2020

Differential Privacy Applications

CS 211 – FINAL PROJECT WRITE-UP

MOLLY EATON

Problem Statement

Using a dataset previously unseen one can perform analyses, comparisons, and contrasts using differential privacy. Implementing varying types of differential privacy allows the analyst to choose differentially private methods to produce accurate query results and make meaningful conclusions, while still adhering to the privatization guidelines.

Implementation

Dataset

The dataset used is a compilation of cities within the US quantifying locations, income data, and physical features (Golden Oak Research Group). The features analyzed were the mean incomes and the latitudes.

Preprocessing

The preprocessing performed on the dataset consisted of clipping the mean incomes. While finding the best clipping parameters uses up some of the privacy budget, it allows the sensitivities of queries to be statically defined and bounded. To find the optimal bounds, a simple function which loops through a defined list of possibilities to find how much of the dataset is included/excluded with each bounding value. The plot confirms that most (if not all) of the mean column lies between 10,000 and 140,000. The column is then clipped to these values before any analysis, defining the sensitivity of the mean as 130,000 (upper bound – lower bound).

Comparing Differential Privacy with K-Anonymity

Before any definitive queries can be run, the method for 'blurring' the data to conform to privacy guidelines needs to be determined. The methods of realizing privacy compared in this project are the Laplace Mechanism, the Gaussian Mechanism, and K-Anonymity. The Laplace Mechanism and Gaussian Mechanism use differential privacy budgets to add noise to the data to keep the raw data private while K-Anonymity generalizes the values so that a certain number of rows (individual cities in this case) are unrecognizable from one another. A goal of 10-K anonymity was placed on the dataset. This implementation is concerned with state's averages versus specific cities. The optimal method for privatizing the dataset is used to compile all cities within a state and find the mean of means.

The Laplace Mechanism and Gaussian Mechanism subscribe to a total privacy cost of (ε) and (ε, δ) respectively. To effectively compare the two methods a series of 100 queries for Vermont's mean income is run with the resulting percent error from the actual mean plotted. The plot definitively shows the Laplace Mechanism performing with greater accuracy and consistency. This mechanism is chosen for comparison with K-Anonymity due to this advantage.

The aim for K-Anonymity was to generalize the mean values to be k = 10 anonymous. By generalizing the 3 least significant digits of the column, this is achieved. Generalizing the 3 least significant digits generalized the values to the thousands, i.e. 74, 555 becomes 74, 000. To compile each state's mean, an essence of randomness is added with the Laplace Mechanism adding noise to the count of cities within the state. To compare K-Anonymity with differential privacy 100

queries for Vermont's mean income is ran with the resulting percent error now plotted against the Laplace Mechanism. This second plot shows the higher accuracy and consistency now attributed to K-Anonymity. K-Anonymity is declared to winning parameter.

With the best method for privacy chosen, additional analyses can be performed. To put the dataset into context the previously found state means are compared with their relative latitudes. Every state's shape and size ranges, so the average latitude for each state is used for comparison. The latitude values were generalized by the 4 least significant digits to achieve k=10 anonymity. As with using K-Anonymity to find the state's mean, the Laplace Mechanism is used to add noise to the city count for each state. Plotting each state's mean income with their average latitude shows a positive correlation. The general trendline attributes higher mean incomes with higher (more northern) latitudes.

Summarized Results

When comparing the Laplace Mechanism with the Gaussian Mechanism, the Laplace Mechanism performs with greater accuracy and consistency. When comparing the Laplace Mechanism with K-Anonymity (k=10) the generalized values perform better. The Laplace Mechanism and K-Anonymity will produce the same answers to the highest earning and lowest earning state. Due to the amount of noise, these answers can periodically be different. When comparing states' incomes with their average latitude a positive correlation is found.

Bibliography

Golden Oak Research Group. "US Household Income Statistics." *Kaggle*, 16 Apr. 2018, www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations?select=kaggle_income.csv.