n iPhones. Let's do this using Python's `range(1, n+1)`, which returns a list of integers from 1 to n. One can then simulate the action of picking phones randomly by shuffling the list using `random.shuffle()`and then choosing elements of the shuffled list one by one.

Write a Python script which has two variables n and N (be sure to define them at the very beginning so that TAs can play with different values) and simulate the above problem. Set $(\mathrm{n}, \mathrm{N}) :=$ $(5, 10\text{^}2), (5, 10\text{^}3), (5, 10\text{^}4), (5, 10\text{^}5)$ and $(\mathrm{n}, \mathrm{N}) := (20, 10^2), (20, 10^3), (20, 10^4), (20, 10^5)$. **Report** the answers for all of them in the hard-copy assignment submission. The program should also print the output as follows:

```
For (n, N) = (.,.), the approximated value for probability is ...
```

Name the script as `deck.py` and **submit the file via email** to the TA Supreeth. Also include all group members' names and SBU IDs in the email; send one email per group to Supreeth, containing two attachments, nba.py for 1(f) and deck.py for 2(b). Use email subject as "CSE 544 A1 code submission". (7 points)

3. Fun with Bayes (Total 10 points)

To give yourself a break after assignment 1, you head to Atlantic City to play slots. Turns out there are only two slot machines, a Red and a Green. You somehow know that one of these machines has a 10% chance of jackpot (bad machine) and the other has 50% (good machine); unfortunately, you do not know which is which. Assume each play of the machines is independent and also conditionally independent of the results of prior plays.

(a) You play the Red machine and don't get a jackpot. What is the probability that Red is the bad machine? (3 points)

(b) What is the probability that the Green machine is good given that you played the Red machine twice and won both times? (3 points)

(c) If you play only the Red machine several times and fail to get a jackpot every time, after how many times would you be at least 95% confident that Red is bad? (4 points)

4. The One Ring? (Total 10 points)

Bilbo Baggins of the Shire has a ring. It is known that there are only 10,000 rings in Middle Earth. Gandalf the Wizard, however, fears that Bilbo's ring may, in fact, be the One Ring!

(a) If the ring is the One Ring, there is a 95% chance that the owner will have an above-average lifespan. If the ring is not the One Ring, there is a 75% 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) To be absolutely sure, Gandalf administers another 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 0.05. Given that writing appears on it, and that Bilbo has an above-average lifespan, what is the probability that this is the One Ring? Assume that the tests are independent conditioned on the ring being the One Ring, and the tests are independent conditioned on the ring not being the One Ring. Do not assume that the tests are independent. (7 points)

5. 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)

6. 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)

7. Pareto distribution (Total 10 points)

The Pareto distribution, X ~ Pareto(α), 1 < α < 2, is a continuous distribution with p.d.f. given by:

$$f_X(x) = \alpha x^{-\alpha-1}, x \geq 1$$

(a) Ensure that the p.d.f. integrates to 1 (2 points)

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

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