Big Data
TD 1 : Impact de de la grande dimension
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 ont pour objectif de *prouver* les phénomènes présentés en cours concernant la grande dimension, à l'aide de calculs.

Exercice 1 : Géométrie en grande dimension

On considère la base canonique $\mathscr{B} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p)$ de l'espace $\mathbb{R}^p$, *i.e.*,

$$\mathbf{e}_i = (0, 0, \dots, 0, \underset{\substack{\uparrow \\ i\text{-ème position}}}{1}, 0, \dots, 0, 0)$$

et on considère le vecteur $\mathbf{a} = (1, 1, \dots, 1) \in \mathbb{R}^p$. Le vecteur $\mathbf{a}$ représente la diagonale principale de l'hypercube $[0, 1]^p$.

Montrer que lorsque p tend vers $+\infty$, le vecteur $\mathbf{a}$ devient orthogonal à l'ensemble des vecteurs de la base $\mathscr{B}$.
Conseil : on pourra utiliser la définition du produit scalaire et calculer le cosinus de l'angle entre les deux vecteurs.

Exercice 2 : Distance entre des points

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*. Au cours de notre première séance, nous avons vu qu'en grande dimension, la distance entre deux exemples a tendance à augmenter avec la dimension du jeu de données avec une impression équi-distance entre les exemples et de vide dans l'espace $\mathbb{R}^p$ étudié.

Dans cet exercice, on va chercher à expliquer ce phénomène en étudiant les distances moyennes entre exemples et on cherchera à voir combien de points sont nécessaires pour combler le vide dans cet espace.

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**.

Distance entre deux points

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.

On peut également montrer (cela est beaucoup plus compliqué [1]) que l'écart-type de la distance entre deux individus est environ égale $0.2\sqrt{p}$.

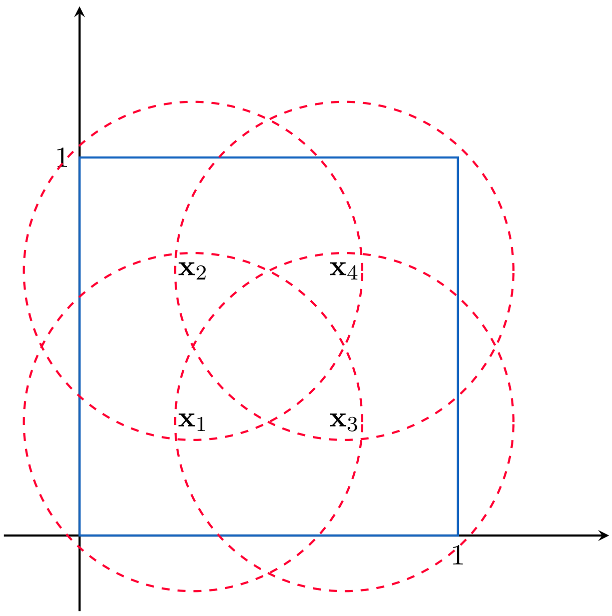
3. Calculer le coefficient de variation associé à la distance entre deux individus et interpréter.

Comment combler le vide ?

La précédente partie a montré, qu'en grande dimension, les exemples sont rapidement éloignés. On souhaite maintenant voir combien de points sont nécessaires pour remplir notre hypercube $[0,1]^p$, *i.e.*, on va chercher à déterminer le nombre n de points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ nécessaires pour s'assurer que n'importe quel point $\mathbf{x}$ de l'espace $[0,1]^p$ se trouve à une distance inférieure à 1 d'un point $\mathbf{x}_i$.

Pour cela, notons $B(\mathbf{x}_i, 1)$ la boule centrée en $\mathbf{x}_i$ et de rayon 1. Si on parvient à recouvrir l'hypercube $[0,1]^p$ par un nombre suffisamment important de boules, alors c'est gagné ! Il faut donc trouver une valeur de n telle que la somme des volumes des n boules soit supérieures au volume de l'hypercube.
Regardons un petit exemple en dimension 2 avec le carré $[0,1] \times [0,1]$ on l'on peut voir que l'espace est bien recouvert à l'aide des boules formées définies par les quatre points $\mathbf{x}_1$, $\mathbf{x}_2$,$\mathbf{x}_3$ et $\mathbf{x}_4$.

1. Pour cela, il faudrait calculer la variance du carré d'une loi uniforme dont la valeur exacte n'est pas calculable mais peut être approchée par une δ-méthode, qui consiste à faire un développement limité.



1. Donner le volume de l'hypercube $[0,1]^p$.
2. En utilisant le fait que le volume de la boule unité peut être approchée, en grande dimension lorsque $p \to \infty$, par
$$V_p(1) \simeq \left(\frac{2\pi e}{p}\right)^{p/2} \times (\pi p)^{-1/2}.$$
   (a) Traduire, par une inégalité le fait que la réunion des volumes des boules doit être plus grande que le volume de l'hypercube.
   (b) En déduire une valeur minimale de n pour que l'on ait un recouvrement.
3. Le recouvrement est-il assuré ? Commenter le résultat précédent. On pourra effectuer un dessin pour s'aider.