TD 1

Exercice 1 – Échauffement : probas discrètes, continues

Q 1.1 Répondez en quelques lignes aux questions suivantes (en français, pas de manière formelle).

Q 1.1.1 Quelle est la relation entre les notions d'*événement*, *événement élémentaire* et *univers* ?

Evénement élémentaire : la plus petite décomposition de la réalisation de l'expérience aléatoire qui nous intéresse (pas de sous-réalisation possible).

Univers Ω=ensemble d'événements élémentaires.
Événement $\in \mathcal{P}(\Omega)$ = sous-ensemble de Ω
Exemple d'événement : tirage pair = $\{2; 4; 6\}$
Si deux tirages alors univers = ensemble de couples

Q 1.1.2 De quoi a-t-on besoin pour construire un espace probabilisé sur un ensemble E dénombrable (pensez à ce que représente un élément de E) ?

Pour construire un espace probabilisé sur un ensemble dénombrable, il suffit de donner les probabilités des événements élémentaires.
$P(E) = \sum_{\omega \in E} P(\omega)$ (car les ω sont des événements élémentaires donc exclusifs / incompatibles / disjoints)
$P(\Omega) = 1$ et $P(\emptyset) = 0$
Toute fonction de $\mathcal{P}(\Omega) \to \mathbb{R}$ tq $P(\Omega) = 1$, $P(x) \geq 0$ et $\forall A, B \in \mathcal{P}(\Omega),\ A \cap B = \emptyset,\ P(A \cup B) = P(A) + P(B)$ est une mesure de probabilité.

Q 1.1.3 Qu'est ce qu'une variable aléatoire et à quoi cela sert ?

Une variable aléatoire est une application (surjective) de l'univers vers un ensemble mesurable (généralement un sous-ensemble de $\mathbb{R}$, elle permet d'affecter une valeur quantitative à tout événement élémentaire) : $X : \mathcal{P}(\Omega) \to \mathbb{R}$
Permet de faire le pont entre la définition de l'univers difficilement manipulable vers un codage plus pratique. Faire un dessin au tableau par exemple sur Ω couple des faces de deux dés (i, j) et l'image de X la somme des deux faces $i + j$. On gagne 1 si $i + j > 10$, -1 sinon.
$P(X = x) = P(X^{-1}(x))$
$P(a < X < b) = P(X^{-1}(]a; b[))$

Q 1.2 On considère que pour une journée, il pleut avec une probabilité p et qu'il ne pleut pas avec une probabilité $(1 - p)$ de manière indépendante des autres jours. Comment s'appelle la loi qui permet de modéliser la variable aléatoire indiquant le nombre de jours consécutifs où il pleut avant un jour sans pluie ? Quel est son espérance et sa variance ?

Rappeler que la loi d'une var. aléatoire c'est la donnée des $P(X = k)$ non nulles. Définir l'univers : $\Omega = \{0, 1\}^{\mathbb{N}}$ (l'ensemble des séquences de tailles finis). Remarquer l'utilité de la variable aléatoire dans ce cas, les séquences $(0, 0, 0, 1, 0, 0, 1)$, $(0, 0, 0, 1, 1, 0, 1, 0, 1, 0)$ se projettent sur la même réalisation qui nous intéressent $(k = 3)$.
La loi géométrique (un peu modifiée car normalement cette loi correspond au rang du premier succès, pas le nombre de succès avant un échec)
Fonction de masse avec X le nombre de jours consécutifs où il pleut : $P(X = x) = p^x(1 - p)$

Espérance de X = $\frac{1}{1-p} - 1$ Variance de X = $\frac{p}{1-p}$

Q 1.3 Dans un jeu de 52 cartes, on prend une carte au hasard : les événements « tirer un roi » et « tirer un pique » sont-ils indépendants ? incompatibles ? quelle est la probabilité de « tirer un roi ou un pique » ?

Indépendants oui car le fait d'avoir tiré un roi ne donne aucune information sur la probabilité d'avoir tiré un pique.
Pas incompatibles, il est possible de tirer un roi de pique
$P(Roi\ ou\ Pique) = P(Roi) + P(Pique) - P(Roi\ et\ Pique)$
$P(Roi\ et\ Pique) = P(Roi) * P(Pique|Roi) = P(Roi) * P(Pique)$ car indépendants. Faire le calcul
$P(Roi\ et\ Pique) = \frac{1}{52},\ P(Roi) * P(Pique) = \frac{4}{52}\frac{13}{52}$ donc oui c'est indépendant.
$P(Roi\ ou\ Pique) = 4/52 + 13/52 - 1/52 = 16/52$

Exercice 2 – Probabilités continues

Considérons dans cet exercice les variables aléatoires X et Y représentant respectivement le niveau d'embouteillage et la météo à un instant donné.

Q 2.1 Définir la notion de densité de probabilité et ses propriétés élémentaires. Qu'appelle-t-on densité jointe ? densité marginale ?

Expliquer différence entre var. aléatoire discrète et continue (sur l'exemple du tir sur une cible par exemple, la proba d'un point est nul. Impossibilité donc de raisonner sur des sommes ponctuelles, il faut passer à des densités : on définit la proba sur un volume de l'univers (fréquence de tomber dans le volume en tirant à l'infini des échantillons), la densité correspond au passage à la limite quand le volume tend vers 0. Equivalence avec la fonction de masse pour les lois discrètes. Si X continue, $P(X = x) = 0$.
Les événements sont les sous-ensembles "normaux" de Ω (en gros dans $\mathbb{R}$ toutes les unions dénombrables d'intervalles $]a, b[$). Pour info, σ-algèbre de l'univers : ensemble d'ensembles stable par intersection fini, union dénombrable, et passage au complémentaire).
Rappeler également qu'on utilise la notation $P(X)$ pour une variable discrète et $p(x)$ pour la densité. Normalement on devrait également indicé P_X pour dire la loi de la var. aléatoire X et p_X pour la densité, mais on simplifie les notations.
Fonction de densité correspond à la Fonction de masse pour les lois continues
Si X continue, $P(X = x) = 0$.
$P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} p(x)dx = 1$ avec $p(x)$ la fonction de densité de X
$P(X \in [a; b]) = \int_a^b p(x)dx \neq \sum_{x \in [a;b]} p(x)$ (indénombrable)
Densité jointe $P(X; Y)$
Densité conditionnelle $P(X|Y)$
Densité marginale $P(X) = \int_{-\infty}^{\infty} p(X, Y)dY$

Q 2.2 Soit X une variable aléatoire réelle de l'espace probabilité $(\Omega, \mathcal{E}, \mathbb{P})$, que vaut $\mathbb{E}(X)$ (en l'exprimant par une intégrale sur Ω et par une intégrale sur $\mathbb{R}$?

$\mathcal{E} \subseteq \mathcal{P}(\Omega)$, l'ensemble des évènements de Ω pour lesquels les probabilités $\mathbb{P}$ sont définies.
$\mathbb{E}(X) = \int_{x \in \mathbb{R}} p(x)xdx = \int_{x \in \mathbb{R}} P(X^{-1}(x))xdx = \int_{\omega \in \Omega} P(\omega)X(\omega)d\omega$

Q 2.3 Exprimer $\mathbb{E}(X|Y = y)$.

$\mathbb{E}(X|Y=y) = \int_{x\in\mathbb{R}} p(x|y)xdx$

Q 2.4 On observe pendant un certain nombre d'années les embouteillages et la météo, on dispose de ces données sous la forme d'un ensemble $E = \{(x_i, y_i)\}$, i indiquant le numéro de l'observation et x_i et y_i respectivement le niveau d'embouteillage et la météo lors de cette observation. Quel rapport entre $\mathbb{E}(X)$ et $\frac{1}{|E|}\sum_i x_i$? où se retrouve le $p(x)$ dans cette somme ? Exprimez de manière analogue $\mathbb{E}(X|Y=y)$

D'après la loi des grands nombres : $\forall\varepsilon > 0, \quad \lim_{n\to+\infty} \mathbb{P}\left(\left|\frac{X_1+X_2+\cdots+X_n}{n} - E(X)\right| \geqslant \varepsilon\right) = 0$
Si E est suffisamment grand, on a donc $\mathbb{E}(X) \approx \frac{1}{|E|}\sum_i x_i$
Où se retrouve le $p(x)$ de $\mathbb{E}(X) = \int_{x\in\mathbb{R}} p(x)xdx$ dans cette somme ? Dans l'échantillonnage des x_i dans E, qui suivent la loi de probabilité de X si les échantillons sont indépandamment et identiquement distribués (i.i.d).

Exercice 3 – Classifieur Bayésien (auteur : F. Rossi)

Q 3.1
On considère la base de données des votes effectués par les membres de la Chambre des représentants des EUA en 1984 sur 16 propositions importantes. Chaque individu est un membre de la Chambre décrit par 17 variables nominales. La variable Parti prend les modalités Démocrate et Républicain. Les autres variables, V1 à V16 représentent les votes et prennent les valeurs OUI, NON et NSP (pour une absence de vote). Il y a 267 représentants démocrates et 168 représentants républicains.

		NON	NSP	OUI
Républicains	V1	134	3	31
	V2	73	20	75
	V3	142	4	22
	V4	2	3	163
	V5	8	3	157
	V6	17	2	149
	V7	123	6	39
	V8	133	11	24
	V9	146	3	19
	V10	73	3	92
	V11	138	9	21
	V12	20	13	135
	V13	22	10	136
	V14	3	7	158
	V15	142	12	14
	V16	50	22	96

Démocrates	V1	102	9	156
	V2	119	28	120
	V3	29	7	231
	V4	245	8	14
	V5	200	12	55
	V6	135	9	123
	V7	59	8	200
	V8	45	4	218
	V9	60	19	188
	V10	139	4	124
	V11	126	12	129
	V12	213	18	36
	V13	179	15	73
	V14	167	10	90
	V15	91	16	160
	V16	12	82	173

Q 3.1.1 Combien de valeurs différentes sont possibles pour le vecteur des votes ?

$3^{16} = 43046721$

Q 3.1.2 Soit le vecteur de vote d'un représentant :
V=(OUI,NON,NSP,OUI,NON,OUI,OUI,OUI,NON,NON,OUI,NON,NON,NON,NON,OUI)
Comment estimer s'il est républicain ou démocrate ?

Rapport de vraissemblance :
$\frac{P(Democrate|V=v)}{P(Republicain|V=v)}$
Si rapport > 1 alors democrate sinon republicain
Quand on passe au log : $\log P(Democrate|V=v) - \log P(Republicain|V=v) > 0$

Selon règle d'inversion de Bayes :

$$\frac{P(Democrate|V=v)}{P(Republicain|V=v)} = \frac{P(V=v|Democrate)P(Democrate)}{P(V=v|Republicain)P(Republicain)}$$

En supposant l'indépendance conditionnelle des composantes de V :

$$\frac{P(Democrate|V=v)}{P(Republicain|V=v)} = \frac{\prod_{i=1}^{16} P(V_i=v_i|Democrate)P(Democrate)}{\prod_{i=1}^{16} P(V_i=v_i|Republicain)P(Republicain)}$$

En passant au log pour éviter les approximations dues aux valeurs trop faibles :

$$\log \frac{P(Democrate|V=v)}{P(Republicain|V=v)} = \sum_{i=1}^{16} \log P(V_i=v_i|Democrate) + \log P(Democrate) - \sum_{i=1}^{16} \log P(V_i=v_i|Republicain) - logP(Republicain)$$

Avec par exemples :
$P(Democrate) = \frac{267}{168+267}$
$P(V1 = Oui|Democrate) = \frac{156}{267}$

Q 3.2 On considère deux populations, les hommes H de taille moyenne 1,74m avec un écart type de 0,07m et les femmes F de taille moyenne 1,62m avec un écart type de 0,065m (chiffres INSEE 2001). La population H contient $|h|$ individus et la population F, $|f|$ individus. On suppose que les répartitions des tailles sont gaussiennes au sein de chaque sous-population.
On choisit aléatoirement uniformément un individu dans la population totale et on veut déterminer en fonction de sa taille uniquement de quelle sous-population il est issu : il s'agit donc de classer les individus en fonction d'une variable continue.

Q 3.2.1 On note G la variable aléatoire indiquant le genre d'une personne choisie au hasard. Donner la loi de G.

$$p(G=h) = \frac{|h|}{|h|+|f|}$$

Q 3.2.2 On note T la variable aléatoire donnant la taille d'une personne choisie au hasard. Donner la densité de T. Donner $P(G=f|T=t)$.

On sait que $p(t|G=h)$ (resp. $p(t|G=f)$) est une loi normale. De plus on sait que :

$$p(T) = p(t|G=h) * p(G=h) + p(t|G=f) * p(G=f)$$

C'est donc une mixture de gaussiennes.
D'après Bayes on a

$$P(G=f|T=t) = \frac{P(t|G=f)P(G=f)}{p(t)}$$

Q 3.2.3 Donner le classifieur bayésien optimal.

Dans le cas de l'erreur de comptage, c'est le classifieur qui minimise la probabilité d'erreur sur chaque exemple, soit celui qui choisit la classe la plus vraissemblable pour chacun : Homme si $P(G = h|T = t) > P(G = f|T = t)$
Soit donc si :
$p(t|G = h) * p(G = h) > p(t|G = f) * p(G = f)$
$\frac{1}{0.07\sqrt{2\pi}} e^{-\frac{(t-1.74)^2}{2*0.07^2}} * |h| > \frac{1}{0.065\sqrt{2\pi}} e^{-\frac{(t-1.62)^2}{2*0.065^2}} * |f|$
En passant au log :
$-\frac{(t-1.74)^2}{2*0.07^2} + \frac{(t-1.62)^2}{2*0.065^2} - \log 0.07 + \log 0.065 > 0$
$16.3t^2 - 28.33t + 1.64 - \log 0.07 + \log 0.065 + log|h| - log|f| > 0$

Q 3.2.4 On suppose que $|h| = |f|$. Préciser les décisions prisent par le classifieur optimal. Comment interpréter cette stratégie de décision ?

On a des priors égaux pour les deux classes. La condition pour classer comme homme est alors :
$16.3t^2 - 28.33t + 1.64 - \log 0.07 + \log 0.065 > 0$
$16.3t^2 - 28.33t + 1.57 > 0$
On trouve les racines par le discriminant $\Delta = b^2 - 4ac = 700.49$
$r1 = \frac{-b + \sqrt{\Delta}}{2a} = 1.68$
$r2 = \frac{-b - \sqrt{\Delta}}{2a} = 0.057$
On est alors positif lorsque la taille est supérieure à 1.68m ou inférieure à 0.057m. Cette borne inférieure peut paraître étrange, elle est due à l'écart-type plus élevé de la taille des hommes.

Q 3.3 De combien de paramètres est constitué le classifieur bayésien ?

2 priors, 2 moyennes, 2 variances : 6 paramètres
qui peuvent se ramener à 5 si on considère que les priors somment à 1

Exercice 4 – Entropie

Q 4.1 On lance un dé truqué n fois de manière indépendante, la probabilité de chaque chiffre étant $p_k, k = 1..6$.

Q 4.1.1 Exprimer la probabilité d'obtenir la suite $(x_1, \ldots, x_n)$ en fonction des p_k et des n_k - le nombre de fois où k apparaît dans le tirage.

Q 4.1.2 Vers quelle expression tend n_k lorsque n tend vers l'infini ? En déduire une expression de la probabilité d'une suite "typique" en fonction de l'entropie $H(p) = -\sum_k p_k log_2(p_k)$. Que remarquez vous pour des valeurs fortes/faibles de H ?

$P((x_1, \ldots, x_n)) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} p_{x_i} = \prod_{k=1}^{6} p_k^{n_k}$
n_k tend vers np_k par la loi des grands nombres, donc $P = \prod_{k=1}^{6} p_k^{np_k} = \prod 2^{np_k log_2(p_k)} = 2^{-nH(p)}$
Pour des valeurs faibles de H, la proba tend vers 1, la suite typique est obtenue de façon très probable (peu de suite typique). Pour des valeurs fortes de H, la proba décroît : plus grosse diversité, plus grand aléa.

Q 4.2 Quelques propriétés de la fonction entropie. On appelle vecteur de probabilité un vecteur qui représente une distribution de probabilité discrète à n modalités : $\mathbf{p} = (p_1, \ldots, p_n)$ tel que $\sum_{i=1}^{n} p_i = 1$ et $p_i \geq 0$.

Q 4.2.1 Montrer que pour $H(\mathbf{p}) \geq 0$

immédiat, $log(p_i) \leq 0$, donc ...

Q 4.2.2 Soit $\mathbf{p}$ et $\mathbf{q}$ deux vecteurs de probabilité de même dimension.
Montrer d'abord que $log(x) \leq x - 1$ en utilisant le fait que la fonction log est concave. En déduire que $-\sum_{i=1}^{n} p_i log(p_i) \leq -\sum_{i=1}^{n} p_i log(q_i)$.

La fonction log est concave, donc toute tangente du log est au dessus de log x.

soit T_{x_0} une tangente du log en x_0, alors $T_{x_0}(x) >= log(x)$

La tangente d'une fonction $f(x)$ en x_0 est une droite d'équation $y = f'(x_0)x + p$

En particulier pour $x_0 = 1$, la tangente $T_1(x)$ passe en $log(1)$ lorsque x vaut 1. Or, $log'(1) = 1$.
On a donc en $x = 1$: $log'(x_0)x + p = 1 \times 1 + p = 1 + p = log(1) = 0$. Donc $p = -1$.
On a alors $T_1(x) = x - 1$. Donc $log(x) \leq x - 1$.

Ce qui nous donne : $\sum_{i=1}^{n} p_i log(\frac{q_i}{p_i}) \leq \sum_{i=1}^{n} p_i(\frac{q_i}{p_i} - 1) = \sum_{i=1}^{n} q_i - p_i = 0$.

On a donc : $-\sum_{i=1}^{n} p_i log(p_i) \leq -\sum_{i=1}^{n} p_i log(q_i)$

Q 4.2.3 En déduire que $H(\mathbf{p}) \leq \log(n)$ pour $\mathbf{p}$ de dimension n.

On a donc pour toutes distributions p et q de même support (inégalité de Gibbs démontrée à la question précédente) : $-\sum_{i=1}^{n} p_i log(p_i) \leq -\sum_{i=1}^{n} p_i log(q_i)$

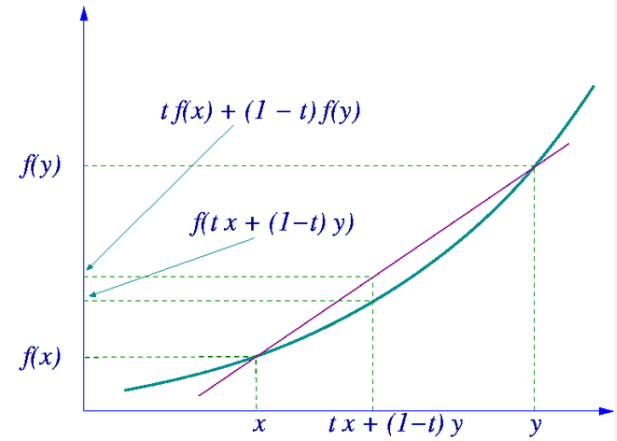
En particulier, en considérant q la loi uniforme, on a : $H_p \leq \sum_{i=1}^{n} p_i log(n) = log(n) \sum_{i=1}^{n} p_i = log(n)$
On a alors une borne sup de l'entropie, qui est atteinte lorsque la distribution est uniforme.

Q 4.3 On appelle distance de Kullback-Leibler la fonction $D(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{n} p_i log(\frac{p_i}{q_i})$

Q 4.3.1 Montrer l'inégalité de Jensen : si f est une fonction convexe dérivable, $\mathbf{p}$ un vecteur de probabilité et t_i des réels quelconques, alors $\sum_{i=1}^{n} p_i f(t_i) \geq f(\sum_{i=1}^{n} p_i t_i)$ (indication : procéder par récurrence et penser à poser $p'_i = \frac{p_i}{1-p_n}$).

On commence par prouver que $f(p_1 t_1 + (1-p_1)t_2) \leq p_1 f(t_1) + (1-p_1) f(t_2)$; une fonction convexe est toujours en dessous de sa tangente, donc en posant $x_0 = p_1 t_1 + (1-p_1)t_2$, on a

$$p_1 f(t_1) + (1-p_1) f(t_2) \geq p_1 T_{x_0}(x_0) + (1-p_1) T_{x_0}(x_0) = T_{x_0}(x_0) = f(x_0)$$



On pose $p'_i = \frac{p_i}{1-p_n}$. On a alors : $f(\sum_{i=1}^{n} p_i t_i) = f((1-p_n)\sum_{i=1}^{n-1} p'_i t_i + p_n t_n) \leq (1-p_n) f(\sum_{i=1}^{n-1} p'_i t_i) + p_n f(t_n)$ (en appliquant l'inégalité ci-dessus pour n=2 en considérant $(1-p_n)\sum_{i=1}^{n-1} p'_i t_i$ à la place de $p_1 t_1$ et $p_n t_n$ pour $p_2 t_2$).

On suppose l'inégalité vraie au rang $n-1$. Il s'en suit alors que $f(\sum_{i=1}^{n} p_i t_i) \leq (1-p_n)\sum_{i=1}^{n-1} p'_i f(t_i) + p_n f(t_n) = \sum_{i=1}^{n} p_i f(t_i)$

Q 4.3.2 En déduire que pour une fonction convexe, $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$. Trouver une autre démonstration pour la question **??**

$log(x)$ est concave, donc $-log(x)$ est convexe.

On a $-log \sum q_i = -log \sum p_i \frac{q_i}{p_i} \leq -\sum p_i log \frac{q_i}{p_i}$

Or $-log \sum q_i = 0$.

Donc $-\sum p_i log \frac{q_i}{p_i} \geq 0$ et alors $-\sum p_i log q_i \geq -\sum p_i log p_i$

Q 4.3.3 Montrer que $D(\mathbf{p}||\mathbf{q}) \geq 0$ avec égalité ssi $\mathbf{p} = \mathbf{q}$. Est-ce que D est une distance ?

On utilise le résultat au dessus : $D(p||q) = -\sum p_i log(\frac{q_i}{p_i}) \geq -log(\sum p_i \frac{q_i}{p_i}) = 0$. Non c'est pas une distance (pas symmétrique).

Q 4.3.4 On considère $x_1, \ldots x_N$ des observations i.i.d. dans $\mathcal{X}$ un ensemble discret fini. La distribution empirique observée est définie par $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)$ avec δ la fonction de dirac (0 partout sauf en 0 où elle vaut 1). Soit p_θ une distribution paramétrisée par θ. Montrer que maximiser la vraisemblance revient à minimiser $D(\hat{p}||p_\theta)$.

$D(\hat{p}||p_\theta) = \sum_{\mathcal{X}} \hat{p}(x) log(\frac{\hat{p}(x)}{p_\theta(x)}) = -H(\hat{p}) - \sum_{\mathcal{X}} \hat{p}(x) log(p_\theta(x)) = -H(\hat{p}) - \frac{1}{N}\sum_{\mathcal{X}}\sum_{i=1}^{n} \delta(x - x_i) log p_\theta(x) = -H(\hat{p}) - \frac{1}{N}\sum_{i=1}^{n} log(p_\theta(x))$