DATA/MSML602: Principles of Data Science
Midterm

November 8th, 2022

Instructions

You have until 9pm on November 9th to finish and submit your solutions for the following problems. Solve and write your solutions by hand unless otherwise indicated. Be sure to read all instructions carefully and present your results in a clear and concise manner. You will be graded on both your correctness and clarity. All of the following questions will be weighted equally.

If you have any questions / if the instructions for a problem are unclear, please contact the TA (Max Springer, mss423@umd.edu).

You are not permitted to work with other students, and any indication of collaboration can result in a zero for the exam.

Problem 1 (Probability & Combinatorics) *An urn contains 30 red balls and 70 green balls. What is the probability of getting exactly* k *red balls in a sample of size 20 if sampling is done* ***with replacement****. How does your solution change when sampling* ***without*** *replacement. Assume* $0 \leq \mathsf{k} \leq 20$. *(You can set up the solution without explicitly calculating the result.)*

Proof: Sampling with replacement:

$$\mathbf{Pr}(\mathsf{k} \text{ red balls}) = \binom{20}{\mathsf{k}}(0.3)^{\mathsf{k}}(0.7)^{20-\mathsf{k}}$$

Sampling without replacement: Let A be the event of getting exactly k red balls. To find $\mathbf{Pr}(A) = \frac{|A|}{|S|}$, we need to find $|A|$ and $|S|$. First, note that $|S| = \binom{100}{20}$. Next, we calculate

$$|A| = \binom{30}{\mathsf{k}}\binom{70}{20-\mathsf{k}}$$

and thus, we have

$$\mathbf{Pr}(A) = \frac{\binom{30}{\mathsf{k}}\binom{70}{20-\mathsf{k}}}{\binom{100}{20}}$$

■

Problem 2 (Probability & Combinatorics) *Suppose you have 10 pairs of socks with each pair being a different color. You put them all in the washing machine. The washing machine eats four socks at random. What is the expected value for the number of complete pairs that make it out alive?*

Proof: Let X be the number of complete pairs that survive the washing machine. Let us number the pairs $1, 2, ..., 10$. Let us write $X = X_1 + ... + X_{10}$, where $X_i = 0$ if the washing machine eats at least one sock in the pair i.

We now compute $\mathbf{E}(X_i) = 0 \times \mathbf{Pr}(X_i = 0) + 1 \times \mathbf{Pr}(X_i = 1)$. To compute $\mathbf{Pr}(X_i = 1)$ look at the first sock in pair i. Its chance of survival is $\frac{16}{20}$. If the first sock is eaten, then we don't care about the second (the pair is no longer in tact), but if the first survives then we have 19 socks left. The washing machine must eat 4 of these 19, so the second sock survives with probability $\frac{15}{19}$. Therefore, the probability both socks survive is $\frac{16}{20} \times \frac{15}{19} = \frac{12}{19}$.

$$\mathbf{E}(X) = \mathbf{E}(X_1) + ... + \mathbf{E}(X_{10}) = 10 \times \frac{12}{19} \approx 6.32$$

■

Problem 3 (Linear Programming) *Calculate a solution of the following linear program. Validate your answer by writing a Python script that solves for the optimal (integer) solution.*

$$\begin{aligned} \max_{x,y} \quad & 5x + 3y \\ s.t. \quad & x + 2y \leq 14 \\ & 3x - y \geq 0 \\ & x - y \leq 2 \end{aligned}$$

Proof: Using the the third constraint, we have that

$$x - 2 \leq y$$

and plugging this into the first constraint we obtain

$$3x - 4 \leq 14 \Rightarrow x \leq 6$$

Since we are seeking to maximize the objective function, we set $x = 6$ and force the constraints to be

$$\begin{aligned} & y \leq 4 \\ & y \leq 18 \\ & y \geq 4 \end{aligned}$$

and the only value of y that abides all constraints is $y = 4$. Therefore, our optimal value is $5(6) + 3(4) = 42$. ■

Problem 4 (Linear Programming) *Suppose there are n companies $(s_1, ..., s_n)$ where each company s_i has expected net worth of c_i (i.e. buying the stocks of the entire company would cost c_i dollars) and investing in the company has expected profit w_i. Suppose we have a budget B and would like to use the budget to buy stocks in a matter that maximizes our expected profit. Write, but do not solve, a linear program that solves this problem. Be sure to clearly define and explain your objective, constraints, and any introduced variables.*

Proof:

$$\begin{aligned}
\max \sum_{i=1}^{n} & w_i \cdot x_i \\
\text{s.t.} \ \sum_{i=1}^{n} & c_i \cdot x_i \leq B \\
& x_i \geq 0 \text{ for all } i \in [n] \\
& x_i \leq \ \text{ for all } i \in [n]
\end{aligned}$$

■

Problem 5 (Graph Metrics) *Using the* SNAP *package in Python, load the LastFM Asia Social Network dataset. Using* NetworkX, *compute the average shortest path length to all other nodes in the network from the first node in the graph. Submit your solution as a Jupyer (or Google CoLab notebook).*

Proof: See Google CoLab notebook for example solution. ■

Problem 6 (Extracting Webpage Data) *For this problem, you will scrape a live webpage for data on COVID cases. Access the website*

https://www.worldometers.info/coronavirus/#countries

Parse the table of countries data on COVID infections and report the average number of active cases, followed by the proportion of the total population that is currently infected. Submit your answer as a Jupyter (or Google CoLab notebook) on ELMS. Your notebook should demonstrate that you have obtained the table of information from the website and subsequently the requested data.

Proof: See Google CoLab notebook for example solution. ■

Extra Credit

Problem 7 *Explain the recent feud between Meta and Apple. What are the motivations of each party?*

Problem 8 *Describe SSP and DSP and their networks. Explain how display ad networks work.*