TD : Arbre de décision et Foret aléatoire

Problème I : Régression Logistique (à la main)

Nous cherchons à expliquer une variable binaire Y (supposée suivre une distribution de Bernoulli $B(p)$) à travers une variable quantitative continue notée X. On suppose que ${}^{X}/_{Y=1} \sim N\left(\mu_1, \sigma^2\right)$ et que ${}^{X}/_{Y=0} \sim N\left(\mu_0, \sigma^2\right)$. Notez que la fonction de densité de probabilité d'une variable aléatoire suivant une loi normale $N\left(\mu, \sigma^2\right)$ est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{avec } x \in \mathbb{R}.$$

1. Déterminez la fonction de densité de probabilité de X (appelée la distribution de mélange gaussien).
2. Calculez la probabilité conditionnelle

$$p(x) = \mathbb{P}\left[{}^{Y=1}/_{X=x}\right].$$

3. En déduire une expression linéaire de $\ln\left[\frac{p(x)}{1-p(x)}\right]$ (de la forme $\beta_0 + \beta_1 x$). Donnez aussi l'expression de $p(x)$ en fonction de $(\beta_0 + \beta_1 x)$.
4. Expliquez pourquoi, en pratique, la détermination de $p(x)$ pour une valeur donnée de x est difficile en utilisant l'expression de la partie (3).
5. Sous l'hypothèse que :

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x.$$

   Proposez une méthode pour estimer les paramètres β_0 et β_1 en se basant sur un ensemble de données de la forme $(x_i, y_i)_{i=1,\ldots,n}$ (n représente le nombre des observations dans la base de données). Veuillez expliquer, en détail, les différentes étapes de la méthode sans résoudre le système à deux équations que vous allez obtenir.
   Indication : Méthode du Maximum de Vraisemblance.
6. Une fois que les estimateurs $\widehat{\beta_0}$ et $\widehat{\beta_1}$ des paramètres β_0 et β_1 sont calculés, expliquez comment, en pratique, le modèle logistique est utilisé pour classifier un nouvel individu x_{new} en 0 ou 1.

Problème II : Classification supervisée et étude de la validité (à la main)

Dans cet exercice, il s'agit d'explorer le processus de la classification supervisée sur les données météorologiques présentées dans le Tableau 1. Le but est d'aider les joueurs d'un jeu sportif donné à décider de jouer ou non selon les conditions météo. Ainsi, la variable à expliquer est « Play ».

Oulook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	No
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Table 1 – Données météo et décision des joueurs

1. À partir des données du Tableau 1 :
   (a) Recopiez et complétez le tableau 2 en indiquant, pour chaque règle, le taux d'erreur de classification de la variable « Play » lorsque cette règle est appliquée aux observations correspondantes.
   (b) Calculez ensuite le taux d'erreur global résultant de l'application simultanée de toutes les règles associées à une même variable.

Variables	Règles	Taux d'erreur	Taux d'erreur Global
Outlook	Sunny ⇒ No		
	Overcast ⇒ Yes		
	Rainy ⇒ Yes		
Temperature	Hot ⇒ No		
	Mild ⇒ Yes		
	Cool ⇒ Yes		
Humidity	High ⇒ No		
	Normal ⇒ Yes		
Windy	False ⇒ Yes		
	True ⇒ Yes		

TABLE 2 – Taux d'erreur des différentes règles sur les différentes variables

2. Le modèle OneR utilise un seul attribut, celui ayant le taux d'erreur le plus faible, pour effectuer la classification. Quelles seront les règles de classification si l'on applique OneR au résultat obtenu à la question (1) ?
3. Explorer le résultat de la classication OneR de la question (2) :
   (a) Construire la matrice de confusion détaillée.
   (b) Calculer : sensibilité, spécificité, valeur prédictive positive (VPP) et valeur prédictive négative (VPN), ainsi que les intervalles de confiance correspondants.

Problème III : Courbe ROC (à la main)

Un nouveau gel hydroalcoolique est en cours de test pour tuer des bactéries. On souhaite découvrir le dosage optimal du gel. On teste ce gel sur 806 milliards de bactéries à des doses variant de 0 µg à 20 µg et on note le nombre de bactéries qui sont éliminées et celles qui demeurent, en considérant des intervalles de dosage de 2 µg. Si le succès est d'éliminer la bactérie et l'échec est que la bactérie survit :

1. Construire une approximation de la courbe ROC en utilisant 5 seuils de votre choix et en se basant sur les données du tableau 3 ci dessus. (Indication : choisir les seuils parmi les bornes supérieures des intervalles du tableau 3)
2. Donner une estimation de la valeur du dosage optimal du gel.

Dosage	Survivants	Décès
$[0, 2[$	34	3
$[2, 4[$	63	7
$[4, 6[$	88	11
$[6, 8[$	105	14
$[8, 10[$	123	23
$[10, 12[$	95	60
$[12, 14[$	9	75
$[14, 16[$	6	41
$[16, 18[$	4	30
$[18, 20]$	0	15
Total	527	279

TABLE 3 – Résultats des tests d'un nouveau gel hydroalcoolique.