Homework Assignment 4 – Due Wednesday Dec 16

CMPSC 190DD/197A Project, Fall 2020

Question 1 (Code) (15 points)

You are given two seemingly unrelated tables:

- `table1.csv` consists of columns: Age, Occupation, Education, Education-Num (how many years of education), Marital Status, Sex, Race
- `table2.csv` consists of columns: Native country, Age, Sex, Occupation, Race, Capital-gain, Capital-loss, Work hours per week

Using **only** Table 1 or **only** Table 2, see if you can identify the education and marital status of a woman from the Philippines (Asian) that is working in "Craft-repair" sector. What if you use **both** Table 1 and Table 2?

Question 2 (Code) (15 points)

See `HW4 Notebook.ipynb` Jupyter notebook for this question.

Question 3 (Pen&Paper or Code) (30 points)

(a) You can solve this question either on paper or through code (on paper is probably easier). Suppose we have 100 records of people with two fields:

- age
- a sensitive information on a trait (modeled as a 0/1 value)

Suppose that the number of records with the sensitive trait = 1 is 20. The sum of ages of all people with trait = 1 is 1000 and the sum of ages of people with trait = 0 is 4500. We wish to publish the average age of the **people with trait = 0**. While doing this, we would like to maintain the privacy of records with a differentially privacy parameter ϵ. We will achieve this by adding Laplacian noise with parameter b onto the average we publish. Remember what differential privacy is trying to achieve: if an observer looks at the average we publish they should not be able to tell if a particular individual's information was used in the computation. Let's go step by step to find what level of noise (determined by b) we need in order to ensure differential privacy with level ϵ. We will assume that ages in this database lie in the range $[40, 100]$.

1. Determine what is the highest effect a single individual can do on this average. You need to consider the following (in all cases, use the age range):
    - What is the minimum/maximum average would be if an individual in the database were excluded?
    - What is the minimum/maximum average would be if an individual outside the database were included?

- What is the minimum/maximum average would be if the age of an individual that is in the database was different?

2. The above hypothetical scenarios each determine a pair of minimum/maximum averages. Essentially, we want to add noise such that by looking at the average we provide (with the noise added), an observer cannot determine which of these scenarios were true within a reasonable doubt. The amount of "doubt" is determined by the differential privacy parameter ϵ. On a piece of paper, mark these 7 scenarios for the average (including the original average) on the real line.

3. Make the real line above the x-axis and draw a perpendicular y-axis. Then draw 7 Laplacian distributions centered around each of these averages. You don't need to worry about the parameter b, just do a qualitative sketch but keep the distributions same (except for their means of course). See Figure 1.

4. Now identify the distribution that is centered at the true average and consider the difference in distribution values between all other distributions and the original distribution. We want to choose our noise level b such that for any value on the x-axis we can say the ratio between a hypothetical scenario and the original scenario is no more than e^{ϵ} and no less than $e^{-\epsilon}$ (see Figure 1). Analytically write out all these ratios and do the necessary cancellations. Then for each ratio, determine a b_i such that the ratio condition is satisfied for $\epsilon = 0.1$.

5. We want all of the conditions to be satisfied, but we also want the minimum such b in order to minimize the pollution in our published statistics. Therefore we take the largest of b_i's, which correspond to the strictest condition, to be the noise level b for our Laplacian noise.

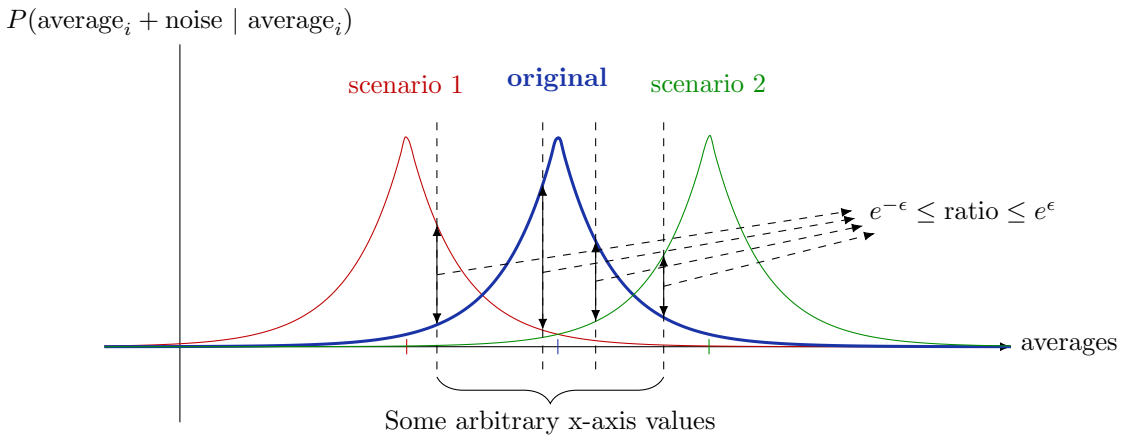

Figure 1: Visual explanation of how noise should be added to ensure differential privacy

Let's look at what we have achieved. When we publish the value of the average (with noise) as x, if a participant says: "Hey, you're revealing information about me. If I wasn't accounted for, the statistics would have been $x - \Delta x$", we can respond, "Umm, actually, due to noise it is not possible to determine for sure which of the scenarios above was true. In fact, these are the edge case scenarios so what you think the information about you being revealed is probably actually even less statistically significant". Essentially, the use of differential privacy allowed us to ensure that no single individiual is too responsible for the given statistics; they all collectively are.

(b) Repeat Part (a) when the number of records is scaled by a factor of 10 and all other statistics remain the same. (So sums are also multiplied by 10). What do you notice?

Question 4 (Code) (20 points)

See `HW4 Notebook.ipynb` Jupyter notebook for this question.

Question 5 (Pen&Paper) (20 points)

Consider the Causal Graphical Network below:

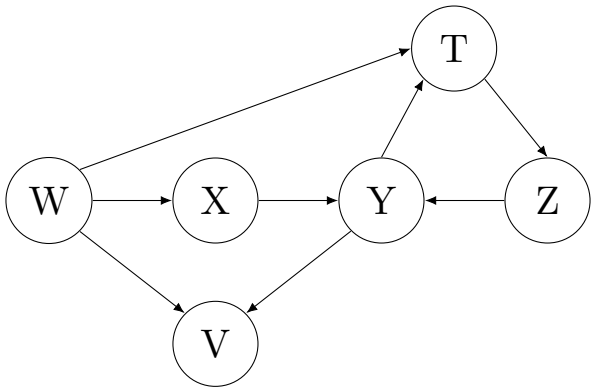

Figure 2: Causal graphical network

(a) Are variables W and Z independent?

(b) Are W and Z conditionally independent given Y?

(c) Are W and Z conditionally independent given X?

(d) Repeat items 1–3 under the condition do(T = 1).