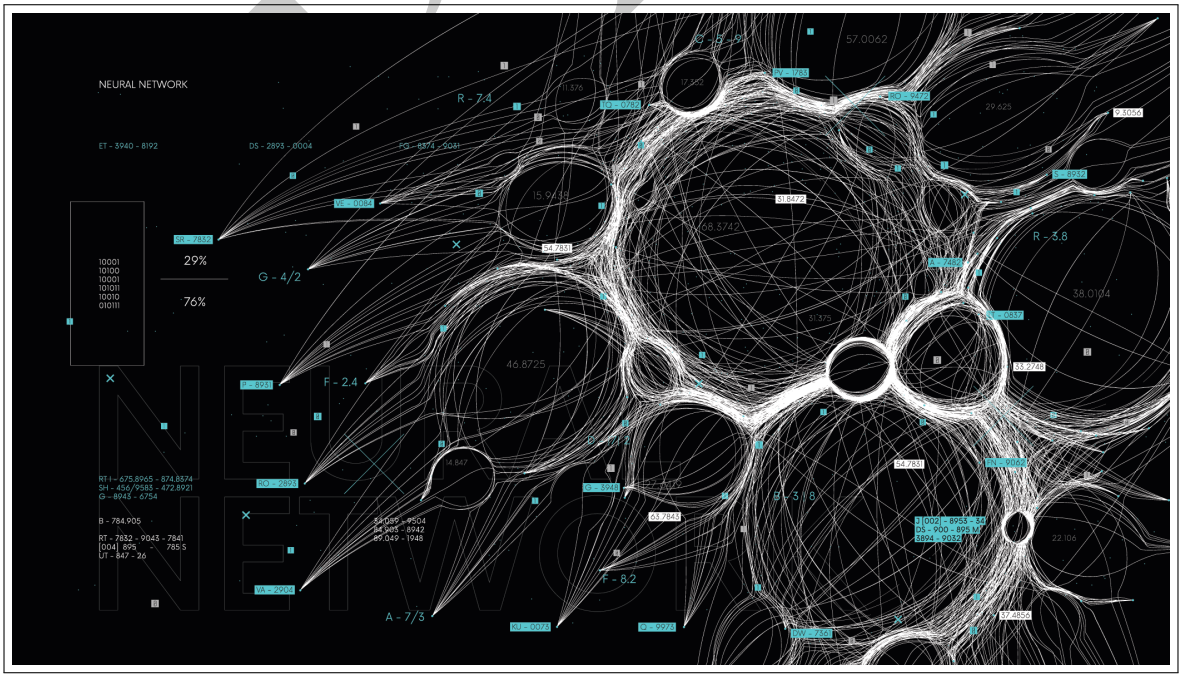
Institut National des Sciences Appliquées
Université de Rennes

Étude d'Attaques par Inférence d'Appartenance (MIA)

Une première approche avec le concours Snake Strikes Back

Auteurs :
Thomas Aubin
Selyan Da Silva
Moussa Ouassou
Émile Peltier

Responsable de projet :
Cédric Eichler
Créateurs de la compétition :
Tristan Allard
Mathias Bernard



Résumé en quelques lignes du projet
Mots-clés : mot1, mot2, mot3, mot4, mot5

3 décembre 2024

Résumé

Table des matières

Table des figures

Liste des tableaux

Liste des Équations

Table des éléments de code

Introduction

Bien que le projet ait pour coeur la participation à la compétition, celui-ci a nécessité un important travail de montée en compétences et de documentation en Machine Learning pour l'ensemble du groupe, ce domaine n'étant que peu abordé à ce stade de la formation. C'est pourquoi la partie opérationnelle et technique du projet est précédée d'une part d'un court travail de bibliographie ayant pour visée la synthèse des connaissances mathématiques et algorithmiques indispensables à la participation au concours, et d'autre part par une présentation des tenants et aboutissants du concours, laquelle prend soin d'expliquer le plus finement possible les données sur lesquelles nous nous entraînons ainsi que le modèle attaqué.

Première partie

Notions d'*Adversarial Machine Learning*

Chapitre I

Concepts utiles d'Intelligence Artificielle

I.1 Algorithmes de classification

I.1.1 Intérêt et fonctionnement de la classification

I.1.2 Exemples d'algorithme de classification

I.1.2.1 Régression logistique

I.1.2.2 Bayésien naïf

I.2 Réseaux de neurones et Deep Learning

I.2.1 Princice du Deep Learning

I.2.2 Un modèle à deux réseaux : le *Generative Adversarial Network* (GAN)

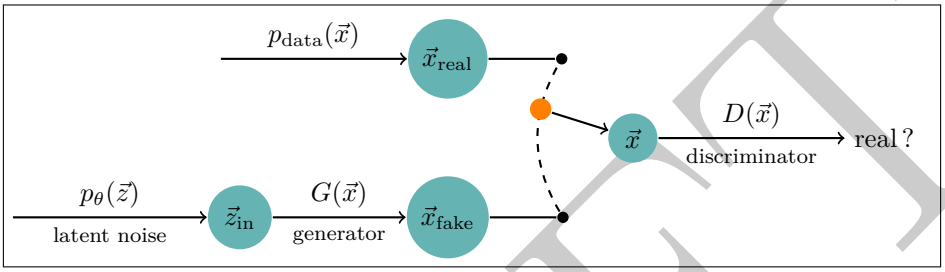


FIGURE I.1 – Schéma haut niveau d'un GAN

Chapitre II

Chapitre III

Attaques par Inférences d'Appartenance : contextualisation du projet

III.1 Types d'attaques

III.2 Conséquences d'une attaque MIA réussie

Deuxième partie

Le concours *Snake Strikes Back* : position du problème

Chapitre IV

Contexte et enjeux de la compétition

IV.1 Principe

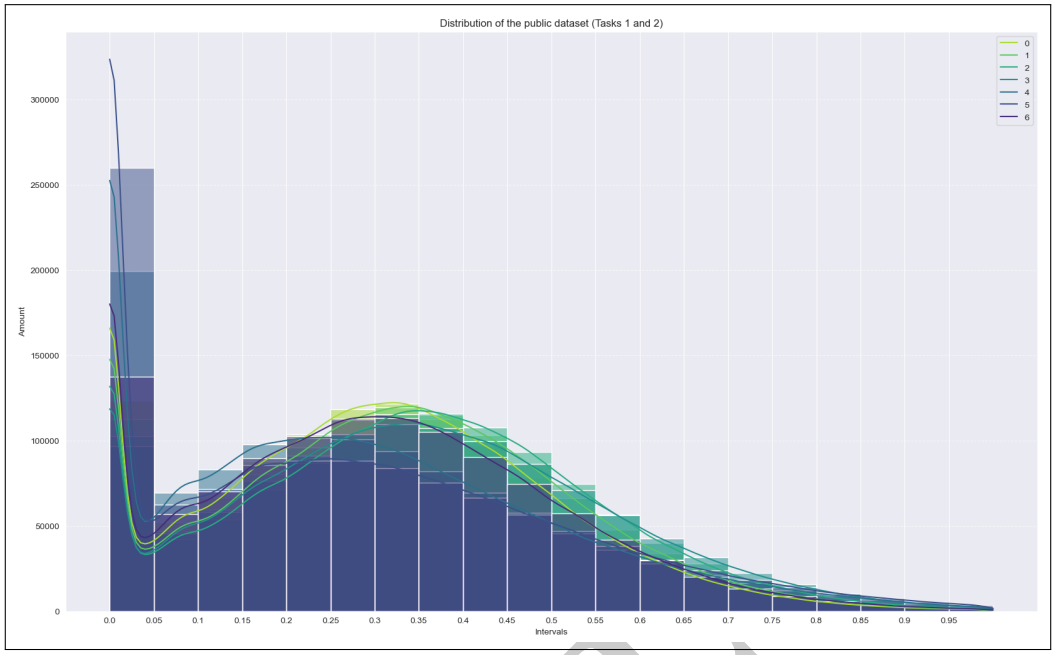
Chapitre V

Parcours des ressources fournies

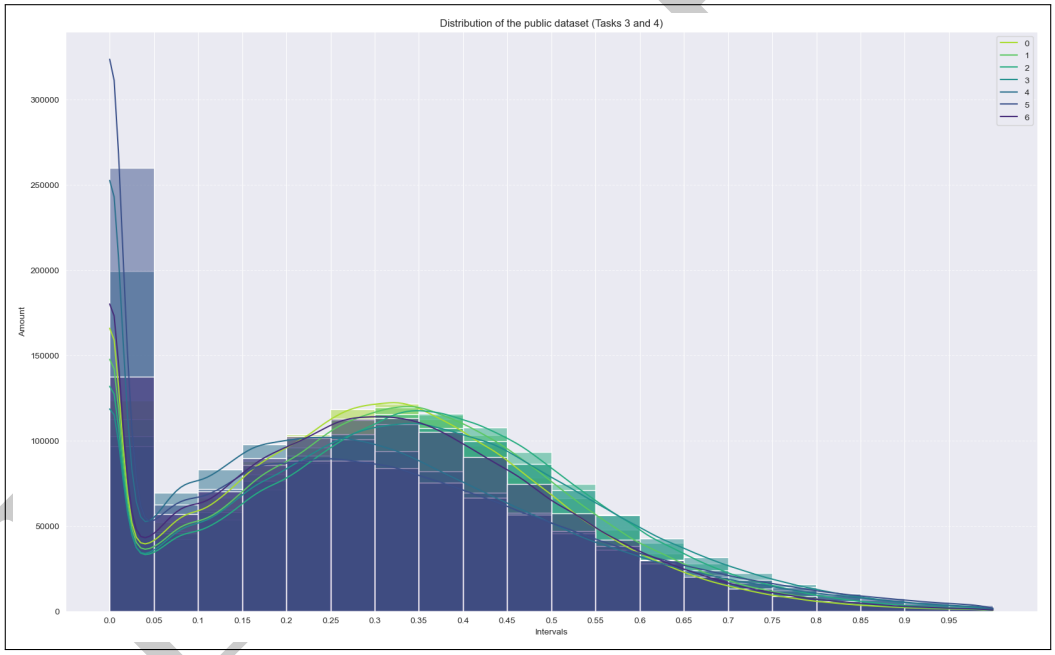
V.1 Processus d'installation : un peu de beta-testing et de documentation d'erreurs

Concours en Beta

V.2 Les datasets publics



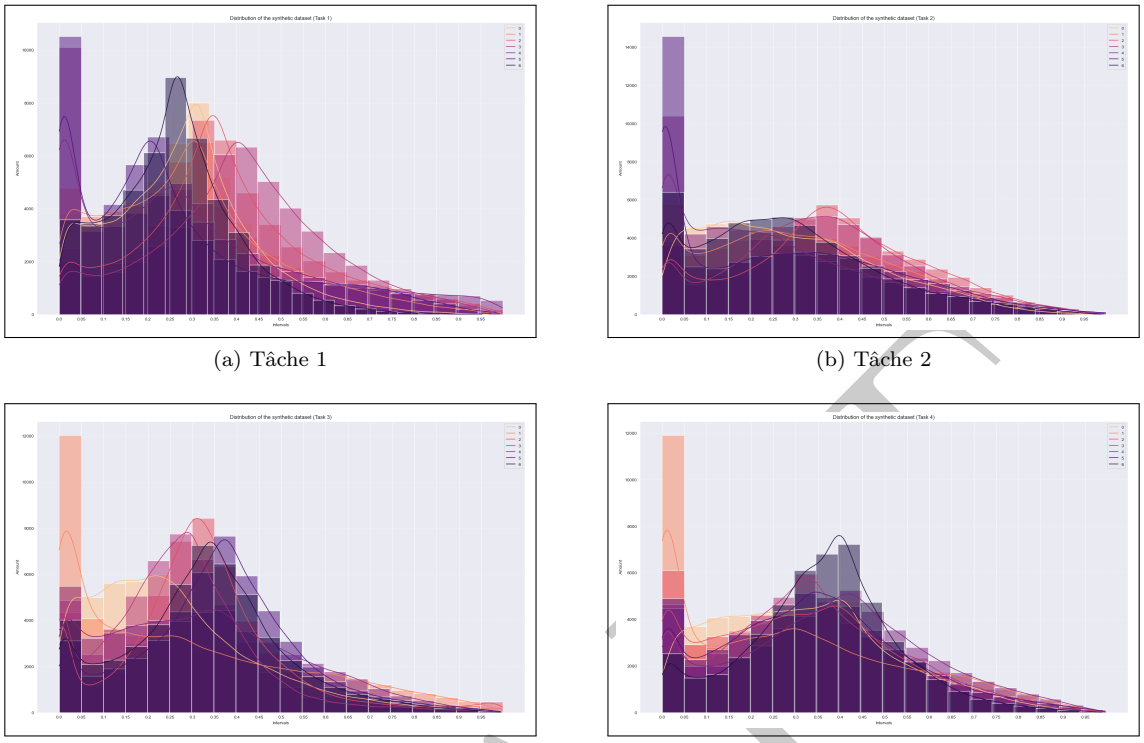
(a) Pour les tâches 1 et 2



(b) Pour les tâches 3 et 4

FIGURE V.1 – Distribution des données des datasets publics, par jour

V.3 Les datasets synthétiques



(a) Tâche 1

(b) Tâche 2

(c) Tâche 3

(d) Tâche 4

FIGURE V.2 – Distribution des données des datasets synthétiques, par jour.

Chapitre VI

DoppelGANger : un générateur de séries temporelles puissant ... mais attaquable

VI.1 Fonctionnement global

VI.2 Les hyperparamètres du modèle

VI.2.1 Batch size

VI.2.2 Generator learning

VI.2.3 Discriminator

VI.2.4 Learning rate

VI.2.5 Generator number of hidden layers

VI.2.6 Generator hidden layer size

VI.2.7 Nombre d'échantillons

Troisième partie

Attaque d'un modèle de Machine Learning par l'utilisation de *Shadow Models*

Chapitre VII

Création de *Shadow Models* pour reproduire le comportement étudié

VII.1 Critères déterminants dans la construction du modèle

VII.1.1 Le problème du surapprentissage ou *overfitting*

VII.1.1.1 Choix de la métrique

VII.1.2 Sous-section

VII.2 Processus de sélection des datasets

VII.2.1 Avec overlap

On casse l'hypothèse des DS disjoints du DS privé

VII.2.2 Sans overlap

VII.2.3 Sous-section

VII.3 Données générées par le modèle

Chapitre VIII

Méthodes de classification des données générées

VIII.1

VIII.2 Régression logistique

VIII.3 Bayésien naïf

VIII.4 Recherche des plus proches voisins (KNN)

Chapitre IX

Synthèse des résultats

IX.1 Tâche 1

IX.2 Tâche 2

IX.3 Tâche 3

IX.4 Tâche 4

Conclusion

Apprentissage non supervisé envisagé puis rejeté car absence d'oracle

Quatrième partie

Annexes

Annexe 1 : Programmes conçus par l'équipe

Annexe 2 : Retour d'expérience et chronologie du projet

Annexe 3 : Framework utilisé

Cinquième partie

Bibliographie

Fondamentaux de mathématiques et de programmation

[6] Chloé-Agathe Azencott. *Introduction au Machine Learning*. (2nd). InfoSup. Dunod, fév. 2022.

[8] Matt Harrison. *Machine Learning - Les Fondamentaux. Exploiter des données structurées en Python*. O'Reilly, 2020.

[9] Benjamin JOURDAIN. *Probabilités et statistiques pour l'ingénieur*. Jan. 2018.

[11] *Machine Learning*. Page Wikipedia du Machine Learning. Nov. 2024. URL : `https://en.wikipedia.org/wiki/Machine_learning`.

Sur le *Machine Learning Antagoniste* (Adversarial Machine Learning)

[1] *Adversarial Machine Learning*. Page Wikipedia de l'Adversarial Machine Learning. Nov. 2024. URL : https://en.wikipedia.org/wiki/Adversarial_machine_learning#Adversarial_attacks_and_training_in_linear_models.

[2] Tristan ALLARD et Mathias BERNARD. « Snakes Strikes Back ». In : (oct. 2024).

[4] AUTHOR. *Membership inference attacks from first principles*. How published. Some note. Month Year. URL : https://www.youtube.com/watch?v=1CNxfhMlk-A.

[7] *Generative adversarial network*. Page Wikipedia du modèle GAN. Nov. 2024. URL : https://en.wikipedia.org/wiki/Generative_adversarial_network.

[10] Zinan LIN et al. « Using GANs for Sharing Networked Time Series Data : Challenges, Initial Promise, and Open Questions ». In : (jan. 2021). Présentation du modèle DoppelGANger. URL : https://arxiv.org/abs/1909.13403.

[14] Reza SHOKRI. *Membership Inference Attacks against Machine Learning Models*. Vidéo de vulgarisation du papier du même nom. Mai 2017. URL : https://www.youtube.com/watch?v=rDm1n2gceJY&t=53s.

[15] Reza SHOKRI et al. « Membership Inference Attacks Against Machine Learning Models ». In : ().

Autres

[3] Tatev Aslanyan. *Machine Learning in 2024 – Beginner's Course*. Fév. 2024. URL : https://www.youtube.com/watch?v=bmmQA8A-yUA&t=1769s.

[5] AUTHOR. *Comparing and Evaluating Datasets : A Simplified Guide*. 24 nov. 2024. URL : https://www.markovml.com/blog/compare-datasets.

[12] Boris Meinardus. *How I'd learn ML in 2024 (if I could start over)*. Youtube. 2024. URL : https://www.youtube.com/watch?v=gUmagAluXpk.

[13] *Overfitting*. Page Wikipedia de l'Overfitting. Nov. 2024. URL : https://en.wikipedia.org/wiki/Overfitting#Machine_learning.

[16] *Training, validation, and test data sets*. Page Wikipedia rappelant la différence entre les datasets d'entraînement et de test. Nov. 2024. URL : https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets.