Compétition Kaggle 2 IFT6390

16 novembre 2022

1 Introduction

Pour ce projet, vous participerez à une 2ème compétition Kaggle basée sur la classification de textes. Vous disposez de plus d'un million de textes annotés comme négatifs, positifs et neutres. La tâche consiste à créer des caractéristiques basées sur les données fournies et à utiliser des algorithmes d'apprentissage automatique pour les classer.

Vous devez implémenter et entraîner plusieurs algorithmes de classification basés sur l'ensemble des données fournies. L'évaluation sera basée sur les performances sur l'ensemble de test fourni et sur un rapport écrit.

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

https://www.kaggle.com/t/6c98b08e131d49abaa8915175de22ced

2 Dates importantes et information

Veuillez prendre en considération les échéances importantes suivantes :

— **16 novembre** Date de publication du concours
— **22 novembre 23 :59** Date limite pour s'inscrire à la compétition sur Kaggle.
— **9 décembre23 :59** Fin du concours. Plus aucune soumission Kaggle n'est autorisée.
— **14 décembre 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 étudiants de **IFT6390** travailleront en équipe de 2.

3.1 Formation d'une équipe Kaggle (étudiants IFT3395 uniquement)

Pour former une équipe :

— Inscrivez-vous à la compétition et créez un compte Kaggle si vous n'en avez pas déjà, en suivant ce lien : https://www.kaggle.com/t/6c98b08e131d49abaa8915175de22ced
— Dans l'onglet "Invite Others", ajoutez les noms de vos coéquipiers.
— Vos coéquipiers vont maintenant devoir accepter l'invitation.
— Remplissez le formulaire google `https://forms.gle/j96iySG3TJAFrXBg8` avec les membres de votre équipe avant le **22 novembre à 23 :59**. Toutes équipes qui ne se sont pas inscrites ou qui se sont inscrites après la date limite ne seront pas évaluées.

Important : Le nombre maximum de soumissions par jour et par équipe est de 3, et par ÉQUIPE. Si au moment de la formation d'une équipe le total des soumissions par les membres est plus que 3, il ne sera pas possible de créer l'équipe ce jour-ci. Par exemple : C'est le premier jour de la compétition. Les étudiants A,B,C veulent former une équipe.

— A a soumis 1 fois.
— B a soumis 2 fois.
— C a soumis 1 fois.

Le maximum autorisé est de 3 évaluations par jour et par équipe, mais la somme des évaluations des futurs membres de l'équipe est déjà de 4. Par conséquent, ils ne pourront pas former une équipe aujourd'hui, et ils devront attendre demain.

Vous pouvez cependant effectuer des évaluations avant de former une équipe, tant que vous prenez bien en compte la limite au jour de la création de l'équipe.

4 Exigences

Pour ce concours, vous êtes autorisé à utiliser n'importe quelle fonction de bibliothèque de votre choix. Pour obtenir des notes complètes, il vous est demandé d'implémenter **les 4 algorithmes suivants**, et de rapporter leurs performances. Cependant, même si vous utilisez une bibliothèque, il est important que vous discutiez de chaque hyperparamètre et de la manière dont vous avez choisi leurs valeurs. De plus, vous ne téléchargerez que le modèle le plus performant sur Kaggle. Mais il est important de rendre compte de la précision de chacun de ces modèles sur l'échantillon de validation et de test (vous pouvez rendre compte du score que vous avez reçu sur Kaggle car vous ne voyez pas l'ensemble de test complet).

Voici les algorithmes à mettre en œuvre :

— un classificateur de Bayes naïf utilisant les caractéristiques du sac de mots.
— SVM kernelisé utilisant des noyaux de chaînes de caractères
— Neural Nets
— Tout autre algorithme de votre choix...

4.1 Explicabilité de l'apprentissage automatique

Les modèles d'apprentissage automatique sont souvent traités comme des boîtes noires où la raison pour laquelle une décision est prise par le classificateur est inconnue. Par conséquent, des algorithmes tels que Local Interpretable Model-Agnostic Explanations(LIME) [1] or Grad-CAM [2].

En plus de la compétition Kaggle, vous devez implémenter un algorithme explicable de votre choix et expliquer la décision prise par le modèle que vous avez formé. Vous êtes autorisé

à utiliser n'importe quelle fonction de la bibliothèque pour cette implémentation.

5 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 essayé.
- 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 8 pages.** Les références et annexes peuvent dépasser les 10 pages.

Vous devez soumettre votre rapport et votre code sur Gradescope avant le **14 décembre 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.
— Vous pouvez soumettre des fichiers *.ipynb mais il est obligatoire de soumettre le fichier *.py associé.
— Veuillez soumettre le fichier *.ipynb associé à l'explicabilité de votre modèle.

6 Critères d'évaluation

Les notes seront attribuées en fonction des critères suivants :

1. Des points vous seront attribués pour chacune des 3 bases de référence que vous battez.
2. Des points vous seront attribués en fonction de votre performance finale à la fin de la compétition, donnée par votre classement par rapport à la classe.
3. Des points vous seront attribués en fonction de la qualité et de la solidité technique de votre rapport final (voir ci-dessus).
4. La répartition complète des notes est la suivante : Compétition : 35 (Algorithme 1-4 : 30, Classement au classement : 5) ; Rapport : 40 (Format : 5, Algorithmes : 10, Méthodologie : 15, Discussion des résultats (y compris l'explicabilité) : 10) ; Explicabilité : 15 ; Code : 10

Références

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you ?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[2] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam : Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.