

Mathématiques pour l'intelligence artificielle

UE MIA, M2 IAAA, AMU, 2022-2023

Thomas Schatz
Vendredi 22 septembre 2022

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques

5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage
9. Putting it all together

Statistique

- Introduction générale
- Statistique inférentielle non asymptotique (classique)
- Statistique inférentielle asymptotique (classique)
- Statistique inférentielle non asymptotique (haute dimension)
- Statistique inférentielle asymptotique (haute dimension)

Introduction

Cadre général

Observations O

Distribution $O \sim P$

$P \in \mathcal{P}$ Modèle statistique

Functional $f : \mathcal{P} \rightarrow E$

Introduction

Cadre général

Observations O

Distribution $O \sim P$

$P \in \mathcal{P}$ Modèle statistique

Functional $f : \mathcal{P} \rightarrow E$

- Estimation ponctuelle $\hat{f} : \Omega \rightarrow E \quad \hat{f}(O) \approx f(P)$
- Estimation par intervalle $R : \Omega \rightarrow 2^{\text{Im} f}$

$$p_{O \sim P}(f(P) \in R(O)) \approx 1 - \alpha$$

Statistique inférentielle non asymptotique (classique)

- Estimation ponctuelle

- biais $b_P(\hat{f}) = E_{O \sim P}[\hat{f}(O)] - f(P) \quad \|b_P(\hat{f})\|_2$

- variance $\text{Var}_P(\hat{f}) = \text{Var}_{O \sim P}[\hat{f}(O)]$

- risque $\ell : \mathbf{R}^d \times \mathbf{R}^d \rightarrow \mathbf{R}$

$$R_P(\hat{f}) = E_{O \sim P}[\ell(f(P), \hat{f}(O))]$$

Outils pour la construction d'estimateurs ponctuels

- Statistique suffisante

Definition 2.4 (Sufficiency). Let X be a sample from an unknown population $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of populations. A statistic $T(X)$ is said to be *sufficient* for $P \in \mathcal{P}$ (or for $\theta \in \Theta$ when $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric family) if and only if the conditional distribution of X given T is *known* (does not depend on P or θ). ■

- Maximum de vraisemblance

Statistique inférentielle non asymptotique (classique)

What about tail probabilities ?

- Concentration inequalities

- Markov's bound
$$p(X \geq t) \leq \frac{E[X]}{t} \quad t \geq 0$$

- Hoeffding's bound
$$p\left(\left|\sum_{i=1}^n (X_i - \mu)\right| \geq t\right) \leq 2 \exp\left(-\frac{2t^2}{n(b-a)^2}\right) \quad t \geq 0$$

- Application : intervalle de confiances non-asymptotique

Statistique

- Introduction générale
- Statistique inférentielle non asymptotique (classique)
- Statistique inférentielle asymptotique (classique)
- Statistique inférentielle non asymptotique (haute dimension)
- Statistique inférentielle asymptotique (haute dimension)

Modes de convergence

- Convergence presque sûre ou presque partout
- Convergence en probabilité
- Convergence en loi
- Convergence en moyenne quadratique

Loi forte des grands nombres

$X_1, X_2, \dots$ variables aléatoires i.i.d.

(ii) (The SLLN). A necessary and sufficient condition for the existence of a constant c for which

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow_{a.s.} c \quad (1.81)$$

is that $E|X_1| < \infty$, in which case $c = EX_1$

Théorème de la limite centrale

(Multivariate CLT). Let $X_1, \dots, X_n$ be i.i.d. random k -vectors with a finite $\Sigma = \text{Var}(X_1)$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - EX_1) \rightarrow_d N_k(0, \Sigma). \quad \blacksquare$$

Transformations continues et théorème de Slutsky

Theorem 1.10. Let $X, X_1, X_2, \dots$ be random k -vectors defined on a probability space and g be a measurable function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. P_X . Then

- (i) $X_n \rightarrow_{a.s.} X$ implies $g(X_n) \rightarrow_{a.s.} g(X)$;
- (ii) $X_n \rightarrow_p X$ implies $g(X_n) \rightarrow_p g(X)$;
- (iii) $X_n \rightarrow_d X$ implies $g(X_n) \rightarrow_d g(X)$. ■

Theorem 1.11 (Slutsky's theorem). Let $X, X_1, X_2, \dots, Y_1, Y_2, \dots$ be random variables on a probability space. Suppose that $X_n \rightarrow_d X$ and $Y_n \rightarrow_p c$, where c is a fixed real number. Then

- (i) $X_n + Y_n \rightarrow_d X + c$;
- (ii) $Y_n X_n \rightarrow_d cX$;
- (iii) $X_n/Y_n \rightarrow_d X/c$ if $c \neq 0$.

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage
9. Putting it all together

Optimisation sans contraintes

$$f : \mathbf{R}^n \mapsto \mathbf{R}$$

$$\min_x f(x) \text{ ?}$$

x^* est un minimum global de f ssi pour tout $x \in \mathbf{R}^n$, $f(x^*) \leq f(x)$

x^* est un minimum local de f ssi il existe un ensemble ouvert $U \subset \mathbf{R}^n$ contenant x^* tel que pour tout $x \in U$, $f(x^*) \leq f(x)$

Analyse

Concept central $\lim_{x \rightarrow a} f(x) = l \text{ ?}$

Notions de base de topologie

Dans $\mathbf{R}^n$, La boule fermée $B_f(x, \epsilon)$, centrée sur x et de rayon ϵ associée à la norme Euclidienne est l'ensemble des points y de $\mathbf{R}^n$ tels que $\|x - y\|_2 \leq \epsilon$

Dans $\mathbf{R}^n$, La boule ouverte $B_o(x, \epsilon)$, centrée sur x et de rayon ϵ associée à la norme Euclidienne est l'ensemble des points y de $\mathbf{R}^n$ tels que $\|x - y\|_2 < \epsilon$

U est un ouvert de $\mathbf{R}^n$ ssi $U \subset \mathbf{R}^n$ et pour tout x dans U , il existe $\epsilon > 0$, tel que $B_f(x, \epsilon) \subset U$.

U est un fermé de $\mathbf{R}^n$ ssi son complément dans $\mathbf{R}^n$ est un ouvert de $\mathbf{R}^n$

Jacobienne, gradient

$$f : U \subset \mathbf{R}^n \rightarrow \mathbf{R}^m, \text{ de classe } C^1$$

Dérivée partielle $f : x \mapsto (f_1(x), f_2(x), \dots, f_m(x))$

$$\frac{\partial f_i}{\partial x_j} : (a_1, \dots, a_n) \mapsto \lim_{h \rightarrow 0} \frac{f_i(a_1, \dots, a_{j-1}, a_j + h, a_{j+1}, \dots, a_n) - f_i(a_1, \dots, a_j, \dots, a_n)}{h}$$

Matrice Jacobienne (transposée du gradient si m=1)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \dots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^T f_1 \\ \vdots \\ \nabla^T f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

“Chain rule”

$$J_{f \circ g}(\mathbf{a}) = J_f(g(\mathbf{a}))J_g(\mathbf{a}),$$

Optimisation sans contraintes

$$f : \mathbf{R}^n \mapsto \mathbf{R}$$

$$\min_x f(x) \text{ ?}$$

x^* est un minimum global de f ssi pour tout $x \in \mathbf{R}^n$, $f(x^*) \leq f(x)$

x^* est un minimum local de f ssi il existe un ensemble ouvert $U \subset \mathbf{R}^n$ contenant x^* tel que pour tout $x \in U$, $f(x^*) \leq f(x)$

Conditions d'optimalité

Conditions nécessaires, premier ordre

Si x^* est un minimum local et f est de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x , alors $\nabla f(x^*) = 0$