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* ?
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) ?

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

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 ?

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 » ?

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 ?

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 $\mathbb{R}$?

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

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 $\mathbb{E}(X|Y=y)$

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 ?

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 ?

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.

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 $P(G = f|T = t)$.

Q 3.2.3 Donner le classifieur bayésien optimal.

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

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

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 ?

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 $H(\mathbf{p}) \geq 0$

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 $-\sum_{i=1}^{n} p_i log(p_i) \leq -\sum_{i=1}^{n} p_i log(q_i)$.

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

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 $p'_i = \frac{p_i}{1-p_n}$).

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 **??**

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 ?

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 $D(\hat{p}||p_\theta)$.