Workshop on privacy for IOT

Héber HWANG ARCOLEZI
heber.hwang_arcolezi[at]univ-fcomte[point]fr

November 12, 2020

Chapter 1

Learning and Anonymization

This TP is largely inspired by the article by J. Brickell[1], which criticizes the syntactic approaches based on k-anonymity and their derivatives.

The main idea is to compare different levels of privacy protection (from the strictest to lighter versions) in terms of learning. Is the gain in learning accuracy worth the cost it imposes on personal information leakage? This is the question to which this TP will try to provide an answer.

1.1 Initialization of the TP

We consider the UCI Adults dataset presenting an extract from census data in 1994/1995 in the USA and whose initial objective was to predict whether such or such person would have a salary higher than 50K $ per year. The attributes that we will keep are: age, workclass , education-num, marital-status, race, sex, native country, occupation.

Exercise 1.1. *Clean up the dataset (excluding missing data ' ?' samples) and keep only the aforementioned attributes. The final dataset must have 45,222 rows and 8 columns; save it in csv format (or download the attached in the email).*

1.1.1 The best learning possible

We imagine here that the scientist responsible for the analysis has access to all the data, *i.e.*, that no data is deleted or transformed. This is the most favorable case for her / him.

Exercise 1.2. *Using the weka tool, try to learn **marital status**, using the three classification approaches with the default parameters for J48, Random Forests, and Naive Bayes. What is the maximum score in terms of learning (U_{Max})?*

1.1.2 The most privacy-friendly learning

We imagine that trust is not in place. All the quasi-identifiers have been generalized to the extreme and we will replay the learning, to assess the precision of the approach, at a minimum. The question immediately arises of defining the quasi-identifiers. In this TP we will consider that the attributes *age*, *education*, *occupation* are the quasi-identifiers.

Exercise 1.3. *Generate a dataset where all the quasi-identifiers have been generalized to the extreme. On this data set, carry out the same training as in the previous section. What is the minimum score (U_{Min})?*

1.2 Anonymization that are useful for learning, but respectful of privacy?

The quasi-identifiers below will be generalized according to the following hierarchies:

- age: grouping by intervals of size 10, 20, 40, 80, from 17.
- education: grouping by intervals of size 2, 4, 8, 16, from 1.
- occupation: total suppression.

[1] Brickell, J., & Shmatikov, V. (2008, August). The cost of privacy: destruction of data-mining utility in anonymized data publishing. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 70-78).

Exercise 1.4. *Implement these hierarchies.*

Exercise 1.5 (Learning based on 10-anonymity).

1. *Generate a 10-anonymity dataset by accepting up to 5% suppression. Export this dataset.*
2. *On this data set, carry out the same training as in the previous section. What score ($U_{k_{10}}$) do you get? As a percentage, express the improvement over U_{Min}.*
3. *Conclude.*

Exercise 1.6 (Learning based on 2-diversity). *Repeat the previous exercise taking the* 2*-diversity as privacy model (don't forget to select marital status as sensitive attribute). U_{l_2} and improvement over U_{Min}? Conclude.*

Exercise 1.7 (Learning based on 0.3-closeness). *Repeat the exercise 1.5 taking the* 0.3*-closeness as privacy model (don't forget to select marital status as sensitive attribute). $U_{t_{0.3}}$ and improvement over U_{Min}? Conclude.*

Exercise 1.8 (Syntactic models). *Give a general conclusion regarding the protection of privacy ensured by syntactic methods (k-anonymity, l-diversity, t-closeness).*