

Sécurité Appliquée-TP 2

Protection de la vie privée-PVP

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

7 décembre 2021

On considère le jeu de données [Adults](#) de l'UCI présentant un extrait des données de recensement en 1994/1995 aux USA et dont l'objectif initial était de prédire si telle ou telle personne allait avoir un salaire supérieur à 50K\$ par an.

On s'intéresse dans ce TP à publier des statistiques sur les statuts maritaux.

Exercice 0.1 (Initialisation).

1. Récupérer le jeu de données.
2. Mémoriser (dans un dictionnaire) les fréquences originales des statuts maritaux.
3. Les représenter l'aide d'un histogramme.

Exercice 0.2 (Histogramme respectueux).

1. On souhaite publier un histogramme à l'aide de l' ϵ -DP. Montrer que la sensibilité d'une telle requête est 1.
2. Implanter un algorithme qui construit un histogramme répondant à la question 1.

Exercice 0.3 (Statut marital le plus fréquent).

1. On souhaite répondre à la question : quel est le statut marital le plus fréquent dans la base, sachant que celle-ci va progressivement évoluer. Quel mécanisme peut-on utiliser ?
2. On propose de choisir la fonction d'utilité $u(D, r)$ qui retourne le nombre de fois où r apparaît dans D . Montrer que Δ_u vaut 1.
3. Montrer que l'on ne va pas pouvoir évaluer $\exp(\frac{\epsilon u(D, r)}{2\Delta_u})$ pour certains couples de valeurs de (ϵ, r) . Pourquoi ?
4. Proposer une fonction qui retourne le statut marital le plus fréquent, tout en respectant l' ϵ -DP pour des valeurs de ϵ (très) petites.
5. Évaluer cette fonction sur 1000 tirages en prenant $\epsilon \in [10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.1]$ et représenter graphiquement le résultat.