Texte. Arbres de décision

Exercice 1 Traitement du texte

On considère les 3 documents texte suivants :

1. document 1 : Le métro est souvent plus pratique et rapide que le vélo
2. document 2 : Le vélo est pratique quand il fait beau
3. document 3 : Il fait beau et souvent chaud en mai

Question 1. Prétraiter chacun de ces documents en éliminant les mots courants (inutiles) : les articles, les verbes courants conjugués, les adverbes (quand, souvent...), les conjonctions,... Puis énoncer un index commun à ces 3 documents pour les représenter sous la forme de vecteur.

Question 2. Pour comparer des textes, on utilise très souvent la "distance cosinus". Cette distance entre 2 textes représentés par des vecteurs est définie par :

$$d_{cos}(t_1, t_2) = 1 - \frac{\langle t_1, t_2 \rangle}{||t_1|| \cdot ||t_2||}$$

Calculer les distances séparant chacun des 3 documents.

Question 3. Ecrire la fonction python qui, étant donné 2 numpy arrays calcule leur distance cosinus.

Question 4. Les 2 premiers documents sont associés à la classe "transport", le dernier document est associé à la classe "climat". On va appliquer l'algorithme Naive Bayes vu en cours pour pouvoir classer des documents.

1. calculer la probabilité de chacune des classes : $p(Y)$
2. calculer la probabilité d'appartenance de chaque mot m de l'index à chacune des 2 classes : $p(m|Y)$
3. pour chaque document d, calculer la probabilité

$$p(d|Y) = \prod_{m \in index} p(m|Y)^{d_m} \, (1 - p(m|Y))^{1-d_m}$$

   où d_m vaut 1 si m appartient à d et 0 sinon.
4. en appliquant la formule de Bayes, calculer $p(Y|d)$ pour les 2 classes et les 3 documents. Rappel : $p(Y|d) = p(Y)p(d|Y)$.

Exercice 2 Algorithme ID3

On considère la base d'apprentissage suivante décrivant des patients d'un ophtalmologue.

Âge	Prescription	Astigmatisme	Classe
jeune	myope	non	pas de lentilles
jeune	hypermétrope	oui	lentilles
âgé	myope	non	lentilles
âgé	myope	oui	pas de lentilles

Question 1. On souhaite construire un arbre de décision pour prédire si un patient doit porter des lentilles. Indiquer le test qui est sélectionné si on utilise l'entropie de Shannon.

Question 2. On souhaite construire un arbre parfait (c'est-à-dire avec des feuilles ne contenant qu'une classe). Indiquer s'il est nécessaire de faire des tests supplémentaires et, le cas échéant, les déterminer et donner l'arbre complet obtenu.

Question 3. Existe-t-il un arbre parfait de taille inférieure ?

Exercice 3 Apprentissage par arbres de décision On considère la base d'apprentissage suivante.

X	Y	Z	Classe
x_1	y_1	z_1	c_1
x_1	y_1	z_2	c_2
x_1	y_2	z_2	c_2
x_2	y_1	z_2	c_1
x_2	y_1	z_1	c_2
x_2	y_2	z_1	c_1

Question 1. Rappeler comment est déterminé le gain d'un attribut à partir de l'entropie pondérée par le nombre d'éléments et expliquer à quoi correspond cette formule.

Question 2. Lors de la construction d'un arbre de décision, si le gain d'information pour chacun des attributs est nul, alors le développement de l'arbre s'arrête. Expliquer pourquoi. Que risque-t-on à continuer le développement ?

Question 3. En détaillant les calculs réalisés, construire l'arbre de décision obtenu par l'application de l'algorithme ID3 sur cette base d'apprentissage. (rem : des valeurs d'entropie sont fournies en annexe).

Question 4. Comment l'algorithme de construction de l'arbre doit-il être adapté dans le cas où l'attribut `X` est un attribut numérique dont les valeurs sont pour chaque ligne, respectivement, 7.1, 6.3, 0, 7.0, 5.2 et 4.4 ? Expliquer.

Annexes

n / d		dénominateur (d)									
		1	2	3	4	5	6	7	8	9	10
numérateur (n)	1	1	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11	0,10
	2		1	0,67	0,50	0,40	0,33	0,29	0,25	0,22	0,20
	3			1	0,75	0,60	0,50	0,43	0,38	0,33	0,30
	4				1	0,80	0,67	0,57	0,50	0,44	0,40
	5					1	0,83	0,71	0,63	0,56	0,50
	6						1	0,86	0,75	0,67	0,60
	7							1	0,88	0,78	0,70
	8								1	0,89	0,80
	9									1	0,90
	10										1

-(n/d) * log(n/d)		dénominateur (d)									
		1	2	3	4	5	6	7	8	9	10
numérateur (n)	1	0	0,50	0,53	0,50	0,46	0,43	0,40	0,38	0,35	0,33
	2		0	0,39	0,50	0,53	0,53	0,52	0,50	0,48	0,46
	3			0	0,31	0,44	0,50	0,52	0,53	0,53	0,52
	4				0	0,26	0,39	0,46	0,50	0,52	0,53
	5					0	0,22	0,35	0,42	0,47	0,50
	6						0	0,19	0,31	0,39	0,44
	7							0	0,50	0,28	0,36
	8								0	0,15	0,26
	9									0	0,14
	10										0