FONDEMENTS DU MACHINE LEARNING

24/09/2024

Programme :

Cours : ACP

Pas de TD !

La semaine prochaine : Cours $13^{h}45 - 15^{h}15$

TD $15^{h}30 - 18^{h}45$

ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

En anglais : Principal Component Analysis

Cadre de travail : Matrice $X \in \mathbb{R}^{m \times n}$

On suppose que X représente des vecteurs de données correspondant à m individus et n attributs

$\Rightarrow$ X s'écrit $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$

$x_i \in \mathbb{R}^n$: représente l'individu i $\qquad$ $v_j \in \mathbb{R}^m$: représente l'attribut j

On peut faire la SVD de X mais

- Pas d'interprétation des valeurs singulières / des vecteurs singuliers
- Pas de liens directs entre la SVD et une distribution de données sous-jacente (ou de manière équivalente, entre la SVD et la géométrie des points $\{x_i\}_{i=1..m}$ et des $\{v_j\}_{j=1..n}$)

$\hookrightarrow$ Approche de l'ACP :
- géométrique (considère les lignes ou les colonnes de la matrice comme des nuages de points)
- statistique (construit des statistiques sur les données)

1 Statistique empirique

↳ Peut être vu comme un processus de pré-traitement de données

Définition : L'individu moyen associé à $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}$

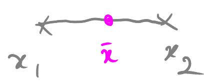
est le vecteur

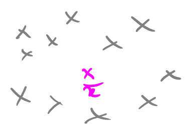
$$\bar{x} := \frac{1}{m} \sum_{i=1}^{m} x_i = \frac{1}{m} X^T 1_m \in \mathbb{R}^n$$

avec $1_m = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^m$

$X^T = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix} \in \mathbb{R}^{n \times m}$ $\quad$ $X^T 1_m$: somme des colonnes de X^T

↳ L'individu moyen représente la tendance centrale des données et le barycentre du nuage de points $\{x_i\}_{i=1..m}$ dans $\mathbb{R}^n$





Déf : La matrice de données centrées relative à X est la matrice

$$X^c := \underbrace{X}_{m \times n} - \underbrace{1_m}_{m \times 1} \underbrace{\bar{x}^T}_{1 \times n} \in \mathbb{R}^{m \times n}$$

Cette matrice s'écrit $X^c = \begin{bmatrix} (x_1^c)^T \\ \vdots \\ (x_m^c)^T \end{bmatrix}$ avec $x_i^c := x_i - \bar{x}$ $\forall i = 1..m$

Remarque : Centrer les données équivaut à :

- mettre l'individu moyen en $0_{\mathbb{R}^n}$;
- enlever la tendance centrale des données, càd une information partagée par les différents individus (et donc potentiellement peu révélatrice des particularités de chaque individu)

Définition : (Matrice de covariance empirique)

Soit $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}$ et $\bar{x}$ son individu moyen.

La matrice de covariance empirique associée à X est défini comme la matrice $\Sigma \in \mathbb{R}^{n \times n}$ avec

$$\Sigma = \frac{1}{m-1} \sum_{i=1}^{m} \underbrace{\overbrace{(x_i - \bar{x})}^{n \times 1} \overbrace{(x_i - \bar{x})^T}^{1 \times n}}_{n \times n}$$

$$= \frac{1}{m-1} \underbrace{(X^c)^T}_{n \times m} \underbrace{X^c}_{m \times n}$$

NB : Certains auteurs définissent

$$\Sigma = \frac{1}{m} (X^c)^T X^c$$

(notamment pour le cas $m=1$, qui a peu d'intérêt ici)

↳ Les éléments diagonaux de la matrice Σ s'appellent les variances empiriques. Pour tout $j = 1..n$, on note :

= notation

$$\sigma_j^2 := [\Sigma]_{jj} = \frac{1}{m-1} \sum_{i=1}^{m} \left([x_i]_j - [\bar{x}]_j \right)^2$$

Variance empirique associée à l'attribut j

$$= \frac{1}{m-1} \sum_{i=1}^{m} \left([v_j]_i - \frac{1}{m} \sum_{k=1}^{m} [v_j]_k \right)^2$$

$$(x_i - \bar{x})(x_i - \bar{x})^T$$

$$\begin{bmatrix} [x_i]_1 - [\bar{x}]_1 \\ \vdots \\ [x_i]_n - [\bar{x}]_n \end{bmatrix} \begin{bmatrix} [x_i]_1 - [\bar{x}]_1 & \cdots & [x_i]_n - [\bar{x}]_n \end{bmatrix} = \begin{bmatrix} ([x_i]_1 - [\bar{x}]_1)^2 & & \\ & \ddots & \\ & & ([x_i]_n - [\bar{x}]_n)^2 \end{bmatrix}$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} = \begin{array}{c} \begin{array}{cccc} v_1 & & & \end{array} \\ \begin{bmatrix} [x_1]_1 & [x_1]_2 & \cdots & [x_1]_n \\ [x_2]_1 & [x_2]_2 & \cdots & \\ [x_m]_1 & \cdots & \cdots & [x_m]_n \end{bmatrix} \end{array}$$

↳ Les coefficients hors diagonale de Σ s'appellent les covariances, parfois notées

$$\sigma_{jk} = \sigma_{kj} = \frac{1}{m-1} \sum_{i=1}^{m} \left([x_i]_j - [\bar{x}]_j \right)\left([x_i]_k - [\bar{x}]_k \right)$$

(mesure la covariance entre l'attribut j et l'attribut k)

Déf : L'inertie d'une matrice $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}$ de covariance empirique $\Sigma \in \mathbb{R}^{n \times n}$ est définie par $I(X) := \text{trace}(\Sigma)$

On a $I(X) = \sum_{j=1}^{n} \sigma_j^2 = \frac{1}{m-1} \sum_{i=1}^{m} \|x_i - \bar{x}\|^2$

↳ variance de la distance des individus à l'individu moyen (mesure de dispersion du nuage de points)

Données centrées réduites

Soit $X \in \mathbb{R}^{m \times n}$, $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix}$, $\bar{x}$ son individu moyen et $\Sigma \in \mathbb{R}^{n \times n}$ sa matrice de covariance empirique. La matrice de données centrées réduites associée à X, et notée $X_{0,1}$, est définie par

$$\underbrace{X_{0,1}}_{m \times n} := \underbrace{X^c}_{m \times n} \underbrace{D_{1/\sigma}}_{n \times n} \quad \text{avec} \quad D_{1/\sigma} = \begin{bmatrix} 1/\sigma_1 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_n \end{bmatrix}$$

(avec la convention que $\frac{1}{\sigma_i} = 1$ si $\sigma_i = 0$)

$\forall i = 1..m, \forall j = 1..n$

$$[X_{0,1}]_{ij} = \frac{[x_i]_j - [\bar{x}]_j}{\sigma_j}$$

⇒ Centrer et réduire les données : normaliser chaque attribut par rapport à la moyenne et à la variance empirique

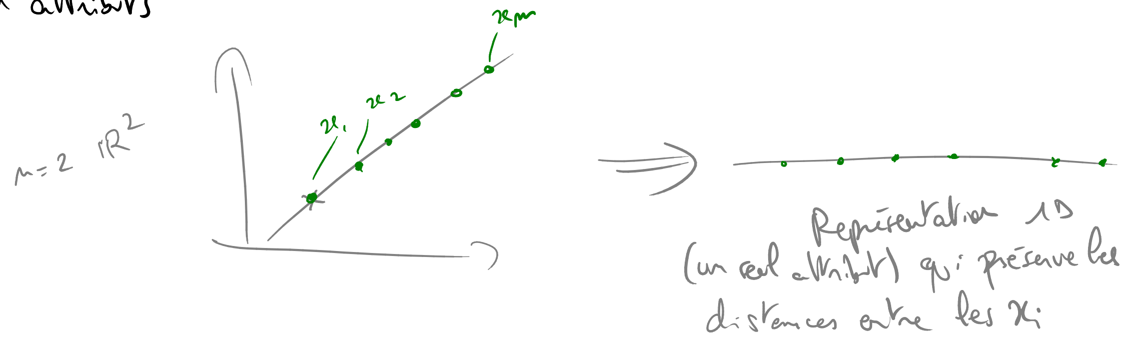
Déf : Matrice de corrélation empirique associée à X

$$R := \frac{1}{m-1} X_{0,1}^T X_{0,1} \in \mathbb{R}^{n \times n}$$

↳ Intérêt de la matrice de corrélation empirique

- Si R_{ij} $(= R_{ji}) = 1$, alors les attributs i et j sont liés par une relation affine, càd $v_j = a v_i + b$ pour $a \in \mathbb{R}$, $b \in \mathbb{R}$

Dans ce cas, on peut représenter ces 2 attributs avec un seul vecteur d'attributs



- Si $R_{ij} = 0$, alors les deux attributs sont indépendants, et on ne peut pas les réduire en un seul dans la représentation des données