Révisions

— Slides : link

Project d'analyse données

Un projet d'analyse de données se découpe en cinq étapes :

1. Définition des objectifs
2. Données
3. Élaboration et validation des modèles
4. Mise en oeuvre
5. Suivi de la performance et amélioration

> Remarque
>
> Dans ce cours, on s'intéressera principalement à l'élaboration et à la validation de modèles.

TODO : Planification d'un projet

Définition des objectifs

Est-ce que l'on veut : visualiser les données ? explorer et émettre des hypothèses ? tester ? regrouper ? comprendre ? prédire ?

Comment fait-on en pratique ? On pose des questions ! Tout d'abord, il faut clarifier les termes. Qui va utiliser le modèle et comment ? Quelle est la population cible ?

> Examples
>
> 1. La Banque National du Canada veut lancer un nouveau produit d'épargne et souhaite mieux connaître ses clients pour prédire s'ils veulent l'acheter.
> 2. L'équipe de hockey des Canadiens de Montréal souhaite mieux connaître ses ad-

versaires pour développer des nouvelles tactiques de jeu.
3. Pharmascience souhaite savoir si son nouveau médicament est efficace.

Données

— Inventaire et qualité
— Constitution de la base de données
— Exploration et traitement préliminaire

Qu'est-ce que l'on veut dire par qualité des données ?

— Est-ce que les données sont représentatives de la population cible ?
— Est-ce que les données permetternt de tirer des conclusions de causalité ?
— Est-ce que les données sont fiables ?

Source de données :

Quelques liens pour récupérer des données.

Nettoyage de données : cf R (importation, nettoyage, tidyverse, types de variables, retirer les doublons, uniformiser les modalités, vérifier le format des valeurs spéciales, pivot, opérateur pipe, jointure).

Exploration des données : modalités rares, modalités trop nombreuses, asymétrie, débalancement des classes, valeurs extrêmes ou aberrantes, variables fortement corrélées, valeurs manquantes.

Statistiques descriptives

Révisions d'algèbre linéaire

Cf. cours MAT-1200. Donner quelques reférences.

Notons M, N et P des matrices de taille $n \times m$, A et B des matrices carrées et I_n la matrice identités, de dimension $n \times n$, et u et v des vecteurs colonnes de taille n.

Propriétés de l'inverse

$$(AB)^{-1} = B^{-1}A^{-1}$$

Propriétés du déterminant

$$
\begin{aligned}
\det(A^\top) &= \det(A) \\
\det(A^{-1}) &= 1/\det(A) \\
\det(AB) &= \det(A)\det(B)
\end{aligned}
$$

Propriétés de la trace

$$
\begin{aligned}
\operatorname{tr}(A+B) &= \operatorname{tr}(A) + \operatorname{tr}(B) \\
\operatorname{tr}(MN) &= \operatorname{tr}(NM)
\end{aligned}
$$

Propriété de matrices :

- Soit A une matrice symmétrique de dimension $n \times n$. A est définie positive si elle est positive et inversible, c'est-à-dire si $u^\top A u > 0$ pour tout $x \in \mathbb{R}^n$ tel que $x \neq 0$.
- Soit A une matrice carrée à valeur dans $\mathbb{R}$. A est orthogonal si $A^\top A = AA^\top = I_n$.

Valeurs et vecteurs propres :

- Soit A une matrice carrée de dimension $n \times n$. On dit que λ est une valeur propre de A si il existe un vecteur $u \neq 0 \in \mathbb{R}^n$ tel que
$$Au = \lambda u.$$
Le vecteur u est appelé vecteur propre correspondant à la valeur propre λ et l'ensemble des nombres réels λ satidfaisant l'équation est appelé spectre de la matrice A et noté $\mathrm{sp}(A)$.
- Si u est un vecteur propre de A correspondant à une valeur propre λ, alors cu, $c \neq 0 \in \mathbb{R}$ sera également un vecteur propre de A correspondant à λ.
- Si A est symmétrique et u_1 et u_2 sont des vecteurs propres correspondant à des valeurs propres différentes de A, alors u_1 et u_2 sont orthogonaux, *i.e.* $u_1^\top u_2 = 0$.
- Si A a comme valeurs propres (réelles, mais pas forcément distinctes) $\lambda_1, \dots, \lambda_n$, alors
$$\mathrm{A} = \prod_{i=1}^{n} \lambda_i \quad \text{et} \quad \operatorname{tr}(A) = \sum_{i=1}^{n} \lambda_i.$$
- Si A est symmétrique, **toutes** ses valeurs propres sont réelles.
- Si A est définie positive, alors toutes ses valeurs propres sont positives.

Diagonalisation de matrices :

— Soit A une matrice carrée de dimension $n \times n$. On dit que A est diagonalisable s'il existe une matrice carrée $n \times n$ non-singulière P et une matrice $n \times n$ diagonale D telles que

$$P^{-1}AP = D \leftrightarrow A = PDP^{-1}.$$

Toute matrice carrée symmétrique est diagonalisable part une matrice orthogonal P.

Théorème de décomposition spectrale :

Soit A une matrice carrée symmétrique de dimension $n \times n$ et ses n valeurs propres $\lambda_1, \dots, \lambda_n$. Alors il existe une matrice orthogonal P telle que

$$A = P\Lambda P^\top, \quad \text{où} \quad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n).$$

Si A admet n valeurs propres positives distinctes, alors on peut prendre P comme la matrice dont la ke coloone est le vacteur propre normé correspondant à la ke valeur propre λ_k.

Révisions de probabilité

Vecteurs aléatoires :

Soit $X = (X_1, \dots, X_p)^\top$, un vecteur aléatoire de taille p.

Espérance :

$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_p) \end{pmatrix} := \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} := \mu.$$

Matrice de variance/covariance :

$$\operatorname{Var}(X) = \begin{pmatrix} \operatorname{Var}(X_1) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_1, X_p) & \cdots & \operatorname{Var}(X_p) \end{pmatrix} := \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{1,p} & \cdots & \sigma_p \end{pmatrix} := \Sigma.$$

Matrice de corrélation :

$$\operatorname{Cor}(X) = \begin{pmatrix} 1 & \cdots & \operatorname{Corr}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \operatorname{Corr}(X_1, X_p) & \cdots & 1 \end{pmatrix} := \begin{pmatrix} 1 & \cdots & \rho_{1,p} \\ \vdots & \ddots & \vdots \\ \rho_{1,p} & \cdots & 1 \end{pmatrix} := R.$$

Propriété des moments :

Properties

Soit X un vecteur aléatoire de moyenne $\mathbb{E}(X) = \mu$ et de variance $\mathrm{Var}(X) = \Sigma$, et soit M une matrice de constantes et v un vecteur de constantes.

1. Σ est définie non-négative et symétrique.
2. $\Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^\top\right] = \mathbb{E}(XX^\top) - \mu\mu^\top$.
3. $\mathbb{E}(MX + v) = \mathbb{E}(X) + v$.
4. $\mathrm{Var}(MX + v) = \mathrm{Var}(MX) = M\Sigma M^\top$.
5. $\Sigma = \Delta R \Delta \iff R = \Delta^{-1}\Sigma\Delta^{-1}$.

Loi normale multivariée : On dit qu'un vecteur aléatoire X de dimension p suit une loi normale multidimensionnelle de moyenne μ et de variance $\Sigma \sim \mathcal{N}_p(\mu, \sigma^2)$, si sa densité est données par

$$f_X(x) = \frac{1}{(2\pi)^{p/2}} \cdot \frac{1}{(\det\Sigma)^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p.$$

Estimation avec un échantillon : En practique, nous ne connaissons pas les valeurs de μ et de Σ et nous voulons les estimer à partir d'un échantillon. Soit $X_1, \ldots, X_n$, n réalisations indépendantes d'un vecteur aléatoire X de moyenne μ et de variance Σ. On estime μ et Σ par :

$$\hat{\mu} = \overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\widehat{\Sigma} = S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^\top.$$

Notons $D = \{\mathrm{diag}(S^2)\}^{1/2}$, la matrice des écarts-types calculés sur l'échantillon. On peut calculer la matrice des corrélations sur l'échantillon par :

$$\widehat{R} = D^{-1}S^2D^{-1} \iff S^2 = D\widehat{R}D.$$