Algorithms in Data Science
Homework3: Privacy

Manuelle Ndamtang
66732000

December 2021

Documents

- use_case1.ipynb: jupyter notebook for the first use case
- use_case1.csv: dataset for the first use case
- use_case1.ipynb: jupyter notebook for the first use case
- use_case2.csv: dataset for the second use case
- report_ndamtang.pdf: the report
- dataset_HW3.csv.csv: the dataset

Introduction

The homework 3 of this course consists in anonymizing a dataset in order to publish it as open-data. The challenge here is to implement anonymization methods while preserving the maximum amount of information important for data analysis. To achieve this, the proposed solutions will be pseudomization, selection of interesting columns for each use case, use of entropy for the choice of columns, k-anonymity and l-diversity.
We will discuss about possible attacks and what could be the solutions to face them.

Pseudonymization

One of the first steps in anonymizing the dataset is to hash the direct identifier. For this, the algorithm used is *sha256*. Having noticed that the result obtained did not give unique keys for only based on the *id* column, a solution would be to add a value increasing the probability that the key is unique. That's why the

result obtained in the *id* column is in fact the result obtained from the hashing of the concatenation between the *id* column and the *dob* column.
In order to avoid the attack thanks to a lookup table, the salt was also added during the hashing of the identifier.

Features selections

Since each dataset corresponds to a specific use-case, it is quite obvious that some features needed for one case, will not necessarily be needed for the other. In the first use case, the objective is to anonymize a dataset for sociologists in order to find stress factors that correlate to the disease or living conditions. Most of the columns obviously correspond to the information needed for such a study, except the *ancestry* column which has been removed. This column does not give any information about the environment of the person and is therefore useless for our dataset.
For the second use case, the dataset should allow the US department of Health and Human services chairman to decide where to build new hospitals, and which department to have in these across country. That's why the choice of some features seemed necessary. The columns chosen are mainly: *id*, *gender*, *zip-code*, *dob*, *employment*, *children*, *commute_time*, *disease*. These columns give enough information to answer the question asked in use case 2.

Anonymization through generalization

Here the objective is to generalize the information in such a way that the data identifies as few individuals as possible in a unique way. Of course we have a loss of information, but the main information is still there.
For the first use-case, we proceeded to the:

- generalization of *children* column:
  Assuming that having children can be a stress factor, it is possible that the number of children does not provide too much information for this study. Especially since the objective is not to find the correlation between the number of children and the impact of stress, but to the living conditions, the environment of these people. That is why we will have to limit ourselves to know if a person has children or not.

- generalization of *dob* column:
  The date of birth in general is a very precise information about people. And for this use case, the objective is not to find correlations between the age of a person and the impact of stress. That's why instead of using the date of birth, we use the age group. Thus, we will classify people according to their age.

- generalization of *zipcode* column:
  Instead of using the zipcode, which gives a lot of information about the locality of a person, the option chosen is to replace it by the state to which the zipcode would belong. For that, the uszipcode library was used because it allows to make the mapping between the zipcode and the states.

- generalization of the *education* feature:
  Having noticed that the proportions of people with a Phd and those with a Less than high school level were the smallest and therefore easily distinguishable, one solution is to join the people with a Less than High School level and those with a High School level so that they are all in one class. We will do the same for masters and Phd in the same group.

- generalization of *commute_times* column:
  Having noticed that in general the column *commute_times* varies between 0 and 3.43 hours, the choice to generalize consists in subdividing the time into groups. Thus there are groups 'less than 1h', 'less than 2h', 'less than 3h', 'more or equal than 3h'

- generalization of *number_vehicles* column:
  Having noticed that the proportion of people with 3 cars was very small, a solution would be to group them with those with 2 cars in order to have a class of people with more than one car.

- generalization of the *accommodation* column:
  What we notice in this column is that in general the houses are either rented or owned. So it is easy to deduce two classes: the rented houses and the houses owned by the people.

For the second use-case, the principle remains the same for the *zipcode*, *children* and *dob* columns. The reason why the column *commute_time* has not been generalized is that it would be interesting for the people handling the dataset to estimate the average time needed to arrive at the hospital and compare it with the average time taken by these people. Knowing whether or not a person has a child could also be interesting in the case that we would like to open a pediatric center. It is also interesting to know if a state has more retired, working or unemployed people in order to make strategic decisions if desired.

Usage of entropy

In order to measure the amount of information lost during data anonymization, the entropy was used. Here it allows the selection of interesting columns that preserve the usefulness of the information for each use case while remaining the least precise.
It turns out that for use case 1, the set of columns consisting of *age_group*,

gender, *education*, *employment*, *children*, *marital_status*, *number_vehicles*, *commute_time* and *accommodation* gives the best entropy (8.56). But when we compare this value with the entropy obtained on the original dataset (10.96), we realize that the generalization had a great impact on the quantity of preserved data. For the use case 2, we also notice a huge difference between the entropy of the original dataset (10.96) and the one of the dataset transformed by the generalization (5.33). In order to avoid that the selection of the best set of features excludes the state column (crucial information for this use-case), this column is removed from the list of features allowing to measure the best entropy, and then added at the end of the feature selection process.

K-diversity

To avoid cases where unique data is found in the dataset, the use of k-anonymity is necessary. The objective here would be to remove the unique data in the dataset. Of course, there will be a significant loss of information, but at least we could make sure that the attacker would not be able to uniquely identify a person based on the quasi-identifiers.
Since the amount of data is very small, we can't afford to use a very large k value, otherwise there would be a huge loss of data. Hence the choice of k=2 for the two use-cases, the minimum value we can afford.

L-diversity

To avoid attacks related to homogeneity, deleting rows that do not meet the l-diversity is also necessary. To avoid the huge impact that an action could have on the dataset, the value of l=2 is a good option for both use-cases, the minimum value we can afford.

Possible improvements

During the manipulation of the data, we did not dwell on intersection attack. Indeed, since there are two datasets, anonymized from the same dataset. So it would be possible for an attacker to have with precision the commute time of a person from the dataset of the use-case 2 and to have an idea of the characteristics of the same person from the dataset of the use-case 1.
Other possible attacks would be the semantic attack. In the dataset, there are 3 types of cancer (skin cancer, breast cancer, prostate cancer) for example, it would be easy for the attacker to know that a person suffers from cancer if this person belongs to a class of 2 people all having cancer. A solution would be to

implement t-closeness.

Conclusion

To succeed in anonymizing a dataset in an optimal way while preserving its usefulness is a real challenge. The solutions proposed for each use-case still show some security flaws described in the previous paragraph. Nevertheless, anonymizing data also sometimes leads to the deletion of features and rows. This would make the already small dataset even smaller and thus a huge loss of information and usefulness for each use case.