TD 02 : Analyse en composantes principales

Fondements du Machine Learning, L3 IM2D

Octobre 2024

Dauphine | PSL
UNIVERSITÉ PARIS

Exercice 2.1: Double ACP (Examen 2020/2021)

Dans cet exercice, on considère deux tableaux de données $\boldsymbol{X} \in \mathbb{R}^{m_1 \times n}$ et $\boldsymbol{W} \in \mathbb{R}^{m_2 \times n}$ représentant des exemples, ou individus, dans un espace à n dimensions. On note ainsi

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_{m_1}^{\mathrm{T}} \end{bmatrix} \quad \text{et} \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{w}_{m_2}^{\mathrm{T}} \end{bmatrix}.$$

a) On considère d'abord les matrices de données individuellement.

   i) Donner la formule de définition des individus moyens des tableaux de données $\boldsymbol{X}$ et $\boldsymbol{W}$, que l'on notera $\overline{\boldsymbol{x}}$ et $\overline{\boldsymbol{w}}$.

   ii) Donner les matrices de covariance empirique respectives des tableaux $\boldsymbol{X}$ et $\boldsymbol{W}$, notées $\boldsymbol{\Sigma}_X$ et $\boldsymbol{\Sigma}_W$.

   iii) Que représentent les vecteurs propres d'une matrice de covariance empirique dans le contexte de l'analyse en composantes principales ?

b) On considère maintenant le tableau de données $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{W} \end{bmatrix} \in \mathbb{R}^{(m_1+m_2)\times n}$.

   i) Exprimer l'individu moyen de ce tableau, noté $\overline{\boldsymbol{v}}$, en fonction des formules de $\overline{\boldsymbol{x}}$ et $\overline{\boldsymbol{w}}$ établies en question a-i).

   ii) Donner une expression de la matrice de covariance $\boldsymbol{\Sigma}_V$ de la matrice $\boldsymbol{V}$.

c) On suppose enfin qu'il existe $\alpha \in \mathbb{R}$ tel que $\boldsymbol{X} = \alpha \boldsymbol{W}$. Proposer une interprétation des valeurs suivantes :

   i) Valeur de $\overline{\boldsymbol{v}}$ lorsque $\alpha = -1$;

   ii) Valeur de $\boldsymbol{\Sigma}_V$ lorsque $\alpha = 1$.

Exercice 2.2: SVD et ACP (Examen 2020/2021)

Soit une matrice $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ de rang $n < m$. On supposera que $\boldsymbol{X}$ est la concaténation de m vecteurs de $\mathbb{R}^n$ notés $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, de sorte que $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} \end{bmatrix}$.

a) On considère une décomposition en valeurs singulières réduite $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$, où les valeurs singulières sont ordonnées par ordre décroissant. Donner alors les tailles des matrices $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ ainsi que leurs caractéristiques.

b) Rappeler la définition de l'individu moyen $\bar{\boldsymbol{x}}$ correspondant au nuage de points $\{\boldsymbol{x}_i\}_{i=1}^n$.

c) On suppose d'abord que $\bar{\boldsymbol{x}} \neq \mathbf{0}_{\mathbb{R}^n}$. Expliquer comment modifier $\boldsymbol{X}$ de sorte à obtenir un nouveau nuage de points (noté $\boldsymbol{X}^c$) d'individu moyen $\mathbf{0}_{\mathbb{R}^n}$.

d) On suppose maintenant que $\bar{\boldsymbol{x}} = \mathbf{0}_{\mathbb{R}^n}$.

   i) Quelle est l'implication de ce résultat pour le nuage de points ?

   ii) Donner la définition de la matrice de covariance empirique associée à $\boldsymbol{X}$, notée $\boldsymbol{\Sigma}_X$, et justifier que $\boldsymbol{\Sigma}_X = \frac{1}{m-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$.

   iii) En utilisant la SVD de $\boldsymbol{X}$ de la question a), écrire une SVD de $\boldsymbol{\Sigma}_X$.

   iv) Justifier que toute décomposition en valeurs propres de $\boldsymbol{\Sigma}_X$ en est également une SVD.

   v) Que représentent les coefficients diagonaux de $\boldsymbol{\Sigma}_X$?

Exercice 2.3: ACP et données augmentées (Examen 2019/2020)

Soit $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m \times n}$ représentant n caractéristiques de m individus $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$.

a) Rappeler la formule de calcul de l'individu moyen, noté $\bar{\boldsymbol{x}}$, ainsi que de la matrice de covariance empirique, notée $\boldsymbol{\Sigma}$, pour le tableau de données $\boldsymbol{X}$.

b) Rappeler le lien entre la première composante principale de $\boldsymbol{X}$ et la matrice $\boldsymbol{\Sigma}$. Comment cette composante peut-elle être utilisée pour approcher le tableau de données $\boldsymbol{X}$?

c) Dans le cas où $1 \leq m \ll n$, la matrice $\boldsymbol{\Sigma}$ peut être de taille trop grande pour pouvoir être stockée et manipulée numériquement. Quelle technique vue en TP peut-on alors utiliser pour obtenir (entre autres) la première composante principale ?

d) On considère maintenant le tableau de données augmenté avec l'individu moyen, c'est-à-dire la matrice $\bar{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} \\ \bar{\boldsymbol{x}}^T \end{bmatrix} \in \mathbb{R}^{(m+1) \times n}$. Calculer l'individu moyen ainsi que la matrice de covariance de ce nouveau tableau en fonction de $\bar{\boldsymbol{x}}$ et $\boldsymbol{\Sigma}$. Comment interprétez-vous ces résultats ?