Hasso-Plattner-Institute, Potsdam, Germany
Privacy Preserving Outlier Detection - Winter Semester 2023/24

Assignment #1: Pseudonymisation Techniques and Considerations – (50 marks)

October 24, 2023

1 General Goal

The growing volumes of personally identifiable information (PII) collected on digital environments and the availability of large language model-based chatbots such as ChatGPT, emphasise the importance of personal data protection. Data processing pipelines are often supported by datasets which must be shared with external clients or with cloud services for high performance processing. Protecting PII from exposure during this process is vital to adhering to privacy legislation such as GDPR.

In this assignment, we will aim to study some data anonymisation techniques that can be used to protect PII by removing (suppression) or obfuscating (e.g. generalisation, pseudonymisation, ...) sensitive data. While the rest of the assignment guide is centered around the python programming language, you are welcome to use other alternatives that you are more familiar with.

2 Dataset

To begin with, please select a dataset to use for your assignment. You will find some proposed on Moodle, just below the description of this assignment. If you have an alternative dataset that you prefer or are more interested in, then please feel free to use this.

3 Tasks...

3.1 Pseudonymisation – 10 marks

As a first task, let's consider how to pseudonymise data. Your goal in this task would be to (1.) Study your dataset to determine which attributes qualify as explicit personally

identifiable information. Explain why you decided on the attributes and show what method was used to identify the attributes (or depending on the dataset, the attribute values); (2.) Using a tool such anonymizedf which builds on pandas and faker generate pseudonymous values to replace the original values.

3.2 Randomisation – 10 marks

Randomisation, is helpful in supporting pseudonymisation by increasing the variability of the data. For example instead of replacing a name with another name, randomisation might use a randomly generated string of characters. Your goal in this task is to use the randomisation technique to generate random strings (that do not necessarily have a meaning) and then to modify the randomisation process to generate random but meaningful replacements. For example instead of replacing "Anne" with **"xyzk"** we might replace instead with **"Amy"**. Python's built-in random library used with pandas might be helpful in generating random replacement values. Finally, create a lookup table to keep track of your changes. This is useful approach to adopt, especially for being able to determine the wider impact of your changes on the data.

3.3 Aggregation – 10 marks

For certain attribute values, such as salary and/or age, combining data to create an aggregated view of the data, so that the data is represented in ranges instead of individually. For example, instead of Age = 35, 23, 20 we could use instead Age = [20 - 35]. Building on your results from the previous tasks, study the data to determine which attributes (and correspondingly attribute values) qualify for aggregation. Define a replacement algorithm to aggregate data, throughout your selected dataset. In Python, the `pandas.cut()` method might be helpful, but you can also achieve this by other techniques such as `groupby()`.

3.4 Perturbation – 10 marks

One of the techniques used to anonymise data is to distort the data by adding noise to make it harder to re-identify individuals within the dataset. For example, a person's age or postcode might be modified slightly to make it harder to tell their age or exact location. Perturbation, typically works well with numerical and/or categorical values. Other well known mechanisms that employ noise additions include Differential Privacy.

In this task, your goal is to design and code a function to add noise to one (1) or two (2) selected attributes of your choice. The amount of noise you add can be determined by metrics such as standard deviation, variance, and mean. Your goal is to preserve the original distribution of values while at the same time distorting the values to minimise disclosure risks.

You must analyse your original data to establish what the distribution is, and the repeat the process once you have applied the noise addition function to the data.

3.5 Data Analysis – 10 marks

Analyse your data to determine the level of information loss. In this case, you will design a function to analyse information loss and discuss its pros and cons relative to your dataset.

4 Submissions

Once you have completed your assignment, please submit a 2-3 page report analysing your results. Code can be uploaded to Github or Gitlab and a link shared. Please note that you should aim to keep the same repository for all five (5) assignments. Your report should contain enough details to ensure that your procedure is repeatable for grading purposes.