

Feuille de TD 6 – Correction

Apprentissage non supervisé

Exercice 1 – Analyse en composantes principales

On note $X_i \in \mathbb{R}^p$ la i -ème ligne de X , identifiée à un vecteur colonne lorsque l'on écrit des produits scalaires. Les colonnes de X étant centrées, la matrice de covariance empirique est $n^{-1}X^\top X$. Le facteur n^{-1} ne change ni ses vecteurs propres ni les sous-espaces principaux.

On note

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

les valeurs propres de $X^\top X$, comptées avec multiplicité.

1. Reformulation du problème en dimension un.

Correction. Pour $\|\alpha\| = 1$, la projection orthogonale de X_i sur $\text{Vect}\{\alpha\}$ vaut

$$\text{Proj}_{\langle \alpha \rangle}(X_i) = \langle X_i, \alpha \rangle \alpha.$$

Le théorème de Pythagore donne

$$\|X_i - \text{Proj}_{\langle \alpha \rangle}(X_i)\|^2 = \|X_i\|^2 - \|\text{Proj}_{\langle \alpha \rangle}(X_i)\|^2 = \|X_i\|^2 - \langle X_i, \alpha \rangle^2.$$

En sommant,

$$\begin{aligned} \sum_{i=1}^n \|X_i - \text{Proj}_{\langle \alpha \rangle}(X_i)\|^2 &= \sum_{i=1}^n \|X_i\|^2 - \sum_{i=1}^n \langle X_i, \alpha \rangle^2 \\ &= \|X\|_F^2 - \|X\alpha\|^2 \\ &= \|X\|_F^2 - \alpha^\top X^\top X \alpha. \end{aligned}$$

Le premier terme ne dépend pas de α . Ainsi,

$$\arg \min_{\|\alpha\|=1} \sum_{i=1}^n \|X_i - \text{Proj}_{\langle \alpha \rangle}(X_i)\|^2 = \arg \max_{\|\alpha\|=1} \alpha^\top X^\top X \alpha.$$

2. Premier axe principal.

Correction. Le quotient de Rayleigh de la matrice symétrique positive $X^\top X$ vérifie

$$\max_{\|\alpha\|=1} \alpha^\top X^\top X \alpha = \lambda_1.$$

On choisit donc pour $\alpha^{(1)}$ un vecteur propre unitaire associé à λ_1 :

$$X^\top X \alpha^{(1)} = \lambda_1 \alpha^{(1)}.$$

Si λ_1 est simple, $\alpha^{(1)}$ est unique au signe près. Si elle est multiple, tout vecteur unitaire de son espace propre est solution : l'axe n'est alors pas unique.

3. Sous-espace principal de dimension d .

Correction. Soit V un sous-espace de dimension d et $(v_1, \dots, v_d)$ une base orthonormée de V . Par Pythagore,

$$\begin{aligned} \sum_{i=1}^n \|X_i - \text{Proj}_V(X_i)\|^2 &= \|X\|_F^2 - \sum_{j=1}^d \sum_{i=1}^n \langle X_i, v_j \rangle^2 \\ &= \|X\|_F^2 - \sum_{j=1}^d v_j^\top X^\top X v_j. \end{aligned}$$

Le principe variationnel de Courant–Fischer, ou le théorème de Ky Fan, donne

$$\max_{\substack{v_1, \dots, v_d \\ v_j^\top v_k = \mathbf{1}_{\{j=k\}}}} \sum_{j=1}^d v_j^\top X^\top X v_j = \sum_{j=1}^d \lambda_j.$$

On peut donc prendre pour $\alpha^{(1)}, \dots, \alpha^{(d)}$ des vecteurs propres orthonormés associés aux d plus grandes valeurs propres, et

$$V_d = \text{Vect}\{\alpha^{(1)}, \dots, \alpha^{(d)}\}.$$

L’hypothèse $d \leq \text{rang}(X)$ assure que $\lambda_1, \dots, \lambda_d > 0$. En présence de valeurs propres multiples à la coupure, le sous-espace optimal peut ne pas être unique; toute base orthonormée d’un sous-espace optimal convient.

4. Orthogonalité des composantes principales.

Correction. Pour $j, k \leq d$, avec $C_j = X\alpha^{(j)}$,

$$\begin{aligned} \langle C_j, C_k \rangle &= (\alpha^{(j)})^\top X^\top X \alpha^{(k)} \\ &= \lambda_k (\alpha^{(j)})^\top \alpha^{(k)} = \lambda_k \mathbf{1}_{\{j=k\}}. \end{aligned}$$

Ainsi les C_j sont deux à deux orthogonaux. En outre,

$$\|C_j\|^2 = \lambda_j.$$

Les vecteurs C_j ne sont donc pas nécessairement unitaires.

5. Reconstruction d’une observation projetée.

Correction. Puisque les $\alpha^{(j)}$ forment une base orthonormée de V_d ,

$$\text{Proj}_{V_d}(X_i) = \sum_{j=1}^d \langle X_i, \alpha^{(j)} \rangle \alpha^{(j)}.$$

Or, par définition de $C_j = X\alpha^{(j)}$,

$$C_j[i] = (X\alpha^{(j)})_i = \langle X_i, \alpha^{(j)} \rangle.$$

Par conséquent,

$$\text{Proj}_{V_d}(X_i) = \sum_{j=1}^d C_j[i] \alpha^{(j)}.$$

Les $C_j[i]$ sont les coordonnées, ou *scores*, de l’individu X_i sur les axes principaux.

6. Part de variance expliquée.

Correction. Pour tout i ,

$$\|\text{Proj}_{V_k}(X_i)\|^2 = \sum_{j=1}^k \langle X_i, \alpha^{(j)} \rangle^2.$$

En sommant sur i et en utilisant $\|C_j\|^2 = \lambda_j$,

$$\sum_{i=1}^n \|\text{Proj}_{V_k}(X_i)\|^2 = \sum_{j=1}^k \lambda_j.$$

De plus,

$$\sum_{i=1}^n \|X_i\|^2 = \|X\|_F^2 = \text{tr}(X^T X) = \sum_{j=1}^p \lambda_j.$$

Le rapport de variance expliquée par les k premières composantes est donc

$$\frac{\sum_{i=1}^n \|\text{Proj}_{V_k}(X_i)\|^2}{\sum_{i=1}^n \|X_i\|^2} = \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^p \lambda_j}.$$

Exercice 2 – Projection aléatoire et plus proche voisin

Dans tout l'exercice, la probabilité porte sur le tirage de la matrice gaussienne $A = m^{-1/2}G$. Les points $x, x_1, \dots, x_N$ sont considérés comme fixés.

1. Loi de l'image d'un vecteur fixé.

Correction. Notons G_r la r -ème ligne de G . Pour $r = 1, \dots, m$,

$$(Au)_r = \frac{1}{\sqrt{m}} \sum_{\ell=1}^d G_{r\ell} u_\ell.$$

C'est une combinaison linéaire de variables gaussiennes indépendantes, donc

$$(Au)_r \sim \mathcal{N}\left(0, \frac{1}{m} \sum_{\ell=1}^d u_\ell^2\right) = \mathcal{N}\left(0, \frac{\|u\|_2^2}{m}\right).$$

Ces coordonnées sont indépendantes car elles dépendent de lignes indépendantes de G . Si $u \neq 0$, les variables $Z_r = \sqrt{m}(Au)_r/\|u\|_2$ sont donc i.i.d. de loi $\mathcal{N}(0, 1)$, et

$$\frac{\|Au\|_2^2}{\|u\|_2^2} = \frac{1}{m} \sum_{r=1}^m Z_r^2 \stackrel{\mathcal{L}}{=} \frac{1}{m} \chi_m^2.$$

Pour $u = 0$, l'égalité de préservation des normes est triviale.

2. Préservation simultanée des distances à la requête.

Correction. Appliquons l'inégalité admise aux N vecteurs fixés $u_i = x - x_i$. Pour chacun d'eux, introduisons l'événement

$$\mathcal{E}_i = \left\{ \left| \|A(x - x_i)\|_2^2 - \|x - x_i\|_2^2 \right| \leq \varepsilon \|x - x_i\|_2^2 \right\}.$$

Comme $A(x - x_i) = Ax - Ax_i$, l'événement $\mathcal{E}_i$ est exactement

$$(1 - \varepsilon) \|x - x_i\|_2^2 \leq \|Ax - Ax_i\|_2^2 \leq (1 + \varepsilon) \|x - x_i\|_2^2.$$

Par l'inégalité de l'union,

$$\begin{aligned} \mathbb{P} \left(\bigcap_{i=1}^N \mathcal{E}_i \right) &= 1 - \mathbb{P} \left(\bigcup_{i=1}^N \mathcal{E}_i^c \right) \\ &\geq 1 - \sum_{i=1}^N \mathbb{P}(\mathcal{E}_i^c) \\ &\geq 1 - 2N \exp(-cm\varepsilon^2). \end{aligned}$$

On n'a pas besoin que les événements $\mathcal{E}_i$ soient indépendants.

3. Choix de la dimension projetée.

Correction. Pour que la probabilité d'échec soit au plus δ , il suffit que

$$2Ne^{-cm\varepsilon^2} \leq \delta.$$

En prenant les logarithmes, une condition exacte est

$$m \geq \frac{1}{c\varepsilon^2} \log \left(\frac{2N}{\delta} \right).$$

En modifiant la constante universelle, cela s'écrit

$$m \geq C\varepsilon^{-2} \log \left(\frac{N}{\delta} \right).$$

La dépendance dans la taille de la base est seulement logarithmique et la borne ne dépend pas de la dimension ambiante d . Ainsi, même si d est très grand, les N distances pertinentes pour une requête fixée peuvent être préservées dans une dimension bien plus faible.

4. Préservation de l'identité du plus proche voisin.

Correction. Plaçons-nous sur l'événement de préservation simultanée de la question 2. Pour le premier voisin,

$$\|Ax - Ax_1\|_2^2 \leq (1 + \varepsilon)d_1^2.$$

Pour tout $i \geq 2$, puisque $\|x - x_i\|_2 \geq d_2$,

$$\|Ax - Ax_i\|_2^2 \geq (1 - \varepsilon)\|x - x_i\|_2^2 \geq (1 - \varepsilon)d_2^2.$$

La condition

$$(1 + \varepsilon)d_1^2 < (1 - \varepsilon)d_2^2$$

implique donc

$$\|Ax - Ax_1\|_2 < \|Ax - Ax_i\|_2 \quad \text{pour tout } i \geq 2.$$

Ainsi Ax_1 reste l'unique plus proche voisin de Ax . Lorsque $d_1 > 0$, la condition précédente équivaut à

$$\frac{d_2}{d_1} > \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}.$$

Elle exprime une hypothèse de *marge multiplicative* entre le premier et le deuxième voisin. Si $d_1 = 0$, la condition directe est automatiquement satisfaite puisque l'unicité impose $d_2 > 0$.

5. Préservation de la prédiction.

Correction. Le classifieur avant projection prédit l'étiquette du plus proche voisin :

$$g_{1\text{NN}}(x) = y_1.$$

Sous la condition de marge et sur l'événement de préservation des distances, la question précédente montre que Ax_1 est encore le plus proche voisin. Le classifieur construit sur les données projetées prédit donc lui aussi y_1 :

$$g_{1\text{NN}}^A(Ax) = y_1 = g_{1\text{NN}}(x).$$

Cette conclusion vaut avec probabilité au moins $1 - 2N \exp(-cm\varepsilon^2)$ sur le tirage de A .

6. Interprétation.

Correction. Pour une requête et une base de données fixées satisfaisant la condition de marge, choisir

$$m = O(\varepsilon^{-2} \log(N/\delta))$$

suffit pour exécuter la recherche du plus proche voisin après projection, tout en conservant la prédiction avec probabilité au moins $1 - \delta$. Le coût géométrique peut ainsi dépendre de m plutôt que de la dimension ambiante d .

Ce résultat ne contredit pas la non-consistance générale du classifieur 1-NN. Les deux affirmations portent sur des questions différentes :

- ici, on compare, pour un échantillon fixé, la règle 1-NN avant et après projection; la projection conserve sous une marge le voisin choisi, et donc aussi son label;
- la consistance étudie, lorsque la taille de l'échantillon tend vers l'infini, si le risque du classifieur converge vers le risque de Bayes.

Préserver exactement la décision du 1-NN préserve aussi ses limites : la projection ne moyenne pas le bruit porté par l'unique label du voisin. Pour obtenir la consistance, on utilise typiquement un classifieur k_N -NN avec $k_N \rightarrow \infty$ et $k_N/N \rightarrow 0$, sous les hypothèses usuelles.