TD 1

Exercice 1 – Classifieur Bayésien (auteur : F. Rossi)

Q 1.1
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 1.1.1 Combien de valeurs différentes sont possibles pour le vecteur des votes ?

Q 1.1.2 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 1.2 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 1.2.1 On note G la variable aléatoire indiquant le genre d'une personne choisie au hasard. Donner la loi de G.

Q 1.2.2 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 1.2.3 Donner le classifieur bayésien optimal.

Q 1.2.4 On suppose que $|h| = |f|$. Préciser les décisions prisent par le classifieur optimal. Comment interpréter cette stratégie de décision ?

Q 1.3 De combien de paramètres est constitué le classifieur bayésien ?

Exercice 2 – Classifieur bayésien

Soit $\mathcal{X}$ un ensemble de description dans $\mathbb{R}^d$ et $\mathcal{Y}$ l'ensemble des labels $\{y_1, ..., y_l\}$.

Q 2.1 Rappeler ce qu'est un classifieur bayésien.

Q 2.2 Exprimer l'erreur faite par le classifieur bayésien à un point $\mathbf{x}$. L'erreur est-elle minimale ?

Q 2.3 Soit $\lambda(y_j, y_i)$ le coût d'une erreur consistant à prédire le label y_j plutôt que y_i. Que valent les λ dans le cas de l'erreur 0-1 ? Donner quelques exemples de coûts asymétrique et des contextes d'utilisation.

Q 2.4 Quelle est l'expression du risque $R(y_i|\mathbf{x})$ de prédire y_i sachant $\mathbf{x}$ en fonction de λ et des probabilités a posteriori ? Dans le cas 0-1 ?

Q 2.5 Donner l'expression du risque sur $\mathcal{X}$ associé au classifieur f, $R(f)$.

Q 2.6 On se place dans le cas binaire. Exprimer le critère de décision en fonction de λ et des probabilités a posteriori, puis donner un critère de décision en fonction de λ, la distribution des classes et la vraissemblance.

Exercice 3 – Estimation de densité

Q 3.1 Donner l'estimation de la densité $p_{\mathcal{B}}$ d'une variable aléatoire X à l'intérieur d'une région d'intérêt $\mathcal{B}$ de volume V, en fonction d'un nombre k d'échantillons observés dans cette zone parmi n échantillons tirés.

Q 3.2 Soit une variable aléatoire $X \in \mathcal{X}$. On souhaite estimer la densité p_X de cette variable à partir d'un ensemble d'observations $\mathcal{X}_o$. Décrire la manière de procéder pour réaliser cette estimation selon la méthode des histogrammes.

Q 3.3 Discuter des méthodes d'estimation de densité à noyaux

Exercice 4 (4 points) – Risque

Soit le résultat d'une analyse médicale exprimé par un réel $x \in \mathbb{R}$, soit deux classes y_-, y_+ pas malade et malade. On sait que $P(y = y_+|x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x \geq 1 \end{cases}$.

On utilise un classifieur de type $f_\theta(x) = \begin{cases} y_+ & \text{si } x > \theta \\ y_- & \text{si } x \leq \theta \end{cases}$.

Q 4.1 Donner l'expression du coût 0-1 et donner le risque en un point x^0 en fonction de θ.

Q 4.2 On suppose le résultat du test x uniformément répartie dans $[-1, 1]$. Quel est le classifieur optimal ? Quelle est la valeur du risque minimale ?

Q 4.3 On sait qu'il est 3 fois plus coûteux de classer un malade en pas malade que l'inverse. Donner une fonction de coût associée, la nouvelle formulation du risque et le classifieur optimal.

Q 4.4 Pourrait-on faire mieux avec un classifieur bayésien ? Un classifieur bayésien naïf ?