Artificial Intelligence for Medical Data with Python

8 SAMPLE SLIDES

UNIVERSITY OF THE
AEGEAN

SCHOOL OF ENGINEERING

DEPARTMENT OF INFORMATION
AND COMMUNICATION
SYSTEMS ENGINEERING

2th session – Privacy

Presenter: Panagiotis Symeonidis

Associate Professor

http://panagiotissymeonidis.com

psymeon@aegean.gr

Anonymization technique based on Generalization

Name	Gender	City	Age	Disease
Petros Petridis	male	Larissa	26	HIV
Yannis Yannou	male	Volos	29	COVID
Maria Marianou	woman	Kilkis	36	HIV
Christina Christidou	woman	Thessaloniki	37	FLU
Vassilis Vassiliadis	man	Karditsa	38	COVID
George Georgiou	man	Trikala	36	HIV

Gender	Geographical Division	Age	Disease
male	Thessaly	20 to 30	HIV
male	Thessaly	20 to 30	COVID
woman	Macedonia	30 to 40	HIV
woman	Macedonia	30 to 40	FLU
man	Thessaly	30 to 40	COVID
man	Thessaly	30 to 40	HIV

Definition Pseudo -identifier :

❖ Pseudo -identifiers are those data that if properly combined with another public data table can identify a person (Gender, City, Age).

Negative of generalization : They limit the ability of systems to make more accurate predictions (loss of information).

K- Anonymity

Definition of k-anonymity : A table T satisfies k-anonymity when every quasi-identifier (QI) satisfies k-anonymity. A QI group satisfies k-anonymity when the QI group size is at least k.

- For a table T we create another table T* so that each individual p will have at least k-1 other individuals that will not be distinguished from p .

Central Differential Privacy

- Algorithms based on " differential privacy " appropriately modify the original data and at the same time provide measurable guarantees of privacy to users.

- Let ε be a positive number and A be a random algorithm with input a data set D.

Algorithm A guarantees ε-differential privacy :

- If the subsets D_1 and D_2 differ by at most one record and
- If for all subsets S of the range of output values A (S ⊆ Range(A)) (where S are the possible different combinations of outputs) holds :

$$\Pr[A(D_1) \in S] \le e^{\varepsilon} \times \Pr[A(D_2) \in S]$$

The probability (Pr) depends on the degree of randomness of the algorithm A.

Differential privacy algorithms using the Laplace Mechanism

Joint Differential Privacy Matrix Factorization algorithm for the factorization of patient-drug interaction matrix

- Provides an ε-differential privacy guarantee
- Laplace Noise is added to the classic Matrix Factorization objective function as follows:

$$\arg\min_{p,q}\left(\sum_{(u,i)}\frac{1}{2}(r_{ui}-\hat{r}_{ui})^2+\frac{\lambda}{2}\left(\sum_{u}\|p_u\|^2+\sum_{i}\|q_i\|^2\right)+\sum_{(u,i)}\eta_{ui}\times p_u^T q_i\right)$$

Where η_{ui} is the Laplace noise corresponding to each interaction r_{ui} of a patient u with an drug i and $\eta_{xy} \sim \mathrm{Lap}(s/\epsilon)$.

We emphasize that the parameter s expresses the sensitivity of the interaction values, i.e., the difference between the maximum and minimum interaction values ($r_{max} - r_{min}$).

Advantage : Guarantees ε-differential privacy while maintaining high predictive value of the data. (low information loss)

Toy example of Local Differential Privacy

We will explain the process with the following example data

name	gender	age	zip	diagnosis
John Johnidis	male	25	56431	Heart Attack
Jack Jackidis	male	25	39100	Diabetes
Liam Liamidis	male	35	56431	Covid
Bob Bobidis	male	45	56431	Diabetes
Maria Mariadou	female	35	39100	Covid
Sofia Sofiadou	female	45	39100	Covid
Luna Luniadou	female	35	39100	Cancer
Elena Eleadou	female	25	56431	Heart Attack

ε-Differential Privacy Guarantee

ε (epsilon) quantifies the privacy guarantee in differential privacy. It measures the difference in the probability of obtaining the same perturbed response given different original responses.

Formula:

$$\epsilon = \log\left(\frac{p(1-q)}{(1-p)q}\right)$$

e.g., For p = 0.75 and q = 0.25, ε = log(9) ≈ 2.197

This formula calculates the privacy loss by considering the probabilities p and q of flipping bits in the encoded response.

A smaller ε value indicates stronger privacy guarantees.

Data Before and After Local Differential Privacy

ε privacy loss = 2.197

Name	Diagnosis
John Johnidis	Heart Attack
Jack Jackidis	Diabetes
Liam Liamidis	Covid
Bob Bobidis	Diabetes
Maria Mariadou	Covid
Sofia Sofiadou	Covid
Luna Luniadou	Cancer
Elena Eleadou	Heart Attack

Name	Diagnosis
John Johnidis	Heart Attack
Jack Jackidis	Heart Attack,Covid
Liam Liamidis	Heart Attack,Covid
Bob Bobidis	Diabetes
Maria Mariadou	Diabetes
Sofia Sofiadou	Covid
Luna Luniadou	Heart Attack, Cancer
Elena Eleadou	Heart Attack