Examen Réparti 1

Durée : 2H

Seuls documents autorisés : Cousrs, notes de cours, notes de TDs, calculatrice
– Barême indicatif –

Exercice 1 (3pts) – Short questions.

Q 1.1 You are given a data set of cellular images of patients with or without a disease. Clinicians ask you to train a classifier that estimates the probability that a subject is ill. What classifier would you train ? Why ?

Q 1.2 A data set has 850 observations from healthy individuals, and 150 from patients having the type 2 diabetes. You trained a classifier and it achieves 85% accuracy. Is it a good classifier ? Why ?

Q 1.3 You applied a linear regression with a regularisation term, and you found out that all the estimated coefficients were equal to 0. Explain what happened and why.

Exercice 2 (4pt) – Canonical Correlation Analysis

Two groups of variables are provided. The first group of parameters is related to exercise (variables associated with exercise, observations such as the climbing rate on a stair stepper, how fast you can run, the amount of weight lifted on bench press, the number of push-ups per minute, etc.) The second group of variables is associated with health variables such as blood pressure, cholesterol levels, glucose levels, body mass index, etc. We would like to reveal existing relations between these two groups of variables.

Q 2.1 Explain briefly how to run the canonical correlation analysis.

Q 2.2 What is the output of the canonical correlation ? What is the maximal and minimal values of the correlation ?

Q 2.3 How to interpret the weights of the canonical correlation model ?

Q 2.4 Is canonical correlation supervised or unsupervised methods ?

Exercice 3 (3 pts) – Gradient descent for k–means clustering

The loss function for k-means clustering with k clusters, sample points $x_1, \dots x_n$ and centers $c_1, \dots, c_k$ is given as follows :

$$\ell = \sum_{j=1}^{K} \sum_{x_i \in S_j} \|x_i - c_j\|^2,$$

where S_j is the set of data points that are closer to c_j than to other clusters means.

Q 3.1 Write down the gradient of the loss with respect to the parameter $\frac{\partial \ell}{\partial c_j}$

Q 3.2 Derive the update formula using the gradient descent

Q 3.3 How would you choose k (number of clusters) in a real application (1 sentence) ?

Exercice 4 (8 points) – **Contingence, probabilités, régression logistique et indépendances**

Soit un problème de classification à une dimension ($d = 1$, classe Y, caractéristique X). La table de contingence de ce problème est donnée par :

effectif	$X = 0$	$X = 1$
$Y = \ominus$	a	b
$Y = \oplus$	c	d

(par exemple : dans la base, le nombre de fois où $X = 0$ et $Y = \oplus$ est c).
On note $N = a + b + c + d$ et on suppose a, b, c, d tous strictement positifs.

Q 4.1 Probabilités
Calculer $P(X = 0, Y = \oplus)$, $P(Y = \ominus)$, $P(X = 1|Y = \ominus)$ en fonction de a, b, c, d.

Q 4.2 Régression logistique
On se place dans le cadre de la régression logistique.

Q 4.2.1 Comment peut-on écrire

$$f(x) = \log \frac{P(Y = \oplus|X = x)}{P(Y = \ominus|X = x)}$$

en fonction de 2 paramètres qu'on nommera w_0 et w_1 ?

Q 4.2.2 Calculer w_0 et w_1 en fonction de a, b, c, d.

Q 4.3 Indépendances

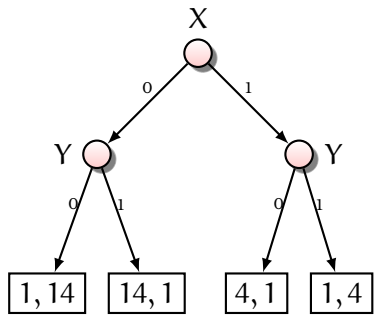
Q 4.3.1 Montrer que si les 2 variables X et Y étaient indépendantes, alors $[a, b] \propto [c, d]$

Q 4.3.2 Que peut-on dire pour les coefficients de la régression logistiques si X et Y sont indépendantes ?

Q 4.4 Test d'indépendances
Dans le cadre d'un test d'indépendance, calculer la valeur de la statistique du χ^2 permettant de vérifier l'indépendance entre les variables X et Y (toujours en fonction de a, b, c, d). Que faut-il que vérifient N, a, b, c, d pour que ce test soit valide ?

Exercice 5 (5 points) – **Arbre de décision**

Soit l'arbre de décision ci-contre. Dans les feuilles, est indiquée la composition de sa région sur les 2 classes $\ominus$ et $\oplus$. Par exemple, dans la région $(X = 0, Y = 0)$, il y a 1 $\ominus$ et 14 $\oplus$.



Q 5.1 Pour chaque nœud (feuilles ou nœuds internes), indiquer la composition de sa région et son entropie. **Q 5.2** Cet arbre est-il optimal en utilisant le critère du gain d'information ?

Q 5.3 On rajouter les variables $U = (X \text{ ou } Y)$ et $V = (X \text{ et } Y)$. Proposer un nouvel arbre de décision (optimal en utilisant le critère du gain d'information) pouvant utiliser X, Y, U et V (S'arrêter a 2 nœuds internes).