PhD Dissertation Work

Evan Shapiro

August 2021

1 Project Proposal

I am interested in performing an uncertainty quantification and machine learning experiment with high dimensional real and synthetic data that has differential privacy techniques applied to it. The point of the experiment is to make a comparison between the statistical results of predictions made on the synthetic and real data with the learned machine learning and UQ. I would like to explore whether ML and UQ performed on synthetic data leads to similar statistical results as the analysis performed on real data, and quantify the difference using different metrics like the KL-divergence. The data set that I have identified is the Adult Data set that is available from the UCI Machine Learning Database. https://archive.ics.uci.edu/ml/datasets/Adult. This is an individual level census data set that allows for binary prediction on income levels. (above or below 50K a year.) Using this data set, a synthetic data set will be generated using the Data Synthesizer python package. Data Synthesizer takes a given data set and generates a new data set that maintains correlation structures between data attributes, and maintains similar distribution and statistical properties of the data. Data Synthesizer uses differential privacy to secure a sensitive, individual level data set, so that malicious actors cannot use the synthetic data set to identify a person of interest. Differential privacy amounts to adding noise from a Laplace distribution to the original data set. The Laplace distribution has scale parameters that are related to the size of the original data set, and the differential privacy parameter . Per the creators of Data Synthesizer "Differential Privacy is a family of techniques that guarantees the output of an algorithm is statistically indistinguishable on a pair of neighboring databases: that is, a pair of databases the differ by only one tuple." ://github.com/DataResponsibly/DataSynthesizer/blob/master/docs/cr-datasynthesizer-privacy.pdf This second step in this project is to use logistic regression on the real and synthetic Adult dataset to make a binary prediction on whether the individuals in the data set make above 50k a year or below 50k a year. The third step is to perform dimension reduction on the real and synthetic datasets using active subspace techniques. The idea behind active subspaces is to use the gradient of a function to identify linear combinations of basis parameter vector directions that cause the function to vary the most. The span of these parameter vectors generates a subspace of the full parameter vector space that can be used

instead of the full parameter vector space. Then, the full data input space is projected down into active subspace, and a new model is constructed based on the inputs from the active subspace. I will need to read further into this to be able to completely understand and explain the math behind active subspaces. I also need to make sure that active subspaces make sense in binary classification problems, as the output function does not appear to have a gradient. Perhaps performing multiple regression makes more sense, as then the regression function will have an explicit gradient. (Source: ACTIVE SUBSPACE METHODS IN THEORY AND PRACTICE: APPLICATIONS TO KRIGING SURFACES)
Once models have been constructed on the full input space and the active subspace, comparisons of predictive abilities of the models will be made between the real data sets and synthetic data sets.

2 Update 8/4/2021

I was successful in generating a synthetic data set using the Data Synthesizer Python package. I then created a logistic regression model with the synthetic training data using the SK Learn in package in Python, and then ran cross validation using the original data to see how well the logistic regression model predicts on the original data. The logistic regression model was used to predict, or classify, whether someone will make more than or less than 50K a year based age and their number of hours worked per week. The predictive ability within the synthetic data is 76%, which is within the realm of expectation, while the predictive ability of the model on the original data set as approximately 0%. I think this was due to the nature of the data though, as the data was not individual level data, but rather weighted population level data. I would like to find some individual level data to determine whether the machine learning model will retain some degree of predictability between the original and synthetic data sets.
Additionally, I would like to setup a machine learning model that takes into account the categorical variables when making predictions and classifications. I plan on having this setup over the next couple of weeks.
Micro Goals - Figure out what went wrong with the logistic regression. Data was not imported properly. Solved this

3 Census Problem

The U.S. Census Bureau is constitutionally required to take a decennial census of the U.S. population in order to accurately apportion congressional seats. While not the primary purpose, this data is used for many other important reasons, such as congressional and state redistricting, and determining the needs of districts when it comes to roads, schools, and healthcare. Businesses use the census data to understand the needs and desires of the communities they are serving.

Simultaneously, the Bureau is legally required to protect the information and identity of all census respondents. Due to the increasing availability of databases containing individual level data, increased computational power, and new algorithms, methods previously used to protect census respondents no longer offer an acceptable degree of protection against participant re-identification.
In the 2010 census, participant privacy was protected by swapping outlier data between census blocks, groups, tracts, or counties. [**?**]. An example of swapping would be if one census block consisting of 49 people has a single family from a particular demographic group, which would make identification of the family from the census data easy. If another census block in the same county with 49 people contains more members of the same demographic group, swapping between census blocks would occur to protect privacy. Research by the census bureau over the following decade has shown that data swapping is no longer effective at protecting census participant privacy. This is because, every time you release a statistic calculated from a private data source you leak a small amount of private information. Given enough statistics from a database, the entire underlying data source will be revealed.
The census bureau used a reconstruction attack on census block data that had swapping applied, with the census block population and exact voting age population counts kept invariant. With their reconstruction attack they were able to exactly reconstruct all 308,745,538 unique records with correct correct block and voting age, and were also able to accurately reconstruct between $46\% - 48\%$ of census participants race, ethnicity, sex and age. It follows that "the 2010 disclosure avoidance techniques used for the 2010 Census no longer meet the acceptable disclosure risk standards that were in place in 2010." The Census bureau has since adopted differential privacy as its disclosure avoidance technique.
I am interested in testing whether or not the statistical properties of a database or dataset that are supposed to be preserved after differential privacy has been applied, are in fact preserved, and to also test the limits of the fidelity of the synthetic data set to the original data set.

4 Differential Privacy Explanation

5 Literature Review

5.1 Summary of Bayesian Network Paper - Idea for Applying Differential Privacy to Databases and Then Applying Differential Privacy

Applying differential privacy to high-dimensional databases is an active area of research. This is due to issues that arise when injecting probabilistic noise into histogram database that is a histrogram, or a database of counts. Consider such a database, where the average count for an individual with a particular set of attributes is 1. Adding noise to each data point to make the database differ-

entially private will swamp the database with noise, yielding an unacceptable signal-to-noise ratio.
Additionally, applying differential privacy to high-dimensional can lead to the curse of dimensionality, even in seemingly low dimensional databases. Many differentially private algorithms represent databases according to the entire domain of the database, resulting in a domain that is the product of the cardinality of the attributes in the database. [**?**]
As an example, consider a database of population, where each member of the population can have 10 attributes, and each attribute has 20 possible values. This yields a total of 20^{10} possible attribute combinations within the problem domain, and would require 20TB of storage. So, although an actual data set may be low dimensional the domain of the problem that is used by a differentially private algorithm is not.