Probabilités pour l'Informatique (PI)

Introduction et Chapitre 1

Pierre Etoré

Ensimag

2020-2021

Pierre Etoré

pierre.etore@univ-grenoble-alpes.fr
http ://www-ljk.imag.fr/membres/Pierre.Etore/
Tel : 04 57 42 17 35

- Chercheur au Laboratoire Jean Kuntzmann, équipe IPS
- Domaine de recherche : étude des processus stochastiques et de leur simulation, lien avec les EDP, statistiques pour les processus stochastiques, ...
- Bureau 142, bâtiment IMAG.
- Enseigne les probas/stats en 1A, en IF 2A et 3A, en MSIAM 2, en 2AA

Objectifs et plan (gros grain) du cours

But du cours : Présenter de nouvelles notions de probabilités (notamment la notion de processus stochastique) et leur lien avec (ou application à) l'informatique.

- Chap. 1 : Rappels de probabilités et premières applications (conditionnement, simulation de variables aléatoires, initiation au machine learning,...)
- Chap. 2 : Chaînes de Markov et leur équilibre (notion de mesure invariante, de convergence en temps long des chaînes de Markov vers une distribution à l'équilibre).
- Chap. 3 : Processus de Poisson, processus de Markov à sauts et files d'attente

Evaluation / Examen

- Examen sur table (si tout va bien...) :
  - Durée 2h
  - Compte pour 2/3 de la note
  - Documents autorisés : slides imprimés, fiches de TD, notes de cours/TD.
- TP
  - Posé dans quelques semaines.
  - A faire en individuel, programmation en R.
  - Compte pour 1/3 de la note.

Chap.1 : Rappels de probabilités et premières applications

But du chapitre :

- Rappeler des éléments de probabilités
- prendre contact (exercices...)
- donner une initiation (légère...) au machine learning.

1.1 Variables aléatoires : rappels

Un espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ est donné.
Soit $(E, \mathcal{E})$ un espace mesurable (i.e. $\mathcal{E}$ est une tribu de parties de E).

Une variable aléatoire (v.a.) c'est une application $X : \Omega \to E$ qui est mesurable (i.e. $\forall B \in \mathcal{E}$ on a
$\{X \in B\} := X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$).

Cette condition est requise car pour tout $B \in \mathcal{E}$ on veut pouvoir calculer la quantité

$$\mathbb{P}(X \in B).$$

Deux cas d'école :

a) Les variables aléatoires "discrètes" : c'est quand E est fini ou dénombrable.
La donnée de $\mathbb{P}(X = k)$, $\forall k \in E$ donne alors la "loi" de X.

Rappel : $\sum_{k\in E} \mathbb{P}(X = k) = 1$ et $\mathbb{E}(X) = \sum_{k\in E} k\mathbb{P}(X = k)$.

Exemple : $X \sim \mathcal{B}(n, p)$ (loi binomiale). On a $E = \{0, \dots, n\}$ et $\forall k \in E$, $\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}$ et $\mathbb{E}(X) = np$.

b) Les variables aléatoires à loi à densité (ou variables aléatoires "à densité") : C'est quand E n'est pas dénombrable (ex : $E = \mathbb{R}$ ou E est une partie de $\mathbb{R}$; imaginons pour fixer les idées que $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et que la loi de X est donnée par la densité de probabilités f_X qui est une fonction qui vérifie

$$\forall B \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}(X \in B) = \int_B f_X(x)dx.$$

NB : Pour tout $x_0 \in \mathbb{R}$ on a $\mathbb{P}(X = x_0) = \int_{\{x_0\}} f_X(x)dx = 0$ (propriété de l'intégrale de Lebesgue).

Rappel : $\int_{\mathbb{R}} f_X(x)dx = 1$ et $E(X) = \int_{\mathbb{R}} x f_X(x)dx$.

Exemple1 : Si $X \sim \mathcal{N}(m, \sigma^2)$ (loi normale) on a

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-m)^2}{2\sigma^2}\Big)$$

et

$$\mathbb{E}(X) = m \quad \text{et} \quad \mathbb{V}\mathrm{ar}(X) = \sigma^2.$$

Exemple2 : Si $X \sim \mathcal{U}([a, b])$ (loi uniforme) on a

$$f_X(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x) \quad \text{et} \quad \mathbb{E}(X) = \frac{a+b}{2}.$$

Noter qu'ici le "support" de la loi de X est le segment $[a, b]$ (la probabilité que X prenne ses valeurs en dehors de ce segment est nulle). On pourrait considérer que $E = [a, b]$ ...

Exemple3 : Si $X \sim \mathcal{E}(\lambda)$ (loi exponentielle) on a $f_X(x) = \lambda e^{-\lambda x}\mathbf{1}_{x>0}$ (loi à support sur $\mathbb{R}_+^*$; en général la loi exponentielle sert à modéliser un temps d'attente...). et $\mathbb{E}(X) = 1/\lambda$.

La fonction de répartition (f.d.r.) de X : c'est la fonction F_X définie par

$$F_X(x) = \mathbb{P}(X \leq x), \quad \forall x \in \mathbb{R}.$$

Elle vérifie :

- F est croissante
- Elle est continue à droite (càd)
- On a $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to\infty} F(x) = 1$.

Pour fixer les idées :

- La f.d.r. d'une v.a. discrète est constante par morceaux (et donc discontinue si non constante).
- La f.d.r. d'une v.a. à densité est continue, car $F(x) = \int_{-\infty}^{x} f_X(z)dz$

La donnée de la f.d.r. est une façon alternative de préciser la loi d'une v.a., en particulier dans le cas continu. En effet on a la relation

$$\frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = f_X(x).$$

1.2 Variables aléatoires bivariées et conditionnement

Dans ce qui suit on considère un couple de variable aléatoires (X, Y), les marginales X et Y étant non nécessairement indépendantes.

On va chercher à donner un sens à $\mathbb{E}[\varphi(Y) \,|\, X]$, l' "espérance de $\varphi(Y)$ sachant X ".

A nouveau il y a deux cas d'école.

a) Cas où X et Y sont deux v.a. discrètes, à valeurs respectivement dans E et E'.
Pour tout $x \in E$ on suppose que $\mathbb{P}(X = x) > 0$. La loi de Y conditionnellement à $\{X = x\}$ est donnée par les quantités

$$\mathbb{P}(Y = y \,|\, X = x) = \frac{\mathbb{P}(Y = y; X = x)}{\mathbb{P}(X = x)}, \quad \forall y \in E'.$$

On peut alors considérer l'espérance de $\varphi(Y)$ sachant $\{X = x\}$:

$$\mathbb{E}[\varphi(Y) \,|\, X = x] = \sum_{y \in E'} \varphi(y) \mathbb{P}(Y = y \,|\, X = x).$$

On définit alors $\Phi : x \mapsto \Phi(x) = \mathbb{E}[\varphi(Y) \,|\, X = x]$ (c'est une fonction de E vers $\mathbb{R}$) puis la *variable aléatoire*

$$\mathbb{E}[\varphi(Y) \,|\, X] := \Phi(X).$$

NB : Il faut bien avoir en tête que l'objet $\mathbb{E}[\varphi(Y) \,|\, X]$ est une v.a., contrairement à $\mathbb{E}[\varphi(Y) \,|\, X = x]$ qui est une quantité déterministe. Nous verrons une propriété fondamentale de $\mathbb{E}[\varphi(Y) \,|\, X]$ dans un instant.

b) Cas bivarié à densité : le couple (X, Y) a une densité de probabilité $f(x, y)$ sur $\mathbb{R}^2$ (i.e. pour toute partie $B = B_1 \times B_2 \subset \mathbb{R}^2$ on a

$$\mathbb{P}[(X, Y) \in B] = \mathbb{P}(X \in B_1; Y \in B_2) = \int\int_B f(x, y)\, dxdy = \int_{B_2} \int_{B_1} f(x, y)\, dxdy.)$$

La loi de Y sachant $\{X = x\}$ est donnée par la densité conditionnelle

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f_X(x)}, \quad \text{avec } f_X(x) = \int_{\mathbb{R}} f(x, y) dy.$$

NB1 : L'objet $f_X(x)$ est la densité de la loi marginale de X : c'est une densité de probabilités sur $\mathbb{R}$.

NB2 : Ici $\mathbb{P}(X = x) = 0$, donc quand on dit "sachant $\{X = x\}$" ce n'est pas sachant l'évènement $\{X = x\}$...

On peut alors considérer l'espérance de $\varphi(Y)$ sachant $\{X = x\}$:

$$\mathbb{E}[\varphi(Y) \,|\, X = x] = \int_{\mathbb{R}} \varphi(y) f_{Y|X=x}(y)\, dy.$$

De façon analogue au cas discret on définit alors
$\Phi : x \mapsto \Phi(x) = \mathbb{E}[\varphi(Y) \,|\, X = x]$ (c'est une fonction de E vers $\mathbb{R}$) puis

$$\mathbb{E}[\varphi(Y) \,|\, X] := \Phi(X).$$

Dans les deux cas d'école on a le résultat suivant.

Théorème (1.2.1)

Pour toute partie $B \subset E$ (t.q. $B \in \mathcal{E}$) on

$$\mathbb{E}\big(\mathbf{1}_{X\in B}\mathbb{E}[\varphi(Y)\,|\,X]\big) = \mathbb{E}\big(\varphi(Y)\mathbf{1}_{X\in B}\big).$$

En particulier en prenant $B = E$ on a

$$\mathbb{E}\big(\mathbb{E}[\varphi(Y)\,|\,X]\big) = \mathbb{E}\big(\varphi(Y)\big) \qquad (1.2.1).$$

Preuve : [au tableau ; au moins dans le cas discret]

Dans la suite du chapitre il nous arrivera de sortir de ces deux cas d'école et de rencontrer la situation mixte où Y est une v.a. discrète et X est une v.a. à densité. La notion d'espérance conditionnelle existe toujours. En particulier on a encore la relation (1.2.1).

1.3 Introduction au machine learning : classification dure et probabiliste

Contexte/objectif : on a des observations $\{(x_k, y_k)\}_{k=1}^N$, où $x_k \in \mathbb{R}^n$, $y_k \in \{0, 1\}$, pour tout k.
Par exemple les x_k's peuvent être des images et les y_k's des étiquettes qui nous disent si x_k est une "image de chat" ($y_k = 0$) ou une "image de chien" ($y_k = 1$)).

Puis on va chercher à construire un "prédicteur" (ou "classifieur") qui, à une image $x \in \mathbb{R}^n$ qui ne fait pas partie a priori de nos observations, associe une quantité qui va nous permettre de décider si c'est une image de chat ou de chien.

Modélisation : Considérons le couple (X, Y) où :

- X a une loi à densité sur $\mathbb{R}^n$.
- Y est à valeurs dans $\{0, 1\}$.
- On ne suppose pas en général l'indépendance de X et Y (bien au contraire : dans les applications on aura accès à $f_{X|Y=0}(x)$ et $f_{X|Y=1}(x)$, qui seront différentes).

On distingue deux types de classifieurs :

a) Un classifieur "dur" c'est une application $c : \mathbb{R}^n \to \{0, 1\}$.

Pour $x \in \mathbb{R}^n$ il renvoie $c(x) \in \{0, 1\}$ qui s'interprète comme une réponse univoque (par exemple "x est une image de chien", si $c(x) = 1$).

Attention : Il peut arriver que le classifieur se trompe ! Si on introduit la fonction de perte

$$L(c(x), y) = \begin{cases} 1 & \text{si } c(x) \neq y \\ 0 & \text{sinon} \end{cases}$$

(qu'on peut évaluer pour (x, y) qui ne fait pas partie de nos observations), cela va être intéressant de trouver un classifieur c qui minimise cette fonction de perte en un certain sens (cf exercice 1, fiche 2).

b) Un classifieur "probabiliste" c'est une application $q : \mathbb{R}^n \to [0, 1]^2$.

Pour $x \in \mathbb{R}^n$ il renvoie $(q_0(x), q_1(x)) \in [0, 1]^2$; par exemple $q_0(x)$ va être proche de 1 si le classifieur estime qu'il y a de grandes chances que la classe y correspondant à x est 0 ...

Exemple de classifieur probabiliste : le classifieur q défini par

$$q_0(x) = \mathbb{P}(Y = 0|x) \quad \text{et} \quad q_1(x) = \mathbb{P}(Y = 1|x).$$

NB : la notation $\mathbb{P}(Y = 0|x)$ est une notation "spécial machine learning" pour $\mathbb{P}(Y = 0 \,|\, X = x)$...

Lien classifieur dur / probabiliste : En pratique on utilise un classifieur probabiliste en l'associant à une règle de décision ... ce qui le transforme en classifieur dur !

Dans l'exercice 1 de la fiche 2 on considèrera le classifieur

$$\begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|x) > \mathbb{P}(Y = 0|x) \\ 0 & \text{sinon,} \end{cases}$$

qui s'avèrera optimal en un certain sens parmi les classifieurs durs.

☛ Place donc à l'exercice 1, fiche 2.

Il y a donc un intérêt à accéder aux quantités $\mathbb{P}(Y = i|x)$, $i = 0, 1$.

En pratique on a une connaissance plus ou moins complète de $f_{X|Y=i}(x)$

- On met souvent un a priori sur la forme de la loi de X sachant $Y = i$, par exemple $X|Y = i \sim \mathcal{N}(m_i, \sigma_i^2)$, $i = 0, 1$,
  i.e. $f_{X|Y=i}(x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-m_i)^2}{2\sigma_i^2}\right)$, $i = 0, 1$.
- Puis on se sert des observations $\{(x_k, y_k)\}_{k=1}^N$ pour trouver les paramètres de ces lois.
  Comment ? Cf question 6), exercice 2 fiche 2.
- On a aussi souvent accès à la loi de Y.

Pour avoir accès aux $\mathbb{P}(Y = i|x)$'s il faut donc en quelque sorte "retourner" le conditionnement.
Nous allons donc utiliser la formule de Bayes.

Rappel : pour A un évènement et $(B_i)_i$ une partition de Ω on a

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j)}, \quad \forall i.$$

... ce qui nous dit que si on connait les $\mathbb{P}(A|B_i)$'s et les $\mathbb{P}(B_j)$'s on peut en déduire les $\mathbb{P}(B_i|A)$'s.

Par analogie on va considérer que

$$\mathbb{P}(Y = i|x) = \frac{f_{X|Y=i}(x)\mathbb{P}(Y = i)}{\sum_{j=0,1} f_{X|Y=j}(x)\mathbb{P}(Y = j)}.$$

Autre rappel pour traiter l'exercice 2 fiche 2 : la loi des grands nombre. Pour (Z_i)'s suite de v.a. i.i.d de loi celle de Z, avec $\mathbb{E}|Z| < \infty$ on a

$$\frac{1}{N}\sum_{k=1}^{N} Z_i \xrightarrow{N\to\infty} \mathbb{E}(Z).$$

☛ Place donc à l'exercice 2, fiche 2, où on va

- Etablir l'expression théorique d'un classifieur optimal.
- Etablir son expression numérique, en estimant les paramètres dont il dépend grâce aux observations $\{(x_k, y_k)\}_{k=1}^{N}$ (la "base d'apprentissage").
- Tester ses performances en utilisant une "base de test".