Apprentissage statistique
Chapitre 1 : Introduction aux statistiques

Lucie Le Briquer

18 février 2018

Table des matières

1 Rappels sur les lois conditionnelles

On se placera dans $(\Omega, \mathcal{F}, \mathbb{P})$.
Soit X, Y deux variables aléatoires dans $(\mathbb{X}, \mathcal{X})$ et $(\mathbb{Y}, \mathcal{Y})$. Soit $A \subset \mathcal{Y}$. $\mathbb{E}[\mathbb{1}_{Y \in A} \mid X]$ définie dans $\mathcal{L}^1(\Omega)$. On aimerait que $A \mapsto \mathbb{E}[\mathbb{1}_A \mid X]$ soit une probabilité $\mathbb{P}-$p.s.

Définition 1 (noyau Markovien)

On définit un noyau Markovien (ou de Markov) sur $(\mathbb{X}, \mathcal{Y})$: P tel que $x \in \mathbb{X} \mapsto P(x, A)$ soit mesurable sur $(\mathbb{X}, \mathcal{X})$, $\forall A \in \mathcal{Y}$ et $A \mapsto P(x, A)$ est une probabilité $\forall x \in \mathbb{X}$.

Définition 2 (loi conditionnelle régulière)

On dit que P un noyau de Markov sur $(\mathbb{X}, \mathcal{Y})$ est une loi conditionnelle régulière pour la loi conditionnelle de $Y|X$ si :

$$\mathbb{E}[\mathbb{1}_A \mid X] = P(X, A) \quad \mathbb{P} - \text{p.s.}$$

Proposition 1

Soit P une loi conditionnelle régulière pour la loi $Y|X$. $\forall g \colon \mathbb{Y} \longrightarrow \mathbb{R}_+$ mesurable :

$$\mathbb{E}[g(Y)|X] = \int g(y) P(X, dy)$$

Remarque. On supposera dans ce cours qu'il existe toujours des versions régulières des lois conditionnelles.

2 Modélisation statistique

$x_1, \ldots, x_n$ tirages de pile-face. On veut savoir si la pièce n'est pas pipée. On suppose que $x_1, \ldots, x_n$ sont des réalisations i.i.d. de variables aléatoires sur $(\Omega, \mathcal{F})$.
On cherche une probabilité sur $(\Omega, \mathcal{F})$ qui explique le mieux nos données $x_1, \ldots, x_n$.

Définition 3 (expérience statistique)

Soit (Ω, F) et $(\mathbb{X}, \mathcal{X})$ des espaces mesurables et $X \colon \Omega \longrightarrow \mathbb{X}$ une variable aléatoire. $\mathcal{P}$ une famille de probabilités sur $(\Omega, \mathcal{F})$.

$$(\Omega, \mathcal{F}, \mathbb{X}, \mathcal{X}, X, \mathcal{P}) \text{ est appelée une expérience statistique}$$

- On se limitera dans ce cours à $\mathbb{X} \subset \mathbb{R}^d$ mesurable assez sympa (ouvert/fermé)
- En général, on prend $\Omega = \mathbb{X}$, $\mathcal{F} = \mathcal{X}$ et $X \colon \omega \mapsto \omega$

- $\mathcal{P}$ et X définissent une famille de lois sur $(\mathbb{X}, \mathcal{X})$. On la note $\mathcal{P}_X$, et pour $\mathbb{P} \in \mathcal{P}$ on note la loi de X suivant $\mathbb{P}$, $\mathbb{P}^X$
- Dans le cas où on se permet un nombre infini d'observations, on est dans le cadre d'une expérience statistique dite "répétée".
$$(\Omega, \mathcal{F}, \mathbb{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}}, X, \mathcal{P}^{\mathbb{N}})$$
où $\mathcal{P}^{\mathbb{N}} = \{$ensemble des probas $\mathbb{P} \mid (X_i)_{i\in\mathbb{N}}$ soit i.i.d. de loi $\mathbb{P} \in \mathcal{P}\}$

Exemple. $\Omega = \mathbb{X} = \{0,1\}^n$, $\mathcal{F} = \mathcal{X} = \mathcal{P}(\Omega)$. $\mathcal{P} = \{\text{Ber}(p) \mid p \in [0,1]\}$.

Définition 4

Soit $\mathcal{P}$ un modèle statistique pour X. $\mathcal{P}$ est dominé par ν mesure $\sigma-$finie sur $(\mathbb{X}, \mathcal{X})$ si $\forall \mathbb{P} \in \mathcal{P},\ \mathbb{P}^X \ll \nu$.

Lemme 1

Si $\mathcal{P}$ est dominé par ν, alors il existe une partie dénombrable de $\mathcal{P}$, $(\mathbb{P}_n)_{n\in\mathbb{N}}$ tel que la mesure $\mathbb{Q} = \sum 2^{-n}\mathbb{P}_n^X$ domine $\mathcal{P}$.

3 Estimation de paramètres

Définition 5 (modèle paramétrique)

On dit que $\mathcal{P}$, un *modèle statistique* pour X, est paramétrique s'il existe un ensemble mesurable $\Theta \subset \mathbb{R}^d$ tel que $\mathcal{P}$ peut être paramétré par Θ i.e. $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$.
On dit que θ est *identifiable* si $\mathbb{P}_\theta = \mathbb{P}'_\theta \Rightarrow \theta = \theta'$.

Remarque. Dans les autres cas on dit $\mathcal{P}$ est non-paramétrique. On dit qu'une quantité $\psi(\mathbb{P})$ est identifiable si elle ne dépend que de $\mathbb{P} \in \mathcal{P}$.

Définition 6 (estimateur)

Soit $\mathcal{P}$ un modèle sur X. On appelle statistique ou estimateur toute v.a. T telle qu'il existe $g\colon \mathbb{X} \longrightarrow \mathbb{R}^d$ mesurable avec $g(X) = T$.

Définition 7 (biais)

Soit T un estimateur $\psi(\mathbb{P})|\theta$. On définit le biais de T par :
$$\mathbb{P} \longmapsto \text{Biais}(T, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[T] - \psi(\mathbb{P})$$

Définition 8 (coût quadratique)

On définit le coût quadratique de l'estimateur pour le paramètre $\psi(\mathbb{P})$ par :
$$\text{MSE}(T, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\|T - \psi(P)\|^2]$$

Propriété 1

$$\mathrm{MSE}(T,\mathbb{P}) = \mathrm{Biais}(T,\mathbb{P})^2 + \mathrm{Var}_{\mathbb{P}}(T)$$

Preuve.

$$\mathbb{E}_{\mathbb{P}}[\|T-\psi(\mathbb{P})\|^2] = \mathbb{E}[\|T \pm \mathbb{E}[T] - \psi(\mathbb{P})\|^2]$$

□

4 Statistiques suffisantes

Définition 9 (statistique suffisante)

$\mathcal{P}$ modèle statistique pour X. On dit que S est une statistique suffisante si la loi conditionnelle de $X|S$ ne dépend pas de $\mathbb{P}$ i.e. s'il existe un noyau de Markov $\mathcal{K}$ tel que $\forall \mathbb{P} \in \mathcal{P}$:

$$\mathbb{P}^{X|S}(.) = \mathcal{K}(S,.)$$

Exemple. (de la modélisation pile-face)

$$\mathcal{P} = \{\mathrm{Ber}(p)^{\otimes n} \mid p \in [0,1]\}$$

n nombre d'observations. $X_1,\ldots,X_n$ i.i.d. $\sim \mathcal{B}(p)$. Un estimateur de p est :

$$T(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Regardons le biais et la variance :

$$\mathrm{Biais}(T,\mathbb{P}) = 0 = \mathbb{E}[T(X_1,\ldots,X_n)] - p$$

$$\mathrm{Var}(T(X_1,\ldots,X_n)) = \frac{1}{n}^2 \sum \mathrm{Var}(X_i) = \frac{p(1-p)}{n}$$

T est-il suffisant ?
Calcul de loi conditionnelle. Soient $A,B \in \mathcal{X},\mathcal{Y}$, on veut trouver ψ tel que :

$$\mathbb{E}[\mathbb{1}_A(X)\mathbb{1}_B(X)] = \mathbb{E}[\mathbb{1}_B(X)\psi_A(X)]$$

On identifie en général la loi conditionnelle par $\psi_A(X)$.

Exemple. Loi conditionnelle de $(X_1,\ldots,X_n)|\sum X_i$ à faire en exo.
Soient X,Y deux variables aléatoires réelles indépendantes et $Z = X+Y$. Soient $A,B \in \mathcal{B}(\mathbb{R})$.

$$\mathbb{E}[\mathbb{1}_A(X)\mathbb{1}_B(Z)] = \mathbb{E}\Big[\mathbb{1}_A(X)\mathbb{E}[\mathbb{1}_B(X+Y)|X]\Big] = \mathbb{E}[\varphi(X)]$$

$\mathbb{E}[\psi(X,Y)|X] = \varphi(X)$ où $\forall x\ \varphi(x) = \mathbb{E}[\psi(x,Y)]$ Ici $\varphi(x) = \int_{\mathbb{R}} \mathbb{1}_B(y+x) d\mathbb{P}^Y(y)$. Si $Y \sim \mathcal{N}(0,1)$:

$$\begin{aligned}
\varphi(x) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathbb{1}_B(y+x) \exp\left(-\frac{y^2}{2}\right) dy \\
&= \int_{\mathbb{R}} \mathbb{1}_B(z) \exp\left(-\frac{-(z-x)^2}{2}\right) dz \\
&= \psi_B(z) \\
&= \mathbb{P}(\tilde{Z} \in B)
\end{aligned}$$

où $\tilde{Z} \sim \mathcal{N}(x,1)$. On en déduit que $Z|X \sim \mathcal{N}(X,1)$.

Théorème 1 (de factorisation de Fisher)

S est une statistique suffisante ssi il existe une fonction mesurable positive $h\colon \mathbb{X} \longrightarrow \mathbb{R}_+$ telle que $\forall \mathbb{P} \in \mathcal{P}$ on ait :

$$\frac{d\mathbb{P}^X}{d\nu} = h \times \rho_{\mathbb{P}} \circ g$$

où $g(X) = S$ et $\rho_{\mathbb{P}}\colon \mathbb{R}^d \longrightarrow \mathbb{R}_+-$mesurable.

Exemple. (cas pile-face)

ν la mesure de comptage domine $\mathcal{P}$. Soit $\mathbb{P}_p \in \mathcal{P}$.

$$\mathbb{P}_p = (p\delta_1 + (1-p)\delta_0)^{\otimes n}$$

Par exemple pour $n=1$, $\nu = \delta_0 + \delta_1$.

$$\mathbb{P}(X_1 \in A) = \sum_{k=0}^{1} \mathbb{P}(X_1 = k,\ k \in A) = \sum \mathbb{P}(X_1 = k, k \in A)\delta_k(A)$$

$$\begin{aligned}
\frac{d\mathbb{P}_p}{d\nu}(x_1,\dots,x_n) = \mathbb{P}_p(X_1 = x_1,\dots,X_n = x_n) &= \prod_{i=1}^{n} \mathbb{P}_p(X_i = x_i) \\
&= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}
\end{aligned}$$

On s'intéresse à

$$S = \sum_{i=1}^{n} X_i = g(X_1,\dots,X_n) = \rho_{\mathbb{P}_p} \circ g$$

où $\rho_{\mathbb{P}_p}(s) = p^s(1-p)^{1-s}$

5 Estimation de paramètres

On suppose que les observations proviennent d'une vraie loi $\mathbb{P}_* \in \mathcal{P}$ et on voudrait l'identifier.

Définition 10 (vraisemblance)

On définit la vraisemblance comme la fonction $\mathbb{P} \longmapsto \rho_{\mathbb{P}} \circ X$ où $\rho_{\mathbb{P}} = \frac{d\mathbb{P}^X}{d\nu}$. Dans le cas paramétrique c'est une fonction $\Theta \longrightarrow \mathbb{R}_+$, $\theta \longmapsto \rho_\theta \circ X$, $\rho_\theta = \frac{d\mathbb{P}_\theta^X}{d\nu}$.

Définition 11

Dans le cas paramétrique, on dit qu'un estimateur $\hat{\theta}$ est un estimateur du maximum de vraisemblance ssi

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \rho_\theta \circ X$$

6 Test statistique

$\mathcal{P} = \{\mathbb{P}_0, \mathbb{P}_1\}$. H_0="X a pour distribution $\mathbb{P}_0$", H_1="X a pour distribution $\mathbb{P}_1$". H_0 est l'hypothèse nulle, H_1 l'hypothèse alternative.

Définition 12

Un test statistique est une statistique δ à valeurs dans $\{0, 1\}$.

Si $\delta = 0$ on dit que H_0 est acceptée, sinon H_0 est rejetée.

On définit deux types de risques pour un test δ.

$$\mathbb{P}_0(\delta = 1) \quad \text{risque de 1ère espèce}$$

c'est le risque de rejeter à tort.

$$\mathbb{P}_1(\delta = 0) \quad \text{risque de 2nde espèce}$$

c'est le risque d'accepter à tort.

Définition 13

On dit δ est de niveau $\alpha \in (0, 1)$ si :

$$\mathbb{P}_0(\delta = 1) \leqslant \alpha$$

soit δ, δ' deux tests de niveau α, alors on dit que δ est plus puissant que δ' si :

$$\mathbb{P}_1(\delta = 1) \geqslant \mathbb{P}_1(\delta' = 1)$$

$$\Leftrightarrow \ \mathbb{P}_1(\delta = 0) \leqslant \mathbb{P}_1(\delta' = 0)$$

Test de Neyman-Pearson (ou du ratio de vraisemblance)

On prend ρ_0 et ρ_1 les densités de $\mathbb{P}_0$ et $\mathbb{P}_1$ par rapport à ν $(= \mathbb{P}_0 + \mathbb{P}_1$ ici). Le ratio de vraisemblance est défini par :

$$T_L(X) = \rho_1(X)/\rho_0(X)$$

On définit le test de N.-P. de seuil $t \in [0, +\infty]$ par :

$$\delta_t = \begin{cases} 1 & \text{si } T_L \geqslant t \\ 0 & \text{sinon} \end{cases}$$

Théorème 2

Soit $\mathcal{P} = \{\mathbb{P}_0, \mathbb{P}_1\}$ modèle statistique de X. Soit $t \in [0, +\infty]$. On définit :

$$\alpha_t = \mathbb{P}_0(\delta_t = 1)$$

Alors, pour tout δ' de niveau α_t, δ_t est plus puissant que δ'.

Preuve.
Soit $y < +\infty$. Soit δ' un test de niveau $\alpha_t = \mathbb{P}_0(\delta_t = 1)$. On veut montrer :

$$\mathbb{P}_1(\delta' = 0) \geqslant \mathbb{P}_1(\delta_t = 0)$$

δ' associée à une fonction $g \colon \mathbb{X} \longrightarrow \{0, 1\}$, $\delta' = g(X)$. $\delta_t = g_t(X)$. Dans le cas $t < +\infty$ $\mathbb{P}_i$−p.s. sur $\delta_t = 1$

$$\rho_1(X) \geqslant t\rho_0(X)$$

par définition de δ_t.

$$\begin{aligned}
\mathbb{P}_1(\delta' = 0) &= \int \mathbb{1}_0(g(x))\rho_1(x)\nu(dx) \pm t \int \mathbb{1}_1(g(x))\rho_0(x)\nu(dx) \\
&= \int_{\mathbb{X}} \mathbb{1}_0(g(x))\rho_1(x)\nu(dx) + \int_{\mathbb{X}} \mathbb{1}(g(x))t\rho_0(x)\nu(dx) - \mathbb{P}_0(\delta' = 1) \\
&\geqslant \int_{\mathbb{X}} \mathbb{1}_0(g(x))\rho_1(x) + \mathbb{1}_1(g(x))t\rho_0(x)\nu(dx) - \alpha_t \text{ car } \delta' \text{ de niveau } \alpha_t
\end{aligned}$$

donc la fonctionnelle :

$$g \longmapsto \int \mathbb{1}_0(g(x))\rho_1(x) + t\mathbb{1}_1(g(x))\rho_0(x)\nu(dx)$$

est optimale pour $g = g_t$.
$x \in \mathbb{X}$, si $\rho_1(x) \geqslant t\rho_0(x)$ alors $\delta_t(x) = 1$

$$\mathbb{1}_0(g_t(x))\rho_1(x) + t\mathbb{1}_1(g_t(x))\rho_0(x) = \min(\rho_1(x), t\rho_0(x))$$

Pareil pour $\delta_t = 0$. Or

$$\mathbb{1}_0(g(x))\rho_1(x) + \mathbb{1}_1(g(x))t\rho_0(x) \geqslant \min(\rho_1(x), t\rho_0(x))$$

et ce min est atteint pour g_t ce qui conclue. □