Rappels Mathématiques et Notations

Machine Learning

Cédric RICHARD
Université Nice Sophia Antipolis

Reconnaissance des formes

Objectifs

L'objectif de l'analyse de données est de synthétiser, structurer, ..., l'information véhiculée par des données multidimensionnelles :

- ▷ n : nombre d'individus
- ▷ p : nombre de variables

Les méthodes mises en œuvre relèvent essentiellement de l'*algèbre linéaire* et de la *théorie des probabilités*. En effet :

- ▷ les données sont vues comme un nuage de points dans un espace vectoriel
- ▷ La statistique inférentielle permet de fournir des résultats relatifs à une population à partir de mesures statistiques réalisées sur des échantillons.

Vocabulaire

Individus et variables

Population : Groupe ou ensemble d'individus que l'on analyse

Recensement : Etude de tous les individus d'une population donnée

Sondage : Etude d'une partie seulement d'une population appelée échantillon

Variables : Ensemble de caractéristiques d'une population

- *quantitatives* : nombres sur lesquels les opérations usuelles ont un sens. Elles peuvent être discrètes ou continues
- *qualitatives* : appartenance à une catégorie donnée. Elles peuvent être nominales, ou ordinales quand les catégories sont ordonnées.

Vocabulaire

Description de données quantitatives

Variable, individu : On appelle variable un vecteur $\boldsymbol{x}$ de taille n. Chaque coordonnée x_i correspond à un individu.

Poids : Chaque individu a éventuellement un poids p_i, tel que $p_1 + \ldots + p_n = 1$. On choisit souvent $p_i = \frac{1}{n}$.

Analyse : On dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, 1$^{\text{er}}$ quartile, ...
Ces indicateurs mesurent principalement la tendance centrale et la dispersion. On utilisera surtout la moyenne, la variance et l'écart type.

Grandeurs statistiques de base

Moyenne arithmétique

Définition 1. *On appelle moyenne arithmétique, que l'on note $\bar{x}$, la quantité suivante*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

ou, dans le cas d'une somme pondérée

$$\bar{x} = \sum_{i=1}^{n} p_i x_i$$

Remarque. La moyenne arithmétique est une mesure de tendance centrale qui dépend de toutes les observations, et est sensible aux valeurs extrêmes.

Grandeurs statistiques de base

Variance et écart-type

Définition 2. *La variance de x est définie par*

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

ou, dans le cas d'une pondération non-uniforme

$$\sigma_x^2 = \sum_{i=1}^{n} p_i(x_i - \bar{x})^2$$

L'écart type σ_x est la racine-carrée de la variance.

Propriété 1. *La variance satisfait la relation suivante*

$$\sigma_x^2 = \sum_{i=1}^{n} p_i x_i^2 - \bar{x}^2$$

L'écart-type, qui a la même unité que x, est une mesure de dispersion.

Grandeurs statistiques de base

Mesure de liaison entre deux variables

Définition 3. *La covariance observée entre deux variables x et y est définie par*

$$\sigma_{xy} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y})$$

Le coefficient de corrélation est donné par

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Grandeurs statistiques de base

Propriétés du coefficient de corrélation

Propriété 2. *D'après l'inégalité de Cauchy-Schwartz, on a*

$$-1 \leq r_{xy} \leq 1$$

Propriété 3. *Le résultat suivant concerne des variables dites linéairement liées.*

$$|r_{xy}| = 1 \Leftrightarrow ax_i + by_i = c, 1 \leq i \leq n$$

En particulier, on a $r_{xx} = 1$.

Remarque. Si $r_{xy} = 0$, les variables sont dites décorrélées. Cela ne signifie pas qu'elles sont indépendantes.

Description des données

Notations matricielles

Matrice : De manière impropre, une matrice à p lignes et n colonnes est un tableau rectangulaire de mn nombres, rangés ligne par ligne.

Vecteur : Un vecteur, ligne ou colonne, est une matrice ne comportant qu'une seule ligne ou qu'une seule colonne.

Transposition : Echange des lignes et des colonnes d'une matrice. On note $\boldsymbol{M}^\top$ la transposée de $\boldsymbol{M}$.

Exemples :

$$\boldsymbol{I} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \qquad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Description des données

Tableau de données

Dans toute la suite, pour n individus et p variables, on s'intéresse aux tableaux de données définis comme suit

$$\boldsymbol{X} = (\boldsymbol{x}^1 \quad \ldots \quad \boldsymbol{x}^p) = \begin{pmatrix} x_1^1 & x_1^2 & & \cdots & & x_1^p \\ x_2^1 & x_2^2 & & & & \\ & & \ddots & & & \\ \vdots & & & x_i^j & & \vdots \\ & & & & \ddots & \\ x_n^1 & & & \cdots & & x_n^p \end{pmatrix} \qquad \boldsymbol{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Description des données

Vecteurs variable et individu

Variable : Une colonne du tableau de données

$$\boldsymbol{x}^j = \begin{pmatrix} x_1^j \\ x_2^j \\ \dots \\ x_n^j \end{pmatrix}$$

Individu : Une ligne du tableau de données, transposées

$$\boldsymbol{x}_i = \begin{pmatrix} x_i^1 & x_i^2 & \cdots & x_i^p \end{pmatrix}^\top$$

Description des données

Matrice de poids

Pourquoi : Elle est nécessaire quand les individus n'ont pas la même importance.

Comment : On associe un poids p_i à chaque individu tel que :

$$p_1 + p_2 + \ldots + p_n = 1$$

On regroupe ces poids dans une matrice diagonale de taille n :

$$\boldsymbol{D} = \begin{pmatrix} p_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n \end{pmatrix}$$

Cas uniforme : Tous les individus ont le même poids $p_i = 1/n$

Pré-traitements

Individu moyen et tableau centré

Individu moyen : L'individu moyen est obtenu à partir de la moyenne arithmétique de chaque variable

$$\boldsymbol{m} = \begin{pmatrix} \bar{x}^1 & \bar{x}^2 & \cdots & \bar{x}^p \end{pmatrix}^\top$$

avec $\bar{x}^j = \sum_{i=1}^n p_i \, x_i^j$. On peut aussi écrire

$$\boldsymbol{m} = \boldsymbol{X}^\top \boldsymbol{D}\mathbf{1}$$

Tableau centré : Il est obtenu en centrant l'ensemble des variables du tableau de données : $y_i^j = x_i^j - \bar{x}^j$. Sous forme matricielle, on écrit

$$\boldsymbol{Y} = \boldsymbol{X} - \mathbf{1}\,\boldsymbol{m}^\top = (\boldsymbol{I} - \mathbf{1}\mathbf{1}^\top \boldsymbol{D})\boldsymbol{X}$$

Pré-traitements

Matrice de variance-covariance

Définition : Il s'agit d'une matrice de dimension p définie par

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & & \cdots & & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & & & & \\ & & \ddots & & & \\ \vdots & & & \sigma_i^2 & & \vdots \\ & & & & \ddots & \\ \sigma_{p1} & & & \cdots & & \sigma_p^2 \end{pmatrix}$$

où σ_{ij} est la covariance des variables x^i et x^j, et σ_j^2 est la variance de $\boldsymbol{x}^j$.

Formulation matricielle :

$$\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X} - \boldsymbol{m}\boldsymbol{m}^\top = \boldsymbol{Y}^\top \boldsymbol{D} \boldsymbol{Y}$$

Pré-traitements
Matrice de corrélation

Définition : Il s'agit d'une matrice de dimension p définie par

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & & \cdots & & r_{1p} \\ r_{21} & 1 & & & & \\ & & \ddots & & & \\ \vdots & & & r_{ij} & & \vdots \\ & & & & \ddots & \\ r_{p1} & & & \cdots & & 1 \end{pmatrix}$$

où $r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$ est le coefficient de corrélation des variables x^i et x^j.

Formulation matricielle :

$$\boldsymbol{R} = \boldsymbol{D}_{1/\sigma} \, \boldsymbol{\Sigma} \, \boldsymbol{D}_{1/\sigma}$$

où $\boldsymbol{D}_{1/\sigma}$ est la matrice diagonale de termes diagonaux $\frac{1}{\sigma_i}$.

Analyse des données

Métrique

Motivation : Il est nécessaire d'introduire une métrique afin de caractériser la topologie du nuage de points.

Définition : On appelle distance sur E une application $d : E \times E \to \mathbb{R}^+$ vérifiant les propriétés suivantes

— symétrie : $\forall\, x, y \in E, \quad d(x, y) = d(y, x)$

— séparation : $\forall\, x, y \in E, \quad d(x, y) = 0 \Leftrightarrow x = y$

— inégalité triangulaire : $\forall\, x, y, z \in E, \quad d(x, z) \leq d(x, y) + d(y, z)$

Exemple : La distance euclidienne entre 2 points $\boldsymbol{u}$ et $\boldsymbol{v}$ de $\mathbb{R}^p$ est définie par

$$d^2(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{p} (u_j - v_j)^2 = \|\boldsymbol{u} - \boldsymbol{v}\|^2$$

Matrice définie positive : Il s'agit d'une matrice symétrique $\boldsymbol{M}$ telle que, pour tout $\boldsymbol{u}$ non nul, on a $\boldsymbol{u}^\top \boldsymbol{M} \boldsymbol{u} > 0$.

Définition : Soit $\boldsymbol{M}$ une matrice définie positive de dimension p. La fonction suivante $d_M : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ définit une métrique

$$d_M^2(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|_M^2 \qquad \text{avec} \qquad \|\boldsymbol{u}\|_M^2 = \sum_{i,j=1}^{p} m_{ij}\, u_i\, u_j$$

Cette distance est appelée distance de Mahalanobis lorsque $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$, où $\boldsymbol{\Sigma}$ est la matrice de variance-covariance des données.

Produit scalaire : La métrique définie ci-dessus dérive du produit scalaire

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_M = \sum_{i,j=1}^{p} m_{ij}\, u_i\, v_j$$

On dit que $\boldsymbol{u}$ et $\boldsymbol{v}$ sont orthogonaux si $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_M = 0$.

Métrique euclidienne : Elle est obtenue pour $\boldsymbol{M} = \boldsymbol{I}$.

L'une des difficultés rencontrées avec la métrique euclidienne est qu'elle privilégie les variables les plus dispersées et dépend donc de leur unité de mesure.

Métrique réduite : Elle consiste à prendre $\boldsymbol{M} = \boldsymbol{D}_{1/\sigma^2}$, où $\boldsymbol{D}_{1/\sigma^2}$ est la matrice diagonale de termes diagonaux les inverses $\frac{1}{\sigma_i}$ des variances des variables.

$$\boldsymbol{D}_{1/\sigma^2} = \begin{pmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_p^2} \end{pmatrix}$$

Analyse des données

Inertie

Définition : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ en un point quelconque $\boldsymbol{a}$ est donnée par

$$I_a = \sum_{i=1}^{n} p_i \|\boldsymbol{x}_i - \boldsymbol{a}\|_M^2$$

Définition : L'inertie totale du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ est donnée par

$$I_m = \frac{1}{2} \sum_{i,j=1}^{n} p_i \, p_j \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2$$

Propriété :

$$I_m = \mathrm{Trace}(\boldsymbol{\Sigma}\boldsymbol{M})$$

Démonstration : On introduit le vecteur moyen $\boldsymbol{m}$, et on déroule le calcul

$$\begin{aligned}
I_m &= \frac{1}{2}\sum_{i,j=1}^{n} p_i\, p_j \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2 \\
&= \frac{1}{2}\sum_{i,j=1}^{n} p_i\, p_j \|(\boldsymbol{x}_i - \boldsymbol{m}) - (\boldsymbol{x}_j - \boldsymbol{m})\|_M^2 \\
&= \underbrace{\sum_{i=1}^{n} p_i \|\boldsymbol{x}_i - \boldsymbol{m}\|_M^2}_{\mathrm{Trace}(\boldsymbol{\Sigma}\boldsymbol{M})} - \underbrace{\sum_{i,j=1}^{n} p_i\, p_j \langle \boldsymbol{x}_i - \boldsymbol{m}, \boldsymbol{x}_j - \boldsymbol{m}\rangle_M}_{0} \\
&= \mathrm{Trace}(\boldsymbol{\Sigma}\boldsymbol{M})
\end{aligned}$$

Métrique euclidienne :

$$I_m = \mathrm{Trace}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_i^2$$

Métrique réduite :

$$\begin{aligned} I_m &= \mathrm{Trace}(\boldsymbol{\Sigma}\boldsymbol{D}_{1/\sigma^2}) \\ &= \mathrm{Trace}(\boldsymbol{D}_{1/\sigma}\boldsymbol{\Sigma}\boldsymbol{D}_{1/\sigma}) \\ &= \mathrm{Trace}(\boldsymbol{R}) = p \end{aligned}$$

Utiliser la métrique $\boldsymbol{M} = \boldsymbol{T}^\top \boldsymbol{T}$ sur le tableau de données $\boldsymbol{X}$ est équivalent à travailler avec la métrique euclidienne sur le tableau transformé $\boldsymbol{X}\boldsymbol{T}^\top$.

Tableau transformé : Lorsqu'on travaille sur le tableau transformé comme ci-dessus, il convient d'utiliser la norme euclidienne. En effet,

$$\langle \boldsymbol{T}\boldsymbol{x}_i, \boldsymbol{T}\boldsymbol{x}_j \rangle = (\boldsymbol{T}\boldsymbol{x}_i)^\top (\boldsymbol{T}\boldsymbol{x}_j) = \boldsymbol{x}_i^\top (\boldsymbol{T}^\top \boldsymbol{T}) \boldsymbol{x}_j = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_M$$

Réciproque : Pour toute matrice définie positive $\boldsymbol{M}$, il existe une matrice définie positive $\boldsymbol{T}$ telle que $\boldsymbol{M} = \boldsymbol{T}^\top \boldsymbol{T}$. On notera improprement $\boldsymbol{T} = \boldsymbol{M}^{\frac{1}{2}}$.

Rappels élémentaires d'algèbre linéaire

Valeurs et vecteurs propres

Définition : Une matrice $\boldsymbol{A}$ à coefficients dans un corps $\mathbb{K}$ est diagonalisable sur ce corps $\mathbb{K}$ s'il existe une matrice inversible $\boldsymbol{P}$ et une matrice diagonale $\boldsymbol{D}$ à coefficients dans $\mathbb{K}$ telles que

$$\boldsymbol{A} = \boldsymbol{P}\,\boldsymbol{D}\boldsymbol{P}^{-1}$$

Chaque colonne $\boldsymbol{p}$ de $\boldsymbol{P}$ est un vecteur propre de $\boldsymbol{M}$, c'est à dire qu'il existe λ sur la diagonale de $\boldsymbol{D}$ tel que

$$\boldsymbol{A}\boldsymbol{p} = \lambda\,\boldsymbol{p}$$

Rappels élémentaires d'algèbre linéaire

Valeurs et vecteurs propres

Propriété : Toute matrice symétrique réelle est diagonalisable sur $\mathbb{R}$ par une matrice orthogonale $\boldsymbol{P}$, c'est à dire telle que

$$\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}$$

Propriété : Toute matrice $\boldsymbol{M}$-symétrique réelle ($\boldsymbol{A}^\top \boldsymbol{M} = \boldsymbol{M}\boldsymbol{A}$) est diagonalisable sur $\mathbb{R}$ par une matrice $\boldsymbol{M}$-orthogonale $\boldsymbol{P}$, c'est à dire telle que

$$\boldsymbol{P}^\top \boldsymbol{M} \boldsymbol{P} = \boldsymbol{I}$$

Rappels élémentaires d'algèbre linéaire

Valeurs et vecteurs propres : cas de la matrice $\boldsymbol{\Sigma M}$

Valeurs propres : La matrice $\boldsymbol{\Sigma M}$ est $\boldsymbol{M}$-symétrique. Elle est donc diagonalisable sur $\mathbb{R}$. Ses valeurs propres sont positives, et l'on note

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

Vecteurs propres : Les vecteurs propres de $\boldsymbol{\Sigma M}$ sont $\boldsymbol{M}$-orthogonaux.

Lien avec l'inertie : On sait que

$$\text{Trace}(\boldsymbol{\Sigma M}) = \lambda_1 + \lambda_2 + \ldots + \lambda_k + \ldots + \lambda_p$$

En conservant l'information relative au sous-espace propre $\{\lambda_1, \ldots, \lambda_k\}$, on conserve l'inertie $\lambda_1 + \lambda_2 + \ldots + \lambda_k$.