Partiel 1

Durée : 2H

Seuls documents autorisés : Cours et notes de cours
– Barême indicatif –

Exercice 1 (4pts) – Cholesterol

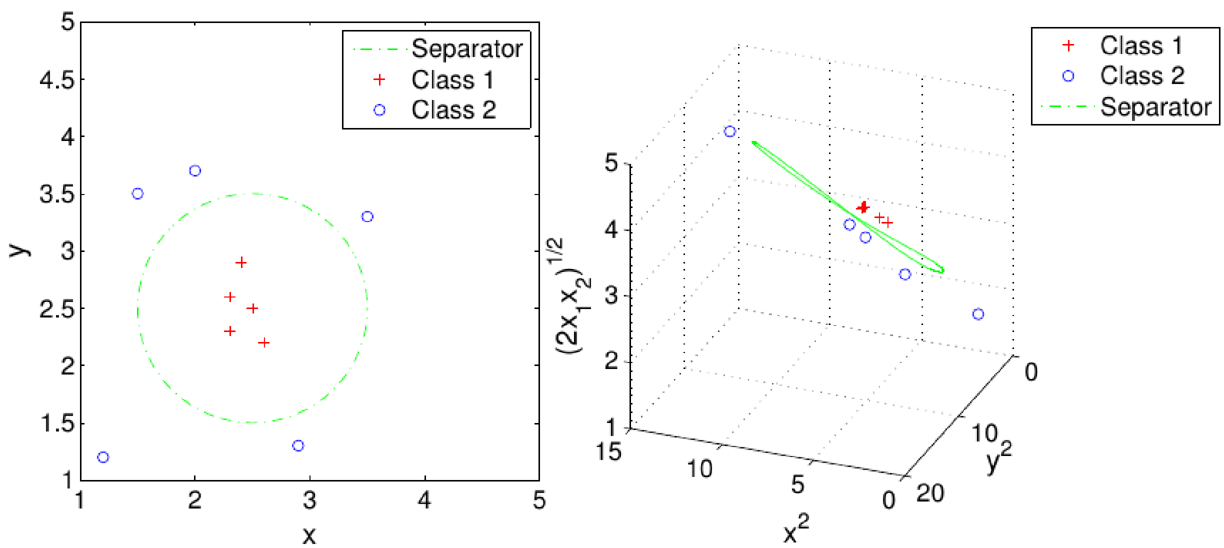
The first lines of your file are as follows

Age	Cholesterol	Class (healthy/not healthy)
48	4.40	0
60	7.89	1
51	3.38	0
66	8.41	1
40	3.05	0

Apply a binary logistic regression to classify the patients. Provide the log-likelihood function and describe (in details) the optimization procedure.

Exercice 2 (4pts) – Classification

Let us imagine that we have the following classification problem. The observations from the class +1 are ones inside the circle, and ones from the class -1 outside (see the figure below, on the left). We know that we can map our data into three-dimensional space, with the attributes (x^2 , y^2, $\sqrt{2xy}$) (see the plot below, on the right).



Q 2.1 Which classifier will you choose ? Why ?

Q 2.2 Provide a brief description on the functioning of the chosen classifier.

Q 2.3 Some patient features are expensive to collect (e.g., biopsy) whereas others are not (e.g., blood test). Therefore, we can decide to first ask our classification algorithm to predict whether a patient has a disease, and if the classifier is 90% confident that the patient is ill, then we will do additional exanimations to collect more patient features. In this case, which classification methods do you recommend ? Justify your answer.

Q 2.4 We have decided to use a neural network to classify patients having heart disease (class 1) or Alzheimer (class 2). We can either train a separate neural network for each disease or to train a single neural network with one output neuron for each disease, but with a shared hidden layer. Which method do you prefer ? Justify your answer.

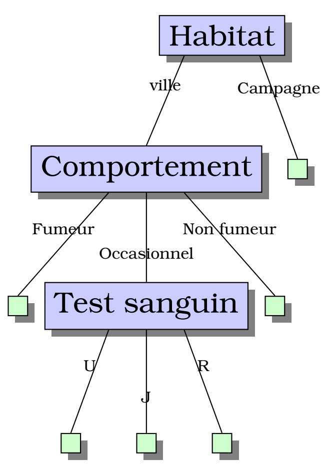
Exercice 3 (7pts) – **Arbres de décision**

Soit la base de données suivantes indiquant si un individus est “à risque” ou non pour une certaine maladie :

Comportement	Test sanguin	Habitat	Antécédent	Résultat
Fumeur	U	Ville	Non	à risque
Fumeur	U	Ville	Oui	à risque
Non fumeur	U	Ville	Non	sans risque
Occasionnel	J	Ville	Non	sans risque
Occasionnel	J	Campagne	Non	sans risque
Occasionnel	J	Campagne	Oui	à risque
Non fumeur	J	Campagne	Oui	sans risque
Fumeur	R	Ville	Non	à risque
Fumeur	J	Campagne	Non	sans risque
Occasionnel	R	Campagne	Non	sans risque
Fumeur	R	Campagne	Oui	sans risque
Non fumeur	R	Ville	Oui	sans risque
Non fumeur	U	Campagne	Non	sans risque
Occasionnel	R	Ville	Oui	à risque

Q 3.1 Construire l’arbre de décision représentant ce problème (il y a 3 nœuds).

Q 3.2 Soit l’arbre suivant obtenu différemment pour la même base de données :



En chaque nœud, le résultat est la classe majoritaire dans ce nœud. On considère cet arbre comme trop adapté à la base de données et on veut donc l’élaguer : retirer certains nœuds.

Q 3.2.1 Calculer en chaque nœud s (feuille ou non de l’abre), l’erreur $e(s)$ d’apprentissage.

Q 3.2.2 On considère qu’un noœud (interne) doit être élagué si l’erreur moyenne de ses enfants est plus grande que sa propre erreur. En notant (s_i) les enfants du nœud s, donner la formule de $e_M(s)$.

Q 3.2.3 Proposer un algorithme qui permet d'élaguer l'arbre de décision. Appliquer cet algorithme sur l'arbre de cette question.

Exercice 4 (6pts) – **Classification bayésienne**

Dans la population composée de champignons il y a deux classes : c_1, les champignons vénéneux ; c_2, les champignons comestibles. La proportion de champignons vénéneux est de 5%.
Un champignon possède ou non une volve : variable V à valeurs $v \in \{volve, pas-de-volve\}$. 90% des champignons vénéneux ont une une, alors que le pourcentage n'est que de 20% parmi les champignons comestibles.

Q 4.1 Estimant les probabilités par les fréquences indiquées, donner les probabilités a priori $\pi(c_k)$ des deux classes et les probabilités conditionnelles $p(v/c_k)$.

Q 4.2 Minimise-t-on la probabilité de se tromper (= d'attribuer un champignon à la classe qui n'est pas la sienne) en classant les champignons à volve dans la classe c_1 (décision d_1) ou dans la classe c_2 (décision d_2) ?

Q 4.3 Prendre la décision d_j alors que la classe vraie est c_k entraîne une perte (psychologique) $w(d_j, c_k)$ variable :

$$w(d_1, c_1) = w(d_2, c_2) = 0 \,; w(d_1, c_2) > 0 \,; w(d_2, c_1) = Mw(d_1, c_2)$$

avec M arbitrairement grand.
Que conseille donc la règle de BAYES ?