Feuille d'exercices

1 S'habituer au formalisme

1.1 Lois à densités – manipulation

On a aborde quelques exercices simples de manipulation des lois à densités. Ces lois nécessitent une certaine aisance dans l'utilisation des intégrales généralisées.

Loi exponentielle et temps d'arrêt. La date de connexion d'un client à votre après 00 : 00 est une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, de loi exponentielle de paramètre λ.

Question 1-1.

1. Rappeler et calculer les moments d'ordre 1 et 2 de d'une loi exponentielle de paramètre λ. En déduire la variance d'une telle loi.
2. Calculer la probabilité $\mathbb{P}(T > 1/\lambda)$.
3. On fixe $\epsilon > 0$. On considère pour $k \in \mathbb{N}$ les plages horaires $I_k = [k\epsilon, (k+1)\epsilon[$. Calculer la probabilité $\mathbb{P}(T \in I_k)$.
4. Soit X la variable aléatoire à valeurs dans $\mathbb{N}$ définie pour tout $\omega in \Omega$ par

$$X(\omega) = k \Leftrightarrow T(\omega) \in I_k.$$

   Quelle est la loi de X ?
5. Calculer pour tout $t > 0$ et tout $h > 0$, la probabilité $\mathbb{P}(t < T)$ ainsi que la probabilité conditionnelle $\mathbb{P}(T > t + h \mid T > t)$. Qu'est-ce que cela signifie ?

Somme de densités. On considère deux variables aléatoires X définies sur Ω, on construit à partir de celles-ci la variable aléatoire sur $\Omega \times \{0, 1\}$ définie par :

$$X(\omega, 0) = X_1(\omega) \quad \text{et} \quad X(\omega, 1) = X_2(\omega)$$

Question 1-2. Exprimer la loi de probabilités de X en fonction de X_1 et X_2. Si X_1 et X_2 sont des lois à densité montrer qu'il en va de même de X et calculer sa densité.

Transfert. On considère la variable aléatoire X à valeurs dans $]0, 1[$ et dont la densité est donnée pour $x \in \mathbb{R}$ par

$$f_X(x) = \frac{1}{\ln(2)} \times \frac{1}{1+x} \chi_{]0,1[}(x).$$

Question 1-3.

1. Déterminer la loi de $Y = 1/X$.
2. On note $E(Y)$ la partie entière de Y, déterminer la loi de $Z = Y - E(Y)$.

1.2 Lois à densités – conditionnement

2 De la probabilité en ML

La ML construit des modèles dont l'objectif est un parmi :

- prédire une caractéristique d'intérêt d'un individu à partir de caractéristiques connues de celui-ci
- identifier des ***patterns*** dans un ensemble de données à l'étude, sans que cela soit conduit par un objectif particulier.

Les situations qu'on décrit par la suite apparaissent dans le premier contexte. Dans ce cadre on suppose qu'on a des données qui contiennent les caractéristiques d'un nombre d'individus ainsi que ***la*** caractéristique qu'on souhaite prédire.

2.1 *Score* d'un classificateur

On chercher dans cette section à quantifier la *qualité* de modèles de ML qu'on appelle les classificateur. Un classificateur est un modèle apparaît dans la situation où la caractéristique qu'on cherche à prédire est discrète ; par exemple des types de plantes, des couleurs de cheveux, des appréciations de goûts etc. C'est une fonction qui étant donné un certaint nombre de caractéristique en entrée renvoie une valeurs discrète, souvent codées entre 0 et le nombre de classes à prédire moins un.

Dans la formalisation qu'on propose ici on considère que les caractéristique en entrée constitue un espace d'états Ω. On considère de plus que celui-ci vient avec une fonction λ qu'on désigne par *label* et qui nous indique la caractéristique qu'on cherche à prédire. On a supposé que le dataset en entrée contient cette information. Dans ce contexte un classificateur est une variable aléatoire $X : \Omega \to F$ où F est l'espace des labels qu'on cherche à attribuer à nos entrées. L'espace Ω est supposé fini probabilisé muni d'une probabilité $\mathbb{P}$. C'est cette probabilité qui fait office de *proportion*.

On suppose que X est donnée et on souhaite étudier trois quantités qui sont liées à l'évaluation de la qualité d'un classificateur.

Cas binaire. On se limite en un premier temps au cas d'un classificateur binaire, c'est-à-dire que $F = \{0, 1\}$.

Question 2-4. On appelle ***précision totale*** d'un classificateur binaire X la proportion des individus bien classés, c'est-à-dire ceux où les réponses de X et λ concident.

1. Exprimer la précision totale de X en s'aidant de la variable aléatoire $X - \lambda$.
2. Déterminer la loi de $X - \lambda$ et interpréter ses deux autres valeurs.

Dans le cas d'un classificateur binaire deux autres quantités apparaissent relativement naturellement à l'évaluation :

— la ***précision*** : proportion des vrais positifs par rapport à l'ensemble des bonnes classifications.
— le ***rappel*** : proportion des vrais positifs par rapport à l'ensemble des positifs (donnés par λ).

Question 2-5.

1. Exprmier précision et rappel de X.
2. Étudier l'interconnexion entre précision et rappel ; qu'arrive-t-il à l'une si l'autre augmente ?

Dans les problématiques de classification on est toujours attentifs au fait d'être confrontés à des modèles dont la dépendance au label λ est très faible. C'est pour cette raison qu'on va souvent chercher à évaluer le ***score*** d'un classificateur X tels que X et λ définissent des variables aléatoires indépendantes.

Question 2-6.

1. Simplifier les expressions de la précision totale, précision et rappel dans le cas où X et λ sont indépendants.
2. On suppose que X et λ sont indépendants et suivent des loi de Bernoulli respectivement de paramètres p et q. Exprimer les différents scores de X dans ce cas.
3. Quels resultats numériques obtenez-vous pour chacun des scores quand $p = q = 0.9$? Qu'en déduisez-vous ?

Cas général. On revient brièvement vers le cas général. On suppose désormais que F est l'ensemble discret $\{0, \ldots, k-1\}$.

Question 2-7.

1. Exprimer la précision totale de X.
2. Étudiez la loi de $X - \lambda$. Est-elle aussi facilement interprétable que dans le cas binaire ?
3. Chercher une variable aléatoire plus adaptée à l'étude des problématiques de classifications qui ne sont pas binaires.