I - Rappels sur les probabilités

Romain HÉRAULT (Alain RAKOTOMAMONJY)

INSA Rouen

Automne 2015

Définitions

Espace des épreuves Ω

Ensemble de tous les évènements possibles issus d'une expérience donnée.

Définition de $P(A)$, approche fréquentiste

Soit A un ensemble d'évènements inclus dans Ω,

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n} \quad \text{si la limite existe,}$$

avec

- n le nombre d'expériences réalisées,
- $n(A)$ le nombre d'expériences où A s'est réalisé.

Axiomes

Les axiomes de probabilité

$$P(\Omega) = 1 \quad P(\emptyset) = 0$$

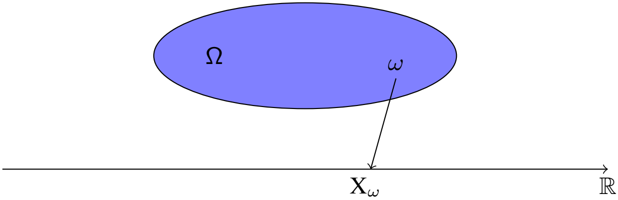
Avec $A \in \Omega,\ B \in \Omega$, nous avons :

$$\begin{aligned} 0 \leq \quad P(A) \quad &\leq 1 \\ P(A \cup B) \quad &= \quad P(A) + P(B) - P(A \cap B) \end{aligned}$$

Si $A \cap B = \emptyset$ alors $P(A \cup B) = P(A) + P(B)$.

Variable Aléatoire

C'est un nombre (réel) X_ω dont la valeur est déterminée par le résultat ω d'une expérience aléatoire.



Exemple

Un dé à 6 faces - l'évènement aléatoire (`e.a.`) est l'apparation d'une face. Si on associe un entier 1 à 6 à chaque face. On associe une `v.a.` à chaque `e.a.`.

Fonction de répartition

La fonction de répartition $F_{\mathrm{X}}(x)$ d'une $\mathtt{v.a.}\mathrm{X}$ est définie comme étant la probabilité que la $\mathtt{v.a.}\mathrm{X}$ soit inférieur ou égale à x,

$$F_{\mathrm{X}}(x) = P(\mathrm{X} \leq x) \quad .$$

Densité de probabilité ($\mathtt{d.d.p.}$)

Elle est définie comme la dérivée de la fonction de répartition,

$$p(x) = \frac{dF(x)}{dx} \quad .$$

Propriétés

Propriétés de la fonction de répartition

$$F_X(-\infty) = 0 \qquad F_X(\infty) = 1$$

$$0 \leq \quad F_X(x) \quad \leq 1$$

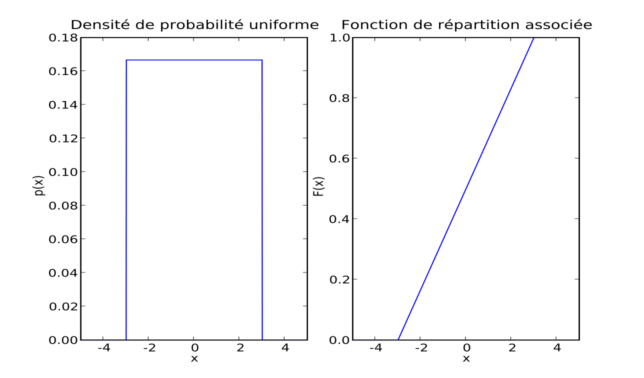
$$P(x_1 \leq x \leq x_2) \quad = \quad F_X(x_2) - F_X(x_1)$$

Propriétés de la densité de probabilité

$$p(x) \geq 0 \qquad \int_{-\infty}^{+\infty} p(x)dx = 1$$

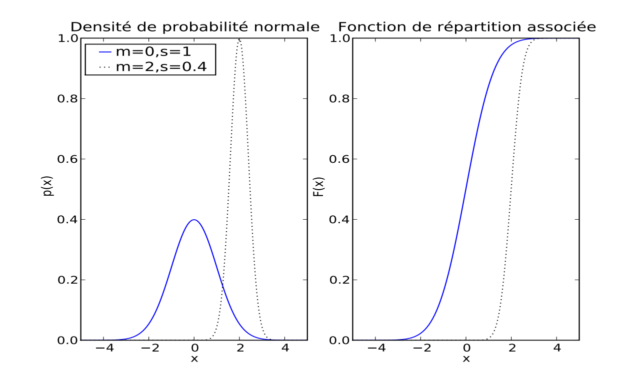
$$P(x \leq x_1) = F_X(x_1) = \int_{-\infty}^{x_1} p(x)dx \qquad P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x)dx$$

Loi uniforme



$$p(x) = \begin{cases} \frac{1}{a-b} & \text{si } x \in [a, b] \ , \\ 0 & \text{ailleurs} \ . \end{cases}$$

Loi normale



$$m = 0, \sigma = 1 \quad m = 2, \sigma = 0.4$$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Moments de v.a.

La connaissance de la d.d.p. est importante, mais n'est pas toujours disponible. On introduit la notion de moment d'une v.a.. Ceux-ci apportent des informations partielles mais intéressantes.

Définition

Le moment $g(x)$ d'une v.a. est donné par l'espérance,

$$E(g(x)) = \int_{-\infty}^{+\infty} g(x)p(x)dx$$

Généralement $g(x) = x^m$, on parle alors de moment d'ordre m,

Moment d'ordre 1	$E(\mathrm{X}) = \int_{-\infty}^{+\infty} x \quad p(x)dx$
Moment d'ordre 2	$E(\mathrm{X}^2) = \int_{-\infty}^{+\infty} x^2 \quad p(x)dx$

Variance

Définition

La variance est l'espérance du carré des écarts par rapport au moment d'ordre 1 $m = E(\mathrm{X})$,

$$\begin{aligned} \nu &= E\left((\mathrm{X}-m)^2\right) = \int_{-\infty}^{+\infty} (x-m)^2 p(x) dx \ , \\ \nu &= E\left(\mathrm{X}^2\right) - E\left(\mathrm{X}\right)^2 \ . \end{aligned}$$

L'écart-type σ est la racine carrée de la variance,

$$\sigma = \sqrt{\nu} \ .$$

Soit x une observation issue d'une loi statistique, loi qui est fonction d'un ou plusieurs paramètres θ.

Définition

La vraisemblance de x est la probabilité d'observer x connaissant *a priori* les paramètres θ,

$$L(x|\theta) = p_\theta(\mathrm{X} = x)$$

Soit $\mathcal{X} = \{x_i | i \in [1..n]\}$ un ensemble de n observations issue de cette même loi.

Définition

La vraisemblance de $\mathcal{X}$ est,

$$L(x_1, \ldots, x_n|\theta) = p_\theta(\mathrm{X} = x_1, \ldots, \mathrm{X} = x_n)$$

Si les observations x_i sont indépendants et identiquement distribués (i.i.d.) entre eux, alors

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p_\theta(\mathrm{X} = x_i)$$

et

$$\log\left(L(x_1, \ldots, x_n|\theta)\right) = \sum_{i=1}^{n} \log\left(p_\theta(\mathrm{X} = x_i)\right)$$

Inférence de paramètres

Soit $\mathcal{X} = \{x_i | i \in [1..n]\}$ un ensemble de n observations issues d'une loi statistique, loi qui est fonction d'un ou plusieurs paramètres θ.

Qu'est-ce que l'inférence ?

On connaît $\mathcal{X}$. On ne connaît pas θ et on cherche à l'estimer.

Comment faire ?

Si on n'a pas de connaissance *a priori* sur les paramètres θ (on ne connaît pas $p(\theta)$), alors on peut donner comme estimation de θ la valeur pour laquelle la vraisemblance (ou la log-vraisemblance) des observations est la plus forte.

$$\hat{\theta}_{ML} = argmax_\theta L(x_1, \ldots, x_n | \theta)$$

Cela s'appelle l'estimation par *Maximum de Vraisemblance* ou *Maximum Likelihood* (ML).

Les paramètres sont

$$\theta = \{\mu, \sigma\} \ .$$

La vraisemblance d'une observation est

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

La vraisemblance d'une série d'observations i.i.d est

$$\begin{aligned} L(\mathcal{X} = \{x_i | i \in [1..n]\}|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ L(\mathcal{X} = \{x_i | i \in [1..n]\}|\mu,\sigma) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \ . \end{aligned}$$

ML pour une loi normale, $\hat{\mu}_{ML}$

La log-vraisemblance d'une série d'observations i.i.d est

$$\log L(\mathcal{X}|\mu,\sigma) \quad = \quad -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \;\; .$$

On cherche le maximum de la log-vraisemblance. On va dériver partiellement par rapport à chacun des paramètres μ et σ.

$$\begin{aligned}
\frac{\partial \log L(\mathcal{X}|\mu,\sigma))}{\partial\mu} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(x_i-\mu) \;\; , \\
\frac{\partial \log L(\mathcal{X}|\mu,\sigma))}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \;\; .
\end{aligned}$$

On annule la dérivée, pour trouver l'estimation $\hat{\mu}$ correspondant au maximum de log-vraisemblance,

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML}) \;\; , \;\; 0 = \left(\sum_{i=1}^{n} x_i\right) - n\hat{\mu}_{ML} \;\; , \;\; \text{soit} \;\; \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \;\; .$$

La meilleure estimation de μ au sens du maximum de vraisemblance est la moyenne des observations !

ML pour une loi normale, $\hat{\sigma}_{ML}$

La log-vraisemblance d'une série d'observations i.i.d est

$$\log L(\mathcal{X}|\mu,\sigma) \quad = \quad -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \ .$$

On cherche le maximum de la log-vraisemblance. On va dériver partiellement par rapport à chacun des paramètres μ et σ.

$$\frac{\partial \log L(\mathcal{X}|\mu,\sigma))}{\partial\sigma} \quad = \quad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \ ,$$

On annule la dérivée, pour trouver l'estimation $\hat{\sigma}$ correspondant au maximum de log-vraisemblance connaissant $\hat{\mu}_{ML}$,

$$0 = -\frac{n}{\hat{\sigma}_{ML}} + \frac{1}{\hat{\sigma}_{ML}^3}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})^2 \ , \ \ 0 = -n\hat{\sigma}_{ML}^2 + \sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})^2 \ , \ \ \hat{\sigma}_{ML}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})^2 \ .$$

La meilleure estimation de σ au sens du maximum de vraisemblance est l'écart-type des observations !

ML pour une loi normale, résumé

L'estimation des paramètres d'une loi normale par le maximum de vraisemblance, à partir des observations $\mathcal{X} = \{x_i | i \in [1..n]\}$, est,

$$\begin{aligned} \hat{\mu}_{ML} &= \frac{1}{n} \sum_{i=1}^{n} x_i \quad , \\ \hat{\sigma}^2_{ML} &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{ML})^2 \quad . \end{aligned}$$

Attention, nous avons pu trouver les paramètres séparément par des dérivées partielles. Ce n'est pas toujours (pas souvent) possible !

Densité de probabilité jointe

Fonction de répartition mutuelle

Soit X et Y deux $\mathtt{v.a.}$alors,

$$F(x, y) = P(\mathrm{X} \leq x, \mathrm{Y} \leq y)$$

Densité de probabilité jointe

Soit X et Y deux $\mathtt{v.a.}$alors,

$$\begin{aligned} p(x, y) &= \frac{\partial^2 F(x, y)}{\partial x \partial y} \\ p(x) &= \int p(x, y) dy \\ p(y) &= \int p(x, y) dx \end{aligned}$$

Pour caractériser l'interdépendance de deux variables, on introduit la notion de covariance.

Définitions

- Moments d'une loi jointe

$$E(g(x,y)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\, p(x,y)dxdy$$

- Corrélation

$$R_{\mathrm{XY}} = E(\mathrm{XY}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, p(x,y)dxdy$$

- Covariance

$$\begin{aligned} C_{\mathrm{XY}} &= \sigma_{\mathrm{XY}} = E\left((\mathrm{X} - m_{\mathrm{X}})(\mathrm{Y} - m_{\mathrm{Y}})\right) \\ C_{\mathrm{XY}} &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x - m_{\mathrm{X}})(y - m_{\mathrm{Y}})\, p(x,y)dxdy \end{aligned}$$

- Coefficient de corrélation

$$r_{\mathrm{XY}} = \frac{C_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}$$

Propriétés

Indépendance

Deux `v.a.`indépendantes sont non-corrélées

$$\begin{aligned} p(x,y) &= p(x)p(y) \\ R_{\mathrm{XY}} &= m_{\mathrm{X}} m_{\mathrm{Y}} \\ C_{\mathrm{XY}} &= 0 \Leftrightarrow r_{\mathrm{XY}} = 0 \end{aligned}$$

Corrélation

Deux variables corrélées sont dépendantes.
Le coefficient de corrélation permet alors de mesurer la dépendance linéaire,

$$E(\mathrm{XY}) = C_{\mathrm{XY}} + m_{\mathrm{X}} m_{\mathrm{Y}} \quad .$$

Proba. conditionnelle : connaissant une des variables, quelle est la loi de probabilité de l'autre variable ?

Règle de Bayes ou inversion du conditionnement

$$
\begin{aligned}
p(x, y) &= p(y|x)p(x) = p(x|y)p(y) \\
p(x|y) &= \frac{p(x, y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \\
p(y|x) &= \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)}
\end{aligned}
$$

Modèles bayésiens

Soit X une variable aléatoire fonction d'un ou plusieurs paramètres θ. On va considérer que θ lui même est issu d'une variable aléatoire dont on connaît la distribution.
On note :

- $p(\theta)$, la distribution *a priori* de θ, le modèle *a priori* ou **Prior model** ;
- $p(x|\theta) = p_\theta(x)$, la distribution des observations x connaissant θ, c'est le modèle de mesure ou **Measurement model**.

Par la règle de Bayes, on peut obtenir $p(\theta|x)$, la distribution *a posteriori* ou **Posterior distribution** de θ connaissant une observation x,

- $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \propto p(x|\theta)p(\theta)$

Si nous disposons de $\mathcal{X} = \{x_i | i \in [1..n]\}$ un ensemble de n observations i.i.d issues de X alors la distribution *a posteriori* est donnée par

- $p(\theta|\mathcal{X}) = \frac{p(\theta)\prod_{i=1}^n p(x_i|\theta)}{\int p(\theta)\prod_{i=1}^n p(x_i|\theta)d\theta} \propto p(\theta)\prod_{i=1}^n p(x_i|\theta)$

Quelle est la distribution d'une nouvelle observation x connaissant un ensemble $\mathcal{X}$ déjà observé ?

- $p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X})d\theta$

C'est la **Predictive posterior distribution**.

Si on dispose de $\mathcal{X} = \{x_i | i \in [1..n]\}$ de n observations i.i.d issues de X paramétrée par un ou plusieurs paramètres θ , comment estimer θ ?

Maximum A Posteriori

Contrairement au maximum de vraisemblance, on considère que l'on dispose d'informations *a priori* sur les paramètres θ, i.e. $p(\theta)$.
On peut donner comme estimation de θ la valeur pour laquelle la distribution *a posteriori* est la plus forte.

$$\hat{\theta}_{MAP} = argmax_\theta p(\theta|\mathcal{X}) = argmax_\theta p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

Cela s'appelle l'estimation par *Maximum A Posteriori* (MAP).

Exemple sur une distribution normale

Soit X une variable aléatoire qui suit une distribution normale $\mathcal{N}(\theta, \sigma_x)$; σ_x est connu, on considère que θ est issu lui-même d'une variable aléatoire qui suit une distribution normale $\mathcal{N}(\mu_\theta, \sigma_\theta)$, avec μ_θ et σ_θ connus.
Alors

$$
\begin{aligned}
p(\theta|x) &\propto p(\theta) \prod_{i=1}^{n} p(x_i|\theta) \ , \\
p(\theta|x) &\propto \frac{1}{\sigma_\theta\sqrt{2\pi}} \exp\left(-\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2}\right) \left(\frac{1}{2\pi\sigma_x^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_x^2}\right) \\
p(\theta|x) &\propto \frac{1}{\sigma_\theta\sqrt{2\pi}} \left(\frac{1}{2\pi\sigma_x^2}\right)^{\frac{n}{2}} \exp\left(-\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2} - \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_x^2}\right) \\
\log p(\theta|x) &\sim -\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2} - \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_x^2} + K
\end{aligned}
$$

Exemple sur une distribution normale

On dérive par rapport à θ et on annule la dérivée pour trouver le maximum.

$$
\begin{aligned}
\frac{\partial \log p(\theta|x)}{\partial \theta} &\propto -\frac{1}{\sigma_\theta^2}(\theta - \mu_\theta) + \frac{1}{\sigma_x^2}\sum_{i=1}^{n}(x_i - \theta) \\
0 &= -\frac{1}{\sigma_\theta^2}(\hat{\theta}_{MAP} - \mu_\theta) + \frac{1}{\sigma_x^2}\sum_{i=1}^{n}(x_i - \hat{\theta}_{MAP}) \\
\left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_x^2}\right)\hat{\theta}_{MAP} &= \frac{1}{\sigma_\theta^2}\mu_\theta + \frac{1}{\sigma_x^2}\sum_{i=1}^{n} x_i \\
\frac{\sigma_x^2 + n\sigma_\theta^2}{\sigma_\theta^2\sigma_x^2}\hat{\theta}_{MAP} &= \frac{1}{\sigma_\theta^2}\mu_\theta + \frac{1}{\sigma_x^2}\sum_{i=1}^{n} x_i \\
\hat{\theta}_{MAP} &= \frac{\sigma_x^2}{\sigma_x^2 + n\sigma_\theta^2}\mu_\theta + \frac{\sigma_\theta^2}{\sigma_x^2 + n\sigma_\theta^2}\sum_{i=1}^{n} x_i \\
\hat{\theta}_{MAP} &= \frac{\sigma_x^2}{\sigma_x^2 + n\sigma_\theta^2}\mu_\theta + \frac{n\sigma_\theta^2}{\sigma_x^2 + n\sigma_\theta^2}\frac{1}{n}\sum_{i=1}^{n} x_i
\end{aligned}
$$

Exemple sur une distribution normale

Soit X une variable aléatoire qui suit une distribution normale $\mathcal{N}(\theta, \sigma_x)$; σ_x est connu, on considère que θ est issu lui-même d'une variable aléatoire qui suit une distribution normale $\mathcal{N}(\mu_\theta, \sigma_\theta)$, avec μ_θ et σ_θ connus.
L'estimation de θ par le maximum *a posteriori* est

$$\hat{\theta}_{MAP} = \frac{\sigma_x^2}{\sigma_x^2 + n\sigma_\theta^2}\mu_\theta + \frac{n\sigma_\theta^2}{\sigma_x^2 + n\sigma_\theta^2}\frac{1}{n}\sum_{i=1}^{n} x_i \quad .$$

Ne pas avoir d'information *a priori* sur θ peut être simulé par une variance σ_θ de la distribution *a priori* $p(\theta)$ tendant vers ∞. Or

$$\lim_{\sigma_\theta \to \infty} \hat{\theta}_{MAP} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad ,$$

on retombe sur l'estimation par le maximum de vraisemblance $\hat{\theta}_{ML}$.