Université de Médéa **Mai 2017**
Faculté des sciences
Départ : Mathématiques & Informatique
M1-Informatique (ISTW) **Durée : 1h30**

Rattrapage –Apprentissage Automatique (ML)-

(Documents et téléphones portables non autorisés)

Exercice01 :(5 points)

En hiver le temps à Médéa est notoirement inconstant. Pour simplifier, nous considérons seulement le soleil et la pluie et nous supposons que le temps change une fois par jour. La météo satisfait les probabilités de transition suivantes:

- Quand il pleut, la probabilité de soleil le lendemain est de 0.6.
- Lorsque le soleil brille, la probabilité de pluie le lendemain est de 0,3.

Q1. Donner la matrice de transition d'état T pour la météo

	p	s
p		
s		

Supposons que nous observons la météo sur une période de dix jours. En particulier, nous observons ce qui suit:

- Le soleil brille le premier jour.
- Il pleut le jour 5.
- Il pleut le jour 7.
- Le soleil brille le jour 10.

Q2. Quelle est la probabilité de soleil au jour 6??
Q3. Quelle est la séquence météorologique la plus probable aux jours 8 et 9 ?

Exercice 02 : :(5 points)

L'image ci-dessous montre un ensemble de huit morceaux Scrabble.

Q1. Quelle est l'entropie de sélectionner en bites les lettres dans cet ensemble?

Q2. Quelle serait la réduction de l'entropie (c'est-à-dire le gain d'information) en bits si on divise ces lettres en deux ensembles, l'un contenant les voyelles et l'autre contenant les consonnes?

Exercice03 : :*(5 points)*

Les modèles de filtrage de spam par courrier électronique utilisent souvent une représentation de sacs de mots pour les emails. Dans une représentation bag-of-words (sac de mots), les caractéristiques descriptives qui décrivent un document (dans notre cas, un email) représentent chacun le nombre de fois qu'un mot particulier se trouve dans le document. Une caractéristique descriptive est incluse pour chaque mot dans un dictionnaire prédéfini. Le dictionnaire est généralement défini comme l'ensemble complet des mots qui se produisent dans l'ensemble de données de formation.

Le tableau ci-dessous répertorie la représentation des sacs de mots pour les cinq emails suivants et une fonctionnalité cible, qu'ils soient des spams ou des emails authentiques:

- "money, money, money"
- "free money for free gambling fun"
- "gambling for fun"
- "machine learning for fun, fun, fun"
- "free machine learning"

	Bag-of-Words							
ID	MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	SPAM
1	3	0	0	0	0	0	0	true
2	1	2	1	1	1	0	0	true
3	0	0	1	1	1	0	0	true
4	0	0	1	0	3	1	1	false
5	0	1	0	0	0	1	1	false

Q1: quel est le type du courrier électronique suivant: «machine learning for free» en utilisant le plus proche voisin (3NN) avec la distance Euclidienne

Q2 : même question 5NN avec la pondération de la distance Euclidienne, pondération égale à l'inverse de la distance euclidienne au carré.

Q3 : Refaire la Q1 en utilisant « Manhattan distance »

Exercice04 : :*(5 points)*

Ecrire un programme python qui calcule la propagation avant d'un réseau de neurone de 3 entrées et une sortie.

بالتـــوفيـــــق
Mazari A.C

Corrigé ML Rattrapage1

Exercice01:

Q1 :-

	p	s
p	0.4	0.6
s	0.3	0.7

Q2:

Given day 5 and day 7 are rainy ;

P(p6/p5,p7)=0.4*0.4=0.16

P(s6/p5,p7)=0.6*0.3=0.18

P(p6)=$\frac{0.18}{0.16+0.18} = \frac{9}{17}$

Q3 :

Given day 7 is rainy and day 10 is sunny, enumerate all possible sequence and evaluate their likelihoods.

P(p8,p9/p7,s10)=0.4*0.4*0.6=0.096

P(p8,S9/p7,s10)=0.4*0.6*0.7=0.168

P(s8,p9/p7,s10)=0.6*0.3*0.6=0.108

P(s8,S9/p7,s10)=0.6*0.7*0.7=0.294

Thereone the most likely sequence is s8→ s9

Exercice02:

Q1 :

We can calculate the probability of randomly selecting a letter of each type from this set:

$$P(O) = \frac{3}{8}, P(X) = \frac{1}{8}, P(Y) = \frac{1}{8}, P(M) = \frac{1}{8}, P(R) = \frac{1}{8}, P(N) = \frac{1}{8}.$$

Using these probabilities, we can calculate the entropy of the set:

$$-\left(\frac{3}{8} \times log_2\left(\frac{3}{8}\right) + \left(\frac{1}{8} \times log_2\left(\frac{1}{8}\right)\right) \times 5\right) = 2.4056\ bits$$

Note that the contribution to the entropy for any letter that appears only once is the same and so has been included 5 times—once for each of X, Y, M, R, and N.

Q2:

Information gain is the reduction in entropy that occurs after we split the original set. We know that the entropy of the initial set is 2.4056 bits. We calculate the remaining entropy after we split the original set using a weighted summation of the entropies for the new sets.
The two new sets are vowels {O,O,O} and consonants {X,Y,M,R,N}.

The entropy of the vowel set is

$$-\left(\frac{3}{3}\times log_2\left(\frac{3}{3}\right)\right)=0\ bits$$

The entropy of the consonant set is

$$-\left(\left(\frac{1}{5}\times log_2\left(\frac{1}{5}\right)\right)\times 5\right)=2.3219\ bits$$

The weightings used in the summation of the set entropies are just the relative size of each set. So, the weighting of the vowel set entropy is 3/8, and the weighting of the consonant set entropy is 5/8.

This gives the entropy remaining after we split the set of letters into vowels and consonants as

$$rem=\frac{3}{8}\times 0+\frac{5}{8}\times 2.3219=1.4512\ bits$$

The information gain is the difference between the initial entropy and the remainder:

$$IG=2.4056-1.4512=0.9544\ bits$$

Exercice03:

ID	MONEY	FREE	FOR	$(q[i[-d_j[i[)^2$ GAMBLING	FUN	MACHINE	LEARNING	Euclidean Distance
1	9	1	1	0	0	1	1	3.6056
2	1	1	0	1	1	1	1	2.4495
3	0	1	0	1	1	1	1	2.2361
4	0	1	0	0	9	0	0	3.1623
5	0	0	1	0	0	0	0	1

Q1:

Based on the distance calculations in part (a) of this question, the three nearest neighbors to the query are instances d5, d3, and d2. The majority of these three neighbors have a target value of SPAM = true. Consequently, the 3-NN model will return a prediction of **SPAM = true**.

Q2:

The weights for each of the instances in the dataset are

ID	Weights	
1	$\frac{1}{3.6056^2}$	$= 0.0769$
2	$\frac{1}{2.4495^2}$	$= 0.1667$
3	$\frac{1}{2.2361^2}$	$= 0.2$
4	$\frac{1}{3.1623^2}$	$= 0.1$
5	$\frac{1}{1^2}$	$= 1$

The total weight for the SPAM = true target level is 0.0769+0.1667+0.2 = 0.4436.
The total weight for the SPAM = false target level is 0.1+1 = 1.1. Consequently, the **SPAM = false** has the maximum weight, and this is the prediction returned by the model.

Q3.

The table below shows the calculation of the Manhattan distance between the query bag-of-words vector and each instance in the dataset:

ID	$abs(\mathbf{q}[i[- \mathbf{d}_j[i[)$ MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	Manhattan Distance
1	3	1	1	0	0	1	1	7
2	1	1	0	1	1	1	1	6
3	0	1	0	1	1	1	1	5
4	0	1	0	0	3	0	0	4
5	0	0	1	0	0	0	0	1

Based on these Manhattan distance calculations, the three nearest neighbors to the query are instances d5, d4, and d3.

The majority of these three neighbors have a target value of SPAM = false. Consequently, the 3-NN model using Manhattan distance will return a prediction of SPAM = false.