TD 3
Arbres de classification

Ici log est le logarithme neperien.

Exercice 1

On considère un jeu de données $v_1, \dots, v_N \in \{1, 2, \dots, k\}$ (les nombres symbolisent k classes). On définit alors *l'entropie empirique* comme

$$E(v_1, \dots, v_N) = -\sum_{\ell=1}^{k} \hat{p}_\ell \log(\hat{p}_\ell)$$

en définissant

$$\hat{p}_\ell = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{v_i = \ell}$$

et en utilisant la convention $0 * \log(0) = 0$. Ici log est en base e avec $\log(\exp(1)) = 1$.

L'entropie empirique représente la variabilité des données.

1) Montrer que l'entropie est ≥ 0 et qu'elle est égale à 0 si et seulement si $v_1 = \cdots = v_N$.

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

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

Exercice 2

On a un jeu de données défini par ce tableau.

i	1	2	3	4	5	6	7	8	9	10
$x^{(i)}$	$(0.1, 0.1)$	$(0.2, 0.2)$	$(0.3, 0.8)$	$(0.4, 0.4)$	$(0.5, 0.7)$	$(0.6, 0.3)$	$(0.7, 0.5)$	$(0.8, 0.9)$	$(0.9, 0.6)$	$(1, 1)$
y_i	1	2	1	1	1	2	2	2	2	2

On note $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$ pour $i = 1, \dots, 10$.

1) Calculer $E(y_1, \dots, y_{10})$.

2) On cherche à séparer les données $(y_1, \dots, y_{10})$ en 2 groupes. Le groupe 1 sera celui des y_i pour lesquels $x_1^{(i)} \leq t$ avec $t = 0.45$ 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 $k + \ell = 10$). Calculer

$$\frac{k}{10} E(u_1, \dots, u_k) + \frac{\ell}{10} E(v_1, \dots, v_\ell)$$

qui est l'entropie empirique moyenne après séparation en 2 groupes. Faire un dessin qui représente cela.

3) On cherche à nouveau à séparer les données $(y_1, \dots, y_{10})$ en 2 groupes. Cette fois le groupe 1 sera celui des y_i pour lesquels $x_2^{(i)} \leq s$ avec $s = 0.45$ 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 $k + \ell = 10$). Calculer

$$\frac{k}{10} E(u_1, \dots, u_k) + \frac{\ell}{10} E(v_1, \dots, v_\ell).$$

Faire un dessin qui représente cela.

4) Quelle séparation en deux groupes préférez-vous dans une optique de classification supervisée, et pourquoi ?

5) La construction d'un arbre de classification se fait selon le principe des questions précédentes. Nous allons illustrer cela en quelques étapes ici (faire un dessin à chaque étape).

1. Matériel de base créé par François Bachoc.

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

$$\frac{k}{10}E(u_1,\ldots,u_k)+\frac{\ell}{10}E(v_1,\ldots,v_\ell)$$

que l'on notera e_1.

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

$$\frac{k}{10}E(u_1,\ldots,u_k)+\frac{\ell}{10}E(v_1,\ldots,v_\ell)$$

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 classification. Dans chacun des rectangles, on classifie un nouveau x selon la classe qui est majoritaire dans le rectangle, parmi les 10 données d'apprentissage.