FONDEMENTS DU MACHINE LEARNING

8 octobre 2024

Cette semaine : CM Intro modèle linéaire 1/2
TD ACP

La suivante CM Intro modèle linéaire 2/2
TD Intro modèle linéaire

PREMIERS PAS AVEC LE MODÈLE LINÉAIRE

1) Introduction

Cours 1 à 5 : Extraire de l'information de $X \in \mathbb{R}^{m \times n}$ (SVD, ACP)

A partir de maintenant : Etablir une relation linéaire entre des entrées et des sorties, représentées par une matrice d'entrées

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}$$ et un vecteur de sorties $y \in \mathbb{R}^m$

X : "attributs"
y : "labels"

Cadre de travail : On cherche une fonction linéaire $h : \begin{array}{l} \mathbb{R}^n \to \mathbb{R} \\ x \mapsto x^T\beta \end{array}$

avec $\beta \in \mathbb{R}^n$ telle que $h(x_i) = x_i^T\beta$ soit égal à y_i $\forall i = 1..m$

paramètres du modèle linéaire

$\Rightarrow$ Réécriture sous forme matricielle

$x_i^T\beta = y_i \quad \forall i = 1..m \iff \boxed{X\beta = y}$ Système linéaire d'équations

⚠ Rien ne garantit qu'il existe un vecteur β vérifiant ce système d'équations, ou qu'un tel vecteur (s'il existe) est unique

Notre plan de bataille : Comprendre les systèmes linéaires tel que celui ci-dessus pour pouvoir proposer les meilleurs modèles linéaires possibles.

NB :

$$X\beta = y \iff \begin{cases} x_1^T\beta = y_1 \\ \vdots \\ x_m^T\beta = y_m \end{cases} \iff \begin{cases} x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1n}\beta_n = y_1 \\ \vdots \\ x_{m1}\beta_1 + x_{m2}\beta_2 + \dots + x_{mn}\beta_n = y_m \end{cases}$$

2) Systèmes linéaires : principes de base

a) Cas d'un système carré

$$X\beta = y \qquad (X \in \mathbb{R}^{n\times n},\ y \in \mathbb{R}^n) \text{ données}$$

$\Rightarrow$ Objectif : Trouver β s'il existe.

3 cas possibles (illustration avec $n=2$)

1) $\begin{cases} \beta_1 + \beta_2 = 0 \\ 3\beta_1 + 2\beta_2 = 1 \end{cases}$ $\quad X = \begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix}$ $y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\Rightarrow$ 1 unique solution $\beta_1 = 1 \quad \beta_2 = -1$

2) $\begin{cases} \beta_1 = 0 \\ \beta_1 = -1 \end{cases}$ $\quad X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ $y = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ $\Rightarrow$ Pas de solution

3) $\begin{cases} \beta_1 + 2\beta_2 = 2 \\ 2\beta_1 + 4\beta_2 = 4 \end{cases}$ $\quad X = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ $y = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ $\Rightarrow$ Une infinité de solutions

$$\left\{ \begin{bmatrix} \alpha \\ 1 - \alpha/2 \end{bmatrix} \;\middle|\; \alpha \in \mathbb{R} \right\}$$

Théorème : Soit $X \in \mathbb{R}^{n\times n}$ une matrice inversible.
Soit $y \in \mathbb{R}^n$.
Alors le système $X\beta = y$ possède une unique solution donnée par $\beta^* = X^{-1}y$

$\hookrightarrow$ Lorsque X n'est pas inversible, alors

- soit $y \in \mathrm{Im}(X) = \{ z \in \mathbb{R}^n \mid \exists \gamma \in \mathbb{R}^n,\ X\gamma = z \}$ et le système a alors une infinité de solutions

- soit $y \notin Im(X)$ et le système n'a pas de solution

b) Systèmes rectangulaires

Si $X \in \mathbb{R}^{m \times n}$ est une matrice rectangulaire, on peut avoir les mêmes cas que précédemment (1 unique solution, pas de solution, une infinité)

$\Rightarrow$ Mais si $m \neq n$, on ne peut pas parler d'inverse de X

Définition: $X \in \mathbb{R}^{m \times n}$ est dite de rang plein (full rank) si $\text{rang}(X) = \min(m,n)$, càd que X n'a pas de valeurs singulières nulles.

Théorème Soient $X \in \mathbb{R}^{m \times n}$ de rang plein et $y \in \mathbb{R}^m$.

a) Si $\text{rang}(X) = m$, alors $\underbrace{X}_{m \times n}\underbrace{X^T}_{n \times m} \in \mathbb{R}^{m \times m}$ est inversible et

$\bar{\beta} = X^T(XX^T)^{-1}y$ est une solution de $X\beta = y$.

b) Si $\text{rang}(X) = n$, alors $X^TX \in \mathbb{R}^{n \times n}$ est inversible et

$\hat{\beta} = (X^TX)^{-1}X^Ty$ est l'unique solution de $X\beta = y$ si $y \in Im(X)$

Démo : a) $\text{rang}(X) = \text{rang}(XX^T) = m$ donc XX^T est bien inversible

Posons $\bar{\beta} = X^T(XX^T)^{-1}y$. Alors,

$$X\bar{\beta} = X\left(X^T(XX^T)^{-1}y\right) = \underbrace{XX^T(XX^T)^{-1}}_{I_m}y = I_m y = y$$

$$\bar{\beta} = \underbrace{\underbrace{X^T}_{n \times m}\underbrace{(XX^T)^{-1}}_{m \times m}\underbrace{y}_{m \times 1}}_{n \times 1}$$

b) $\text{rang}(X) = \text{rang}(X^TX) = n$ donc X^TX est inversible

Soit une solution de $X\beta = y$, notée $\hat\beta$.

$$X\hat\beta = y \Rightarrow X^TX\hat\beta = X^Ty$$ (multiplication à gauche par X^T)

X^TX inversible $\rightarrow$ $\Longleftrightarrow$ $$\hat\beta = (X^TX)^{-1}X^Ty$$

Donc toute solution de $X\beta = y$ est nécessairement égale à $\hat\beta$.

Réciproquement, $\hat\beta$ n'est solution de $X\beta = y$ que lorsque ce système possède une solution, c'est-à-dire quand $y \in \mathrm{Im}(X)$

$\hookrightarrow$ Les résultats ci-dessus ne couvrent qu'un sous-ensemble des cas possibles. On cherche une approche plus systématique, que l'on va construire via deux ingrédients : la pseudo-inverse et les moindres carrés linéaires.

③ Pseudo-inverse (aka le retour de la SVD)

Pour une matrice carrée $X \in \mathbb{R}^{n\times n}$, l'inverse est définie comme $X^{-1} \in \mathbb{R}^{n\times n}$ telle que $X^{-1}X = XX^{-1} = I_n$, et n'existe pas toujours.

__Définition__ Pour toute matrice $X \in \mathbb{R}^{n\times m}$, il existe une unique matrice $M \in \mathbb{R}^{m\times n}$ telle que

Équations de Moore-Penrose $\longrightarrow$
$$\begin{cases} XMX = X \\ MXM = M \\ (XM)^T = XM \\ (MX)^T = MX \end{cases}$$

On l'appelle la pseudo-inverse de X, et on la note X^+.

Remarque : La pseudo-inverse est aussi appelée inverse généralisée ou inverse de Moore-Penrose

Q) Comment calculer la pseudo-inverse ?

Proposition : Soit une matrice $\Sigma \in \mathbb{R}^{m \times n}$ de la forme

$$\Sigma = \begin{array}{c} \\ \begin{bmatrix} \begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{matrix} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} \updownarrow r \\ \updownarrow m-r \end{matrix} \\ \overleftrightarrow{\;r\;} \quad \overleftrightarrow{\;n-r\;} \end{array}$$

avec $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$

et $r \leq \min(m,n)$

Alors la pseudo-inverse de Σ est donnée par

$$\Sigma^+ = \begin{array}{c} \overleftrightarrow{\;r\;} \quad \overleftrightarrow{\;m-r\;} \\ \begin{matrix} r \updownarrow \\ m-r \updownarrow \end{matrix} \begin{bmatrix} \begin{matrix} 1/\sigma_1 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_r \end{matrix} & 0 \\ 0 & 0 \end{bmatrix} \end{array} \in \mathbb{R}^{n \times m}$$

Démo : Vérifier les équations de Moore-Penrose avec $X = \Sigma$ et $M = \Sigma^+$

Par exemple,

$$XM = \underbrace{\overbrace{\Sigma \Sigma^+}^{m \times n}}_{n \times m} = \begin{bmatrix} \begin{matrix} \sigma_1 & & \\ 0 & \ddots & \sigma_r \end{matrix} & 0 \\ 0 & 0 \end{bmatrix} \begin{array}{c} \overleftrightarrow{\;r\;} \quad \overleftrightarrow{\;m-r\;} \\ \begin{bmatrix} \begin{matrix} 1/\sigma_1 & & 0 \\ 0 & \ddots & 1/\sigma_r \end{matrix} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} \updownarrow r \\ \updownarrow m-r \end{matrix} \end{array}$$

$$= \begin{array}{c} \overleftrightarrow{\;r\;} \quad \overleftrightarrow{\;m-r\;} \\ \begin{matrix} r \updownarrow \\ m-r \updownarrow \end{matrix} \begin{bmatrix} \begin{matrix} 1 & & \\ & \ddots & 1 \end{matrix} & 0 \\ 0 & 0 \end{bmatrix} \end{array} \in \mathbb{R}^{m \times m}$$

$\Rightarrow$ Matrice diagonale carrée donc symétrique $(XM)^T = XM$

Théorème : Soit $X \in \mathbb{R}^{m \times n}$ et $X = U \Sigma V^T$ une SVD de X.

avec $\Sigma = \begin{bmatrix} \begin{matrix} \sigma_1 & & \\ & \ddots & \sigma_r \end{matrix} & 0 \\ 0 & 0 \end{bmatrix}$ $\quad \sigma_1 \geq \dots \geq \sigma_r > 0 = \sigma_{r+1} = \dots = \sigma_{\min(m,n)}$

$r = \text{rang}(X)$

Alors, la pseudo-inverse de X est donnée par

$$X^+ = \underbrace{V}_{n \times n} \underbrace{\Sigma^+}_{n \times m} \underbrace{U^T}_{m \times m} \in \mathbb{R}^{n \times m}$$

<u>Remarques</u>

- Le calcul de X^+ à partir de la SVD se fait très facilement
- $X = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ et $X^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T$
- La méthode via la SVD est la méthode de calcul numérique de pseudo-inverse la plus usuelle (et la plus stable numériquement, càd la plus robuste aux erreurs numériques)

<u>Corollaire</u> : Pour toute $X \in \mathbb{R}^{m \times n}$, on a :

- $X^+ = X^T (X X^T)^{-1}$ si $\text{rang}(X) = m$
- $X^+ = (X^T X)^{-1} X^T$ si $\text{rang}(X) = n$
- $X^+ = X^{-1}$ si $\text{rang}(X) = n = m$ (X carrée inversible)

(et dans ce dernier cas $X^+ = X^{-1} = (X^T X)^{-1} X^T = X^T (X X^T)^{-1}$)

<u>Théorème</u> : Soit $X \in \mathbb{R}^{m \times n}$ et $y \in \text{Im}(X)$.

Alors $\hat{\beta} = X^+ y$ est une solution de $X\beta = y$.

$\longrightarrow$ Généralisation des cas X inversible et X de rang plein (avec $y \in \text{Im}(X)$)

Démo Identique au cas $\text{rang}(X)=m$, $\text{rang}(X)=n$ et X inversible

Dans le cas où $\text{rang}(X)<\min(m,n)$, on utilise la décomposition en valeurs singulières de X

Si $y\in \text{Im}(X)$, $y = X\gamma$ pour $\gamma\in\mathbb{R}^n$

$$= U\Sigma V^T\gamma \quad \text{où } X=U\Sigma V^T \text{ est une SVD}$$

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \quad \text{avec} \quad r<\min(m,n)$$

$$\left(y = \sum_{i=1}^{r} \sigma_i u_i v_i^T \gamma \right)$$

Par ailleurs, $X^+y = V\Sigma^+U^Ty$ $\quad \left[\begin{array}{ccc|c} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ \hline & 0 & & 0 \end{array}\right]$

$$= V\Sigma^+U^TU\Sigma V^T\gamma = V\Sigma^+\Sigma V^T\gamma$$

$$\left(= \sum_{i=1}^{r} v_i v_i^T \gamma \right)$$

et donc $X(X^+y) = U\Sigma V^T(V\Sigma^+\Sigma V^T\gamma)$

$$= U\Sigma \underbrace{V^TV}_{I_m} \Sigma^+\Sigma V^T\gamma$$

$$= U\underbrace{\Sigma\Sigma^+\Sigma}_{=\Sigma \text{ par propriété de } \Sigma^+} V^T\gamma \quad \left(= \sum_{i=1}^{r} \sigma_i \times \frac{1}{\sigma_i} \times \sigma_i \, u_i v_i^T \gamma \right)$$

$$= U\Sigma V^T\gamma = y$$