INF8215 - Intelligence artif.: méthodes et algorithmes

TP3 - Phishing

Remise le 13 décembre sur Moodle.

1 Introduction

Selon Wikipedia, « le phishing est la tentative frauduleuse d'obtenir des informations ou des données sensibles, telles que des noms d'utilisateur, des mots de passe et des détails de carte de crédit, en se déguisant en entité digne de confiance dans une communication électronique. Généralement réalisé par l'usurpation d'adresse électronique, la messagerie instantanée et la messagerie textuelle, le phishing incite souvent les utilisateurs à saisir des informations personnelles sur un faux site web qui correspond à l'apparence du site légitime. » Aujourd'hui, cette menace commence à être courante. Wandera [2] a indiqué que des nouveaux sites internet de *phishing* sont créés toutes les 20 secondes. De plus, les criminels ont amélioré leurs attaques : de nombreux sites de *phishing* utilisent le protocole SSL [1]. Il est donc crucial de concevoir une technique permettant de détecter automatiquement les attaques de *phishing*.

2 Instructions

Pour ce TP, vous participerez à un concours InclassKaggle dont le but est de développer une approche d'apprentissage automatique (ML) pour détecter si une adresse (url) correspond à un site de *phishing* ou légitime. Pour cette classification binaire, vous avez accès à 87 attributs sur le lien. Vous pouvez trouver plus d'informations sur les données à cette adresse `https://www.kaggle.com/c/tp3-inf8215-a20/data`.

Vous êtes libre d'utiliser les méthodes et les bibliothèques de ML de votre choix. La seule contrainte est le langage de programmation. **Vous devez implémenter votre solution en python 3**. Pour le concours Kaggle, vous avez accès à 3 fichiers : *train.csv*, *test.csv*, et *sample_submission.csv*. Vous utiliserez les exemples

du premier fichier pour entraîner votre modèle. Le deuxième fichier contient des exemples dont vous devez prédire les étiquettes. Vous devez soumettre vos prédictions en utilisant le même format que celui qui se trouve dans sample_submission.csv. Pour s'assurer que les équipes ne puissent pas tricher, les lignes du jeu de test ont été échantillonnées en lignes publiques et privées. Les scores publics sont affichés sur le classement public, tandis que seuls les scores privés sont utilisés pour déterminer le vainqueur du concours. Vous pouvez sélectionner deux de vos contributions pour être éligible au classement privé final.

Vous pouvez utiliser toutes les données externes au concours pour extraire plus d'informations sur les urls. La seule contrainte est que vous ne pouvez utiliser aucune règle manuelle concernant le domaine ou le lien. Par exemple, il n'est pas permis de créer une liste des domaines sécurisés.

Le concours est disponible à l'adresse suivante : `https://www.kaggle.com/t/a963f0fcf33d4ef79b020eeb3f5144be`.

3 Rapport

En plus d'implémenter votre méthode, vous devez rédiger un rapport qui détaille votre méthodologie pour résoudre ce problème et fournir les résultats de votre méthode. Précisément, votre rapport doit contenir au minimum les informations suivantes :

- Titre du projet
- Nom de l'équipe sur Kaggle, ainsi que la liste des membres de l'équipe (nom complet et matricule)
- Prétraitement des attributs (*Feature Design*) : décrivez et justifiez vos étapes de prétraitement des attributs et indiquez ceux que vous avez sélectionnés dans votre modèle.
- Méthodologie : décrivez et justifiez toutes les décisions concernant la répartition des données en ensemble d'entraînement et de validation, ainsi que les techniques utilisées pour gérer le déséquilibre entre les classes (*Unbalanced data*), la stratégie de régularisation, le réglage des hyperparamètres, etc. Ajoutez également toutes les informations que vous jugez nécessaires pour la compréhension de votre modèle.
- Résultats : Présentez une analyse détaillée de vos résultats à l'aide de tableaux ou de graphiques. En plus des meilleures performances obtenues, vous devez décrire et justifier l'impact de vos choix de conception (par exemple, les étapes de prétraitement, ou la régularisation) sur les performances du modèle.
- Discussion : Discutez vos résultats et indiquez quels sont les avantages et les inconvénients de votre approche et de votre méthodologie.
- Références (si applicable)

Le rapport ne doit pas dépasser 5 pages et doit être rédigé sur une ou deux colonnes, à simple interligne, avec une police de caractères de 10 points ou plus

(des pages supplémentaires pour les références et le contenu bibliographique sont autorisées). Vous êtes libre de structurer le rapport comme vous le souhaitez. Nous vous suggérons de vous concentrer sur l'explication de votre solution et des résultats (les sections d'introduction et de conclusion ne sont pas obligatoires).

4 Exigences de soumission

— Vous devez soumettre le code développé lors du TP.

— Vous devez soumettre ou présenter les liens pour toutes les données externes.

— Vous devez inclure un fichier README contenant des instructions sur la façon d'exécuter le code. De plus, dans ce même fichier, vous devez présenter les étapes nécessaires pour générer le fichier de prédiction soumis sur Kaggle. Assurez-vous que tous les fichiers de données nécessaires à l'exécution de votre code se trouvent dans le dossier et sont chargés avec le chemin d'accès relatif. Nous devons pouvoir exécuter votre code sans faire de modifications.

— Le fichier de prédiction doit être soumis en ligne sur le site de Kaggle.

— Vous devez soumettre un rapport écrit (**Format PDF**) selon la présentation générale décrite plus haut.

5 Critères d'évaluation

Les points seront attribués selon la répartition suivante : 10 points pour la performance à l'épreuve privée fixée dans le cadre du concours et 10 points pour le rapport écrit. L'équation suivante sera utilisée pour noter votre performance au concours Kaggle :

$$points = 10 \times \begin{cases} 0.0 \text{ if } P \leq Random \\ 0.7 * \frac{P - Random}{Baseline - Random} \text{ if } Random < P \leq Baseline \\ 0.7 + 0.3 * \frac{P - Baseline}{Group1 - Baseline} \text{ if } Baseline < P \leq Group1 \\ 1.0 \text{ if } P = Group1 \end{cases},$$

où $Random$, $Baseline$, $Group5$ et P sont les valeurs de la précision (accuracy) obtenues respectivement par une prédiction aléatoire, la méthode *baseline*, la première équipe du classement privé, et votre équipe du classement privé.

Pour le rapport écrit, les critères d'évaluation comprennent :

— L'efficacité de votre méthodologie (pré-traitement, sélection des caractéristiques, validation, algorithmes utilisés, etc.).

— La rigueur technique de la description de vos algorithmes.

— La clarté du rapport : la méthodologie, le dispositif expérimental, les résultats et les figures. Par exemple, n'oubliez pas d'indiquer à quoi correspondent les axes, et de mettre les légendes des figures. Expliquez également les résultats des figures dans le texte principal.

— La qualité de l'analyse des résultats finaux et intermédiaires.

— La qualité des descriptions, des graphiques, des figures et des tableaux.

— L'organisation générale et la qualité rédactionnelle.

Veuillez noter que le point essentiel du rapport (là où il y a le plus de points en jeu) se situe sur vos méthodes de pré-traitement des données, des techniques d'optimisation et vos modèles.

6 Règles spécifiques au concours Kaggle

— Ne trichez pas !

— Vous devez soumettre un code qui peut reproduire les numéros de votre solution de classement. Votre groupe recevra une note de 0 si nous ne pouvons pas reproduire vos résultats.

— Vous ne devez pas tenter de tricher en cherchant des informations sur le jeu de test.

— Ne faites pas plus d'une équipe pour votre groupe (par exemple, pour faire plus de soumissions). Vous recevrez une note de 0 pour avoir créé intentionnellement de nouveaux groupes dans le but de faire plus de soumissions Kaggle.

— Toutes règles de l'école concernant le plagiat et les TPs s'appliquent aussi.

Références

[1] Anti-phishing working group (2020). phishing activity trends, jul 2020.

[2] Wandera. Mobile threat landscape 2020 : Understanding the key trends in mobile enterprise security in 2020.