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

Assignments #2-3: Anonymisation – (100 marks)

November 16, 2023

1 General Goal

This assignment builds on the previous one. In this case, our goal is to anonymise a dataset, to minimise the risk of privacy disclosure.

2 Dataset

You may use the same dataset that you used for the first assignment.

3 Tasks...

3.1 Anonymisation: Bare Bones – 10 marks

As a first task, formulate a k-anonymity inspired anonymisation algorithm of your choice. Explain the reasoning behind your decisions. You should discuss the pros and cons of your algorithm with respect to the dataset. However, at this stage no analysis of information loss etc... is required.

3.2 Anonymising your dataset – 30 marks

Select any two (2) of the anonymisation algorithms (mechanisms) that we discussed in class and apply both of these as well as the algorithm you proposed in the previous section, to your dataset. Discuss your results in terms of comparing the output generated by each of the algorithms you selected.

3.3 Compartmentation and Clustering – 20 marks

Design an anonymisation algorithm that is inspired by either by the principle of unique column combination discovery and/or clustering and apply your algorithm to your dataset.

3.4 Testing Data Utility – 50 marks

Design experimental scenarios to test the following:

1. The utility of the data that you have generated from the anonymisation schemes (algorithms) above.
2. Compare the relative utility of the data obtained from each anonymisation scheme, and discuss the implications of your results with respect to a case study of your choice related to privacy preserving machine learning.
3. Degree of information loss, time to anonymise, and per query accuracy with respect to changing dataset sizes. Note that this should be done for each of the anonymised dataset you generated. So you should end up with a couple of plots.

4 Submissions

Once you have completed your assignment, please submit a 5-6 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.