Big Data

TD 3 : Gestion des données et traitements statistiques

BUT 3

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; antoine.rolland@univ-lyon2.fr

Les exercices de cette fiche sont des exercices présentées lors du cours introductif et permettent de revenir sur des notions de calculs sur plusieurs serveurs.

Exercice 1 : Petit échauffement

Dans cet exercice, on suppose que l'on dispose de N serveurs sur lesquels sont stockées des données (plus ou moins sensibles) relatives à un ensemble de n individus.

Calcul de grandeurs statistiques et confidentialité

Nous faisons l'hypothèse que les informations relatives à un individu $i \in [\![1, n_s]\!]$ est stockée sur un seul et unique serveur et que chaque s, $s \in [\![1, N]\!]$ contient un nombre n_s d'informations.

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 ?
   Indice : penser à l'ANOVA et à la décomposition de la variance

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 serveurs 2 et ainsi de suite jusqu'aux plus grandes dans le serveur N).

Optimisation du stockage des données

On regarde maintenant un problème pour stocker des données de façon optimale pour des traitements pré-définis. Une commanditaire vient me voir car elle souhaite stocker des données temporelles représentant la température toutes les 30 secondes sur une année de la façon la plus pertinente possible pour avoir accès à la moyenne entre deux moments t_1 et t_2 dans l'année. Quelle(s) information(s) doit-elle stocker pour que le calcul soit rapide. Quel(s) problème(s) numérique(s) risque-t-elle de rencontrer ? Proposer une solution.

Exercice 2 : Données manquantes

Soit $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ un n échantillon avec une loi ayant un moment d'ordre 2 et d'espérance μ dont $n - m$ valeurs sont manquantes avec $m \in \{1, \ldots, n-1\}$. Comme nous avons des variables indépendantes et identiquement distribuées, l'ordre des valeurs est arbitraire et nous supposons que les m premières valeurs sont celles qui sont connues et les $n - m$ suivantes sont manquantes. Nous étudions le vecteur

$$\tilde{\mathbf{Y}}_n = \big(Y_1, \ldots, Y_m, \underbrace{\bar{Y}_m, \ldots, \bar{Y}_m}_{n-m \text{ fois}}\big) \text{ où } \bar{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i. \tag{1}$$

1. Explicitez la forme de la moyenne empirique $\tilde{\bar{y}}_n$ basée sur le vecteur $\tilde{\mathbf{y}}_n$ de l'équation (1) en fonction de la moyenne $\bar{Y}_m$.
2. Montrez que la moyenne empirique $\tilde{\bar{Y}}_n$ basée sur le vecteur $\tilde{\mathbf{Y}}_n$ de l'équation (1) est presque sûrement égale à la moyenne empirique $\bar{Y}_m$ basée uniquement sur les observations présentes dans le n échantillon $\mathbf{Y}_n$.
3. Donnez la forme de l'estimateur de la variance en prenant :
   (a) toutes les valeurs du vecteur $\tilde{\mathbf{Y}}_n$ de l'équation (1).
   (b) uniquement les valeurs connues du n échantillon $\mathbf{Y}_n$.
4. Démontrez que l'estimation (a) faite avec le vecteur $\tilde{\mathbf{Y}}_n$ de l'équation (1) est presque sûrement strictement plus petite que l'estimation (b) faite uniquement avec les valeurs connues du n l'échantillon $\mathbf{Y}_n$.