Perte, risque et erreur en apprentissage automatique

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 le nombre d'erreurs de classement réalisées par chaque modèle.

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

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

où p désigne la valeur prédite et v la vraie valeur. On rappelle que le risque empirique d'un modèle g pour la perte l_1 est donné ici par

$$\widehat{L}_1(g) = \frac{1}{10}\sum_{i=1}^{10} l_1(g(x_i),y_i).$$

Déterminer le meilleur modèle (entre g_1 et g_2) au sens de $\widehat{L}_1$.

Question 3 On appelle matrice de confusion d'un modèle la matrice carrée dont les termes résument le comportement du modèle en fonction des valeurs à prédire. Chaque colonne correspond à une vraie valeur de y et chaque ligne à une valeur prédite[1]. À l'intersection d'une ligne et d'une colonne, on indique le nombre d'objets de la vraie valeur en colonne pour lesquels le modèle prédit la valeur en ligne. Dans le cas présent, il faut donc remplir une matrice de la forme suivante

g	$v=-1$	$v=1$
$p=-1$	nombre d'objets $y=-1$ pour lesquels g donne -1	...
$p=1$	...	...

Calculer les matrices de confusion de g_1 et de g_2.

Question 4 Quel lien simple peut on faire entre $\widehat{L}_1(g)$ d'une part et la matrice représentant l_1 et la matrice de confusion de g d'autre part ?

1. On peut prendre la convention inverse, il faut juste se tenir à une convention unique.

Question 5 On peut estimer diverses probabilités à partir de la matrice de confusion d'un modèle. Le faire pour les probabilités suivantes :

$$\hat{\mathbb{P}}(g_2(X) = 1, Y = -1) = ?$$
$$\hat{\mathbb{P}}(g_2(X) = 1 | Y = 1) = ?$$
$$\hat{\mathbb{P}}(Y = 1 | g_1(X) = 1) = ?$$

Exercice 2

On étudie des modèles définis sur $\mathcal{X}$ à valeurs dans $\mathcal{Y} = \{-1, 1\}$. On évalue ces modèles sur un ensemble de test comportant 100 observations. Deux modèles g_1 et g_2 sont comparés. Ils ont les matrices de confusion suivantes (en ligne la valeur prédite par le modèle, en colonne la valeur réelle) :

	$Y = -1$	$Y = 1$
$g_1(X) = -1$	40	15
$g_1(X) = 1$	5	40

	$Y = -1$	$Y = 1$
$g_2(X) = -1$	50	5
$g_2(X) = 1$	10	35

Question 1 Déterminer le modèle optimal (entre g_1 et g_2) au sens de la fonction de perte $l_0(p,v) = \mathbb{I}_{p \neq v}$ (p est la prévision, v la vraie valeur).

Question 2 Même question avec la fonction de perte l_1 définie par :

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

où p désigne la valeur prédite et v la vraie valeur.

Exercice 3

On étudie un ensemble d'observations $(x_i, y_i)_{1 \leq i \leq 10}$, avec $x_i \in \mathcal{X}$ et $y_i \in \{-1, 1\}$. On étudie des modèles construits en deux parties, à partir d'une fonction f de $\mathcal{X}$ dans $\mathbb{R}$ et d'un seuil λ. Le modèle g est défini par

$$g(x) = \begin{cases} -1 & \text{si } f(x) \leq \lambda, \\ 1 & \text{si } f(x) > \lambda. \end{cases}$$

On étudie tout d'abord f_1 donnée par le tableau suivant :

i	1	2	3	4	5	6	7	8	9	10
$f(x_i)$	-3	-2	$-1{,}5$	-1	-1	$-1{,}2$	0	$-0{,}5$	1	2
y_i	-1	-1	-1	-1	1	1	1	1	1	1

Question 1 Déterminer g_1 en utilisant $\lambda = -1$.

Question 2 Calculer la matrice de confusion de g_1.

Question 3 On utilise la fonction de perte l_1 définie par :

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

où p désigne la valeur prédite et v la vraie valeur. Déterminer le risque empirique de g_1 par rapport à l_1.

Question 4 Déterminer la valeur de λ qui conduit au modèle g avec le plus petit risque empirique possible pour l_1.

Exercice 4

On étudie un ensemble d'observations $(x_i,y_i)_{1\leq i\leq N}$, avec $x_i \in \mathcal{X}$ et $y_i \in \{-1,1\}$. Comme dans l'exercice précédent, on considère des modèles construits en deux temps. On a donc deux fonctions f_1 et f_2 de $\mathcal{X}$ dans $\mathbb{R}$ à partir desquelles on construit deux modèles, g_1 et g_2 grâce à des seuils λ_1 et λ_2 par

$$g_k(x) = \begin{cases} -1 & \text{si } f_k(x) \leq \lambda_k, \\ 1 & \text{si } f_k(x) > \lambda_k. \end{cases}$$

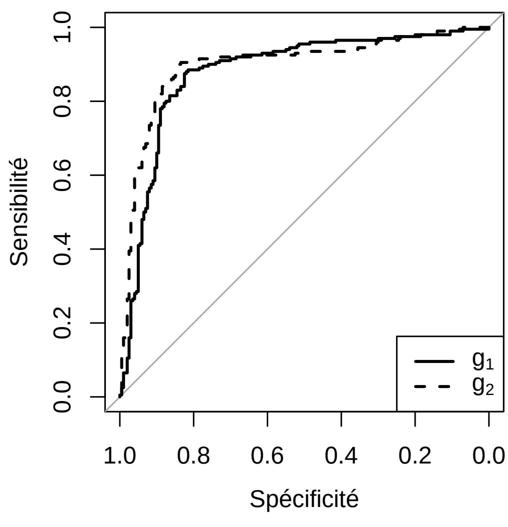
Question 1 On suppose fixée une fonction de perte l. Écrire le problème d'optimisation sur λ_k associé à la minimisation du risque empirique de g_k pour l. On pourra par exemple passer par des ensembles de la forme

$$\{i|y_i = 1 \text{ et } g_k(x_i) \leq \lambda_k\}.$$

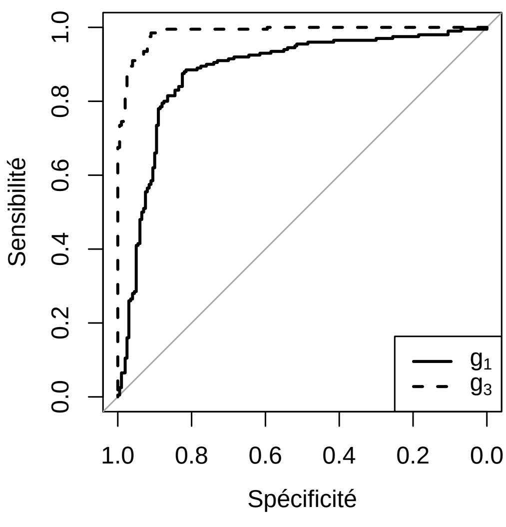
Attention, on ne peut pas résoudre ce problème explicitement.

Question 2 Écrire la version simplifiée de ce problème pour $l = l_{0/1}$ donnée par $l_{0/1}(p,v) = \mathbb{1}_{v=p}$.

Pour analyser les modèles g_k sans fixer de valeur à λ_k, on étudie leur courbe ROC (*Receiver Operating Characteristic*) : pour chaque $\lambda_k \in \mathcal{R}$, on calcule la *spécificité* et la *sensibilité* de g_k, et on trace la courbe obtenue ainsi. La spécificité est le *taux* de vrais négatifs (ici les x_i tels que $y_i = -1$ pour lesquels $f_k(x_i) \leq \lambda_k$) alors que la sensibilité est le *taux* de vrais positifs (ici les x_i tels que $y_i = 1$ pour lesquels $f_k(x_i) > \lambda_k$). La figure 1a représente les courbes ROC de g_1 et g_2.



(a) Courbes ROC de g_1 et g_2.



(b) Courbes ROC de g_1 et g_3.

Question 3 On suppose que

$$|\{i|y_i = -1\}| = |\{i|y_i = 1\}|.$$

Quel point de la courbe ROC de g_1 correspond (approximativement) au meilleur λ_1 pour la fonction de perte $l_{0/1}$? Que peut-on dire si l'hypothèse ci-dessus n'est pas vérifiée ?

Question 4 On suppose toujours que l'hypothèse de la question précédente est vérifiée. Comment trouver (approximativement) le point de la courbe ROC de g_1 correspondant au meilleur λ_1 pour la fonction de perte l_1 donnée par

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

où p désigne la valeur prédite et v la vraie valeur.

Question 5 Peut-on dire que g_1 est meilleur que g_2 (ou moins bon) *de façon générale* en utilisant la figure 1a ? Même question entre g_1 et le nouveau modèle g_3, en s'appuyant sur la figure 1b.