Intelligence artificielle – Exercices – Devoirs

Exercice 1 corrigé disponible

Avant d'être mesurées par des puissances de 10 d'octets (ko, Mo, Go, etc), les données informatiques étaient mesurées en puissances de 2 d'octets. Pendant longtemps, un kilooctet a désigné 2^{10} octets, un mégaoctet a désigné 2^{20} octets, un gigaoctet a désigné 2^{30} octets et ainsi de suite. Cependant, cette tradition propre au domaine de l'informatique entrait en conflit avec les normes internationales (selon lesquelles 1 kilo correspond à 10^3, un méga à 10^6, etc). Ainsi, en 1988, la Commission électrotechnique internationale a normalisé les unités en introduisant des préfixes spécifiques pour les puissances de 2. Ainsi, sont apparus les termes *kilo binaire*, *méga binaire*, *giga binaire*, *téra binaire* et *péta binaire* abrégé en kibi, mébi, gibi, tébi et pébi. On a alors le tableau suivant :

unité	kibioctet (Kio)	mébioctet (Mio)	gibioctet (Gio)	tébioctet (Tio)	pébioctet (Pio)
en octets	2^{10}	2^{20}	2^{30}	2^{40}	2^{50}

1. Les premières disquettes 3'½ construites par Sony avaient une capacité de 400 Kio. Déterminer la capacité en ko d'une telle disquette.
2. Un disque dur S-ATA Hitachi de fin 2005 avait une capacité de stockage de 76,688 Gio. Convertir cette capacité en Go.

Exercice 2 corrigé disponible

1. Quelle la taille en octet d'un fichier texte codé en ASCII et contenant le texte suivant ?

```
Je dois déterminer la taille de ce texte.
Pour cela, je ne dois pas oublier les espaces et la ponctuation,
ni les retours à la ligne.
```

2. Un fichier texte codé en ASCII compte 12 lignes. Chaque ligne compte 30 caractères (espaces et ponctuation compris). Quelle est la taille en octet de ce fichier ?
3. Un fichier texte codé en ASCII a un taille de 236 ko. Déterminer le nombre maximum de caractères qu'il peut contenir.

Exercice 3 corrigé disponible

Dans un dossier, on trouve les fichiers suivants :

`fichier1.jpg` `fichier2.txt` `fichier3.mov` `fichier4.exe` `fichier5.png`
`fichier6.mp4` `fichier7.doc` `fichier8.avi` `fichier9.wav` `fichier10.mp3`

Regrouper ces fichiers en 5 catégories : fichiers texte, fichiers image, fichier son, fichiers vidéo et fichiers exécutables.

Exercice 4 corrigé disponible

On dispose 3 fichiers `fichier1`, `fichier2` et `fichier3`. On sait que la taille de `fichier1` est 840 Mo, la taille de `fichier2` est 53 Mo et la taille de `fichier3` est 14 ko. On sait également que les extensions de ces fichiers sont .txt, .wav et .avi.

En se référant aux ordres de grandeurs standards, déterminer l'extension de chaque fichier.

Exercice 5 corrigé disponible

Lorsqu'on numérise une image en niveau de gris, on stocke cette image sous la forme d'un tableau constitué de petits carrés appelés pixels. À chaque pixel, on va associer un nombre entre 0 et 255 correspondant à un certain niveau de gris comme sur l'image suivante :

Image A

Chaque nombre est ensuite codé sur un octet comme dans le cas du code ASCII.

1. L'image A a un résolution de 240 par 240, c'est-à-dire qu'elle correspond à un tableau ayant 240 lignes et 240 colonnes. Combien de pixels constituent cette image ? Déterminer sa taille en ko.

2. Pour gagner de la place, on peut soit diminuer le nombre de pixels soit diminuer le nombre de niveaux de gris.

Image B

Image C

a. Déterminer la taille de l'image B sachant qu'elle a été obtenue à partir l'image A en ne conservant qu'une ligne sur deux et qu'une colonne sur deux.

b. Déterminer la taille de l'image C sachant qu'elle a été obtenue à partir l'image A en ne conservant que 16 niveaux de gris au lieu de 256.

Exercice 6

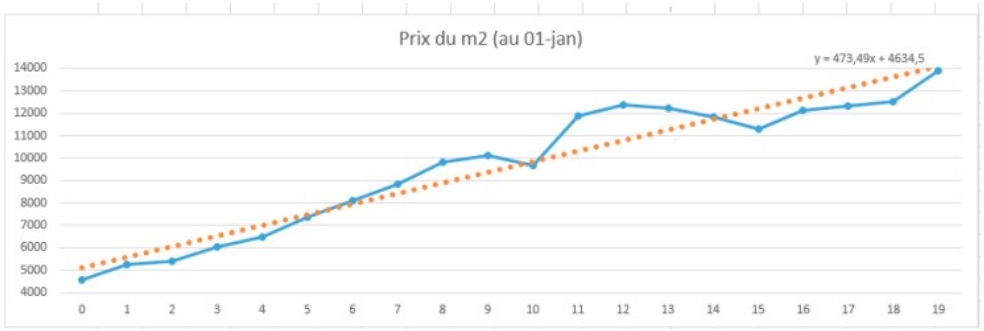
1. Ecrire une fonction Python **def** lire(fichier) qui réalise la lecture d'un fichier *txt
2. Ecrire une fonction Python **def** ecrire(fichier) qui réalise l'enregistrement d'une chaîne de caractère dans un fichier *txt
3. Ecrire une fonction Python **def** ascii(chaine) qui réalise la conversation d'un mot binaire vers des caractères ascii

Exercice 7

Le tableau suivant donne le prix moyen du m^2 dans le 6e arrondissement de Paris au 1er janvier entre 2000 et 2019. Ces données sont représentées sur le graphique en dessous et on a également tracé en pointillés la droite d'ajustement de ces données.

Evolution du prix du mètre carré dans le 6ème arrondissement de Paris (année 0 en 2000)

Année (x_i)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Prix du m2 y_i (au 01-jan)	4562	5270	5400	6020	6460	7360	8090	8840	9830	10100	9690	11870	12400	12250	11820	11280	12150	12320	12530	13880



Donner une estimation du prix du mètre carré dans le 6e arrondissement au 1er janvier 2025.

Exercice 8

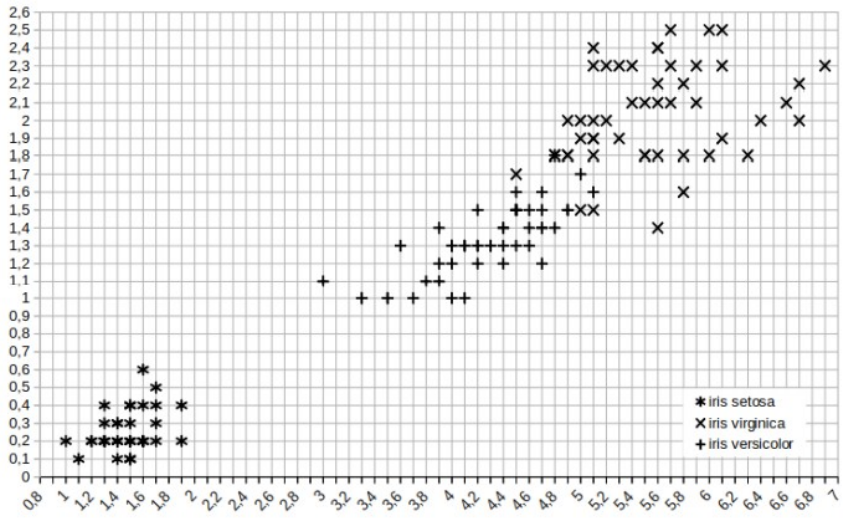
On se propose de répartir en deux catégories (Maligne, Bénigne) des tumeurs dont on connaît un certain nombre de caractéristiques. Dans la réalité, un tel diagnostic automatique est effectué à partir de données d'apprentissage portant sur une cinquantaine de caractéristiques mesurées sur un échantillon d'un millier de tumeurs déjà étiquetées « Maligne » ou « Bénigne ». Dans un souci de simplification, on se restreint ici à 2 caractéristiques (diamètre et concavité) mesurées sur un panel de 10 patientes.

Diamètre moyen (en mm)	Concavité moyenne	Catégorie
13,2	8,3	Bénigne
18,7	19,7	Bénigne
8,2	15,9	Maligne
13,2	9	Bénigne
13,5	4,8	Maligne
11,8	1,7	Maligne
13,6	1,9	Maligne
12	2	Maligne
18,2	17,7	Bénigne
12	6,6	Maligne

1. À quel type d'apprentissage a-t-on affaire ici ?
2. Placer les points correspondant aux données du tableau dans un repère (on mettra le diamètre en abscisse et la concavité en ordonnées).
3. En utilisant la méthode du plus proche voisin, une tumeur de 10 mm et ayant une concavité de 12 doit-elle être considérée comme bénigne ou maligne ?

Exercice 9

En 1936, Edgar Anderson a collecté des données sur 3 espèces d'iris : iris setosa, iris virginica et iris versicolor. Pour chaque espèce, Anderson a mesuré (en cm) différents paramètres. Sur le graphique ci-dessous, on a représenté la longueur (en abscisse) et la largeur (en ordonnée) mesurées lors de ces relevés.



Déterminer à quelle espèce appartient un iris dont les pétales mesure 2,4 cm de long et 0,8 cm de large :

1. en utilisant le méthode du plus proche voisin ;
2. en utilisant la méthode des 5 plus proches voisin.

Exercice 10

Chercher des arguments et des situations montrant l'importance du choix des données d'entraînement pour une IA.

Chercher des arguments et des situations montrant les problèmes juridiques et éthiques posés par l'IA.

Exercice 11

Le virus de l'immmunodéficience humaine (VIH) est responsable du SIDA. Il existe aujourd'hui des tests rapides appelés « Tests rapides d'orientation diagnostique » (TROD) qui ont l'avantage de pouvoir être réalisés à partir d'un échantillon de salive

1. On a testé avec le TROD deux populations : une première composée de 10 000 personnes infectées par le VIH et une seconde composée de 100 000 personnes non infectées. On a obtenu les résultats suivants.

	Personnes infectées	Personnes non infectées
Tests salivaires positifs	9 803	260

Calculer la sensibilité et la spécificité de ce test

2. Calculer la valeur prédictive positive (VPP) et la valeur prédictive négative de ce test (VPN).

Exercice 12

Parmi les femmes de 40 ans ayant effectué une mammographie, 1% a un cancer du sein. À la suite de mammographies sur un échantillon, on a établi que :

- pour 82% des femmes ayant un cancer du sein, la mammographie détecte une anomalie ;
- pour 9% des femmes n'ayant pas de cancer du sein, la mammographie détecte une anomalie.

On suppose que 10 000 femmes de 40 ans ont effectué une mammographie.

1. Déterminer la sensibilité et la spécificité d'une mammographie.
2. Compléter le tableau suivant.

	Anomalie détectée	Pas d'anomalie détectée	Total
Personnes malades			
Personnes non malades			
Total			10 000

3. Une femme de 40 ans a effectué une mammographie qui a permis de détecter une anomalie. Quelle est la probabilité qu'elle soit atteinte d'un cancer du sein ?
4. Calculer les valeurs prédictives positive et négative d'une mammographie chez les femmes de 40 ans.

Exercice 13

La dengue est une maladie virale transmise à l'être humain par un moustique du genre Aedes. Ses symptômes les plus fréquents sont de la fièvre et des douleurs articulaires. Originaire des régions tropicales, la dengue a fait son apparition en France métropolitaine en 2010 et progresse depuis (51 départements touchés en 2019 selon Santé Publique France).

On s'intéresse aux méthodes de dépistage et de prévention de cette maladie.

Partie 1- Le dépistage de la dengue dans une population humaine.

Tout test de dépistage est caractérisé par :

- sa sensibilité : probabilité qu'un test soit positif quand la personne est atteinte ;
- sa spécificité : probabilité qu'un test soit négatif quand une personne n'est pas atteinte (on dit aussi que la personne est saine).

Un test de dépistage de la dengue est basé sur la détection de l'antigène NS1 dans le sang. La notice du test indique que sa sensibilité est de 97,7 %.

Document 1 : tableau de contingence pour le test de détection de l'antigène NS1.

	Personnes atteintes de la dengue	Personnes saines	Effectif total
Test positif			
Test négatif		8 990	
Effectif total	365	9 635	10 000

Source : Haute autorité de santé

1- Calculer, à partir du tableau de contingence, la spécificité du test de dépistage de la dengue.

2- Recopier et compléter le tableau de contingence (arrondir au besoin à l'unité).

3- Une personne vient de se faire tester et son résultat est positif, calculer la probabilité que cette personne soit effectivement atteinte de la dengue.

Exercice 14

Le 10 Juillet 2020, une application de streaming musical a été perturbée par un problème de bug logiciel.

1- Après avoir rappelé ce qu'est un bug, indiquer ses conséquences sur un programme informatique.

Au moment de se connecter au service de streaming musical, on proposait à l'utilisateur de se connecter soit avec le réseau social R, soit avec un compte de messagerie M, soit en s'inscrivant à l'aide d'un autre compte.
Le résultat du choix de l'utilisateur est stocké dans la variable « resultatclic », puis est passé en paramètre de la fonction prête-à-l 'emploi « connexionavec( ) ».

Voici un extrait de l'algorithme qui devait permettre de gérer cette opération. Cependant l'algorithme ne pouvait pas fonctionner car cet extrait contient un ou des bugs.

```
L1  if resultatclic == "R":
L2      connexionavec(R)
L3  else resultatclic == "M":
L4      connexionavec(M)
L5  else resultatclic == "autre compte":
L6      connexion(autre_compte)
```

2- Pointer le(s) bug(s) en citant la (ou les) ligne(s) suspecte(s) et en la (ou les) réécrivant.

Chaque fois qu'un utilisateur se connecte à cette application de streaming musical en utilisant un compte R, un fichier texte est enregistré sur les serveurs de ce dernier. Il indique le jour et l'heure de sa connexion, son identifiant, le lieu où il se trouve et le système d'exploitation qu'il utilise.

Voici un exemple de fichier enregistré, il contient 30 caractères :

```
08/12/2020
8 pm
Élise
Paris
Système
```

En moyenne, pour chaque utilisateur, le fichier texte enregistré a la taille du fichier texte donné en exemple.
Le réseau R compte 2,7 milliards d'utilisateurs. Dans la même journée 3% d'entre eux se connectent à cette application de streaming musical en utilisant leur compte R.
3- Calculer la taille moyenne de l'ensemble des fichiers textes enregistrés sur le serveur durant cette journée, liés à la connexion à cette application.

Cette application possède une intelligence artificielle, notée IA, que l'on souhaite entraîner afin qu'elle identifie les goûts musicaux des utilisateurs. Par exemple, on décide de l'entraîner pour identifier un utilisateur qui écoute ou qui n'écoute pas du rap.

4- Choisir en le justifiant, parmi les deux jeux de données proposés, celui qui permettra à l'intelligence artificielle de distinguer un utilisateur écoutant du rap, d'un autre utilisateur.

1er jeu de données
Rap conscient
Reggae
Rock
Rap égotrip
Rap poétique
Rap hardcore
Jazz
Rap commercial
Blues

2ème jeu de données
Rap poétique
Jazz
Rap conscient
Blues

Après avoir fourni un grand nombre de profils d'utilisateurs d'entraînement à l'intelligence artificielle, ses résultats sont les suivants :
Sur 100 utilisateurs écoutant du rap, l'IA a reconnu le profil utilisateur de 98 d'entre eux.
Sur 150 utilisateurs n'écoutant pas de rap, l'IA n'a pas reconnu le profil utilisateur de 5 d'entre eux.

5- Recopier et compléter le tableau de contingence associé à cette expérience à cette étape de l'entraînement.

		Réponse de l'IA		
		Rap	Autres styles	Total
Réponse de l'utilisateur	Rap			
	Autres styles			
	Total			

Un nouvel utilisateur est présenté à l'IA. L'IA qualifie ce nouvel utilisateur d'amateur de Rap.

6- Calculer la probabilité, arrondie au centième, que ce résultat de l'IA soit correct.

Exercice 15

Un groupe de lycéens discute de l'intérêt d'acheter et de pratiquer un autotest de dépistage du VIH vendu sans ordonnance en pharmacie. Ils décident de consulter la notice disponible sur Internet.

Document 1 : extrait de la notice d'un autotest de détermination du VIH

Performances diagnostiques du test :

Sensibilité = probabilité d'un résultat positif du test chez un patient malade (infecté par le VIH)	96,70 %
Spécificité = probabilité d'un résultat négatif du test chez un patient non-malade (non infecté par le VIH)	99,42 %

Prévalence (probabilité qu'une personne soit malade dans la population) du VIH en France : 0,30 %

Document 2 : Tableau de contingence pour un groupe de 10 000 personnes de la population française testées avec l'autotest de détermination du VIH du document 1.

	Malade	Non malade
Test positif	29	58
Test négatif	1	9912

1. Sur les 10 000 personnes testées dans le document 2, combien sont des « vrais positifs » ? Combien sont des « faux positifs » ?

2. En déduire, pour le groupe testé, la fréquence de vrais positifs, c'est-à-dire le pourcentage de personnes réellement malades parmi les résultats positifs au test.
3. Montrer que seules 0,01 % des personnes ayant un résultat négatif au test sont en réalité malades (fréquence de faux négatifs).

4. En Afrique du Sud, la prévalence du VIH est de 18,9 % : sur un groupe de 10 000 personnes, combien sont malades ?

5. Recopier et compléter le tableau de contingence pour ce groupe de 10 000 personnes de la population sud-africaine testées avec l'autotest de détermination du VIH du document 1 (on arrondira les résultats à l'unité).

	Malade	Non malade
Test positif		
Test négatif		

6. Montrer que la fréquence de vrais positifs, c'est-à-dire le pourcentage de personnes réellement malades quand le test est positif, est supérieure à 97 % en Afrique du Sud.

7. Comparer les fréquences de vrais positifs entre la France et de l'Afrique du Sud, en lien avec la prévalence du VIH dans les populations considérées.

8. En France, on recommande de réserver la pratique de ces autotests aux personnes ayant eu une situation à risques (rapport sexuel non protégé, exposition au sang, …) pour lesquelles la prévalence est alors plus forte. Expliquer cette recommandation.

Exercice 16

Identifier les différents types d'apprentissage automatiques :

Il y a deux grandes familles d'algorithmes d'apprentissage : l'apprentissage "supervisé" ou et l'apprentissage "non supervisé".

Pour comprendre la différence, on peut prendre un exemple : on a une nouvelle base de photos à catégoriser. On a des données d'exemples (training set) préalables pour entraîner le modèle.

En apprentissage supervisé, on récupère des données dites annotées de leurs sorties pour entraîner le modèle, c'est-à-dire qu'on a déjà associé un label et on veut que l'algorithme devienne capable, après entraînement, de prédire ce label sur de nouvelles données non annotées.

En apprentissage non supervisé, les données d'entrées ne sont pas annotées. L'algorithme d'entraînement doit dans ce cas trouver seul les similarités et distinctions au sein des données, et effectuer les regroupements judicieux.

Comment savoir quel type d'apprentissage automatique utiliser ?

Dans le cas où le problème posé est tel qu'on peut d'ores et déjà déterminer précisément pour chaque situation la sortie, on utilise l'apprentissage supervisé.
Dans le cas où on doit mieux comprendre les situations auxquelles on est confronté ou bien si on doit identifier des comportements intéressants, on utilise l'apprentissage non supervisé

Voici un exemple.

Des chercheurs de Google Brain ont appliqué des algorithmes d'apprentissage il y a quelques années à des vidéos YouTube, afin de voir ce que cet algorithme réussirait à apprendre comme information.

Aujourd'hui la technologie du projet Google Brain est utilisée dans le système de reconnaissance vocale du système d'exploitation Android, dans la recherche de photos de Google + et pour les recommandations vidéos de YouTube.

De quel type sont les algorithmes d'apprentissage utilisé pour google Brain ?