Contrôle Terminal

Le contrôle est composé de trois exercices indépendants et est noté sur 20 points. Si un ajustement est nécessaire, votre note finale ne sera pas inférieure à celle calculée en faisant simplement la somme des points.

Ici log est le logarithme népérien.

Exercice 1

Un sondage a receuilli des informations sur le prix (au kilogramme) de certains fruits et legumes. Le tableau suivant donne les effectifs pour chaque paire type / classe de prix.

	pomme	poire	courgette	aubergine
2 €	12	24	54	23
3 €	45	26	72	16
4 €	34	63	34	33

Par exemple, dans le sondage il y a 12 pommes qui coutent 2 € (au kilogramme).

1) (1 point) Calculer la probabilité que le fruit/légume du sondage soit une pomme sachant que son prix est 3 € (Avec justification).

2) (1 point) Calculer la probabilité que le fruit/légume du sondage coute 2 € ou plus sachant que c'est une aubergine (Avec justification).

3) (1 point) Calculer la probabilité que le fruit/légume du sondage soit un fruit sachant qu'il coute 3 € ou moins (Avec justification).

4) (1 point) Calculer l'espérance conditionnelle du prix d'un fruit/légume du sondage sachant que c'est une courgette (Avec justification).

5) (1 point) Calculer la variance conditionnelle du prix d'un fruit/légume du sondage sachant qu'on est un légume et que notre prix est pair (Avec justification).

Exercice 2

On considère un ensemble de points $v_1, \dots, v_N \in \mathbb{R}^d$ (d étant l'entier correspondant à la dimension des points). On définit alors *la variance empirique* comme

$$E(v_1, \dots, v_N) = \frac{1}{N} \sum_{i=1}^{N} (v_i - \hat{\mu})^2$$

où

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} v_i \ .$$

La variance empirique est utilisée comme une mesure de non-uniformité dans les données.

1) (2 points) Montrer que la variance empirique est ≥ 0 et qu'elle est égale à 0 si et seulement si $v_1 = \cdots = v_N$.

2) (1 point) Justifier à l'aide d'une phrase pourquoi la variance empirique est une bonne mesure de non-cohésion.

3) (1 point) Calculer $E(1, 2, 3, 2, 3, 2, 2, 2, 1, 2)$.

On a un jeu de données défini par ce tableau.

i	1	2	3	4	5	6	7	8
$x^{(i)}$	$(0.1, 0.1)$	$(0.2, 0.2)$	$(0.7, 0.2)$	$(0.8, 0.3)$	$(0.2, 0.6)$	$(0.3, 0.7)$	$(0.7, 0.7)$	$(0.8, 0.9)$
y_i	-1	-3	-4	-3	2	4	2	3

On note $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$ pour $i = 1, \dots, 8$.

4) (1 point) Représenter ces points sur un dessin en numérotant bien chacun

5) (2 points) Calculer $E(y_1, \dots, y_8)$.

On cherche à séparer les données $(y_1, \dots, y_8)$ en 2 groupes. Le groupe 1 sera celui des y_i pour lesquels $x_1^{(i)} \leq t$ avec $t = 0,5$ et le groupe 2 sera celui des autres y_i. On note $u_1, \dots, u_k$ les éléments du groupe 1 et $v_1, \dots, v_\ell$ les

éléments du groupe 2 (on a $k+\ell=8$). Calculer

$$\frac{k}{8}E(u_1,\dots,u_k)+\frac{\ell}{8}E(v_1,\dots,v_\ell)$$

qui est la variance empirique moyenne après séparation en 2 groupes. Faire un dessin qui représente cela.

6) (1 point) On cherche à nouveau à séparer les données $(y_1,\dots,y_8)$ en 2 groupes. Cette fois le groupe 1 sera celui des y_i pour lesquels $x_2^{(i)} \leq s$ avec $s=0,5$ et le groupe 2 sera celui des autres y_i. On note $u_1,\dots,u_k$ les éléments du groupe 1 et $v_1,\dots,v_\ell$ les éléments du groupe 2 (on a $k+\ell=8$). Calculer

$$\frac{k}{8}E(u_1,\dots,u_k)+\frac{\ell}{8}E(v_1,\dots,v_\ell).$$

Faire un dessin qui représente cela.

7) (1 point) Quelle séparation en deux groupes préférez-vous dans une optique de régression supervisée, et pourquoi ?

La construction d'un arbre de régression se fait selon le principe des questions précédentes. La fin de cet exercice décrit l'algorithme, mais il n'y a plus de questions.

— Faire la séparation de la question 2, mais cette fois, trouver la valeur de t qui minimise la quantité

$$\frac{k}{8}E(u_1,\dots,u_k)+\frac{\ell}{8}E(v_1,\dots,v_\ell)$$

que l'on notera e_1.

— Ensuite, faire la mème chose mais selon la question 3 en trouvant la valeur de s qui minimise la quantité

$$\frac{k}{8}E(u_1,\dots,u_k)+\frac{\ell}{8}E(v_1,\dots,v_\ell)$$

que l'on notera e_2.

— Garder celle des deux séparations qui correspond à la plus petite valeure entre e_1 et e_2. Cela revient à diviser le carré $[0,1]^2$ en 2 rectangles.

— Dans chacun des deux rectangles faire la même chose que toutes les étapes d'avant (si il y a encore deux classes représentées). A la fin, on a divisé le carré $[0,1]^2$ en 3 ou 4 rectangles. Cela correspond aux premières étapes de construction d'un arbre de régression. Dans chacun des rectangles, on classifie un nouveau x selon la moyenne empirique des points qui sont dans ce rectangle.

Exercice 3

Soit (X,Y) un couple de variables aléatoires telles que $X \in \mathbb{R}^d$ un espace de features et $Y \in \{0,1\}$ est une variable catégorielle.

Pour une fonction de classification $g:\mathbb{R}^d \to \{0,1\}$, nous définissons $L(g)=\mathbb{E}(\mathbb{1}_{g(X)\neq Y})$ le risque théorique de g pour la perte du $0-1$.

1) (2 points) Si g et g' sont deux fonctions de classification, montrer que

$$|L(g)-L(g')| \leq \mathbb{P}(g(X)\neq g'(X))\ .$$

2) (2 points) Si g est une fonction de classification, montrer que

$$L(g)=\mathbb{E}(\mathbb{1}_{g(X)\neq 1}(2\eta(X)-1)+(1-\eta(X)))$$

où $\eta(x)=\mathbb{E}(Y|X=x)=\mathbb{P}(Y=1|X=x)$.

3) (2 points) En déduire que si g et g' sont deux fonctions de classification,

$$|L(g)-L(g')|=\mathbb{E}(\mathbb{1}_{g(X)\neq g'(X)}|2\eta(X)-1|)\ .$$