DS-UA 202, Responsible Data Science, Spring 2024

Homework 3: Privacy and Data Protection
Due at 11:59pm EDT on Monday, April 22

Objectives and Learning Outcomes

This assignment consists of written problems and programming exercises on the data science lifecycle and data protection. In the programming part of the assignment you will use the DataSynthesizer library for privacy-preserving synthetic data generation.

After completing this assignment, you will:

1. explore the interaction between the complexity of the learned model (a summary of the real dataset) and the accuracy of results of statistical queries on the derived synthetic dataset, under differential privacy
2. understand the variability of results of statistical queries under differential privacy, by generating multiple synthetic datasets under the same settings (model complexity and privacy budget), and observing how result accuracy varies
3. explore the trade-off between privacy and utility, by generating and querying synthetic datasets under different privacy budgets, and observing the accuracy of the results

You must work on this assignment individually. If you have questions about this assignment, please post a private message to all instructors on Piazza.

IMPORTANT: Make sure to include all figures, graphics, etc. in the report. If a question is just asking you to execute coding instructions, include a short description in plain English of what your code does in the report – do not include any code in the report.

When uploading to Gradescope, be sure to tag your pages!

Grading

The homework is worth 50 points, or 10% of the course grade. Your grade for the programming portion will be significantly impacted by the quality of your written report for that portion. In your report, you should explain your observations carefully.

You cannot use any late days for this assignment. We will discuss solutions in class / lab on Tuesday April 23/25/26, so you can start preparing for the final exam that will take place the following week. **If you submit HW3 late, you will receive no credit.**

Submission instructions

Provide written answers to all problems in a single PDF file created using LaTeX. (If you are new to LaTeX, Overleaf is an easy way to get started.) Provide code in answer to Problems 3 in a Google Colaboratory notebook. On **Gradescope**, please submit the PDF in the "Homework 3 PDF" submission link, and your notebook in the "Homework 3 Code" link. Please clearly label each part of each question. Name the files in your submission *abc123*_hw3.pdf and *abc123*_hw3.ipynb (replace *abc123* with your UNI). **You must include figures and results from your notebook in your main submission PDF in order to receive credit for them. You also must tag your pages in Gradescope!**

Problem 1 (10 points): Randomized response

(a) (5 points) The simplest version of randomized response involves flipping **a single fair coin** (50% probability of heads and 50% probability of tails). As in the example we saw in class, an individual is asked a potentially incriminating question, and flips a coin before answering. If the coin comes up tails, he answers truthfully, otherwise he answers "yes".

Is this mechanism differentially private? If so, what epsilon (ε) value does it achieve? *Carefully justify your answer.*

(b) (5 points) Next**,** consider a randomized response mechanism that uses a pair of fair 6-sided dice D_1 and D_2, with faces labeled 1, 2, 3, 4, 5, 6. The mechanism works as follows:

- Roll D_1. If it comes up with a face whose value is smaller than 4 (i.e., 1, 2 or 3), **respond truthfully**
- Otherwise roll D_2. If D_2 comes up with a face whose value is smaller than 3 (i.e., 1 or 2), **respond yes**. Otherwise, **respond no**.

What is the value of the differential privacy parameter epsilon (ε) achieved by this mechanism? *Carefully justify your answer.*

Problem 2 (10 points): Differentially private statistical queries

You are given a database *D* of college applicants, with attributes $gender \in \{F, M, X\}$ and $SAT \in \{high, med, low\}$, and the admissions outcome $admit \in \{yes, no\}$. For convenience, we will use the following notation: g^F, g^M, g^X for gender groups; s^{high}, s^{med}, s^{low} for SAT scores; y^+ for $admit = yes$ and y^- for $admit = no$. You are required to compute the values of several fairness measures explained below, and to release **differentially private** versions of the answers.

- Mean difference: $m_1 = p(y^+ | g^M) - p(y^+ | g^F \vee g^X)$

  It computes the **difference** between the proportion of positive outcomes among the male applicants and the proportion of positive outcomes among the applicants of other genders.

- Disparate impact: $m_2 = \; p(y^+ \mid g^F \vee g^X) \; / \; p(y^+ \mid g^M)$

  It computes the **ratio** between the proportion of positive outcomes among the applicants who are either female or non-binary, and the proportion of positive outcomes among the male applicants.
- Elift ratio: $m_3 = \; p(y^+ \mid g^M \wedge s^{high}) \; / \; p(y^+ \mid s^{high})$

  It computes the **ratio** between the proportion of positive outcomes among the male applicants, and the proportion of positive outcomes overall, both **conditioned on high GPA**.

Your overall privacy budget is $\varepsilon = 1$. You should use this budget to compute noisy versions of **all** fairness measures we are asking you to compute in each question below (**not** one measure at a time). You may assume that neighboring databases D and D' have the same number of tuples, but that they may differ in an assignment of values to attributes of one tuple.

Identify the queries needed to compute these fairness measures and explain how answers to these queries will be used to compute the measures. You may use SQL notation or explain in words what the queries compute.

Use **sequential and parallel composition** to allocate portions of the privacy budget to each query as appropriate, and explain how you allocated the privacy budget. Be specific: write down an ε value for each query as appropriate. Carefully justify your answers, you will receive no credit for this question without a proper justification.

(a) (5 points) Answer the question for mean difference and disparate impact only. For this part, you can use the entire budget on these two fairness measures:

- Mean difference: $m_1 = \; p(y^+ \mid g^M) \; - \; p(y^+ \mid g^F \vee g^X)$
- Disparate impact: $m_2 = \; p(y^+ \mid g^F \vee g^X) \; / \; p(y^+ \mid g^M)$

(b) (5 points) Answer the question for all three metrics. For this part, you can use the entire budget on these three fairness measures:

- Mean difference: $m_1 = \; p(y^+ \mid g^M) \; - \; p(y^+ \mid g^F \vee g^X)$
- Disparate impact: $m_2 = \; p(y^+ \mid g^F \vee g^X) \; / \; p(y^+ \mid g^M)$
- Elift ratio: $m_3 = \; p(y^+ \mid g^M \wedge s^{high}) \; / \; p(y^+ \mid s^{high})$

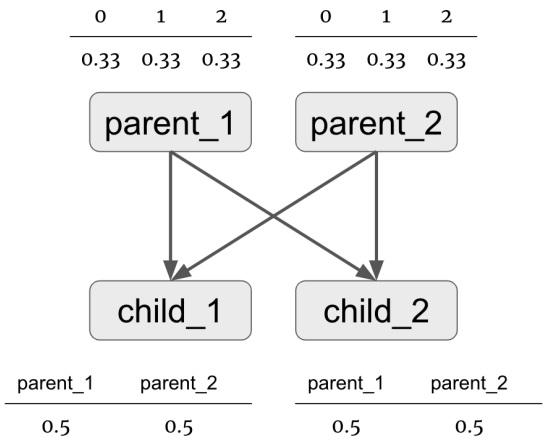
Problem 3 (30 points) : Privacy-preserving synthetic data

In this problem, you will take on the role of a data owner, who owns two sensitive datasets, called **hw_compas** and **hw_fake**, and is preparing to release differentially private synthetic versions of these datasets.

The first dataset, **hw_compas** is a subset of the dataset released by ProPublica as part of their COMPAS investigation. The **hw_compas** dataset has attributes age, sex, score, and race, with the following domains of values: age is an integer between 18 and 96, sex is one of 'Male' or 'Female', score is an integer between -1 and 10, race is one of 'Other', 'Caucasian', 'African-American', 'Hispanic', 'Asian', 'Native American'.

The second dataset, **hw_fake**, is a synthetically generated dataset. We call this dataset "fake" rather than "synthetic" because you will be using it as *input* to a privacy-preserving data generator. We will use the term "synthetic" to refer to privacy-preserving datasets that are produced as *output* of a data generator.

We generated the **hw_fake** dataset by sampling from the following Bayesian network:



In this Bayesian network, **parent_1**, **parent_2**, **child_1**, and **child_2** are random variables. Each of these variables takes on one of three values {0, 1, 2}.

- Variables **parent_1** and **parent_2** take on each of the possible values with an equal probability. Values are assigned to these random variables independently.
- Variables **child_1** and **child_2** take on the value of one of their parents. Which parent's value the child takes on is chosen with an equal probability.

To start, use the Data Synthesizer library to generate 4 synthetic datasets for each sensitive dataset **hw_compas** and **hw_fake** (8 synthetic datasets in total), each of size N=10,000, using the following settings:

- A: random mode
- B: independent attribute mode with **epsilon = 0.1**.
- C: correlated attribute mode with **epsilon = 0.1**, with Bayesian network degree **k=1**
- D: correlated attribute mode with **epsilon = 0.1**, with Bayesian network degree **k=2**

For guidance, you can use the HW3 Code Template here. Please make sure to duplicate this file rather than put your code directly here

(a) (15 points): Execute the following queries on synthetic datasets and compare their results to those on the corresponding real datasets:

- **Q1** (**hw_compas** only): Execute basic statistical queries over synthetic datasets.

  The **hw_compas** has numerical attributes **age** and **score.** Calculate **Median, Mean, Min, Max** of **age** and **score** for the synthetic datasets generated with settings A, B, C, and D (described above). Compare to the ground truth values, as computed over **hw_compas**. Present results in a **table**. Discuss the accuracy of the different methods in your report. Which methods are accurate and which are less accurate? If there are substantial differences in accuracy between methods - explain these differences.

- **Q2 (hw_compas** only): Compare how well random mode (A) and independent attribute mode (B) replicate the original distribution.

  Plot the distributions of values of **age** and **sex** attributes in **hw_compas** and in synthetic datasets generated under settings A and B. Compare the **histograms** visually and explain the results in your report.

  Next, compute cumulative measures that quantify the difference between the probability distributions over age and sex in **hw_compas** vs. in privacy-preserving synthetic data. To do so, use the Two-sample Kolmogorov-Smirnov test (KS test) for the numerical attribute and Kullback-Leibler divergence (KL-divergence) for the categorical attribute, using provided functions **ks_test** and **kl_test**. Discuss the relative difference in performance under A and B in your report.

  For Two-sample Kolmogorov-Smirnov test and Kullback-Leibler divergence, you might find functions such as '*entropy*' and *'ks_2samp*' from *scipy.stats* useful.

- **Q3** (**hw_fake** only): Compare the accuracy of correlated attribute mode with k=1 (C) and with k=2 (D).

  Display the pairwise mutual information matrix by heatmaps, showing mutual information between all pairs of attributes, in **hw_fake** and in two synthetic datasets (generated under C and D). Discuss your observations, noting how well / how badly mutual information is preserved in synthetic data.

  To compute mutual information, you can use functions from https://github.com/DataResponsibly/DataSynthesizer/blob/master/DataSynthesizer/lib/utils.py
  For heatmaps, we suggest considering functions (*heatmap*) provided in the seaborn library (see example: https://seaborn.pydata.org/examples/many_pairwise_correlations.html) and remember to set up *vmax* and *vmin* when plotting.

(b) (5 points, hw_compas only**)**: Study the variability in accuracy of answers to Q1 under part (a) for A, B, and C for attribute **age**.

To do this, fix epsilon = 0.1, generate 10 synthetic databases (by specifying different seeds). Plot **median, mean, min, max** as a **box-and-whiskers** plot of the values for all 10 databases, and evaluate the accuracy of the synthetic data by comparing these metrics to the ground truth median, mean, min, and max from the real data. Carefully explain your observations: which mode gives more accurate results and why? In which cases do we see more or less variability?

Specifically for the box-and-whiskers plots, we expect to see four subplots: one for each of the **median, mean, min, max,** with the three parameter settings (A, B and C) along the X-axis and age on the Y-axis.

(c) (10 points, both datasets): Study how well statistical properties of the data are preserved as a function of the privacy budget. To see robust results, execute your experiment with 10 different synthetic datasets (with different seeds) for each value of epsilon, for each data generation setting (B, C, and D). Compute the following metrics, visualize results as appropriate with box-and-whiskers plots, and discuss your findings in the report.

- KL-divergence over the attribute **race** in **hw_compas**. Vary epsilon from 0.01 to 0.1 in increments of 0.01, generating synthetic datasets under B, C, and D.

  Specifically, the epsilons are [0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1] and in total, you should have 3*10*10 datasets generated. Please plot the distributions of KL-divergence scores (10 samples each) with box-and-whiskers plots where you treat epsilon as the X-axis and generation settings as subplots.

- The difference in pairwise mutual information, aggregated (summed up) over all pairs of attributes, for both **hw_compas** and **hw_fake**, computed as follows:

  Suppose that m_ij represents the mutual information between attributes and i and j derived from sensitive dataset D, and m'_ij represents the mutual information between the same two attributes, i and j, derived from some privacy-preserving synthetic counterpart dataset D'. Compute the sum, over all pairs i, j, with i<j, of the absolute value of the difference between m_ij and m'_ij: $\sum_{i<j} \left| m_{ij} - m'_{ij} \right|$

  Run these experiments for the following epsilon values: 0.0001. 0.001, 0.01, 0.1, 1, 10, and 100, generating synthetic datasets under B, C and D. Specifically, you should have 3*7*10 datasets generated for each hw_compas and hw_fake.

  You should generate 3 plots, one for each data generation method (i.e., one plot for B, one for C, and one for D). The y-axis in all cases should start at 0. All plots should have the same range of y-axis values, so that the values are comparable across experiments.