ISIFC 3, Crypto
Premières analyses respectueuses de la vie privée

Jean-François Couchot
Université de Franche-Comté, UFR-ST

Plan

Généralités : big data et vie privée

Apprentissage Machine

Publication non sure de données

Un premier modèle de PVP : le k-anonymat

Plan

Plan

PVP : Exacerbé par le Big Data

Big Data & Data mining

- L'exploration de données (Data Mining) : inférence de connaissances intéressantes à partir de grandes quantités de données (Big Data)
- Tendances générale : exploration des données en croissance // Big Data
- Analyse/extraction de connaissance : techniquement réalisable aujourd'hui

Big Data : application et volume

- Domaines : veille économique, découverte scientifique, santé, profilage
- Marché du Big Data en santé : ≈ 67,82 G$ en 2025 (Globe News Wire)
- 90% de toutes les données : créées au cours des deux dernières années (IBM)
- 97,2% des organisations : investissent en Big Data et IA. (New Vantage)
- Offres d'emploi dans le domaine : ≈ 2,7M en 2020. (Forbes)

Plan

Vie Privée ? [3]

Historique

- Expression du “droit d’être laissé tranquille” (the right to be let alone) [1]
- “Nul ne sera l’objet d’immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, [. . .]. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes.” [2]
- A l’heure d’Internet et des données (personnelles) transmises par : smartphones, messageries, GPS, appareils de fitness, moteur de recherche. . .

Des exemples d’inférences problématiques

- Réfrigérateur commandant des produits consommés : $\rightsquigarrow$ nbre. de présents/absents au domicile, risque sanitaire ? assurance ?
- Application de suivi de la santé : positions, fréquences card. partagées avec Apple/Google seulement ?

1. Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. Harvard law review, 193-220.
2. Organisation des Nations Unies, (1949), Déclaration universelle des droits de l’homme.
3. https ://openclassrooms.com/fr/courses/5280946-protegez-les-donnees-personnelles

Scandales de non protection de la vie privée

National Security Agency (NSA) PRISM (2007-2013)

- En 2007, création par la NSA du programme PRISM de surveillance des communications échangées sur les services en ligne des GAFAM notamment
- Transmission des données brutes à des pays tiers
- Espionnage de leaders politiques internationaux (dont A.Merkel)

Cambridge Analytica (CA) (2014-2018)

- “This is your digital life” : application pour Facebook qui permet l’aspiration de données personnelles présentes du réseau
- 87M comptes utilisateurs Facebook aspirés dès 2014
- Ciblage politique pour convaincre de voter pour D. Trump en 2016
- ”Sans CA, il n’y aurait pas eu de Brexit” selon C. Wylie

Plan

Rôles des personnes accédant à la donnée - 1

Fournisseur de contenu

- Exemple d'une photo : les personnes sur la photo + photographe ev.

Autres entités connues du fournisseur avec lesquelles il souhaite la partager

- Exemple d'une photo : amis pour partager, fournisseur de service
- Moyens utilisés : application, mécanisme de groupes et de permissions
- Règlement Général sur la Protection des Données (RGPD) :
  - A. 6 : "Données minimales collectées et traitées de manière loyale et licite"
  - A. 12 : "Transparence des informations et des communications et modalités de l'exercice des droits de la personne concernée"
  - A. 16 : "Droit à la rectification"
  - A. 17 : "Droit à l'effacement, à l'oubli"
  - A. 20 : "Droit à la portabilité" : données récupérables (format lisible)

Rôles des personnes accédant à la donnée - 2

Gestionnaire de cette donnée

- Exemple d'une photo : employés de réseau social, d'entreprises de stockage dans le nuage ;
- Catégorie critique :
  - Modèle économique du service : extraction d'information
  - Administrateurs de BD : souvent honnête mais à risque
  - Chiffrement des données : indispensable,. . ., mais intérêt si on veut faire des analyses ?

Reste du monde

- Exemple d'une photo : celles/ceux avec lequel on ne veut pas la partager

Données de santé ? [4]

Europe : RGPD, cons. 35

- "Les données à caractère personnel concernant la santé devraient comprendre les données [...] qui révèlent des informations sur l'état de santé physique ou mentale passé, présent ou futur d'une personne."
- Ensemble contenant au moins les informations :
  - collectées lors d'une inscription à un services de santé (NSS)
  - obtenues lors de tests/examens sur le corps : résultats d'analyses biologiques, génétiques
  - concernant une maladie, un handicap, des antécédents, un traitement clinique ou l'état physiologique ou biomédical d'une personne
  - qui deviennent des données de santé, du fait de croisement avec d'autres données : nombre de pas quotidien croisé avec une mesure de poids par exemple

Question

- Pourquoi laisser tant de liberté à l'interprétation ? Avantages, inconvénients ?

4. https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante

Un régime juridique particulier[4]

Aperçu des différentes législations susceptibles de s'appliquer

- loi Informatique et Libertés (art. 8 et chapitre IX) ;
- dispositions sur le secret (art. L. 1110-4 du CSP) ;
- dispositions relatives aux référentiels de sécurité et d'interopérabilité des données de santé (art. L. 1110-4-1 du CSP) ;
- dispositions sur l'hébergement des données de santé (art. L. 1111-8 et R. 1111-8-8 et s. du CSP) ;
- dispositions sur la mise à disposition des données de santé (art. L. 1460-1 et s. du CSP) ;
- interdiction de procéder à une cession ou à une exploitation commerciale des données de santé (art. L. 1111-8 du CSP, art. L 4113-7 du CSP)...

Et ailleurs ?

Aux USA (HIPAA ?), au Canada ?

RGPD : Considérant 26

Applicabilité du RGPD aux données

- Applicable : sur des données à caractère personnel, susceptibles de permettre d'identifier (directement ou non) la personne.
- Non applicable : aux données initialement anonymes, ou rendues au moyen d'une démarche

« **Il n'y a dès pas lieu** d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. **Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche**. »

RGPD : Considérant 26

Identification d'une personne physique

« il convient de prendre en considération l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne pour identifier la personne physique. Pour établir si des moyens sont raisonnablement susceptibles d'être utilisés pour identifier une personne physique, il convient de prendre en considération l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci. »

Conséquences

- Nécessite d'évaluer régulièrement la qualité des fichiers rendus anonymes par le passé qui ont été diffusés (dont les acteurs sont responsables), de s'appuyer sur des méthodes éprouvé, publiées, dont les faiblesses sont connues pour réaliser cet anonymat, enfin de documenter l'anonymisation et les analyses de celle-ci.
- Repose sur la robustesse de la méthode d'anonymisation mise en place

Plan

Plan

Apprentissage suppervisé probabiliste

Exemple [5] de classification

Preg.	Gluc.	BloodP.	SkinThick.	Insul.	BMI	DPF	Age	Outcome
6	148	72	35	0	33.6	0.627	50	YES
1	85	66	29	0	26.6	0.351	31	NO
8	183	64	0	0	23.3	0.672	32	YES
1	89	66	23	94	28.1	0.167	21	NO
				...				

- Connaissant une valeur pour chaque attribut de mesure : predire Outcome
- Comparer les probabilités suivantes et conclure :

$$\Pr[\text{OutCome} =' \text{YES}' | \text{Preg} = p, \text{Gluc} = g, \ldots, \text{Age} = a]$$
$$\Pr[\text{OutCome} =' \text{NO}' | \text{Preg} = p, \text{Gluc} = g, \ldots, \text{Age} = a]$$

- A évaluer : $\Pr[y_1 | X_1, X_2, \ldots, X_d]$ et $\Pr[y_2 | X_1, X_2, \ldots, X_d]$
- Remarque : attributs discrets (Preg., Gluc, ...) ou réels (BMI, DPF)

5. https ://www.kaggle.com/uciml/pima-indians-diabetes-database

Bayes : théorème et conséquences

Théorème (Bayes)

$$\Pr[Y|X_1,\ldots,X_d] = \frac{\Pr[Y] \times \Pr[X_1,\ldots,X_d|Y]}{\Pr[X_1,\ldots,X_d]}$$

Simplifications immédiates

$$\left.\begin{array}{l} \Pr[y_1|X_1,\ldots,X_d] = \frac{\Pr[y_1]\times\Pr[X_1,\ldots,X_d|y_1]}{\Pr[X_1,\ldots,X_d]} \\ \Pr[y_2|X_1,\ldots,X_d] = \frac{\Pr[y_2]\times\Pr[X_1,\ldots,X_d|y_2]}{\Pr[X_1,\ldots,X_d]} \end{array}\right\} \begin{array}{l} \text{ṁ dénom. } \Pr[X_1,\ldots,X_d] \\ \rightsquigarrow \text{ son calcul : inutile} \end{array}$$

A évaluer : $\Pr[y_j]$, et $\Pr[X_1,\ldots,X_d|y_j]$

- $\Pr[y_j]$: fréquence d'apparition de la valeur y_j pour l'attribut y
- $\Pr[X_1,\ldots,X_d|y_j] = \Pr[X_1|y_j] \times \cdots \times \Pr[X_d|y_j] = \prod_{i=1}^{d} \Pr[X_i|y_j]$:
  - Hypothèse naïve : indépendance de chaque X_i p.r. aux autres $X_{i'}$, $i' \neq i$, conditionnellement à y_j

Attention

- Publication des proba. $\Pr[X_i|y_j] \leftrightarrow$ fuite d'information sur X

Exemple avec des données discrètes-1

Données [6], question et premières probabilités

Age	Income	Gender	Missed Payment
Young	Low	Male	Yes
Young	High	Female	Yes
Medium	High	Male	No
Old	Medium	Male	No
Old	High	Male	No
Old	Low	Female	Yes
Medium	Low	Female	No
Medium	Medium	Male	Yes
Young	Low	Male	No
Old	High	Female	No

- Q : défaut de payement pour une jeune femme avec un revenu moyen ?
- $y_1 \equiv'$ Missed Payment$' ='$ YES$'$, $y_2 \equiv'$ Missed Payment$' ='$ NO$'$
- Comparer $\Pr[y_1|\text{Age} = \text{Young}, \text{Income} = \text{Medium}, \text{Gender} = \text{Female}]$ et $\Pr[y_1|\text{Age} = \text{Young}, \text{Income} = \text{Medium}, \text{Gender} = \text{Female}]$
- $\Pr[y_1] = \frac{4}{10}$ et $\Pr[y_2] = \frac{6}{10}$

Probabilités pour chaque attribut conditionnellement à y_j

- $\Pr[\text{Age} = \text{Young}|y_1] = \frac{2}{4}$, $\Pr[\text{Age} = \text{Young}|y_2] = \frac{1}{6}$ ...
- $\Pr[\text{Income} = \text{Medium}|y_1] = \frac{1}{4}$, $\Pr[\text{Income} = \text{Medium}|y_2] = \frac{1}{6}$ ...
- $\Pr[\text{Gender} = \text{Female}|y_1] = \frac{2}{4}$, $\Pr[\text{Gender} = \text{Female}|y_2] = \frac{2}{6}$ ...

6. Yilmaz, E., Al-Rubaie, M., & Chang, J. M. (2019). Locally differentially private naive bayes classification. arXiv preprint arXiv:1905.01039.

Exemple avec des données discrètes-2

Evaluation de $\Pr[y_j|\text{Age} = \text{Young}, \text{Income} = \text{Medium}, \text{Gender} = \text{Female}]$

- $Pr[y_1] \times \Pr[\text{Age} = \text{Young}|y_1] \times \Pr[\text{Income} = \text{Medium}|y_1] \times \Pr[\text{Gender} = \text{Female}|y_1] = \frac{4}{10} \times \frac{2}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{1}{40} \approx 0.025$
- $Pr[y_2] \times \Pr[\text{Age} = \text{Young}|y_2] \times \Pr[\text{Income} = \text{Medium}|y_2] \times \Pr[\text{Gender} = \text{Female}|y_2] = \frac{6}{10} \times \frac{1}{6} \times \frac{1}{6} \times \frac{2}{6} = \frac{1}{180} \approx 0.0056$

Réponse

La probabilité qu'elle ratte son paiement est beaucoup plus importante que celle opposée.

Exemple avec des données continues-1

Données[7], question et premières probabilités

sexe	taille (cm)	masse (kg)	point. (cm)
masc.	182	81.6	30
masc.	180	86.2	28
masc.	170	77.1	30
masc.	180	74.8	25
fém.	152	45.4	15
fém.	168	68.0	20
fém.	165	59.0	18
fém.	175	68.0	23

- Q : sexe d'une personne mesurant 183cm, pesant 59kg et dont les pieds mesurent 20cm ?
- $y_1 \equiv'$ Sexe$' ='$ masc.$'$, $y_2 \equiv'$ Sexe$' ='$ fém.$'$
- Comparer $\Pr[y_1|\text{Taille} = 183, \text{Masse} = 59, \text{Point.} = 20]$ et $\Pr[y_2|\text{Taille} = 183, \text{Masse} = 59, \text{Point.} = 20]$ et
- $\Pr[y_1] = \Pr[y_2] = \frac{1}{2}$

Probabilités pour chaque attribut X_i conditionnellement à y_j : $\mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})$

1. Calcul des paramètres de $\mathcal{N}$: moyene $\mu_{i,j}$ et variance $\sigma^2_{i,j}$

Sexe	μ_{taille}	σ^2_{taille}	μ_{masse}	σ^2_{masse}	$\mu_{point.}$	$\sigma^2_{point.}$
masc.	178	29.3	79.9	25.5	28.25	5.58
fém.	165	92.7	60.1	114	19	11.3

2. Avec la densité de probabilité $\frac{1}{\sqrt{2\pi\sigma^2_{i,j}}} \exp\left(-\frac{1}{2\sigma^2_{i,j}}(x-\mu_{i,j})^2\right)$, calcul de

$$\Pr[\text{taille} = 183|y_1] = \frac{1}{\sqrt{2\pi \times 29.3}} \exp\left(\frac{-1}{2\times 29.3}(183-178)^2\right) \approx 0.0481$$

7. Classification naïve bayésienne, Wikipedia

Exemple avec des données continues-2

Valeurs numérique des robabilités pour chaque attribut X_i conditionnellement à y_j

- $\Pr(taille = 183|y_1) = 0.0481$, $\Pr(poids = 59|y_1) = 0.0000146$ et $\Pr(point. = 20|y_1) = 0.000381$
- $\Pr(taille = 183|y_2) = 0.00721$, $\Pr(poids = 59|y_2) = 0.0372$ et $\Pr(point. = 20|y_2) = 0.114$

Evaluation de $\Pr[y_j|taille = 183, poids = 59, point. = 20]$

- $Pr[y_1] \times \Pr(taille = 183|y_1) \times \Pr(poids = 59|y_1) \times \Pr(point. = 20|y_1) \approx 1.3404 \times 10^{-10}$
- $Pr[y_2] \times \Pr(taille = 183|y_2) \times \Pr(poids = 59|y_2) \times \Pr(point. = 20|y_2) \approx 1.52 \times 10^{-5}$

Réponse

La personne est probablement une femme.

Traitement du jeu de données

```
import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LinearRegression

from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score

from google.colab import drive
drive.mount('/content/drive')
path = "..."

dataset = pd.read_csv(path+'diabetes.csv')
dataset['Outcome'].replace({'NO': 0, 'YES': 1},inplace=True)

X = dataset.iloc[:, 0:8]
y = dataset.iloc[:, 8]

# Replace Zeroes,NaN with the median value of the column
for column in ['Glucose', 'BloodPressure', 'SkinThickness', 'BMI', 'Insulin']:
    X[column] = X[column].replace(0, np.NaN)
    mean = int(X[column].mean(skipna=True))
    X[column] = X[column].replace(np.NaN, mean)

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.20)
```

Apprentissage naïf bayésien gaussien (code)

```
gnb = GaussianNB()
gnb.fit(X_train, y_train)
y_pred = gnb.predict(X_test)

print(accuracy_score(y_test, y_pred))
print(f1_score(y_test, y_pred,pos_label='YES'))

#0.7857142857142857
#0.6373626373626374
```

Plan

Introduction puis formalisation

Même exemple que pour NB

Preg.	Gluc.	BloodP.	SkinThick.	Insul.	BMI	DPF	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
...								

- Connaissant une valeur pour chaque attribut numérique $X_1, \ldots, X_d$, de mesure : predire Outcome, $Y = Y_i$

Formalisation

- Base $D = \{(X_{11}, \ldots, X_{1d}, Y_1), \ldots, (X_{n1}, \ldots, X_{nd}, Y_n)$ de n tuples $t_1, \ldots t_n$
- Hypothèse : $Y_i = \omega_0 + \omega_1 X_{i1} + \omega_2 X_{i2} + \cdots + \omega_d X_{id} + e_i$, avec $i = 1, \ldots, n$
- e_i : erreur dans l'explication linéaire de Y_i à partir des $(X_{i1}, \ldots, X_{id})$
- $\omega = (\omega_0, \omega_1, \ldots, \omega_d)$: paramètres à estimer en minimisant $e_1, \ldots, e_n$

Notation matricielle

Formalisation

$$\begin{cases} y_1 = \omega_0 + \omega_1 x_{1,1} + \ldots + \omega_d x_{1,d} + e_1 \\ y_2 = \omega_0 + \omega_1 x_{2,1} + \ldots + \omega_d x_{2,d} + e_2 \\ \ldots \\ y_n = \omega_0 + \omega_1 x_{n,1} + \ldots + \omega_d x_{n,d} + e_n \end{cases} \Leftrightarrow$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_d \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

De manière compacte $y = X\omega + e$ avec :

- y de dimension : $(n, 1)$
- X de dimension : $(n, d+1)$
- ω de dimension : $(d+1, 1)$
- e de dimension : $(n, 1)$

Estimateur des moindres carrés ordinaires (MCO)

Objectif

- Modèle complet initial : $y_i = \omega_0 + \omega_1 x_{i,1} + \cdots + \omega_d x_{i,d} + e_i$
- Estimation finale des paramètres : $\hat{y}_i = \hat{\omega}_0 + \hat{\omega}_1 x_{i,1} + \cdots + \hat{\omega}_d x_{i,d}$
- Résidus estimés : différence entre la valeur de y observée et estimée

$$\hat{e}_i \equiv y_i - \hat{y}_i$$

Quelles valeurs de $\omega_0, \ldots, \omega_d$ minimisent la somme des carrés des résidus ?

- $f_D(\omega) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - (\omega_0 + \omega_1 x_{i,1} + \cdots + \omega_d x_{i,d})\right)^2$ à minimiser
- Rechercher des solutions de $\frac{\partial(\sum \hat{e}_i^2)}{\partial \omega_j} = 0$, ($j = d + 1$ équations)
- Solution (facile) : $\hat{\omega} = (X^T X)^{-1} X^T Y$, X^T la transposée de X

Attention

- Publication de $\hat{\omega} \leftrightarrow$ fuite d'information sur X

Exemple

A partir des données $D = \{(1, 0.4), (0.9, 0.3), (-0.5, -1)\}$

- $y = X\omega + e$ avec $y = \begin{pmatrix} 0.4 \\ 0.3 \\ -1 \end{pmatrix}$, $X = \begin{pmatrix} 1 & 1 \\ 1 & 0.3 \\ 1 & -0.5 \end{pmatrix}$, $\omega = \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix}$ $e = \begin{pmatrix} e_0 \\ e_1 \end{pmatrix}$
- Fonction MCO : $f_D(\omega) = 3\omega_0^2 + 2.8\omega_0\omega_1 + 0.6\omega_0 + 2.06\omega_1^2 - 2.34\omega_1 + 1.25$
- Estimateur MCO : $\hat{\omega} = \begin{pmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{pmatrix} = (X^T X)^{-1} X^T Y = \begin{pmatrix} -\frac{564}{1055} \\ \frac{393}{422} \end{pmatrix}$
- Vérifications :
  - $\hat{y}_1 = -\frac{564}{1055} + 1\frac{393}{422} = \frac{837}{2110} \approx 0.397.$
  - $\hat{y}_2 = -\frac{564}{1055} + 0.9\frac{393}{422} = \frac{1280}{4220} \approx 0.306.$
  - $\hat{y}_3 = -\frac{564}{1055} - 0.5\frac{393}{422} = \frac{4221}{4220} \approx -1.0002.$
- Valeur minimum de $f_D(\omega)$: $\approx 2.36E - 5$

Regression Linéaire multiple (code)

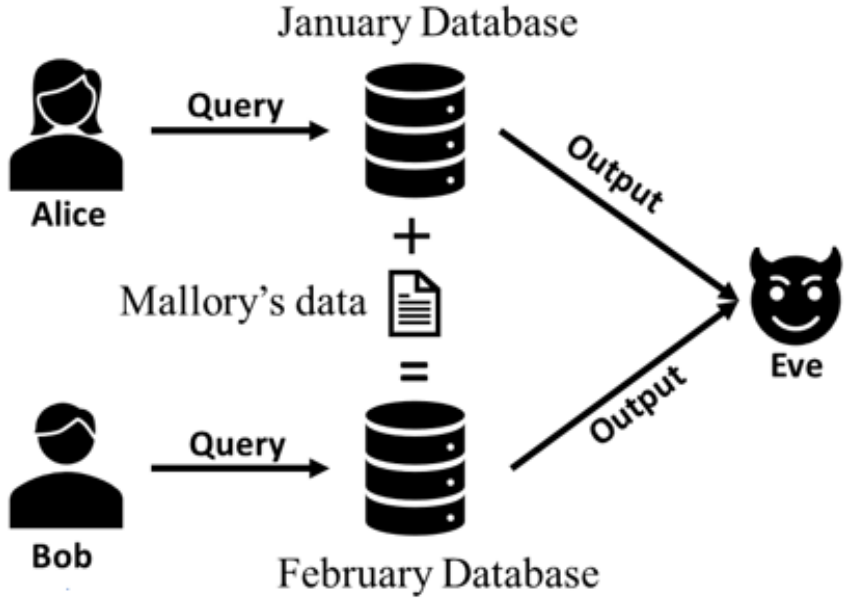
```
reg = LinearRegression()
reg.fit(X_train, y_train)
y_pred = np.array([0 if a_ <0.5 else 1 for a_ in reg.predict(X_test)])

print(accuracy_score(y_test, y_pred))
print(f1_score(y_test, y_pred))

#0.8051948051948052
#0.6428571428571429
```

Plan

Plan

Attaques sur des bases statistiques [8]

Publication de données statistiques



- Requête mensuelle : (nb empl., salaire moyen).
- Res. :$\{$jan. : $(100, \$55000)$, fev. : $(101, \$56000)\}$.

Pas toujours robuste

- Connaissance supplémentaire : 0 sortie + Mallory en fev..
- $\rightsquigarrow$ salaire de Mallory : \$156000.

8. Privacy-Preserving Machine Learning. Manning Early Access Program Publications, 2021.

Plan

Pseudonymisation

- Champs identifiants : supprimés et remplacés par un id (H(NSS)).

	Non-sensibles				Sensibles
H	CP	Age	Genre	Nationalité	Pathologie
c4ca4238a0b923820dcc509a6f75849b	13053	28	H	russe	trouble cardiaque
c81e728d9d4c2f636f067f89cc14862c	13068	29	H	américaine	trouble cardiaque
eccbc87e4b5ce2fe28308fd9f2a7baf3	13068	21	F	japonaise	infection virale
a87ff679a2f3e71d9181a67b7542122c	13053	23	H	américaine	infection virale
e4da3b7fbbce2345d7772b0674a318d5	14853	49	H	indienne	cancer
1679091c5a880faf6fb5e6087eb1b2dc	14853	48	F	russe	trouble cardiaque
8f14e45fceea167a5a36dedd4bea2543	14850	47	H	américaine	infection virale
c9f0f895fb98ab9159f51fd0297e236d	14850	49	F	américaine	infection virale
45c48cce2e2d7fbdea1afc51c7c6ad26	13053	31	H	américaine	cancer
d3d9446802a44259755d38e6d163e820	13053	37	H	indienne	cancer
6512bd43d9caa6e02c990b0a82652dca	13068	36	F	japonaise	cancer
c20ad4d76fe97759aa27a0c99bff6710	13068	35	F	américaine	cancer

- Avantage : calculs identiques à ceux sur la base de données initiale (Age moyen/cancer=37,8)

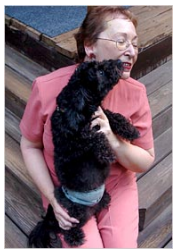
Pseudonymisation : attaque par intersection [9]

2006, diffusion par AOL de 20M requêtes, 658K utilisateur sans nom

AnonID	Query	QueryTime
1326	"holiday mansion houseboat"	2006-03-29
1326	"back to the future"	2006-04-01
591476	"english spanish translator"	2006-03-20
591476	"panama vacations"	2006-03-20
591476	"breast reduction"	2006-03-23
591476	"volunteer work at hospitals in brooklyn"	2006-05-24
591476	...	...
591476	"how to secretly poison your ex"	2006-03-12

Thelma Arnold, 62 ans, veuve vivant à Lilburn, Ga., réidentifiée en 3 j.

AnonID	Query
4417749	"people with last name 'Arnold"'
4417749	"landscapers in Lilburn, Ga"
4417749	"60 single men"
4417749	"dog that urinates on everything"
4417749	dog-related queries



⇝ Suppression hâtive des données sur le site d'AOL.

9. BARBARO, Michael, ZELLER, Tom, et HANSELL, Saul. A face is exposed for AOL searcher no. 4417749. New

Plan

Plan

Les quasi-identifiants (QID)

Intuition et définition

- QID, intuition [10] : "des éléments qui ne sont pas en eux-mêmes des identificateurs uniques, mais qui sont suffisamment bien corrélés avec une entité pour pouvoir être combinés avec d'autres quasi-identifiants afin de créer un identificateur unique"
- QID, définition [11] : Les attributs de $Q \subseteq \{A_1, \ldots, A_M\}$ sont quasi-identifiants de la relation T si la requête suivante retourne au moins un résultat

```
1  SELECT Q FROM T GROUP BY Q HAVING COUNT(*)=1
```

Exemple

- (CP, genre, date de naissance) : triplets uniques dans 87% des cas $\rightsquigarrow$ quasi-identifiants

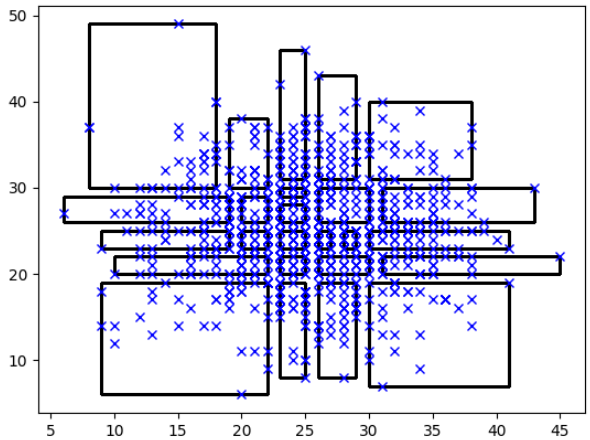
10. https://en.wikipedia.org/wiki/Quasi-identifier

11. Nguyen, B., & Castelluccia, C. (2020). Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. arXiv preprint arXiv :2001.02650.

Le k-anonymat [12]

Intuition : regrouper les QID pour casser l'unicité

- Niveau de détail des valeurs des QID : à réduire pour qu'il y ait au moins k individus différents dont les QIDs sont égaux
- Individus avec mêmes QIDs : font partie de la même classe d'équivalence



Définition de la propriété

Un jeu de données D est k-anonyme si les informations relatives à chaque personne dans celui-ci ne peuvent être distinguées d'au moins $k-1$ individus dont les informations figurent dans D. Aucun résultat ne doit être retourné par :

```
1  SELECT Q, COUNT(*) AS C FROM T GROUP BY Q HAVING C > 0 AND C < k
```

12. Sweeney, L. (2002). k-anonymity : A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

k-anonymat par généralisation

Données avant et 4-anonymes

	Non-sensibles				Sensibles
H	CP	Age	Genre	Nationality	Pathologie
1	13053	28	H	russe	trouble cardiaque
2	13068	29	H	américaine	trouble cardiaque
3	13068	21	F	japonaise	infection virale
4	13053	23	H	américaine	infection virale
5	14853	49	H	indienne	cancer
6	14853	48	F	russe	trouble cardiaque
7	14850	47	H	américaine	infection virale
8	14850	49	F	américaine	infection virale
9	13053	31	H	américaine	cancer
10	13053	37	H	indienne	cancer
11	13068	36	F	japonaise	cancer
12	13068	35	F	américaine	cancer

$\rightsquigarrow$

CP	Age	Genre	Nationalité	Pathologie	
{13053 13058}	[20; 30[	*	*	trouble cardiaque	4 ind.
	[20; 30[	*	*	trouble cardiaque	
	[20; 30[	*	*	infection virale	
	[20; 30[	*	*	infection virale	
{14850 14853}	[40; 50[	*	*	cancer	4 ind.
	[40; 50[	*	*	trouble cardiaque	
	[40; 50[	*	*	infection virale	
	[40; 50[	*	*	infection virale	
{13053 13058}	[30; 40[	*	*	cancer	4 ind.
	[30; 40[	*	*	cancer	
	[30; 40[	*	*	cancer	
	[30; 40[	*	*	cancer	

Hiérarchie de généralisation (après avoir enlevé H)

- CP : laisser, **regroupements par 2**, suppr.
- Age : laisser, par intervalles d'amplitudes 10, 20,**suppr.**
- Genre : laisser, **suppr.**
- Nationalité : laisser, par continent, **suppr.**

Loss : Perte dûe à la généralisation

Définition pour $|T|$ enregistrements, n QID et propriétés

$$Loss = \frac{1}{n|T|} \sum_{j=1}^{|T|} \sum_{i=1}^{n} \frac{R_{i_j}}{R_i}$$

- R_{i_j}/R_i : rapport acquis/acquérables
  - discret : $(|\text{généralisation } i_j| - 1)/(|\text{Qid } i| - 1)$
  - continu : $(U_{i_j} - L_{i_j})/(U_i - L_i)$
- Moyenne des acquis/acquérables, utilité décroissante // *Loss* croît

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
{13053 13058}	[20; 30[	*	*	trouble cardiaque
	[20; 30[	*	*	trouble cardiaque
	[20; 30[	*	*	infection virale
	[20; 30[	*	*	infection virale
{14850 14853}	[40; 50[	*	*	cancer
	[40; 50[	*	*	trouble cardiaque
	[40; 50[	*	*	infection virale
	[40; 50[	*	*	infection virale
{13053 13058}	[30; 40[	*	*	cancer
	[30; 40[	*	*	cancer
	[30; 40[	*	*	cancer
	[30; 40[	*	*	cancer

- Nat. $= \{\text{russe}, \text{am.}, \text{jap.}, \text{ind.}\} \rightsquigarrow$ $R_{\text{Nat}} = 3$, $R_{\text{CP}} = 3$, $R_{\text{Genre}} = 1$,
- $R_{\text{Age}} = 50 - 20 = 30$
- $Loss = \frac{1}{4 \times 12}\left(12 \times \left(\frac{1}{3} + \frac{10}{30} + \frac{1}{1} + \frac{3}{3}\right)\right) = \frac{2}{3}$

k-anonymat : attaques

Exemple

CP	Age	Genre	Nationalité	Pathologie
	[20; 30[	*	*	trouble cardiaque
{13053	[20; 30[	*	*	trouble cardiaque
13058}	[20; 30[	*	*	infection virale
	[20; 30[	*	*	infection virale
	[40; 50[	*	*	cancer
{14850	[40; 50[	*	*	trouble cardiaque
14853}	[40; 50[	*	*	infection virale
	[40; 50[	*	*	infection virale
	[30; 40[	*	*	cancer
{13053	[30; 40[	*	*	cancer
13058}	[30; 40[	*	*	cancer
	[30; 40[	*	*	cancer

4 individus

4 individus

4 individus

Attaques

- Homogénéité :
  - $\oplus$ Patient de 35 ans connu $\rightsquigarrow$ cancer.
  - $\ominus$ Patient de 29 ans connu $\rightsquigarrow$ ~~cancer~~.
- Connaissance supplémentaire : un japonais de 21 ans, P(trouble cardiaque|japonais)=faible $\rightsquigarrow$ infection virale.