Sécurité Appliquée
Protection de la vie privée-PVP

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

Examen de janvier 2021

Organisation

Cette épreuve est personnelle. Vous n'avez pas le droit de vous faire aider ni de communiquer avec quiconque. Par contre, vous pouvez utiliser toutes les sources d'information possibles : votre cours, vos TDs, et notes personnelles, des exemples en ligne. Vous n'avez pas le droit de partager des éléments avec quiconque.

Le jeu de données est téléchargeable sur le site

https://www.kaggle.com/kingabzpro/heart-disease-from-cleveland.

Une description des attributs et d'une méthode d'apprentissage sur ces données est accessible à l'url

https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7.

Les réponses textuelles à cet examen doivent être consignées dans un fichier nommé "nom_prenom.odt". A la fin de de la séance, construire une archive avec tous les fichiers demandés et l'envoyer par mail.

1 S'approprier le jeu de données.

Dans cet examen, vous avez la charge travailler sur ce jeu de données en imaginant que vous n'êtes pas habilités à lire les données brutes qu'il contient. Votre rôles va consister à répondre à des questions tout en protégeant la vie privée des patients de cette base.

On commence par faire un renommage des attributs de ce jeu de données.

Exercice 1.1. *Attributs.*

1. *Renommer les attributs de ce tableau en des termes en langue française.*
2. *Cette tâche de renommage porte-t-elle atteinte à la vie privée ?*
3. *Quels attributs personnels pourraient être considérés comme quasi-identifiants ? Justifier.*

2 Calculer l'âge moyen des personnes à risque

On considère qu'une personne risque d'avoir une attaque cardiaque si elle est dans la classe 3 ou dans la classe 4. On vous pose la requête :

Q = *"quel est l'âge moyen des patients de cette classe ?"*

2.1 Par des méthodes syntaxiques

Pour répondre à cette question, vous demandez que le jeu de données soit rendu k-anonyme et vous calculerez ensuite les moyennes sur ces jeux de données pour différentes valeurs de k. En pratique, dans cet examen, c'est vous qui allez rendre k-anonyme le jeu de données

Exercice 2.1. *Hiérarchies de généralisation.*

1. *Proposer une hiérarchie de généralisation fine pour l'âge. L'exporter sous le nom "hierarchie_age.csv".*

2. *Proposer une hiérarchie de généralisation simple pour le genre. L'exporter sous le nom "hierarchie_genre.csv".*

Exercice 2.2. 6-*anonymité.*

1. *Demander à l'outil ARX de générer un jeu de données* 6*-anonyme en autorisant 5% de suppression.*
2. *Interpréter le score obtenu pour la/les généralisation(s) la/les plus utile(s).*
3. *Exporter le jeu données minimisant ce score sous le nom "k6.csv".*
4. *Ce jeu de données* 6*-anonyme contient les lignes reproduites ci-dessous :*

age	sex	cp	$trestbps$	$chol$	fbs	$restecg$	$thalach$	$exang$	$oldpeak$	$slope$	ca	$thal$	$class$
$[69, 77[$	1	4	145	174	0	0	125	1	2.6	3	0	7	4
$[69, 77[$	1	4	130	322	0	2	109	0	2.4	2	3	3	1
$[69, 77[$	1	3	160	269	0	0	112	1	2.9	2	1	7	3
$[69, 77[$	1	3	140	254	0	2	146	0	2.0	2	3	7	2
$[69, 77[$	1	1	160	234	1	2	131	0	0.1	2	1	3	0
$[69, 77[$	1	2	156	245	0	2	143	0	0.0	1	0	3	0

Vous connaissez un homme de 70 ans souffrant d'angine de poitrine due à l'effort (attribut exang). Montrer que la publication de cette base de données permet d'acquérir d'autres informations critiques sur ce patient.

5. *On décide de prendre une valeur de k-anonymat plus grande. Expérimentez plusieurs valeur pour ce k et montrer qu'il est nécessaire de prendre au moins $k = 16$ pour ce prémunir de ce genre d'attaque par connaissance supplémentaire. On pourra se concentrer sur l'attribut anglais fbs par exemple.*

Exercice 2.3. 16-*anonymité.*

1. *Exporter le jeu données* 16 *anonyme minimisant le score Loss sous le nom "k16.csv".*
2. *A l'aide d'un tableur, montrer qu'il est possible d'estimer l'âge moyen des patients dans les classes 3 ou 4. Le mettre en place dans une feuille de calcul et sauvegarder l'ensemble dans un fichier "k16.ods".*

2.2 Par confidentialité différentielle

Dans une démarche de confidentialité différentielle, une des premières tâches consiste à évaluer la sensibilité numérique de la requête.

Exercice 2.4. *Sensibilité numérique de Q.*

1. *Évaluer la sensibilité de la requête Q (donnée à la page précédente) en considérant que les patients peuvent avoir un âge entre 25 ans et 100 ans. Détailler les calculs.*

Exercice 2.5. *Réponse aseptisée à Q.*

1. *Quel mécanisme allez-vous retenir ? Justifier.*
2. *Donner une réponse à Q qui protège la vie privée des patients de la base en considérant que la fuite autorisée ϵ vaut 1. Justifier tous les calculs. Vous pouvez enregistrer vos calculs dans un tableur "DP.ods".*

2.3 Synthèse

Exercice 2.6. *Comparaison entre 16-anonymat et et la 1-DP pour la requête Q.*

1. *Entre les deux approches, quelle est celle qui retourne la réponse la plus précise ? Justifier.*
2. *Entre les deux approches, quelle est celle qui protège le mieux la vie privée des patients ? Justifier.*