Big Data

Examen Théorique - Juin 2025

BUT 3

Durée : 2h00

Guillaume Metzler et Antoine Rolland
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

L'usage de la calculatrice, de l'ordinateur ou de tout autre matériel électronique n'est pas autorisé pendant la durée de cet examen. L'usage des notes des fiches de cours et des fiches personnelles est également prohibé.

Les exercices cet examen sont indépendants et peuvent être traités dans n'importe quel ordre. Il est simplement demander de préciser l'exercice ainsi que la question traitée.

Questions de cours

1. Expliquez quel est le problème de la *volumétrie* des données en Big data. Donner un exemple où ce cas est rencontré.
2. Expliquez quel est le problème de la *variété* en Big data. Donner un exemple pour illustrer.
3. Expliquez quel est le problème de la *vélocité* en Big data. Donner un exemple pour illustrer.
4. Donner les trois autres "V" vus en cours.

Exercice 1 : Complexité

1. Classer les complexités suivantes, de la plus complexe à la moins complexe :
   - $\mathcal{O}(n!)$
   - $\mathcal{O}(n\log(n))$
   - $\mathcal{O}(\sqrt{n})$
   - $\mathcal{O}(\log(n))$
   - $\mathcal{O}(n)$
   - $\mathcal{O}(\log(n^2))$
   - $\mathcal{O}(\exp(n))$

2. On considère les algorithmes suivants :

```
## Algorithme 1 ##

x <- rnorm(n)

## Algorithme 2 ##

Fibo = function(n)
  if (n <= 1){
    return(n)
  } else {
    return(Fibo(n-1) + Fibo(n-2))
}

## Algorithme 3 ##

Fibo = function(n){
a = 0
b = 1
for (i in 1:n){
    Fib = a+b
    a = b
    b = Fib
    }
return(Fib)
}

## Algorithme 4 ##

myfunc = function(n){
    x = rnorm(n)
    for (i in 1:length(x)){
        for (j in 1:n){
            print(x[i]*x[j])
        }
    }
}
```

Donner la complexité de ces différents algorithmes.

3. On suppose maintenant que l'on dispose d'un jeu de données $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, où $\mathbf{x}_i \in \mathbb{R}^p$.

   (a) Quelle est la complexité de l'algorithme qui consiste à calculer toutes les distances deux à deux entre les exemples de l'ensemble S ?

   (b) Quelle est le coût de stockage (complexité de stockage) de cette matrice des distances ?

   (c) Si on considère un nouvel exemple $\mathbf{x}'$ que l'on cherche à classer à l'aide des données de l'ensemble S, à l'aide d'un algorithme k-NN. Quelle est la complexité de la procédure en fonction de k, m, p ?

Exercice 2 : Distance

L'algorithme du k plus proche voisin est un algorithme permettant de faire de la classification et de la régression. Le fonctionnement de cet algorithme repose sur la notion de *distance entre individus*.

On considère deux individus **indépendants** X_i et X_j appartenant à l'hypercube $[0,1]^p$. Nous faisons aussi l'hypothèse que les coordonnées des vecteurs X_i et X_j suivent une distribution uniforme dans $[0,1]$, *i.e.*,

$$\forall\, k \in [\![1,p]\!],\ X_i^{(k)} \sim \mathcal{U}([0,1]) \text{ et } X_j^{(k)} \sim \mathcal{U}([0,1]),$$

et que les coordonnées sont toutes **indépendantes**. On commence par étudier la distance entre les deux individus X_i et X_j.

1. Rappeler la densité de la variable aléatoire X suivant une loi uniforme sur $[0,1]$. Calculer son espérance et sa variance.

2. On cherche à calculer la distance moyenne entre les individus X_i et X_j. Pour cela, on utilisera le norme L_2, dite *norme euclidienne*, entre deux vecteurs.

   (a) Rappeler la définition de norme L_2 d'un vecteur $\mathbf{x} \in \mathbb{R}^p$.
   (b) Donner l'expression de la distance euclidienne au carré entre les individus X_i et X_j.
   (c) On considère deux variables aléatoires indépendantes U et U' qui suivent une loi uniforme sur $[0,1]$.
   Déterminer l'espérance de la variable aléatoire $(U-U')^2$.
   (d) En déduire la distance moyenne entre les deux points X_i et X_j, en fonction de la dimension p.

3. Que peut-on dire sur la distance entre les différents points lorsque la dimension du jeu de données augmente ?

Exercice 3 : Calcul sur des serveurs

1. On s'intéresse à une variable aléatoire Y qui est attribut sensible relative à des individus, cependant, nous souhaiterions connaître quelques grandeurs statistiques à son sujet. Pour cela, chaque serveur s ne peut retourner que deux informations : (i) la valeur moyenne de Y, notée $\bar{Y}_s$ sur le serveur, ainsi que le nombre n_s d'individus.

   (a) Est-il possible de déterminer la valeur moyenne de la variable Y sur l'ensemble des serveurs ? Si oui, expliquer comment.
   (b) Est-il possible de déterminer la valeur variance de la variable Y sur l'ensemble des serveurs ? Si oui, expliquer comment. Si non, que nous faudrait-il comme information ? Justifier votre réponse.

2. Comment calculer une médiane si les données sont stockées sur N serveurs **ordonnés** (c'est-à-dire que les plus petites valeurs sont dans le serveur 1 et les suivantes dans le serveur 2 et ainsi de suite jusqu'aux plus grandes valeurs qui se trouve, elles, dans le serveur N).