Réduction de dimension par projection aléatoire

On se place dans $\mathbb{R}^d$ muni de la norme euclidienne standard, notée $\|\cdot\|$. On cherche à réduire la dimension ambiante sans trop distordre les distances entre les données. Pour cela on va démontrer le résultat suivant, dû à Johnson et Lindenstrauss :

Théorème 0.1. *Soit $\varepsilon \in]0,1[$ un réel, et k, n deux entiers strictement positifs tels que*

$$k \geq 4\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)^{-1} \ln n.$$

Alors, pour tout nuage de points $P = \{p_1, \cdots, p_n\} \subset \mathbb{R}^d$, il existe une application $f : \mathbb{R}^d \to \mathbb{R}^k$ telle que

$$(1-\varepsilon)\|p_i - p_j\|^2 \leq \|f(p_i) - f(p_j)\|^2 \leq (1+\varepsilon)\|p_i - p_j\|^2 \tag{1}$$

pour toute paire de points $p_i, p_j \in P$.

Question 0.1. Pour quelles valeurs de n ce résultat est-il utile, sachant d et ε fixés ? Justifiez votre réponse.

La preuve du théorème emploie un argument probabiliste : on choisit f aléatoirement parmi une famille particulière de fonctions, et on montre que la probabilité que f vérifie (1) est non nulle, ce qui implique l'existence d'au moins une fonction f vérifiant (1).

On prend f parmi la famille des projections orthogonales sur des sous-espaces vectoriels aléatoires de $\mathbb{R}^d$ de dimension k, avec probabilité uniforme sur le choix du sous-espace. Pour toute paire de points $p_i \neq p_j \in P$, on s'intéresse alors au ratio $\|f(p_i) - f(p_j)\|^2 / \|p_i - p_j\|^2$. L'astuce est que, comme on choisit le sous-espace avec probabilité uniforme, la distribution de probabilité de ce ratio est la même que celle du ratio $\|\pi_k(X)\|^2/\|X\|^2$, où $X = (X_1, \cdots, X_d)$ est un vecteur aléatoire dont les coordonnées sont indépendantes et identiquement distribuées selon la loi normale $\mathcal{N}(0,1)$ dans $\mathbb{R}$, et où π_k désigne la projection orthogonale de X sur le sous-espace vectoriel engendré par les k premières coordonnées dans $\mathbb{R}^d$. En d'autres termes, la distortion métrique liée à la projection d'un vecteur fixé sur un sous-espace aléatoire suit la même loi que la distortion métrique liée à la projection d'un vecteur aléatoire sur un sous-espace fixé. Ce fait sera admis ici.

Question 0.2. Montrez que $\mathbb{E}\left[\|\pi_k(X)\|^2\right] = \frac{k}{d}\,\mathbb{E}\left[\|X\|^2\right]$.

Le point clé de la preuve du théorème est que la variable aléatoire $\|\pi_k(X)\|^2/\|X\|^2$ est relativement concentrée autour de sa moyenne k/d. Plus précisément, en supposant $k < d$ (ce que l'on fera dans toute la suite de l'exercice), on a les inégalités de concentration suivantes (admises ici) :

$$\beta < 1 \implies \Pr\left[\frac{\|\pi_k(X)\|^2}{\|X\|^2} < \beta\,\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(1-\beta+\ln\beta)\right) \tag{2}$$

$$\beta > 1 \implies \Pr\left[\frac{\|\pi_k(X)\|^2}{\|X\|^2} > \beta\,\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(1-\beta+\ln\beta)\right) \tag{3}$$

Question 0.3. En utilisant les éléments ci-dessus, montrez que, pour toute paire de points $p_i \neq p_j \in P$ fixée, on a

$$\Pr\left[(1-\varepsilon)\,\frac{k}{d} \leq \frac{\|f(p_i) - f(p_j)\|^2}{\|p_i - p_j\|^2} \leq (1+\varepsilon)\,\frac{k}{d}\right] \geq 1 - \frac{2}{n^2}.$$

Question 0.4. Déduisez de ce qui précède que, avec probabilité au moins $1/n$, la fonction f vérifie

$$(1-\varepsilon)\,\frac{k}{d} \leq \frac{\|f(p_i)-f(p_j)\|^2}{\|p_i-p_j\|^2} \leq (1+\varepsilon)\,\frac{k}{d}$$

pour toutes les paires de points $p_i \neq p_j \in P$.

Question 0.5. Complétez la preuve du théorème 0.1.

L'approche adoptée dans la preuve du théorème suggère une méthode pour réduire la dimension d'un nuage de points $P = \{p_1, \cdots, p_n\} \subset \mathbb{R}^d$ sans distordre trop les distances entre les points : étant donné une dimension cible $k < d$, on tire au hasard un sous-espace vectoriel de dimension k dans $\mathbb{R}^d$, puis on projette les points de P orthogonalement dessus, enfin on leur fait subir une homothétie de rapport $\sqrt{d/k}$. Pour le tirage du sous-espace, on tire aléatoirement et indépendamment k vecteurs de $\mathbb{R}^d$, puis on prend le sous-espace engendré par ces k vecteurs. Le tirage aléatoire d'un vecteur X se fait comme décrit plus haut, à savoir : on échantillonne iid ses coordonnées $X_1, \cdots, X_d$ selon la loi normale $\mathcal{N}(0,1)$ dans $\mathbb{R}$.

Question 0.6. Quelle est la probabilité que les k vecteurs de $\mathbb{R}^d$ ainsi tirés au hasard engendrent un sous-espace vectoriel de dimension k ? Justifiez votre réponse.

Question 0.7. En admettant que le sous-espace de dimension k ainsi obtenu est bien tiré selon la loi uniforme, donnez une borne inférieure sur la probabilité que la réduction de dimension telle que décrite vérifie (1). Que proposez-vous pour améliorer cette probabilité, par exemple pour la rendre supérieure ou égale à une constante $c \in]0,1[$ indépendante de n ? Justifiez votre réponse.

Réduction de dimension par projection aléatoire

Question 0.1. Ce résultat est utile à condition que l'on puisse choisir $k < d$. Pour cela, étant donnés $\varepsilon \in]0,1[$ et d, il faut et il suffit que

$$4\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)^{-1} \ln n < d,$$

soit (en utilisant le fait que la fonction ln est croissante)

$$n < \exp\left(\frac{d}{4}\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)\right).$$

Ainsi, le nombre de points de données ne doit pas être plus qu'exponentiel en la dimension ambiante. En fait, le résultat n'a vraiment d'utilité que pour un nombre de points sous-exponentiel, voire même polynômial — ce qui permet de réduire la dimension à $O(\ln d)$.

Question 0.2. Rappelons que $X = (X_1, \cdots, X_d)$, donc $\|X\|^2 = \sum_{i=1}^d X_i^2$ et $\|\pi_k(X)\|^2 = \sum_{i=1}^k X_i$. Par linéarité de l'espérance et le fait que les X_i sont iid, on obtient :

$$\mathbb{E}\left[\|X\|^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[X_i^2\right] = d\,\mathbb{E}\left[X_1^2\right]$$
$$\mathbb{E}\left[\|\pi_k(X)\|^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[X_i^2\right] = k\,\mathbb{E}\left[X_1^2\right],$$

ce qui donne le résultat.

Question 0.3. Fixons une paire de points $p_i \neq p_j \in P$. Prenons un sous-espace aléatoire de dimension k et appelons q_i, q_j les projetés orthogonaux respectifs de p_i, p_j sur ce sous-espace. En posant $\beta = (1-\varepsilon)$, on a alors, d'après l'équation (2) de l'énoncé et le fait (admis dans l'énoncé) que projeter le vecteur fixe $p_i - p_j$ sur un sous-espace aléatoire revient à projeter un vecteur aléatoire sur un sous-espace fixe :

$$\Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} < (1-\varepsilon)\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(\varepsilon + \ln(1-\varepsilon))\right)$$

En utilisant l'inégalité $\ln(1-\varepsilon) \leq -\varepsilon - \varepsilon^2/2$ pour $\varepsilon < 1$, on obtient :

$$\Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} < (1-\varepsilon)\frac{k}{d}\right] \leq \exp\left(-\frac{k\varepsilon^2}{4}\right) \leq \exp\left(-2\ln n\right) = 1/n^2$$

De manière similaire, en posant $\beta = (1+\varepsilon)$, on a, d'après l'équation (3) de l'énoncé :

$$\Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} > (1+\varepsilon)\frac{k}{d}\right] \leq \exp\left(\frac{k}{2}(-\varepsilon + \ln(1+\varepsilon))\right)$$

En utilisant l'inégalité $\ln(1+\varepsilon) \leq \varepsilon - \varepsilon^2/2 + \varepsilon^3/3$ pour $\varepsilon \geq 0$, on obtient :

$$\Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} > (1+\varepsilon)\frac{k}{d}\right] \leq \exp\left(-\frac{k(\varepsilon^2/2 - \varepsilon^3/3)}{2}\right) \leq \exp\left(-2\ln n\right) = 1/n^2$$

On applique maintenant l'inégalité de Boole, ce qui nous donne :

$$\Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} < (1-\varepsilon)\frac{k}{d} \;\vee\; \frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} > (1+\varepsilon)\frac{k}{d}\right] \leq \Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} < (1-\varepsilon)\frac{k}{d}\right] + \Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} > (1+\varepsilon)\frac{k}{d}\right] \leq \frac{2}{n^2}$$

Le résultat en découle.

Question 0.4. D'après la question précédente, on a

$$\Pr\left[\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} < (1-\varepsilon)\frac{k}{d} \ \vee \ \frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} > (1+\varepsilon)\frac{k}{d}\right] \leq \frac{2}{n^2}$$

pour toute paire de points $p_i \neq p_j \in P$ fixée. En appliquant l'inégalité de Boole sur l'ensemble des paires de points $p_i \neq p_j \in P$, on obtient :

$$\Pr\left[\bigvee_{p_i \neq p_j \in P}\left(\frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} < (1-\varepsilon)\frac{k}{d} \ \vee \ \frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} > (1+\varepsilon)\frac{k}{d}\right)\right] \leq \sum_{p_i \neq p_j \in P}\frac{2}{n^2} = \binom{n}{2}\frac{2}{n^2} = 1 - \frac{1}{n}$$

En passant à l'événement contraire, on obtient :

$$\Pr\left[\bigwedge_{p_i \neq p_j \in P}\left((1-\varepsilon)\frac{k}{d} \leq \frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} \leq (1+\varepsilon)\frac{k}{d}\right)\right] \geq \frac{1}{n}$$

D'où le résultat.

Question 0.5. D'après la question précédente, avec une probabilité non nulle la projection orthogonale sur un sous-espace vectoriel de dimension k aléatoire vérifie

$$\bigwedge_{p_i \neq p_j \in P}\left((1-\varepsilon)\frac{k}{d} \leq \frac{\|q_i - q_j\|^2}{\|p_i - p_j\|^2} \leq (1+\varepsilon)\frac{k}{d}\right)$$

Il existe donc au moins un sous-espace de dimension k dont la projection orthogonale vérifie cette propriété. En post-composant cette projection par une homothétie de rapport $\sqrt{d/k}$, on obtient une fonction $f\colon \mathbb{R}^d \to \mathbb{R}^k$ qui vérifie l'équation (1) de l'énoncé pour toute paire de points $p_i \neq p_j \in P$. Enfin, l'équation (1) est automatiquement vérifiée pour $p_i = p_j$, ce qui conclut la preuve du théorème.

Question 0.6. Comme $k < d$, la probabilité que les k vecteurs tirés aléatoirement engendrent un sous-espace de dimension k est 1. En effet, on peut compléter la famille de vecteurs en tirant $d - k$ vecteurs supplémentaires, toujours iid selon la même loi. On peut alors former la matrice $d \times d$ des coordonnees de ces d vecteurs, représentés chacun en ligne, avec les k vecteurs de l'énoncé sur les k premières lignes. Or, on sait que les matrices inversibles forment un ensemble générique (i.e. ouvert et dense) dans l'espace des matrices $d \times d$ à coefficients réels. Par conséquent, la matrice obtenue, dont les entrées sont iid selon la loi normale (qui est à densité par rapport à la mesure de Lebesgue dans $\mathbb{R}$), est inversible avec probabilité 1, et donc le sous-espace engendré par ses k premières lignes est de dimension k avec probabilité 1 également.

Question 0.7. D'après la question 0.5, la probabilité que la réduction de dimension ainsi obtenue vérifie (1) est bornée inférieurement par $1/n$. Cette borne diminuant lorsque n augmente, elle n'est pas très utile en pratique.

Une approche possible pour remédier à cela est de répéter l'expérience m fois indépendamment, ce qui nous donne m applications $f_1, \cdots, f_m\colon \mathbb{R}^d \to \mathbb{R}^k$ indépendantes, puis de garder celle qui minimise la distortion métrique :

$$f := \operatorname*{argmin}_{1 \leq l \leq m} \max_{p_i \neq p_j \in P} \max\left\{\frac{\|f_l(p_i) - f_l(p_j)\|^2}{(1+\varepsilon)\|p_i - p_j\|^2}, \frac{(1-\varepsilon)\|p_i - p_j\|^2}{\|f_l(p_i) - f_l(p_j)\|^2}\right\}$$

Comme les expériences sont indépendantes, d'après la question 0.4 la probabilité qu'aucune fonction f_l ne satisfasse (1) est au plus $(1-1/n)^m$. Pour rendre cette borne plus petite qu'une constante $1-c$, il suffit de prendre

$$m \geq \frac{\ln(1-c)}{\ln(1-1/n)} \overset{n\to\infty}{\sim} -\ln(1-c)\, n$$

Ainsi, un nombre linéaire (en n) de répétitions de l'expérience suffit pour atteindre la probabilité de succès constante c souhaitée.