Devoir surveillé de classification supervisée

Vandewalle/Grimonprez

Durée : 2 heures
Tous documents et calculettes autorisés

Exercice 1

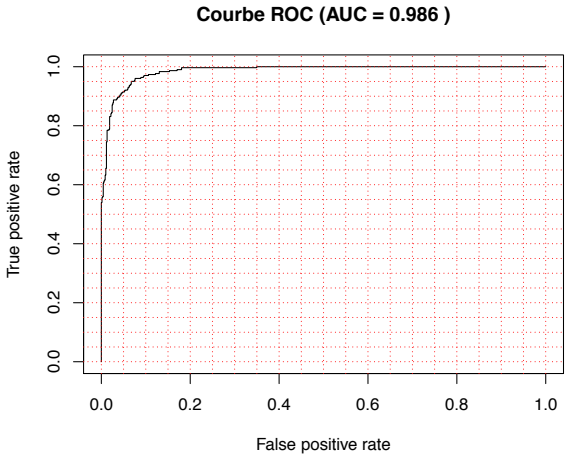
Soit Y une variable aléatoire binaire représentant la rechute d'un patient atteint d'une maladie cardiovasculaire ($Y = 1$ si rechute, $Y = 0$ sinon), on estime à 30% la probabilité de rechute d'un patient. Soit X_1 la variable aléatoire égale au nombre d'heures de sport par semaine pratiquées par le patient. On a :

- la distribution de X_1 sachant $Y = 1$ est une loi normale univariée d'espérance 2 et d'écart-type $0,5$.
- la distribution de X_1 sachant $Y = 0$ est une loi normale univariée d'espérance $3,5$ et d'écart-type $0,5$.

1. La modélisation du nombre d'heures de sport par une loi normale vous semble-t-elle raisonnable ?
2. Quelle est la densité de probabilités marginale de X_1 (toutes valeurs de Y confondues) ?
3. Donner l'allure de la densité de probabilités de X_1.
4. Donner la formule permettant de calculer $P(Y = 1|X_1 = x)$ pour tout $x \in \mathbb{R}$.
5. En déduire la probabilité de rechute d'un patient effectuant 1 heure de sport par semaine ? pour un patient effectuant 2 heures de sport par semaine ?
6. Calculer le odds pour 1 heure de sport par semaine, pour 2 heures de sport par semaine, ainsi que le odds-ratio quand le patient passe de 1 heure à 2 heures de sport par semaine. Commenter.
7. Donner l'expression de $\ln \frac{P(Y=1|X_1=x)}{1-P(Y=1|X_1=x)}$. A quoi cette expression vous fait-elle penser ?
8. Donner l'allure de $P(Y = 1|X_1 = x)$ en fonction de x.
9. Sur un échantillon de 1 000 patients, en appliquant la régle de classement du maximum a posteriori (classification de l'individu dans la classe la probable), nous obtenons le tableau de confusion suivant :

		pr ?dit		total
		$Y = 0$	$Y = 1$	
r ?el	$Y = 0$	674	24	698
	$Y = 1$	33	269	302
total		707	293	1000

   Déduire de ce tableau la sensibilité, la spécificité, et le taux de bon classement.
10. En faisant varier le seuil de classement on obtient la courbe ROC suivante :



A partir de ce graphique, quand la sensibilité est égale ? 95%, quelle est la valeur approximative de la spécificité ? En d ?duire le taux de bon classement.

On dispose de la variable X_2 qui représente la distance moyenne par jour. On a :

— la distribution du couple X_2 sachant $Y = 1$ est une loi normale univariée d'espérance 4 et d'écart-type $0,5$.

— la distribution de X_2 sachant $Y = 0$ est une loi normale univariée d'espérance 3 et d'écart-type $0,5$.

Par ailleurs dans chacune des classes le couple (X_1, X_2) suit une loi normale bivariée dont le coefficient de corrélation linéaire est égale à $0,6$.

11. Ecrire la matrice de variance covariance intraclasse
12. En utilisant la régle d'affectation basée sur la distance de Mahalanobis avec la métrique

$$W^{-1} = \begin{pmatrix} 3,125 & -1,875 \\ -1,875 & 3,125 \end{pmatrix}$$

? quelle classe affectez-vous un patient effectuant 5 heures de sport par semaine et marchant 1 km par jour ?

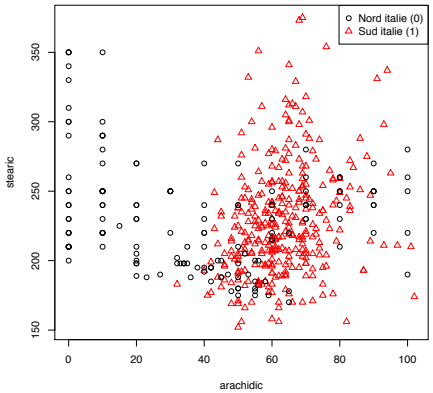
13. A quelle(s) méthode(s) la régle de classement basée sur la métrique de Mahalonobis est-elle équivalente ?

Exercice 2

On souhaite prédire l'origine d'une huile (nord italie (0) ou sud italie (1)) en fonction de sa concentration en acide stearic et en acide arachidic (cf. figure).

En ajustant la régression logistique de la région en fonction de sa concentration en acide stearic et en acide arachidic, on obtient les sorties suivantes :

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.602e+00  7.882e-01  -3.302 0.000961 ***
stearic     -1.096e-05  3.131e-03  -0.004 0.997205
arachidic    6.400e-02  6.858e-03   9.332  < 2e-16 ***
---
```



1. Commenter les sorties et donner l'équation de la régression logistique.
2. Que feriez-vous dans un second temps pour améliorer les performances du modèle ?