Sécurité Appliquée-PVP TD2
Distances entre bases, mécanismes ϵ-DP

Jean-François COUCHOT
couchot[arobase]femto-st[point]fr

12 novembre 2023

1 Distances entre bases de données, bases voisines

Sx	Age	Disease	Salary
F	35	bronchitis	16k
M	38	stomach cancer	12k
M	38	stomach cancer	12k
F	54	gastritis	12k

TABLE 1 – Extrait d'une base de données médicale

Dans cette section, on nomme D_1 la base de données donnée dans le tableau 1.

Exercice 1.1 (Insertion d'une ligne). *Soit D_2 la base obtenue en ajoutant à D_1 la ligne suivante concernant* Alice

Sx	*Age*	*Disease*	*Salary*
F	*35*	*bronchitis*	*14k*

1. *Definir $\mathcal{X}$ l'univers des valeurs possibles pour ce genre de base.*
2. *Exprimer D_1 comme un vecteur de $\mathbb{N}^{|\mathcal{X}|}$. Idem avec D_2.*
3. *Calculer $\|D_1 - D_2\|_1$. Que dire de D_1 relativement à D_2 ?*
4. *Refaire les calculs en considérant que le salaire d'Alice n'est pas de 14k, mais de 16k. La conclusion change-t-elle ?*

Exercice 1.2 (Modification d'une ligne). *Soit D_3 la base identique à D_1, sauf pour la $3^{ème}$ ligne où 12k est remplacé par 14k.*

1. *Exprimer D_3 comme un vecteur de $\mathbb{N}^{|\mathcal{X}|}$.*
2. *Calculer $\|D_1 - D_3\|_1$. Comparer cette dernière valeur à $\|D_1 - D_2\|_1$ obtenue dans l'exercice précédent. D_3 et D_1 sont-elles voisines ?*

2 Mise en place de mécanismes sur une base

Dans cette section, on reprend le tableau 1.

Exercice 2.1 (Sensibilités de requêtes). *Pour chacune des requêtes suivantes, évaluer sa sensibilité, ou la sensibilité de la fonction d'utilité associée.*

— *Q_1 : « nombre de patients de moins de 40 ans atteints d'un cancer » ;*
— *Q_2 : « répartition des âges (histogramme/camembert) de l'ensemble des patients ».*
— *Q_3 : « maladie la plus fréquente » ;*
— *Q_4 : « âge moyen des patients » ;*
— *Q_5 : « âge moyen des patients arrondi à l'entier le plus proche » ;*

Exercice 2.2 (Choisir un mécanisme naïvement). *Pour chacune des requêtes précédentes, choisir le mécanisme vérifiant la ϵ-DP qui vous paraît le plus adapté. Donner ses paramètres.*

Exercice 2.3 (Mécanisme géométrique pour valeurs entières). *On considère le mécanisme suivant qui retourne toujours une valeur entière.*

$$\mathcal{M}_G(D) = Q(D) + K \text{ tel que } \Pr[K = k] = \frac{1-\alpha}{1+\alpha}\alpha^{|k|}$$

avec $\alpha = exp(-\epsilon/\Delta_Q)$, $k \in \mathbb{Z}$. Ce mécanisme est connu sous le nom de mécanisme géométrique [1].

$D = (26, 36, 77, 157, 274, 610, 1082, 1517, 1807, 829)$ *donne le nombre d'hospitalisés en Île de France* [2] *pour Covid-19 le 18-11-2020, par classe d'âge d'amplitude 10*

1. *Soit la requête $Q(D)$ qui retourne le nombre de personnes hospitalisées qui ont moins de 30 ans dans D. Quelle est sa sensibilité ?*
2. *Remplir le tableau suivant en exploitant le mécanisme géométrique*

r	...	135	136	137	138	139	140	141	142	143	...
$\Pr[\mathcal{M}_G(D) = r]$											

3 Théorie

Exercice 3.1. *Preuves de confidentialité différentielle de certains mécanismes.*
Pour chacun des mécanismes suivants, montrer qu'il vérifie la ϵ-DP.

1. *Mécanisme laplacien (relire le transparent présentant cette preuve dans le CM).*
2. *Mécanisme exponentiel.*
3. *Mécanisme géométrique (vu à l'exercice précédent).*

Exercice 3.2. *Choix des mécanismes en fonction de leur utilité.*
Pour une requête numérique, on dispose de trois mécanismes : laplacien, géométrique, exponentiel.

1. *Pour un ϵ donné, que peut-on dire de ces trois mécanismes en terme de confidentialité différentielle ?*
2. *Quel(s) indicateur(s) pourrait-on calculer pour nous aider à choisir entre ces trois mécanisme en étant guidé par l'utilité ?*
3. *Pour ϵ positif, on admet que*
   — $\mathrm{Var}[V] = 2\left(\frac{\Delta}{\epsilon}\right)^2$ *si V suit une loi de Laplace*$(0, \frac{\Delta}{\epsilon})$
   — $\mathrm{Var}[G] = 2\dfrac{e^{-\frac{\epsilon}{\Delta}}}{\left(1-e^{-\frac{\epsilon}{\Delta}}\right)^2}$ *si G est la variable aléatoire réelle (de bruit géométrique) définie comme ci-dessus.*
   Que faudrait-il comparer le mécanisme pour choisir le mécansime maximisant l'utilité ?

1. Ghosh, A., Roughgarden, T., & Sundararajan, M. (2012). Universally utility-maximizing privacy mechanisms. SIAM Journal on Computing, 41(6), 1673-1693.
2. www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/