LELEC2770 Practical Session 6: Privacy

1. Download the datasets *diabetes1.csv* and *diabetes2.csv*. It corresponds to medical data for 384 persons. The datasets are decomposed as follow:

diabetes1.csv	*diabetes2.csv*
Cholesterol	Height [inches]
Location	Frame
Age [years]	Waist [inches]
Gender	

   (a) Compute the K-anonymity metric for both datasets.

   (b) Apply a protection mechanism on some variables in the datasets in order to maximize K-anonymity. For the sake of simplicity, consider only numerical values.

   (c) Evaluate your transformed data with an utility function based on the distortion, which will be computed as

$$d = \sum_{p \in P} \sum_{a \in A} d_a(p_a, p'_a),$$

   where P is the set of persons in the databases, $A = \{C, L, Y, G\}$ (or $A = \{H, F, W\}$) is the set of attributes (cholesterol, location, age and gender, or height, frame and waist), and p_a is the value of the attribute a for person p (before applying the protection mechanism) and p'_a is the same value, after applying the protection mechanism. d_a is a distance:

   - $d_L(x, x') = d_G(x, x') = \delta(x - x')$ where $\delta(\cdot)$ is 0 if its argument is 0, otherwise it is 1.
   - $d_C(x, x') = |x - x'|/100$.
   - $d_Y(x, x') = |x - x'|/50$.
   - $d_H(x, x') = |x - x'|/15$.
   - $d_F(small, medium) = d_F(medium, small) = d_F(medium, large) = d_F(large, medium) = 1$
   - $d_F(small, large) = d_F(large, small) = 2$
   - $d_W(x, x') = |x - x'|/20$.

   (d) Do you think the resulting data set would be useful? (bonus:) What if you apply a protection mechanism to get a K-anonymity of 5? How would the data looks like?

(e) What's the advantage of differential privacy and how could it help here?

2. Download the dataset *basket.csv*. It corresponds to purchases in a supermarket for 10 users. Each observation is a set of products resumed as a basket.

   Observations are in rows while variables are in columns. The columns are:

   Milk / Meat / Apple / Bread / Pizza / Beer / Banana / Fish / Sugar / Corn Flakes / ID

   Notice that the last column is for the users IDentification.

   (a) Assuming U is a random variable with uniform distribution over the set of users $\mathcal{U}$, what is $Pr[U = u]$ for any user $u \in \mathcal{U}$? What is the Shannon entropy of U ($H[U]$)? Explain the notion of "amount of information" tied to the Shannon entropy.

   (b) Let use consider first a reduced dataset where only the columns Milk and ID are available. Let O_1 be a random boolean variable representing whether or not a user buys milk (the possible issues for O_1 are thus 0 (false) and 1 (true): $\mathcal{O} = \{0, 1\}$). We want to compute:

   $$HI[U; O_1] = H[U] + \sum_{u \in \mathcal{U}} Pr[U = u] \sum_{o \in \mathcal{O}} \tilde{Pr}[O_1 = o | U = u] \log_2 \left(\tilde{Pr}[U = u | O_1 = o] \right).$$

   To perform it, applying the following steps would help:

   - Implement the function $f_1 : \mathcal{U} \times \mathcal{O} \to [0, 1] : (u, o) \mapsto f_1(u, o)$ where $f_1(u, o) = \tilde{Pr}[O_1 = o | U = u]$. (Hint: Compute an histogram of the empirical distribution using the database.)
   - Give an analytical expression of $\tilde{Pr}[U = u | O_1 = o]$, then implement a function $f_2(u, o)$ that computes it. (Hint: Use Bayes formula, then the Law of total probabilities.)
   - You have now all the building blocks to compute $HI[U; O_1]$. What is its value? Would an adversary having access to one leakage of the reduced database (one row, where the ID is erased) be able to gain information about the identity of the user associated to the leakage? Looking at the dataset, would you have expected this result?
   - Compute $HI[U; O_2]$, $HI[U; O_3]$, etc. where O_2 is the purchase of Meat, etc. What is the most informative column ? Give a bound on the re-identification success rate for an adversary that observes only that column.

- A better adversary can exploit all the column simultaneously. Compute $HI[U; (O_1, \ldots, O_{10})]$ and give a bound on the success rate of the adversary.

(c) Download the dataset *basket_ test.csv* and re-identify the observations with regards to the previous dataset and users.

(d) Compute the success rate of your re-identification with the help of the file *user_ test.csv* which corresponds to the true *basket_ test* users.

(e) Compute the Perceived Information with *basket.csv* the training set to build the model and *basket_ test.csv* the test set.