DS-GA 3001.009, Responsible Data Science, Spring 2019

Homework 2: Privacy-preserving synthetic data generation with the DataSynthesizer
Due at 5 pm on Thursday, March 28

Objectives

This assignment focuses on differential privacy, and specifically on privacy-preserving synthetic data generation. You will use the open-source DataSynthesizer library to complete this assignment. We encourage you to do your work using JupyterHub, in which the environment is already pre-configured for you for this assignment.

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 (goal 1)
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 (goal 2)
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 (goal 3)
4. learn several useful methods for comparing probability distributions (goals 2 and 3)

Grading

This homework is worth 10 points. You will be graded on your execution of Goal 1 (4 points), Goal 2 (2 points) and Goal 3 (4 points) of the homework, and on the quality of your written analysis. You should submit a jupyter notebook implementing all parts of the homework, and an accompanying written report in PDF format, not to exceed 3 pages. (See detailed instructions under **What to submit** below.)

As in homework 1, your grade will be significantly impacted by the quality of your report. In your report, you should explain your observations carefully. Details about what we expect you do discuss are given below, in the description of each goal.

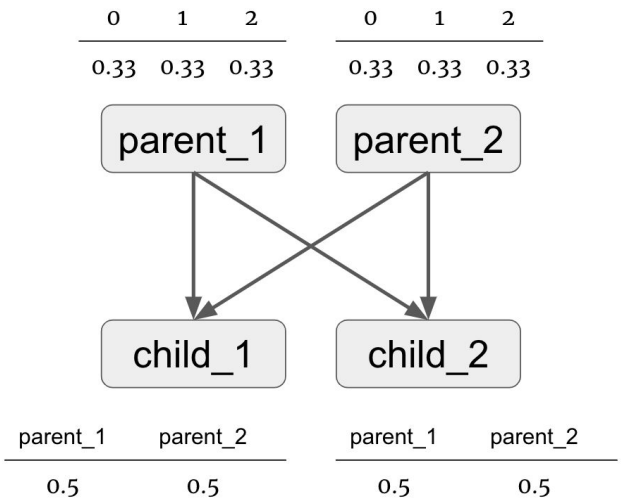
Datasets: In this assignment, you will take on the role of a data owner, who owns two sensitive datasets, called **hw_compas** and **hw_fake**, and is preparing to release differentially private synthetic versions of these datasets using the Data Synthesizer library.

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 the Data Synthesizer. We will use the term “synthetic” to refer to privacy-preserving datasets that are produced as the output of the Data Synthesizer.

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.

Detailed description and goals

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**

Goal 1 (4 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 2 numerical attributes **age** and **score**. Calculate **Median, Mean, Min, Max** of the 2 numerical attributes for 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 the relative performance (accuracy) of random mode (A) and of independent attribute mode (B).

  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, 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.

- **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.

Goal 2 (2 points, hw_compas only**)**: Study the variability in accuracy of answers to Q1 under goal 1 for A, B, and C for attribute **age**. To do this, fix epsilon = 0.1, generate 10 synthetic databases (by specifying different seeds). Plot accuracy as a box-and-whiskers plot. Carefully explain your observations: which mode gives more accurate results and why? In which cases do we see more or less variability?

Goal 3 (4 points, both datasets): Study how well statistical properties of the data are preserved in 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.
- 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 D, and m'_ij represents the mutual information between the same two attributes, i and j, derived from some D'. (For our purposes, D is the sensitive dataset and D' is its privacy-preserving synthetic counterpart.) 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. 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.

In both parts of this goal: If you see a trend, discuss it in your report. If you don't see a trend, discuss in your report why that may be the case.

What to submit

1. Submit one zip file including:
   a. a pdf file of your report.
   b. your .ipynb notebook with code
   c. If it is not obvious how to run your program, you should include a README file.
2. You have to **work alone on this homework.**
3. Please mention your name and net_id in your homework files.
4. Naming convention to follow for your files: RDS_HW2_<net_id>
5. You may *discuss* this homework with other students but **YOU MAY NOT SHARE WRITTEN ANSWERS OR CODE WITH ANYONE.**
6. This homework is worth 10 points.
7. No credit for late homework. 2 late days over the term, no questions asked. If homework is submitted late - a day is used in full.
8. Feel free to look at the example notebooks in the Data Synthesizer Github and Lab 7 notebook. They can help you in doing the homework.
9. Who to contact:
   a. Instructor: Julia Stoyanovich (stoyanovich@nyu.edu) office hours Mondays 1:30 - 3 pm or by appointment, at 60 5th Avenue, room 605. **Prof. Stoyanovich is on**

travel March 25-29, and will not be holding office hours in person. She is available on Skype or similar by appointment during that week.

b. Section Leader: Udita Gupta (ung200@nyu.edu) office hours Thursdays 4 - 5pm or by appointment, at 60 5th Avenue, room 663