Apprentissage statistique

Sujet 1

Exercice 1

On étudie un ensemble de 10 observations, $(x_i,y_i)_{1\leq i\leq 10}$, avec $x_i \in \mathcal{X}$ et $y_i \in \{-1,1\}$. Grâce à un algorithme d'apprentissage automatique, on construit deux modèles, g_1 et g_2. Le tableau suivant donne les valeurs de $g_1(x_i)$, $g_2(x_i)$ et y_i pour tout i :

x_i	$g_1(x_i)$	$g_2(x_i)$	y_i
x_1	1	1	1
x_2	1	-1	1
x_3	-1	1	1
x_4	1	1	1
x_5	1	-1	1
x_6	1	-1	-1
x_7	-1	-1	-1
x_8	-1	-1	-1
x_9	-1	1	-1
x_{10}	1	-1	-1

Question 1 Calculer les matrices de confusion des deux modèles.

Question 2 On choisit la fonction de perte l_1 définie par :

$l_1(v,p)$	$p=-1$	$p=1$
$v=-1$	0	2
$v=1$	1	0

où p désigne la valeur prédite et v la vraie valeur. Déterminer le meilleur modèle (entre g_1 et g_2) au sens du risque empirique construit à partir de l_1.

Question 3 Quel modèle choisir au sens de la fonction de perte $l_0(v,p) = \mathbb{I}_{p\neq v}$?

Exercice 2

On étudie des données distribuées selon le modèle $Z = (X,Y)$ suivant :

— Y est une variable aléatoire à valeurs dans $\{-1,1\}$ avec $\mathbb{P}(Y=-1) = \frac{2}{3}$;

— X est une variable aléatoire à valeurs dans $\{a,b,c\}$ dont la loi conditionnelle est donnée par :

x	a	b	c
$\mathbb{P}(X=x\|Y=-1)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
$\mathbb{P}(X=x\|Y=1)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Question 1 Déterminer le modèle optimal de Y sous la forme d'une fonction de X au sens de la fonction de perte l_0 de l'exercice précédent.

Question 2 Calculer le risque du modèle optimal obtenu à la question précédente.

Question 3 Mêmes questions avec la fonction de perte l_1 de l'exercice précédent.

Exercice 3

On étudie des données distribuées selon le modèle $Z = (X,Y)$ suivant :

— Y est une variable aléatoire à valeurs dans $\{-1, 1\}$ avec $\mathbb{P}(Y = 1) = \frac{1}{2}$;

— $X = (X_1,X_2)$ est une variable aléatoire à valeurs dans $\{S, M, L\} \times \{1, 2\}$ (il s'agit donc de couples). La loi jointe de X conditionnellement à Y est donnée par les tableaux suivants :

$\mathbb{P}(X_1 = t, X_2 = p \mid Y = -1)$	$p = 1$	$p = 2$
$t = S$	0	$\frac{1}{12}$
$t = M$	$\frac{3}{12}$	$\frac{1}{6}$
$t = L$	$\frac{1}{6}$	$\frac{1}{3}$

$\mathbb{P}(X_1 = t, X_2 = p \mid Y = 1)$	$p = 1$	$p = 2$
$t = S$	$\frac{5}{12}$	$\frac{1}{6}$
$t = M$	$\frac{1}{6}$	$\frac{1}{12}$
$t = L$	$\frac{1}{12}$	$\frac{1}{12}$

Question 1 Les variables X_1 et X_2 sont elles conditionnellement indépendantes sachant Y ?

Question 2 Déterminer le modèle optimal de Y comme une fonction de X au sens de l'erreur l_0 (cf exercice 2).

Question 3 Calculer le risque de ce modèle optimal.

Question 4 Calculer la loi $\mathbb{P}(X_1 = t|Y = 1)$.

Question 5 En généralisant les calculs de la question précédente, déterminer le modèle obtenu selon le principe du classifieur bayésien naïf.

Question 6 Calculer le risque du classifieur bayésien naïf.

Exercice 4

On considère la classe $\mathcal{F}$ des fonctions de $\mathbb{R}$ dans {0,1} de la forme

$$f_{a,b}(x) = \begin{cases} 0 & \text{si } ax + b < 0, \\ 1 & \text{sinon.} \end{cases}$$

On cherche à déterminer la dimension de Vapnik-Chernonenkis de cette classe.

Question 1 Donner un exemple d'ensemble réduit à un point $\{x\}$ pulvérisé par $\mathcal{F}$: on donnera explicitement 2 éléments de $\mathcal{F}$ qui réalisent cette pulvérisation.

Question 2 Montrer que pour $n \geq 2$, le coefficient de pulvérisation $\mathcal{S}(n)$ vaut $n+1$. On pourra raisonner sur un ensemble de n points $\{x_1, \ldots, x_n\}$ rangés par ordre strictement croissant $x_1 < x_2 < \ldots < x_n$ sans perte de généralité, puis donner explicitement les valeurs possibles de $(f(x_1), \ldots, f(x_n))$ pour $f \in \mathcal{F}$.

Question 3 En déduire la dimension de Vapnik-Chernonenkis de $\mathcal{F}$.