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 ?

Solutions

Solution de l'exercice 2.1: Double ACP

a) Question de cours. Pour le iii), il s'agit des axes principaux.

b) On a

$$\bar{\boldsymbol{v}} = \frac{1}{m_1+m_2}\left(\sum_{i=1}^{m_1}\boldsymbol{x}_i + \sum_{i=1}^{m_2}\boldsymbol{w}_i\right) = \frac{m_1}{m_1+m_2}\bar{\boldsymbol{x}} + \frac{m_2}{m_1+m_2}\bar{\boldsymbol{w}}.$$

c)

$$\boldsymbol{\Sigma}_V = \frac{1}{m_1+m_2-1}\left(\sum_{i=1}^{m_1}(\boldsymbol{x}_i-\bar{\boldsymbol{v}})(\boldsymbol{x}_i-\bar{\boldsymbol{v}})^{\mathrm{T}} + \sum_{i=1}^{m_2}(\boldsymbol{w}_i-\bar{\boldsymbol{v}})(\boldsymbol{w}_i-\bar{\boldsymbol{v}})^{\mathrm{T}}\right)$$

d) Pour le cas i), on obtient que $\bar{\boldsymbol{v}} = \mathbf{0}$, car $m_1 = m_2$ et $\boldsymbol{x}_i = -\boldsymbol{w}_i$ pour tout i. Le jeu de données est centré en l'origine, ce qui est logique puisqu'il ne consiste que de paires de vecteurs opposés. Pour ii), on obtient que $\Sigma_V = \frac{m-1}{2m-1}\Sigma_X$, ce qui veut notamment dire que les variances seront plus faibles. Là encore, c'est logique parce qu'on a dupliqué le jeu de données en combinant $\boldsymbol{X}$ et $\boldsymbol{W}$.

Solution de l'exercice 2.2: SVD et ACP (Examen 2020/2021)

a) $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ orthogonale, $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ est nulle sauf potentiellement sur la diagonale, où les coefficients sont $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$.

b) $\bar{\boldsymbol{x}} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$.

c) En centrant les données, on définit un nouveau jeu de données $\boldsymbol{X}^c = \boldsymbol{X} - \mathbf{1}_m\bar{\boldsymbol{x}}^{\mathrm{T}}$.

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

   i) Le jeu de données est centré en $\mathbf{0}_{\mathbb{R}^n}$.

   ii) Par définition, la matrice de covariance empirique est égale à $\frac{1}{m-1}\boldsymbol{X}^c(\boldsymbol{X}^c)^{\mathrm{T}}$, où $\boldsymbol{X}^c$ est le jeu de données centrées. Comme on a ici $\boldsymbol{X}^c = \boldsymbol{X}$, on en conclut que $\frac{1}{m-1}\boldsymbol{X}^c(\boldsymbol{X}^c)^{\mathrm{T}} = \frac{1}{m-1}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$.

   iii) Immédiat, la seule différence avec le cas général étant que les deux dimensions sont égales.

   iv) $\boldsymbol{\Sigma}_X$ est symétrique semi-définie positive, donc il existe une base orthonormée $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ telle que $\boldsymbol{\Sigma}_X = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{\mathrm{T}}$, où $\boldsymbol{\Lambda}$ est une matrice diagonale à coefficients positifs ou nuls (ordonnés par ordre décroissant sans perte de généralité). Cette décomposition définit aussi une SVD de $\boldsymbol{\Sigma}_X$.

   v) Les coefficients diagonaux de $\boldsymbol{\Sigma}_X$ représentent les variances empiriques des attributs.

Solution de l'exercice 2.3: ACP et données augmentées (Examen 2019/2020)

a) Cours

b) Cours

c) Au lieu de considérer $\boldsymbol{\Sigma} = \frac{1}{m-1}\boldsymbol{X}^c(\boldsymbol{X}^c)^{\mathrm{T}}$, on peut considérer la matrice $\frac{1}{m-1}(\boldsymbol{X}^c)^{\mathrm{T}}\boldsymbol{X}^c$, qui est une matrice de taille $m \times m$, donc avec beaucoup moins de coefficients. Pour autant, les deux matrices ont les mêmes valeurs propres.

d) On trouve que rajouter la moyenne réduit les valeurs propres de la matrice de covariance, et biaise les données vers $\bar{\boldsymbol{x}}$. L'individu moyen ne change pas quand on ajoute l'individu moyen. En effet, l'individu moyen associé à $\bar{\boldsymbol{X}}$ est

$$\frac{1}{m+1}\left(\sum_{i=1}^{m}\boldsymbol{x}_i + \bar{\boldsymbol{x}}\right) = \frac{m}{m+1}\bar{\boldsymbol{x}} + \frac{1}{m+1}\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}.$$

Par ailleurs, la matrice de covariance associée à $\bar{\boldsymbol{X}}$ est

$$\begin{aligned}
\boldsymbol{\Sigma}_{\bar{\boldsymbol{X}}} &= \frac{1}{(m+1)-1}\left(\sum_{i=1}^{m}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} + (\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}})^{\mathrm{T}}\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} \\
&= \frac{m-1}{m}\boldsymbol{\Sigma}_{\boldsymbol{X}}.
\end{aligned}$$

Si l'on considère les valeurs propres de $\boldsymbol{\Sigma}_{\bar{\boldsymbol{X}}}$, on voit que celles-ci sont plus petites (et ≥ 0) que celles de $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, et donc l'inertie du nuage de points est réduite.