FONDEMENTS DU MACHINE LEARNING

1er octobre 2024

Aujourd'hui :

Cours (13h45 - 15h15) : ACP

TD (15h30 - 18h45) : Fin exercices SVD

TD/TP ACP

Analyse en composantes principales (ACP) pour la réduction de dimension

Problème : $X \in \mathbb{R}^{m \times n}$ matrice de données

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix}, \quad x_i \in \mathbb{R}^n$$ vecteur représentatif de l'individu i, à n attributs

But : Réduire X à une matrice de taille $m \times k$ avec $k < n$ en conservant le maximum d'information pour chacun des individus $x_1, \dots, x_m$

$\Rightarrow$ Analyse en k composantes principales

(1) Analyse en 1 composante principale

↳ Objectif : Remplacer les n attributs par une nouvelle variable que l'on appelle composante principale.

$\Rightarrow$ Géométriquement, cela revient à représenter le nuage de points $\{x_1, \dots, x_m\} \subseteq \mathbb{R}^n$ par $\{y_1, \dots, y_m\} \subseteq \mathbb{R}^n$

- tels que les $\{y_i\}$ appartiennent à une même droite

Deux points de vue : Géométrique : On reste dans $\mathbb{R}^n$ et on passe d'un nuage de points quelconque à un nuage de points sur une droite

Algébrique $X \in \mathbb{R}^{m \times n} \longrightarrow c \in \mathbb{R}^{m \times 1}$

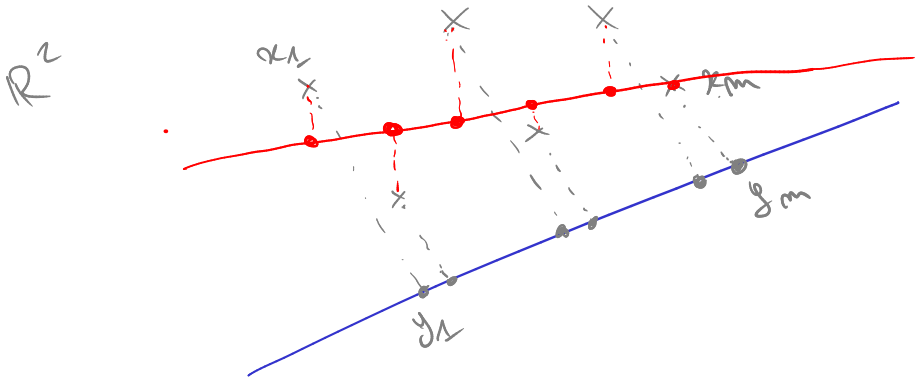
$\Rightarrow$ La meilleure façon de réduire les n attributs en 1 consiste à déterminer la meilleure projection des $\{x_i\}$ sur une droite dans $\mathbb{R}^n$

$\Rightarrow$ La "meilleure droite possible" (au sens de fournir la représentation des données la plus proche des points de départ) est celle qui maximise l'inertie du nuage de points projetés

$X = \{x_1, \dots, x_m\}$: nuage de points de départ

Inconnue $Y = \{y_1, \dots, y_m\}$: nuage des points projetés, contenus dans une droite / colinéaires.

(Inertie de Y : $\frac{1}{m-1} \sum_{i=1}^{m} \|y_i - \bar{y}\|^2$ $\quad \bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i \in \mathbb{R}^n$)



$\hookrightarrow$ Comment calcule-t-on les y_i / la droite en pratique ?

- Une "bonne" droite devrait passer par l'individu moyen $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$ afin de conserver l'information moyenne du nuage de points $\Rightarrow$ Les coordonnées des projetés du nuage de points ne prendront pas en compte la tendance centrale des données (mais les y_i s !)

- Les $\{y_i\}$ appartiennent à une droite d'équation

$$\{y = \bar{x} + \gamma u \mid \gamma \in \mathbb{R}\}$$

où u est le vecteur directeur de la droite à déterminer

On a alors : $y_i = \bar{x} + c_i u$ avec $c_i \in \mathbb{R}$

↑ Composante principale de x_i

$X \in \mathbb{R}^{m \times n} \longrightarrow c \in \mathbb{R}^{m \times 1}$, $c = \begin{bmatrix} c_1 \\ | \\ c_m \end{bmatrix}$ vecteur des premières composantes principales

Théorème : Pour tout $x \in \mathbb{R}^n$, la projection de x sur la droite $\{y = \bar{x} + \gamma u\}$ est donnée par

$$\bar{x} + u^T (x - \bar{x})\, u$$

Corollaire : Les projections $\{y_1, \dots, y_m\}$ du nuage de points $\{x_1, \dots, x_m\}$ sont données par

$$y_i = \bar{x} + u^T (x_i - \bar{x})\, u \qquad \forall i = 1..m$$

↳ On cherche le vecteur u tel que l'inertie du nuage de points projetés $Y = \{y_1, \dots, y_m\}$ soit maximale

$$I(Y) = \frac{1}{m-1} \sum_{i=1}^{m} \|y_i - \bar{y}\|^2$$

$$= \frac{1}{m-1} \sum_{i=1}^{m} \|\bar{x} + u^T (x_i - \bar{x})\, u - \bar{y}\|^2$$

On $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\bar{x} + u^T(x_i - \bar{x})\,u\right)$$

$$= \left(\frac{1}{m}\sum_{i=1}^{m}\bar{x}\right) + \left(\frac{1}{m}\sum_{i=1}^{m} u^T(x_i - \bar{x})\,u\right)$$

$$= \bar{x} + \frac{1}{m}\, u^T\left(\sum_{i=1}^{m}(x_i - \bar{x})\right)u$$

$$\sum_{i=1}^{m} u^T(x_i - \bar{x})\,u = \left[\sum_{i=1}^{m} u^T(x_i - \bar{x})\right]u = \left[u^T\left(\sum_{i=1}^{m} x_i - \bar{x}\right)\right]u$$

Donc

$$I(y) = \frac{1}{m-1}\sum_{i=1}^{m}\left\| \bar{x} + u^T(x_i - \bar{x})\,u - \bar{y}\right\|^2$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}\left\| \bar{x} + u^T(x_i - \bar{x})\,u - \bar{x} - \frac{1}{m}u^T\sum_{i=1}^{m}(x_i - \bar{x})\,u\right\|^2$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}\left\| u^T(x_i - \bar{x})\,u - \frac{1}{m}u^T\sum_{i=1}^{m}(x_i - \bar{x})\,u\right\|^2$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}\left\| u^T\left[(x_i - \bar{x}) - \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})\right]u\right\|^2$$

$$x_i - \bar{x} - \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x}) = x_i - \bar{x} - \frac{1}{m}\sum_{i=1}^{m} x_i + \bar{x} = x_i - \frac{1}{m}\sum_{i=1}^{m} x_i = x_i - \bar{x}$$

$$I(y) = \frac{1}{m-1}\sum_{i=1}^{m}\left\| u^T(x_i - \bar{x})\,u\right\|^2 \qquad \|v\|^2 = v^T v$$

$$= \frac{1}{m-1}\sum_{i=1}^{m} u^T(x_i - \bar{x})^T u\, u^T(x_i - \bar{x})\,u$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}\left[(x_i - \bar{x})^T u\right]^2 u^T u$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}\left[u^T(x_i - \bar{x})\right]^2 \|u\|^2$$

$$= \frac{1}{m-1} \sum_{i=1}^{m} u^T (x_i - \bar{x})(x_i - \bar{x})^T u \times \|u\|^2$$

$$(a^T b)^2 = (a^T b)(a^T b) = (a^T b)(b^T a) = a^T (b b^T) a$$

$$I(y) = \frac{1}{m-1} \|u\|^2 \sum_{i=1}^{m} u^T (x_i - \bar{x})(x_i - \bar{x})^T u$$

$$= \|u\|^2 u^T \left(\frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(x_i - \bar{x})^T \right) u$$

$$= \|u\|^2 u^T \Sigma u$$

Σ : matrice de covariance de $\{x_1, \dots, x_m\}$

↳ Pour maximiser cette inertie par rapport à u, on a 2 cas :

- Soit $x_i = \bar{x}$ $\forall i$ et dans ce cas $u = 0$
  ↳ $\Sigma = [0]$
- Sinon, on prend u de norme 1 et égal à un vecteur propre de Σ associé à la plus grande valeur propre de Σ

Formule à retenir

ACP avec 1 composante principale

$$\{x_1, \dots, x_m\} \subseteq \mathbb{R}^n \longrightarrow \{y_1, \dots, y_m\} \subseteq \mathbb{R}^n$$

avec $y_i = \bar{x} + u^T (x_i - \bar{x}) u$, $\forall i = 1..m$

u vecteur propre de norme 1 associé à la plus grande valeur propre de $\Sigma = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(x_i - \bar{x})^T$

c_i : (1re) composante principale de x_i

u : (1er) vecteur principal

$\{\bar{x} + \gamma u\}$: (1er) axe principal

Toutes ces notions sont déterminées par le nuage de points X

En pratique, on calcule les points projetés via

$$\underbrace{Y}_{m\times m} = \begin{bmatrix} y_1^T \\ \vdots \\ y_m^T \end{bmatrix} = \underbrace{\left(X - 1_m \bar{x}^T\right) u u^T}_{m\times m} + 1_m \bar{x}^T$$

(X: $m\times n$; $1_m \bar{x}^T$: $m\times 1$ · $1\times n$; $u u^T$: $n\times 1$ · $1\times n$ = $n\times n$; $\left(X - 1_m \bar{x}^T\right) u u^T$: $m\times n$)

(2) Analyse en k composantes principales, $k \geq 1$

But : $X \in \mathbb{R}^{m\times n} \longrightarrow C = \begin{bmatrix} c_1^T \\ \vdots \\ c_m^T \end{bmatrix} \in \mathbb{R}^{m\times k}$

$$\mathcal{X} = \{x_1, \dots, x_m\} \subset \mathbb{R}^n \longrightarrow \mathcal{Y} = \{y_1, \dots, y_m\} \subset \mathbb{R}^n$$

avec $\{y_i\} \subset \{\bar{x} + U_k \gamma \mid \gamma \in \mathbb{R}^k\}$

sous-espace affine de dimension k

et $U_k = [u_1 \dots u_k] \in \mathbb{R}^{n\times k}$ formé par k vecteurs linéairement indépendants et de norme 1

$\Rightarrow$ On écrira : $y_i = \bar{x} + U_k c_i$

$c_i \in \mathbb{R}^k$ est le vecteur des composantes principales

↳ Le meilleur choix possible pour les U_k consiste à prendre $u_1, \dots, u_k$ orthogonaux ($u_i^T u_j = 0$ $i \neq j$) et à les prendre comme k vecteurs propres de Σ associés aux k plus grandes valeurs propres

NB: On exclut le cas $\Sigma = [0]$, pour lequel on peut prendre $U_k = [0]$

↳ On a alors $y_i = \bar{x} + U_k c_i$, avec $c_i = \begin{bmatrix} u_1^T(x_i - \bar{x}) \\ \vdots \\ u_k^T(x_i - \bar{x}) \end{bmatrix}$

c_i : vecteur des k premières composantes principales de x_i

U_k : k premiers vecteurs principaux

$\{\bar{x} + U_k \gamma \mid \gamma \in \mathbb{R}^k\}$: sous-espace associé aux k premiers axes principaux

$\{\bar{x} + \gamma u_1 \mid \gamma \in \mathbb{R}\}, \{\bar{x} + \gamma u_2 \mid \gamma \in \mathbb{R}\} \dots, \{\bar{x} + \gamma u_k \mid \gamma \in \mathbb{R}\}$

Propriété importante : Calculer les composantes principales peut se faire en calculant les composantes principales une à une (notamment parce Σ possède une base orthonormée de vecteurs propres)

(3) Le cas k = m

$X \in \mathbb{R}^{m \times m} \longrightarrow C \in \mathbb{R}^{m \times m}$

$x_i \in \mathbb{R}^m \longrightarrow y_i = \bar{x} + U_m c_i = x_i$ $\quad c_i \in \mathbb{R}^m$

$[u_1, \dots, u_m]$ base orthonormée de vecteurs propres de Σ

c_i : coordonnées de $x_i - \bar{x}$ dans la base $(u_1, \dots, u_m)$

Avec $k = m$, on fait un changement de base (+ une translation par rapport à $\bar{x}$) relativement à notre jeu de données $x_1, \dots, x_m$

$\Rightarrow$ Dans cette base, les coordonnées des $x_i - \bar{x}$ sont décorrélées

Représentation naturelle de X	Représentation avec l'ACP C
Base canonique $\{e_1, \dots, e_m\}$	Base de vecteurs propres de Σ
$X^c = X - 1_m \bar{x}^T$ (Données centrées)	$C^c = \underbrace{X^c}_{m \times m} \underbrace{U_m}_{m \times m}$
$\Sigma = \frac{(X^c)^T X^c}{m-1}$	$\Sigma' = \frac{(C^c)^T (C^c)}{m-1} = U_m^T \Sigma U_m$ $= \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix}$ où $\lambda_1 \geq \dots \geq \lambda_m$ sont les valeurs propres de Σ $\Rightarrow$ Représentation dans laquelle les m attributs sont décorrélés