

Analyse, Classification et Indexation des données (ACID)

Rappels de probabilités et de statistiques

Akka Zemmari

LaBRI, Université de Bordeaux

2026 - 2027

Pourquoi ces rappels dans ACID ?

- ▶ **Décrire** un jeu de données : position, dispersion, dépendances.
- ▶ **Préparer** les variables : centrage et standardisation.
- ▶ **Comprendre** les critères utilisés par les algorithmes : erreur quadratique, variance, entropie, distance.
- ▶ **Mesurer** la variabilité due à l'échantillonnage.

À noter

Ces rappels sont volontairement centrés sur les outils utilisés dans le cours, hors du chapitre consacré à la théorie bayésienne.

Population, échantillon et données

On observe un échantillon de n individus :

$$\mathcal{D} = \{x_1, \dots, x_n\}, \quad x_i \in \mathbb{R}^d.$$

Les observations sont regroupées dans une matrice :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix} \in \mathbb{R}^{n \times d}.$$

- ▶ une **ligne** représente un individu ;
- ▶ une **colonne** représente une variable (ou feature) ;
- ▶ les statistiques calculées sur l'échantillon servent à estimer les caractéristiques de la population.

Variables quantitatives et qualitatives

- ▶ **Quantitative continue** : taille, température, score, durée.
- ▶ **Quantitative discrète** : nombre de clics, fréquence d'un terme.
- ▶ **Qualitative nominale** : couleur, catégorie, classe.
- ▶ **Qualitative ordinale** : faible, moyen, élevé.

Exemples dans le cours

Un label de classe est qualitatif ; un pixel, une coordonnée ou un embedding sont quantitatifs ; un document peut être représenté par des fréquences.

Mesures de position

Pour une variable observée $x_1, \dots, x_n$, la moyenne empirique est

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

- ▶ La **moyenne** utilise toutes les observations, mais elle est sensible aux valeurs extrêmes.
- ▶ La **médiane** partage les observations ordonnées en deux groupes de même effectif ; elle est plus robuste aux valeurs extrêmes.

Exemple

Pour 1, 2, 3, 4, 100, la moyenne vaut 22, tandis que la médiane vaut 3.

Mesures de dispersion

La variance empirique corrigée est

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_x = \sqrt{s_x^2}.$$

- ▶ s_x^2 mesure la dispersion quadratique autour de la moyenne.
- ▶ L'écart-type s_x s'exprime dans la même unité que x .

À retenir

Moyenne et variance doivent être interprétées ensemble.

Valeurs atypiques et robustesse

Une valeur atypique peut être :

- ▶ une erreur de mesure ou de saisie ;
- ▶ une observation rare mais légitime ;
- ▶ le signal recherché, par exemple une fraude ou une anomalie.

Attention

Détecter une valeur atypique ne signifie pas qu'il faut la supprimer.
Son traitement dépend du problème et de la provenance des données.

Covariance et corrélation

Pour deux variables x et y , on définit

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

La corrélation linéaire normalise la covariance :

$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}, \quad -1 \leq r_{xy} \leq 1.$$

- ▶ $r_{xy} > 0$: les deux variables ont tendance à évoluer dans le même sens.
- ▶ $r_{xy} < 0$: elles ont tendance à évoluer en sens opposé.
- ▶ $r_{xy} \simeq 0$: absence de liaison *linéaire* détectable.

Attention

Corrélation ne signifie causalité.

Matrice de covariance

On note $\bar{x} \in \mathbb{R}^d$ le vecteur moyen et Z la matrice des données centrées :

$$z_i = x_i - \bar{x}, \quad Z = X - \mathbf{1}_n \bar{x}^T.$$

La matrice de covariance empirique est

$$\hat{\Sigma} = \frac{1}{n-1} Z^T Z.$$

- ▶ $\hat{\Sigma}_{jj}$ est la variance de la variable j ;
- ▶ $\hat{\Sigma}_{jk}$ est la covariance entre les variables j et k ;
- ▶ $\hat{\Sigma}$ est symétrique et positive semi-définie.

Lien avec le cours

En ACP, les directions principales sont les vecteurs propres de $\hat{\Sigma}$ et leurs valeurs propres mesurent les variances conservées.

Centrer et standardiser

Pour chaque variable j :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}.$$

Après standardisation, chaque colonne a approximativement :

moyenne 0, écart-type 1.

- ▶ Le **centrage** retire l'effet de position.
- ▶ La **réduction** rend les échelles comparables.
- ▶ Elle est importante pour k -NN, k -means, SVM et ACP.

Fuite de données

La moyenne et l'écart-type sont calculés sur l'ensemble d'entraînement, puis réutilisés tels quels sur validation, test et nouvelles données.

Pourquoi l'échelle change les distances

Considérons deux individus décrits par l'âge et le revenu :

$$x = (20, 20000), \quad y = (40, 21000).$$

La distance euclidienne brute vaut

$$\|x - y\|_2 = \sqrt{20^2 + 1000^2} \simeq 1000,2.$$

La variable revenu domine presque entièrement la distance, simplement parce que son unité produit des valeurs plus grandes.

Conséquence

Sans standardisation, les voisins de k -NN, les centroïdes de k -means et les directions de l'ACP peuvent être dominés par une seule échelle de mesure.

Échantillonnage et hypothèse i.i.d.

On suppose souvent que

$X_1, \dots, X_n$ sont indépendantes et identiquement distribuées.

- ▶ **Identiquement distribuées** : les observations proviennent du même mécanisme statistique.
- ▶ **Indépendantes** : connaître une observation n'informe pas directement sur une autre.

Dans les données réelles

Cette hypothèse peut être violée par le temps, les groupes, les doublons ou plusieurs mesures d'un même individu. Le découpage entraînement–test doit respecter cette structure.

La loi normale : le minimum utile

Une variable X suit une loi normale de moyenne μ et de variance σ^2 , notée $X \sim \mathcal{N}(\mu, \sigma^2)$, si sa densité est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

La variable standardisée

$$Z = \frac{X - \mu}{\sigma}$$

suit la loi normale centrée réduite $\mathcal{N}(0, 1)$.

Pourquoi la rencontrerons-nous ?

Elle sert notamment à modéliser un bruit de mesure et les résidus de certains modèles, et à construire des intervalles ou des tests sous des hypothèses précises.

Ce qu'il faut retenir

- ▶ **Décrire** : moyenne, médiane, quantiles, variance et écart-type.
- ▶ **Relier** : covariance, corrélation et matrice de covariance.
- ▶ **Préparer** : centrer et standardiser sans fuite de données.
- ▶ **Modéliser** : distinguer signal, bruit, prédiction et résidu.

Et Bayes ?

Probabilités conditionnelles, vraisemblance, règles ML/MAP et Naive Bayes sont développés dans le chapitre dédié à la théorie bayésienne.