Compétition Kaggle 1 IFT 3395/6390

3 octobre 2023

1 Introduction

Pour ce projet, vous participerez à un concours Kaggle sur la détection d'événements météorologiques extrêmes à partir de données atmosphériques. L'objectif est de concevoir un algorithme d'apprentissage automatique capable de classer automatiquement si un ensemble de variables climatiques correspondant à un moment et un emplacement (latitude et longitude) est associé à i) des conditions standards (rien à signaler), ii) un cyclone tropical ou iii) une rivière atmosphérique. Les modèles capables de détecter avec précision de tels événements sont cruciaux pour notre compréhension de la façon dont ils peuvent évoluer dans divers scénarios de changement climatique.

L'ensemble de données que nous avons préparé pour ce concours est un sous-ensemble relativement petit d'un ensemble de données plus important, ClimateNet. L'ensemble de données complet s'élève à près de 30 Go car il contient des variables climatiques à près de 900 000 endroits dans le monde. Pour ce concours, nous avons réduit le nombre d'emplacements à 120, tout en conservant les données sur l'axe du temps. L'ensemble de données est divisé en 44 760 points de données pour l'entrainement (données de 1996 à 2009) et 10 320 points de données pour les tests (données de 2010 à 2013). Chaque exemple se compose de 16 variables atmosphériques telles que la pression, la température et l'humidité. Plus de détails sur ces variables sont disponibles sur Kaggle.

L'objectif de ce projet est de vous faire implémenter et entraîner plusieurs algorithmes de classification. L'évaluation sera basée sur la performance sur l'ensemble de test et sur un rapport écrit.

La compétition, y compris les données, est disponible ici :

https://www.kaggle.com/t/dd37e4ab8ada4123a077b0da52b8a72c.

2 Dates importantes et information

Veuillez prendre en considération les échéances importantes suivantes :

— **11 octobre 23 :59** Date limite pour les étudiants IFT3395 pour s'inscrire à la compétition et former des équipes.

— **1er Novembre 23 :59** Fin du concours. Plus aucune soumission Kaggle n'est autorisée.

— **6 novembre 23 :59** Les rapports et le code sont dûs sur Gradescope.

Note sur le partage et le plagiat : Vous êtes autorisé à discuter des techniques générales avec les autres équipes. Vous n'êtes PAS autorisé à partager votre code. Ce comportement constitue du plagiat et il est très facile à détecter. Toutes les équipes impliquées dans le partage de code recevront une note de 0 dans la compétition de données.

3 Participation à la compétition et formation de l'équipe

Les élèves de **IFT3395** travailleront en équipe de 2 ou 3. (Les étudiants de **IFT6390** doivent faire la compétition seul).

3.1 Formation de l'équipe Kaggle (étudiants IFT3395 uniquement)

Pour former une équipe :

— Participez à la compétition et créez un compte Kaggle si vous n'êtes pas encore inscrit en suivant le lien : https://www.kaggle.com/t/dd37e4ab8ada4123a077b0da52b8a72c
— Dans la section « Inviter d'autres personnes », entrez les noms de vos coéquipiers ou le nom de l'équipe.
— Votre coéquipier a la possibilité d'accepter.
— Remplissez le formulaire google https://forms.gle/DcxkTMYndmWwAjZB7 avec les informations de votre équipe avant le **11 octobre à 23h59**. **UNE SEULE PERSONNE PAR EQUIPE DOIT REMPLIR LE FORMULAIRE.**

Remarque importante : Le nombre maximum de soumissions par jour et par équipe est de 3. Toute équipe dont les membres individuels ont un nombre de soumissions supérieur à ce qui est autorisé à ce jour ne pourra pas former une équipe. Exemple : Aujourd'hui est le premier jour de la compétition. A, B et C sont trois coéquipiers qui n'ont pas encore formé d'équipe.

— A soumis 0 fois.
— B soumis 3 fois.
— C soumis 1 fois.

Étant donné que le nombre maximum de soumissions est de 3 par équipe et par jour, le nombre total de soumissions possibles pour une équipe est de 3. Cependant, le nombre cumulé de soumissions pour A,B et C est de 4. Par conséquent, ils ne pourront pas former une équipe. Ils devront attendre demain et ne soumettre aucune soumission pour le lendemain.

Vous pouvez commencer à soumettre des solutions avant de former une équipe, tant que vous faites attention à la limitation ci-dessus lors de la formation d'équipes.

4 Références

Nous vous demandons de créer un classifieur **régression logistique** et de battre les baselines mises en évidence dans le classement. Ces baselines sont :

— un classificateur fictif initialisé avec des valeurs aléatoires.

— un classificateur de régression logistique.
— la meilleure référence du TA.

Pour participer au concours, vous devez fournir une liste de prédictions pour les instances sur le site web de Kaggle. Vous pouvez soumettre 3 prédictions par jour au cours de la compétition, nous vous suggérons donc de commencer tôt, en vous accordant suffisamment de temps pour soumettre plusieurs fois et avoir une idée de vos performances.

Pour chacune des 3 baselines que vous battez, vous obtenez des points supplémentaires.

Votre classificateur de régression logistique doit être implémenté à partir de zéro. En dehors des bibliothèques Python standard, les seules bibliothèques autorisées sont `numpy`, les matrices creuses de scipy (`scipy.sparse`) et `pandas`.

5 Autres modèles

Vous devez essayer au moins **1 autre modèle** en plus de la régression logistique, et comparer leurs performances. Vous êtes encouragés à mettre en œuvre les techniques étudiées pendant le cours et à rechercher d'autres moyens de résoudre cette tâche. Voici quelques idées :

— Machine à vecteurs de support
— Naïve Bayes
— Forêts d'arbres décisionnels
— Utiliser des features fait-mains qui prennent en compte la nature des données.

L'objectif est de concevoir la méthode la plus performante telle que mesurée en soumettant des prédictions pour l'ensemble de test sur Kaggle. Votre performance finale sur Kaggle comptera comme critère d'évaluation (voir Section 6). Si un modèle testé ne fonctionne pas bien, vous pouvez toujours l'ajouter dans votre rapport et expliquer pourquoi vous pensez qu'il n'est pas approprié pour cette tâche. Ce type de discussion est un élément important que nous utiliserons pour évaluer votre rapport final de concours.

Pour cette partie, vous êtes libre d'utiliser n'importe quelle bibliothèque de votre choix.

6 Rapport

En plus de vos méthodes, vous devez rédiger un rapport qui détaille les techniques de pré-traitement, de validation, d'algorithmique et d'optimisation, ainsi que des résultats qui vous aident à comparer différentes méthodes/modèles.

Le rapport doit contenir les sections et éléments suivants :

— Titre du projet
— Nom de l'équipe sur Kaggle, ainsi que la liste des membres de l'équipe, y compris leurs noms complet et leurs matricules.
— Introduction : décrivez brièvement le problème et résumez votre approche et vos résultats.
— Feature Design : Décrivez et justifiez vos méthodes de pré-traitement, et comment vous avez conçu et sélectionné vos features.
— Algorithmes : donnez un aperçu des algorithmes d'apprentissage utilisés sans entrer dans trop de détails, sauf si nécessaire pour comprendre d'autres détails.

— Méthodologie : inclure toutes les décisions concernant la division de l'ensemble d'entrainement et de validation, les stratégies de régularisation, les astuces d'optimisation, le choix des hyper-paramètres, etc.
— Résultats : présentez une analyse détaillée de vos résultats, y compris des graphiques et des tableaux. Cette analyse doit être plus large que le simple résultat de Kaggle : inclure une courte comparaison des hyper-paramètres les plus importants et de toutes les méthodes (au moins 2) que vous avez mises en œuvre.
— Discussion : discutez des avantages/inconvénients de votre approche et de votre méthodologie et proposez des idées d'amélioration.
— Division des contributions : décrire brièvement les contributions de chaque membre de l'équipe vers chacune des composantes du projet (par exemple, définir le problème, développer la méthodologie, coder la solution, effectuer l'analyse des données, rédiger le rapport, etc.) À la fin de l'énoncé des contributions, ajouter la mention suivante : "Nous déclarons par la présente que tous les travaux présentés dans ce rapport sont ceux des auteurs".
— Références : très important si vous utilisez des idées et des méthodes que vous avez trouvées dans un papier ou en ligne ; c'est une question d'intégrité académique.
— Annexe (facultatif) : Ici, vous pouvez inclure des résultats supplémentaires, plus de détails sur les méthodes, etc.

Vous perdrez des points si vous ne suivez pas ces directives. **Le texte principal du rapport ne doit pas dépasser 6 pages.** Les références et annexes peuvent dépasser les 6 pages.

Vous devez soumettre votre rapport et votre code sur Gradescope avant le **6 novembre 23 :59**.

Instructions de soumission

— Vous devez soumettre le code développé pendant le projet. Le code doit être bien documenté. Le code doit inclure un fichier README contenant des instructions sur comment exécuter le code.
— Le fichier de prédiction contenant vos prédictions sur l'ensemble de test doit être soumis en ligne sur le site Web de Kaggle.
— Le rapport au format pdf (écrit selon les critères définit au-dessus) et le code doivent être téléchargés sur Gradescope.

7 Critères d'évaluation

Les notes seront attribuées en fonction des critères suivants :

1. Des points vous seront attribués pour chacune des 3 baselines que vous battez.
2. Des points vous seront attribués en fonction de votre performance finale à la fin de la compétition. Le classement que vous pouvez voir est calculé en utilisant la moitié des données test. L'autre moitié des données test permet de réaliser un classement privé (qui n'est pas visible). Vous pouvez donc être premier au classement public, mais pas

au classement privé. La note est calculée par pondération du classement public et du classement privé.

3. Des points vous seront attribués en fonction de la qualité et de la solidité technique de votre rapport final (voir ci-dessus).