Fiche TD N°2

Méthodes de l'IA pour les NIDS

Questions de cours :

1. Un NIDS, ou système de détection d'intrusion réseau, peut être basé sur l'apprentissage automatique (Machine Learning) ou ne pas l'être. Quelle est la principale différence entre les deux ?
2. C'est quoi le rôle de la phase d'entraînement ?
3. Qu'est-ce qui permet de résoudre le problème de biais dans les données à passer en entrée dans un système d'apprentissage automatique ?
4. Expliquer en quoi la réduction de dimensionnalité est importante dans les NIDS.

Exercice 01 :

Un arbre de décision a été utilisé pour détecter les attaques réseau à partir d'un jeu de données de trafic. L'arbre est le suivant :

Nœud racine :

- *Caractéristique : Nombre de connexions /s*
- *Seuil : 20*
- *Groupe 1 (Gauche) : Trafic Normal*
- *Groupe 2 (Droite) : Attaque*

Nœud de gauche (Groupe 1) :

- *Caractéristique : Taille des paquets (en octets)*
- *Seuil : 800*
- *Groupe 1 (Gauche) : Trafic Normal*
- *Groupe 2(Droite) : Attaque*

Nœud de droite (Groupe 2) :

- *Caractéristique : Protocole*
- *Seuil : HTTP*
- *Groupe 1 (Gauche) : Trafic Normal*
- *Groupe 2 (Droite) : Attaque*

1. Analyser l'arbre de décision fourni et décrire comment il prend des décisions pour classer le trafic réseau.
2. Si une nouvelle observation arrive avec les caractéristiques suivantes, expliquer comment elle serait classée :
 - Nombre de connexions par seconde : 15
 - Taille des paquets : 900 octets
 - Protocole : HTTPS
3. Quelles seraient les étapes pour améliorer la performance de cet arbre de décision dans la détection d'attaques réseau ?

Exercices :

Les modèles de filtrage de spam par courrier électronique utilisent souvent une représentation de sacs de mots (bag-of-words) pour les emails. Dans une représentation bag-of-words, les caractéristiques descriptives qui décrivent un document (dans ce cas, un email) représentent chacune 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"

	Sac de mots							Spam
	Money	Free	For	Gambling	Fun	Machine	Learning	
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

Exercice 1 : kNN

1. Quel est le type du courrier « *machine learning for free* » en utilisant l'algorithme k plus proches voisins (kNN) avec $k = 3$, avec la distance Euclidienne.
2. Refaire la question (1) en utilisant la distance de "Manhattan".

NB.

$$d_{Euclidienne}(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

$$d_{Manhattan}(p, q) = |q_1 - p_1| + |q_2 - p_2| + \dots + |q_n - p_n|$$

Exercice 2 : Apprentissage Bayésien

1. Construire un modèle de décision bayésien en utilisant l'estimateur de Laplace.
2. Calculer la décision obtenue par ce modèle pour le courrier de l'exercice précédent.

Exercice 3 : Réseau de Neurones Artificiels

Quel est le type du courrier « *machine learning for free* » en utilisant un réseau de neurones artificiels.

Exercice 4 : Méthode ensemble

Trouver le type du courrier « *machine learning for free* » en utilisant la méthode de combinaison *Bagging* en utilisant les trois modèles construits dans les exercices 1, 2 et 3.