CHAPITRE 9

Introduction à la théorie de l'information

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Automne 2019

La théorie de l'information est utile en informatique notamment pour la compression, le stockage et la transmission des données. Bien qu'à première vue ça puisse vous sembler déconnecté de la théorie des probabilités, la théorie de l'information est basée sur des concepts probabilistes. Nous verrons qu'un encodage efficace attribue des mots-codes petits aux données communes et des mots-codes plus longs aux données rares. Nous verrons également une mesure permettant de comparer la similarité de deux lois de probabilité. Cette mesure est très utile en apprentissage machine, par exemple pour approximer une loi de probabilité complexe par une autre plus simple ou pour construire un arbre de classification. Ce chapitre n'est qu'une très courte introductions aux concepts essentiels de la théorie de l'information. Nous illustrons la théorie qu'avec une variable aléatoire Y catégorielle. La théorie peut bien sûr être étendue ux variables aléatoire continues. Pour en savoir plus, vous pourriez consulter le livre de MacKay.

À la fin de ce chapitre, vous devriez être en mesure de

- Relier les concepts de surprise et d'entropie à l'information et à l'incertitude concernant la réalisation d'une variable aléatoire.
- Comparer les efficacités de plusieurs encodages entre eux et par rapport au code optimal.
- Calculer la divergence de Kullback-Leibler.
- Calculer et interpréter l'information mutuelle.

9.1 Surprise !

La notion fondamentale de la théorie de l'information repose sur la notion de surprise d'un événement. On utilise parfois le terme *information propre* mais je préfère le terme *surprise* car il est intuitif. La surprise S que l'on peut associer à un événement dépend de sa probabilité d'occurrence p, donc la surprise $S(p)$ est une fonction de p. Apprendre qu'un événement inattendu s'est réalisé provoque un grand niveau de surprise tandis qu'apprendre la réalisation d'un événement anticipé ne provoque que très peu de surprise.

Dans cette section, nous allons quantifier mathématiquement le concept de surprise S d'un événement en fonction de sa probabilité d'occurrence $0 \leq p \leq 1$. En se basant sur une suite d'axiomes que le concept de surprise devrait raisonnablement satisfaire, nous serons en mesure d'établir la forme fonctionnelle de la fonction $S(p)$ permettant de quantifier le niveau de surprise.

Le premier axiome consiste à supposer que la réalisation d'un événement certain n'entraîne aucune surprise :

Axiome 1 : $S(1) = 0$.

Le deuxième axiome stipule que plus un événement est rare, plus la surprise qu'il provoque est grande. Formellement, l'axiome s'énonce ainsi :

Axiome 2 : $S(p)$ est une fonction strictement décroissante. Autrement dit, si $p < q$ alors $S(p) > S(q)$.

Le troisième axiome stipule qu'un petit changement de probabilité d'occurrence d'un événement devrait produire un petit changement de surprise. Mathématiquement, on impose alors que la fonction $S(p)$ soit continue :

Axiome 3 : $S(p)$ est une fonction continue.

Le quatrième et dernier axiome est un peu plus subtil. Pour l'introduire, considérons les événements indépendants A et B. La probabilité que A se réalise est $0 \leq p \leq 1$ et la probabilité que B se réalise est $0 \leq q \leq 1$. Puisque les événements sont indépendants, la probabilité que A et B se réalisent est $\mathbb{P}(A \cap B) = pq$. La surprise associée à la réalisation de l'événement $(A \cap B)$ correspond à $S(pq)$. Considérons le scénario suivant. Supposons que nous apprenons d'abord que l'événement A s'est réalisé. Cet événement provoque la surprise $S(p)$. Supposons que l'on apprenne ensuite que B s'est réalisé, ce qui provoque la surprise additionnelle $S(q)$. Puisque les événements sont indépendants, la surprise totale provoquée par les deux événements $S(p) + S(q)$ doit être égale à la surprise $S(pq)$ de l'événement $(A \cap B)$. De façon formelle, l'axiome s'exprime de la façon suivante :

Axiome 4 : Si $0 < p \leq 1$ et $0 < q \leq 1$, alors $S(pq) = S(p) + S(q)$.

Les quatre axiomes précédents permettent de déterminer la forme fonctionnelle de $S(p)$. Le résultat est énoncé dans le théorème suivant.

Théorème. *Si la fonction $S(\cdot)$ satisfait les axiomes 1 à 4, alors*

$$S(p) = -C \log_2(p),$$

où C est une constante arbitraire positive.

La fonction logarithme est en effet la seule fonction permettant de satisfaire les axiomes 1 à 4. La base du logarithme n'est cependant pas unique. En effet, la constante C permet de changer la base. Si on prend $C = \ln(2)$, alors $S(p) = -\ln(p)$ qui est une forme fonctionnelle qui satisfait les axiomes 1 à 4. C'est aussi le cas si on prend $C = \log_{10}(2)$, alors $S(p) = -\log_{10}(p)$. Lorsque $C = 1$, on a que $S(p) = -\log_2(p)$. Dans ce dernier cas, on dit que la surprise est exprimée en *bits*. C'est l'unité que nous retiendrons pour le reste du chapitre.

9.2 Entropie

Soit la variable aléatoire Y catégorielle pouvant prendre les valeurs dans l'ensemble $\{1, \ldots, m\}$ définie de la façon suivante :

$$Y = \begin{cases} 1 \text{ avec probabilité } p_1; \\ \vdots \\ m \text{ avec probabilité } p_m. \end{cases}$$

La réalisation j pour $\{1, \ldots, m\}$ de la variable aléatoire Y provoque la surprise $S(p_j) = -\log_2(p_j)$. L'espérance de la surprise provoquée par la réalisation de la variable aléatoire Y, dénotée par $H(Y)$, est calculée de la façon suivante :

$$\begin{aligned} H(Y) &= \sum_{j=1}^{m} S(p_j)\, p_j \\ &= -\sum_{j=1}^{m} \log_2(p_j)\, p_j. \end{aligned}$$

En théorie de l'information, $H(Y)$ est appelé l'**entropie de la variable aléatoire** Y. Plus l'entropie est grande, plus l'incertitude est grande sur les valeurs possibles que Y peut prendre. Il est donc difficile de prédire la réalisation future de Y si son entropie est grande. Au contraire lorsque l'entropie est faible, l'incertitude sur les valeurs que Y peut prendre est petite. Autrement dit, on pourra prédire plus aisément les réalisations futures que Y prendra.

Remarque. *Par continuité mathématique, on a que*

$$\lim_{p \to 0} p \, \log_2(p) = 0.$$

Exemple 1

Soit les variables aléatoires X et Y suivantes :

$$X = \begin{cases} 1 & \text{avec probabilité } 1/4; \\ 2 & \text{avec probabilité } 1/4; \\ 3 & \text{avec probabilité } 1/4; \\ 4 & \text{avec probabilité } 1/4; \end{cases}$$

et

$$Y = \begin{cases} 1 & \text{avec probabilité } 1/2; \\ 2 & \text{avec probabilité } 1/4; \\ 3 & \text{avec probabilité } 1/8; \\ 4 & \text{avec probabilité } 1/8. \end{cases}$$

On a que l'entropie de X est égale à

$$H(X) = 2$$

et l'entropie de Y est égale à

$$H(Y) = 1.75.$$

L'entropie de la variable aléatoire Y est plus petite que celle de la variable aléatoire X. Cela indique qu'on a plus d'incertitude sur les réalisations futures de X que sur les réalisations futures de Y. Autrement dit, il est plus facile de prédire le résultat de Y que de X. Est-ce bien le cas ?

9.3 Entropie conjointe

Soit les variables aléatoires discrètes X et Y possédant les fonctions de masse marginales $p_X(x)$ et $p_Y(y)$ et la fonction de masse conjointe $p_{(X,Y)}(x, y)$. Soit $x_1, \ldots, x_n$ les résultats possibles de X et $y_1, \ldots, y_m$ les résultats possibles de Y. L'**entropie conjointe** du vecteur aléatoire (X, Y) est définie par l'équation suivante :

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \log_2 p(x_i, y_j)\, p_{(X,Y)}(x_i, y_j)$$

L'interprétation de l'entropie demeure la même dans le cas d'un vecteur aléatoire. Elle représente l'espérance de la surprise provoquée par la réalisation du vecteur aléatoire (X, Y). On peut montrer (très difficile) que l'entropie conjointe est plus petite ou égale à la somme des entropies marginales :

$$H(X, Y) \leq H(X) + H(Y).$$

L'égalité survient lorsque les variables aléatoires X et Y sont indépendantes.

Exemple 2

On lance une pièce de monnaie bien balancée 10 fois de façon indépendante. Soit Y_i la variable aléatoire suivante :

$$Y_i = \begin{cases} 0 & \text{si le résultat du lancer } i \text{ est face}\,; \\ 1 & \text{si le résultat du lancer } i \text{ est pile.} \end{cases}$$

Puisque les lancers sont indépendants, l'entropie des 10 lancers est égale à la somme de l'entropie des 10 lancers. On a que

$$H(Y_1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1.$$

Alors,

$$H(\boldsymbol{Y}) = 10 \times H(Y_1) = 10.$$

9.4 Encodage et entropie

Supposons que nous voulions transmettre une ou plusieurs observations de Y par un canal de transmission où la source ne génère que des «0» et des «1». Chaque résultat possible de Y doit donc être encodé par une suite binaire. La performance du code optimal, mesurée en nombre de bits moyens requis pour transmettre le résultat de Y, est reliée à l'entropie de Y.

Définition. *Un encodage est dit* ***déchiffrable*** *si pour chaque message codé, il existe au plus un message source. Autrement dit, la fonction $C(y)$ doit être injective pour que le code soit déchiffrable.*

Il existe plusieurs façon d'assurer la déchiffrabilité d'un code. Voici quelques exemples classiques :

Longueur fixe des mots codés. La façon la plus évidente d'assurer la déchiffrabilité consiste à attribuer la même longueur de caractères à tous les mots-codes. Les encodages ASCII et UTF utilisent cette procédure.

Séparateur. Une autre façon consiste à utiliser un caractère spécial pour séparer les mots-codes. L'encodage CSV en est un exemple.

Condition du préfixe. La condition du préfixe consiste à imposer aux mots d'un code la propriété suivante : aucun mot-code doit être le préfixe d'un autre mot-code. C'est le cas de l'encodage $\mathcal{C}_2$ présenté à l'exemple 3. Un code qui satisfait la condition du préfixe est dit **instantané** puisqu'il est possible d'effectuer le décodage pas à pas aussitôt qu'un mot-code est reconnu.

Exemple 3

Supposons les encodages suivants pour la variable aléatoire X définie à l'exemple 1.
Soit l'*encodage binaire instantané* $\mathcal{C}_1$ suivant :

$$(X = 1) \leftrightarrow 00$$
$$(X = 2) \leftrightarrow 01$$
$$(X = 3) \leftrightarrow 10$$
$$(X = 4) \leftrightarrow 11$$

Cet encodage est aussi un code où la longueur des mots-codes est fixe.
Un autre *encodage binaire instantané* est le suivant, dénoté par $\mathcal{C}_2$:

$$(X = 1) \leftrightarrow 0$$
$$(X = 2) \leftrightarrow 10$$
$$(X = 3) \leftrightarrow 110$$
$$(X = 4) \leftrightarrow 111$$

L'encodage $\mathcal{C}_3$ suivant ne respecte pas la condition du préfixe :

$$(X = 1) \leftrightarrow 0$$
$$(X = 2) \leftrightarrow 1$$
$$(X = 3) \leftrightarrow 00$$
$$(X = 4) \leftrightarrow 01.$$

Un encodage efficace pour transmettre les réalisations de la variable aléatoire X minimise le nombre moyens de bits à transmettre. L'idée générale consiste à encoder avec un code instantané la variable aléatoire X en utilisant des mots-codes courts pour les réalisations les plus probables.

Exemple 4

Soit la variable X définie précédemment. La quantité espéré de bits à transmettre avec l'encodage $\mathcal{C}_1$ est égale à

$$L_{\mathcal{C}_1}(X) = 2 \times \frac{1}{4} + 2 \times \frac{1}{4} + 2 \times \frac{1}{4} + 2 \times \frac{1}{4} = 2.$$

Avec l'encodage $\mathcal{C}_2$, la quantité espéré de bits à transmettre est égale à

$$L_{\mathcal{C}_2}(X) = 1 \times \frac{1}{4} + 2 \times \frac{1}{4} + 3 \times \frac{1}{4} + 3 \times \frac{1}{4} = 2.25.$$

L'encodage $\mathcal{C}_2$ est donc moins performant pour la variable aléatoire X que l'encodage $\mathcal{C}_1$.

Exemple 5

La performance d'un encodage dépend de la variable aléatoire à transmettre. Par exemple pour la variable aléatoire Y définie précédemment, la quantité espéré de bits à transmettre est égale à

$$L_{\mathcal{C}_1}(Y) = 2 \times \frac{1}{2} + 2 \times \frac{1}{4} + 2 \times \frac{1}{8} + 2 \times \frac{1}{8} = 2.$$

Avec l'encodage $\mathcal{C}_2$, la quantité espéré de bits à transmettre est égale à

$$L_{\mathcal{C}_2}(Y) = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = 1.75.$$

L'encodage $\mathcal{C}_2$ est donc plus performant pour la variable aléatoire Y que l'encodage $\mathcal{C}_1$.

L'exemple précédent soulève la question suivante : existe-t-il un code source optimal pour une variable aléatoire Y ? Un résultat fondamental en théorie de l'information stipule que le nombre moyen de bits requis pour transmette le résultat de Y ne peut pas être plus petit que l'entropie de Y. Il est énoncé dans le théorème suivant :

Théorème (Premier théorème de Shannon ou Théorème du codage de source non bruitée.)**.** *Soit la variable aléatoire Y prenant des valeurs possibles dans l'ensemble $\{1, 2, \ldots, m\}$ avec les probabilités $\{p_1, p_2, \ldots, p_m\}$. Alors pour chaque code source $\mathcal{C}$ qui attribue ℓ_i bits au résultat $(Y = i)$, on a que*

$$L = \sum_{i=1}^{m} \ell_i \, p(y_i) \geq H(Y).$$

Autrement dit, le nombre moyen de bits requis pour transmettre une réalisation de Y, dénoté L, est supérieur ou égal à l'entropie de la source :

$$L \geq H(Y).$$

L'encodage binaire instantané optimal pour la variable aléatoire Y possède une longueur moyenne de bits correspondant à l'entropie. Dans la plupart des cas, il n'existe pas de code source pour lequel le nombre moyen de bits requis atteint cette borne inférieure. Toutefois, il est toujours possible développer un code dont le nombre moyen de bits requis se situe entre $H(Y)$ et $1 + H(Y)$:

Lemme. *Pour la variable Y d'entropie $H(Y)$, il existe au moins un code tel que le nombre moyen de bits requis L pour transmettre un résultat de Y est contenu dans l'intervalle $[H(Y), 1 + H(Y)]$.*

Exemple 6

Selon le premier théorème de Shannon, l'encodage $\mathcal{C}_1$ est optimal pour la variable aléatoire X puisque la moyenne du nombre de bits requis est égal à l'entropie de X. Pour la même raison, l'encodage $\mathcal{C}_2$ est optimal pour la variable aléatoire Y.

9.5 Divergence de Kullback-Leibler

La divergence de Kullback-Leibler mesure la proximité entre deux distributions. Soit $p(y)$ et $q(y)$ deux fonctions de masse définies sur le même support, c'est-à-dire sur le même ensemble de valeurs possibles. La divergence de Kullback-Leibler entre les fonction de masse $p(y)$ et $q(y)$, dénotée $D(p||q)$, est définie par l'équation suivante :

$$D(p||q) = \sum_{\{y:p(y)>0\}} p(y) \log_2 \frac{p(y)}{q(y)}.$$

On peut montrer que

$$D(p||q) \geq 0,$$

où l'égalité survient lorsque les fonctions de masse $p(y)$ et $q(y)$ sont identiques. La divergence de Kullback-Leibler n'est pas symétrique. Autrement dit, on peut avoir que

$$D(p||q) \neq D(q||p).$$

La divergence de Kulback-Leibler peut être interprétée comme la différence moyenne du nombre de bits requis par rapport à l'entropie de Y pour encoder un échantillon généré de la distribution $p(y)$ avec un code optimisé pour une source de distribution $q(y)$.

Exemple 7

La divergence Kullback-Leibler entre la distribution de X et Y est la suivante :

$$\begin{aligned} D\{p_Y(y)||p_X(x)\} &= \frac{1}{2}\log_2\frac{1/2}{1/4} + \frac{1}{4}\log_2\frac{1/4}{1/4} + \frac{1}{8}\log_2\frac{1/8}{1/4} + \frac{1}{8}\log_2\frac{1/8}{1/4} \\ &= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 1 + \frac{1}{8}\log_2\frac{1}{2} + \frac{1}{8}\log_2\frac{1}{2} \\ &= \frac{1}{4}. \end{aligned}$$

L'entropie de Y est égale à 1.75 bits. Si on utilise le code $\mathcal{C}_1$ optimisé pour X, le nombre de bits additionnels requis par rapport à l'entropie et de 0.25. Le nombre moyen de bits nécessaires est donc égal à $1.75 + 0.75 = 2$, ce qui est égal à $L_{\mathcal{C}_1}(Y)$. Ce résultat est cohérent avec l'exemple 4.

La divergence de Kullback-Leibler peut être utilisée comme fonction objectif à minimiser pour identifier la meilleure distribution pour la variable aléatoire Y. Cette procédure peut remplacer la méthode du maximum de la vraisemblance.

9.6 Information mutuelle

L'information mutuelle mesure la dépendance entre les variables aléatoires X et Y. L'information mutuelle $I(X,Y)$ entre X et Y est définie comme la divergence de Kullback-Leibler entre la loi conjointe $p_{(X,Y)}(x,y)$ du couple (X,Y) et le produit des lois marginales $p_X(x)$ et $p_Y(y)$:

$$I(X,Y) = D\left\{p_{(X,Y)}(x,y) \,||\, p_X(x)\, p_Y(y)\right\}$$

Si les variables aléatoires X et Y sont indépendantes, alors la densité conjointe du couple (X,Y) est égale au produit des fonctions de masse marginales et l'information mutuelle est nulle. Contrairement au coefficient de corrélation linéaire de Pearson ρ, l'information mutuelle ne se limite pas à mesure la dépendance linéaire entre les variables aléatoire Xet Y.

Exemple 8

Revenons au cas du TD8 sur la classification des messages électroniques d'un des employés de la compagnie Enron. Soit les variables suivantes :

$$X_i = \begin{cases} 0 & \text{si le message } i \text{ ne contient pas le mot } \textit{enron}\,; \\ 1 & \text{si le message } i \text{ contient le mot } \textit{enron}\,; \end{cases}$$

et

$$Y_i = \begin{cases} 0 & \text{si le message } i \text{ est un pourriel}\,; \\ 1 & \text{si le message } i \text{ est un courriel.} \end{cases}$$

On obtient le tableau de contingence suivant pour les variables X et Y des 3448 messages de l'ensemble d'entraînement :

	$Y=0$	$Y=1$
$X=0$	1000	1449
$X=1$	0	999

L'information mutuelle est égale à 0.176 qui n'est pas nulle suggérant une dépendance entre X et Y.

Exemple 9

Reprenons l'exemple précédent avec le mot *fine* plutôt que *enron*. On obtient alors le tableau de contingence suivant pour les variables X et Y des 3448 messages de l'ensemble d'entraînement :

	$Y=0$	$Y=1$
$X=0$	992	2427
$X=1$	8	21

L'information mutuelle est égale à 6.01×10^{-6}. Cette valeur très près de 0 suggère que les variables X et Y sont

indépendantes.
L'information mutuelle peut être utilisée pour identifier les mots les plus discriminant. Il suffit d'identifier les mots pour lesquelles l'information mutuelle est la plus élevée.

9.7 Application aux arbres de classification

Les arbres de classification constituent un cas particulier des arbres de décision, appelés *CART* en anglais pour *classification and regression trees*. Soit un échantillon aléatoire constitué de n observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, où pour l'observation Y_i, le vecteur des p variables explicatives suivantes $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ sont recueillies. Les arbres de régression sont très utile pour **prédire** la valeur non-observée de $\tilde{Y}$ correspondant à une valeur connue du vecteur de variables explicatives $\boldsymbol{X} = \tilde{\boldsymbol{x}}$.

Les arbres de classification utilisent le partitionnement récursif de l'espace des variables explicatives en plusieurs rectangles pour prédire $\tilde{Y}$. L'idée plutôt simple et intuitive consiste à partitionner l'espace des variables explicatives afin d'obtenir des catégories homogènes pour les observations $\boldsymbol{Y}$. Une fois ce partitionnement obtenu, on trouve la région dans laquelle se trouve le vecteur $\tilde{\boldsymbol{x}}$ puis on attribue à $\tilde{Y}$ la classe la plus nombreuse dans cette région.

Pour les arbre de classification, l'homogénéité du partitionnement peut être mesurée par l'entropie des régions. Posons $\hat{p}_{jk}$, la proportion de la classe k dans la région $\mathcal{R}_j$:

$$\hat{p}_{jk} = \frac{1}{n_j} \sum_{\{\boldsymbol{x}_i \in \mathcal{R}_j\}} \mathbf{1}_{\{k\}}(y_i);$$

où n_j est le nombre d'observations dans la région $\mathcal{R}_j$. On cherche donc le partitionnement qui minimise l'entropie observée e_j pour chacune des région $\mathcal{R}_j$:

$$e_j = -\sum_{k=1}^{m} \hat{p}_{jk} \log_2 \hat{p}_{jk}.$$

Le nombre de régions est un paramètre à ajuster avec ce modèle. Un nombre trop grand de régions peut résulter en sur-apprentissage tandis qu'un nombre trop petit de régions risque de négliger des caractéristiques importantes du jeu de données. Une stratégie souvent utilisée consiste à construire le plus grand arbre possible puis ensuite à le tailler en fonction des résultats de la validation croisée.

Les arbres de classification ne suppose pas de modèle statistique comme c'était le cas notamment pour la régression. On dit alors que c'est une approche non-paramétrique. Les approches non-paramétriques sont généralement plus coûteuses en temps de calcul que les approches paramétriques. C'est le prix à payer pour s'affranchir des hypothèses statistiques.

Le défaut majeur des arbre de classification concerne l'instabilité partitionnement de l'espace des variables explicatives. Souvent un petit changement dans les données résulte en un arbre complètement différent. L'utilisation de forêts aléatoires permet de prendre en compte cette instabilité dans les prédictions. Nous en parlerons en classe si le temps le permet.

9.8 Exercices

1. On lance 10 fois une pièce de monnaie possédant une probabilité de pile égale à 1/4 de façon indépendante. Posons Y_i le résultat du lancer i et $\boldsymbol{Y} = (Y_1, \ldots, Y_{10})$ le vecteur aléatoire des 10 lancers.
   (a) Quelle est l'entropie du premier lancer $H(Y_1)$?
   (b) Quelle est l'entropie de l'expérience aléatoire $H(\boldsymbol{Y})$?
   (c) Existe-t-il un encodage plus efficace que le suivant :

   $$(Y_i = 0) \leftrightarrow 0 \quad \text{et} \quad (Y_i = 1) \leftrightarrow 1$$

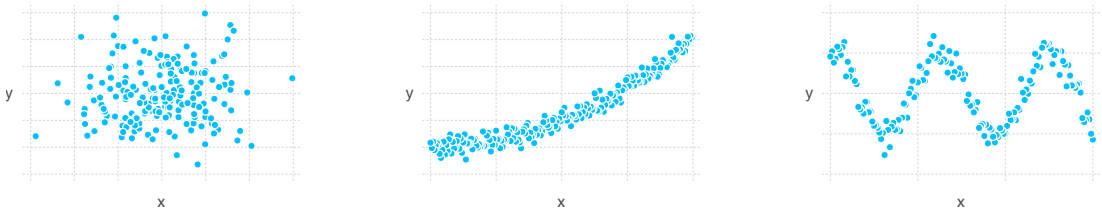
   pour encoder les 10 réalisations du jet de la pièce de monnaie ?
   (d) Le cas échéant, pouvez-vous développer un code plus efficace ?

2. Laquelle des lois suivantes est une meilleure approximation de la loi

   $$Y \sim \mathcal{C}ategorielle(1/3, 1/3, 1/3)?$$

   — La loi $X_1 \sim \mathcal{C}ategorielle(1/4, 1/4, 1/2)$.
   — La loi $X_2 \sim \mathcal{C}ategorielle(1/2, 1/3, 1/6)$.
   Justifez votre réponse.

3. Pour chacune des trois figures suivantes, dites si le coefficient de corrélation entre X et Y est nul ou non-nul et si l'information mutuelle $I(X, Y)$ est nulle ou non-nulle.



(a) (b) (c)

Bibliographie

[1] J.-M. Brossier. Théorie de l'information. Technical report, Notes de cours de INP Grenoble, Cours Ensimag 1A, 2014.

[2] David J.C. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2005.

[3] Kevin P. Murphy. *Machine Learning : A Probabilistic Perspective*. The MIT Press, first edition, 2012.