Big Data

Correction 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 $\mathscr{B} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p)$ de l'espace $\mathbb{R}^p$, *i.e.*,

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

et on considère le vecteur $\mathbf{a} = (1, 1, \ldots, 1) \in \mathbb{R}^p$. 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.

Notons θ_i l'angle entre les vecteurs $\mathbf{a}$ et $\mathbf{e}_i$. Par définition du produit scalaire, nous avons

$$\cos(\theta_i) = \frac{\langle \mathbf{a}, \mathbf{e}_i \rangle}{\|\mathbf{a}\| \|\mathbf{e}_i\|} = \frac{1}{\sqrt{p}} \underset{p \to +\infty}{\longrightarrow} 0.$$

Ainsi, le cosinus de l'angle entre les deux vecteurs tend vers 0 en grande dimension, cela signifie que l'angle formé entre les deux vecteurs est égal à $\pm\pi/2$.

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 équidistance 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.

   La densité f de la loi uniforme sur $[0,1]$ est définie par

   $$f(x) = \mathbb{1}_{\{x\in[0,1]\}},$$

   où $\mathbb{1}_{\{x\in[0,1]\}}$ est la fonction indicatrice du segment $[0,1]$ qui prend la valeur 1 si $x \in [0,1]$ et 0 sinon.

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 $\mathbf{x} \in \mathbb{R}^p$.

   La norme du vecteur $\mathbf{x}$ est donnée par

   $$\|\mathbf{x}\|_2 = \sqrt{\sum_{k=1}^{p} (x^{(k)})^2}.$$

   (b) Donner l'expression de la distance euclidienne au carré entre les individus X_i et X_j.

   En reprenant la définition de norme de la question précédente, nous obtenons directement

$$\|X_i - X_j\|_2^2 = \sum_{k=1}^{p} \left(X_i^{(k)} - X_j^{(k)}\right)^2.$$

(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 $(U-U')^2$.

On va simplement développer le carré et calculer l'espérance.

$$\begin{aligned}
\mathbb{E}[(U-U')^2] &= \mathbb{E}[U^2 - 2UU' + U'^2], \\
&\quad \downarrow \text{ par linéarité de l'espérance} \\
&= \mathbb{E}[U^2] - 2\underbrace{\mathbb{E}[UU']} + \mathbb{E}[U'^2], \\
&\quad \downarrow \text{ U et U' sont indépendantes} \\
&= \mathbb{E}[U^2] - 2\mathbb{E}[U]\,\mathbb{E}[U'] + \mathbb{E}[U'^2], \\
&\quad \downarrow U \text{ et } U' \text{ suivent la même loi} \\
&= 2\,\mathbb{E}[U^2] - 2\,\mathbb{E}[U]^2, \\
&\quad \downarrow \text{ on reconnaît la définition de la variance de } U \\
&= 2\,\mathrm{Var}[U], \\
&= \frac{1}{6}.
\end{aligned}$$

(d) En déduire la distance moyenne entre les deux points X_i et X_j, en fonction de la dimension p.

On sait que les coordonnées $X_i^{(k)}$ et $X_j^{(k)}$ suivent toutes deux des lois uniformes sur $[0,1]$ et indépendantes. Il nous suffit donc de mobiliser les deux questions précédentes.

$$\begin{aligned}
\mathbb{E}[\|X_i - X_j\|^2] &= \mathbb{E}\left[\sum_{k=1}^{p} \left(X_i^{(k)} - X_j^{(k)}\right)^2\right], \\
&\quad \downarrow \text{ par linéarité de l'espérance} \\
&= \sum_{k=1}^{p} \mathbb{E}\left[\left(X_i^{(k)} - X_j^{(k)}\right)^2\right], \\
&\quad \downarrow X_i^{(k)} \sim \mathcal{U}([0,1]) \text{ et } X_i^{(k)} \sim \mathcal{U}([0,1]), \text{ à l'aide de la question précédente} \\
&= \sum_{k=1}^{p} \frac{1}{p}, \\
&= \frac{p}{6}.
\end{aligned}$$

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.

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.

Le coefficient de variation v est défini comme par le rapport de l'écart-type à la moyenne :

$$v = \frac{\sqrt{\mathrm{Var}[\|X_i - X_j\|^2]}}{\mathbb{E}[\|X_i - X_j\|^2]} \simeq \frac{1.2}{\sqrt{p}}.$$

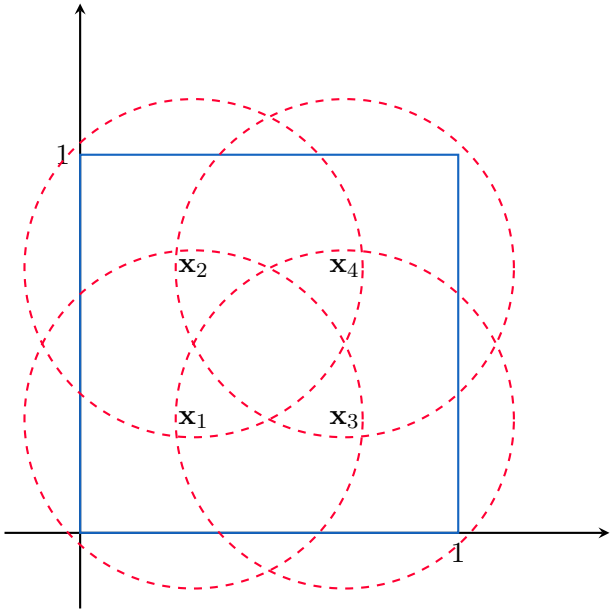
C'est une mesure de dispersion relative qui permet de tenir compte de l'ordre de grandeur des prises par la quantité étudiée.
Si on remarque que la moyenne et l'écart-type des distances augmentent avec la dimension, la dispersion relative, elle, tend vers 0 lorsque p tend vers l'infini montrant que l'écart-type des distances devient négligeable devant la distance moyenne entre les exemples.

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 que la condition suivante soit vérifiée :

$$\forall\, \mathbf{x} \in \mathbb{R}^p;\ \exists\, i \in [\![1, n]\!] \text{ t.q. } \|\mathbf{x} - \mathbf{x}_i\| \leq 1.$$

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. Donner le volume de l'hypercube $[0,1]^p$.

   Le volume est directement égal 1.

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.

Si on souhaite que la réunion de n boules ait un volume supérieur au volume de l'hypercube, il faut que n vérifie l'inégalité suivante

$$1 \leq nV_p(1).$$

(b) En déduire une valeur minimale de n pour que l'on ait un recouvrement.

On va simplement résoudre l'inéquation précédente pour trouver une valeur approchée de n.

$$\begin{aligned} & 1 \leq nV_p(1), \\ \iff & n \geq V_p(1)^{-1}, \\ \iff & n \geq \left(\frac{2\pi e}{p}\right)^{-p/2} \times (\pi p)^{1/2}, \\ \iff & n \geq \left(\frac{p}{2\pi e}\right)^{p/2} \times (\pi p)^{1/2} \end{aligned}$$

Regardons le nombre d'exemples nécessaires n pour différentes valeurs de p.

p	1	2	3	4	5	10	20	50	100
n	4	22	141	1035	8500	8×10^8	1.7×10^{20}	2.2×10^{59}	5.9×10^{132}

Bien évidemment, cela n'a pas beaucoup de sens pour les petites valeurs de p ! Il faut vraiment s'intéresser au comportement au grande dimension, donc pour de grandes valeurs de p. Il est néanmoins intéressant de voir que pour de "petites valeurs" de p, il faille déjà énormément de points pour effectuer notre recouvrement.

A titre comparaison, il n'y a que 3.1×10^7 secondes dans une année ou encore entre 10^{80} et 10^{85} particules dans l'univers !

3. Le recouvrement est-il assuré ? Commenter le résultat précédent. On pourra effectuer un dessin pour s'aider.

Si la condition est respectée, oui le recouvrement est assuré. En revanche, si on tire aléatoirement les points, il n'y aucune garantie de recouvrir l'hypercube ! (Je me demande comment faire d'ailleurs ...)