Instructions pour Compétition Kaggle 2024

IFT3395/6390A

October 16, 2024

1 Description

Dans ce projet, vous participerez à un compétition Kaggle de classification de textes. L'objectif est de concevoir un algorithme d'apprentissage automatique capable de trier automatiquement de courts documents textuels dans un ensemble de catégories prédéterminé. Afin de préserver l'anonymat de l'ensemble de données et d'assurer une concurrence équitable, nous ne divulguons pas le nom de l'ensemble de données et ne fournissons pas d'entrées de documents originales. Au lieu de cela, vous recevez des vecteurs de comptage de mots (termes) par document comme caractéristiques, où chaque index est le nombre de fois qu'un terme donné est présent dans le document. Parallèlement à ces vecteurs, vous recevez également une carte de vocabulaire qui correspond à chaque index à un terme. Votre objectif est d'exploiter cette matrice de comptage de termes pour résoudre une tâche de classification de texte binaire. La métrique d'évaluation est le macro F1 score sur l'ensemble de test. En résumé, vous recevez les données suivantes:

- `data_train.npy` - Il s'agit d'un tableau NumPy représentant le nombre de termes vectoriels pour l'entraînement, où chaque ligne correspond à un document et chaque colonne représente un terme du vocabulaire. Les valeurs (principalement des 0) indiquent le nombre de chaque terme dans le document respectif, formant une matrice creuse.

- `data_test.npy` - Un tableau NumPy similaire pour les tests. Vous devez créer des étiquettes pour cet ensemble de tests et le soumettre pour évaluation.

- `vocab_map.npy` - Contient un mappage entre les termes (mots) et leurs indices correspondants dans la matrice vectorielle des termes.

- `labels_train.npy` - Contient les étiquettes ou les valeurs cibles pour l'ensemble de données d'entraînement (0 ou 1).

2 Participation

Pour la section IFT6390A, la tâche doit être résolue individuellement et sans l'aide d'autres étudiants. Les étudiants de premier cycle (IFT3395) peuvent participer individuellement ou en équipes de deux. Pour participer au compétition, vous devez:

- Créer un compte Kaggle si vous n'en avez pas déjà un.
- Participez au compétition en utilisant le lien d'invitation suivant: `https://www.kaggle.com/t/b156d192b9374a549cc887d465431873`.
- (Réservé aux étudiants de premier cycle) Remplissez le Google form suivant `https://forms.gle/Y1d7FTFVXAjt9TjL8` avec les informations de votre équipe avant le **13 octobre à 23:59**. Si vous ne remplissez pas le Google form avant cette date, vous devrez participer individuellement. Après avoir rempli le formulaire, vous pouvez ajouter un autre participant à votre équipe via l'onglet "Team" sur la page du compétition.
- Désormais, vous pouvez accéder au compétition via `https://www.kaggle.com/competitions/classer-le-text/`.

Remarque importante: Le nombre maximum d'évaluations test sur Kaggle est de 2 par jour, et par ÉQUIPE. Si au moment de la formation d'une équipe le total des évaluations par les membres de cette future équipe est supérieur à 2, il ne sera pas possible de créer une équipe ce jour. Par exemple: C'est le premier jour de la compétition. Les étudiants A,B,C veulent former une équipe.

- A a effectué 0 évaluation.
- B a effectué 2 évaluations.
- C a effectué 1 évaluation.

Le maximum autorisé est de 2 évaluations par jour et par équipe, mais la somme des évaluations des futurs membres de l'équipe est déjà de 3. Par conséquent, ils ne pourront pas former une équipe aujourd'hui, et ils deveront 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,

3 Première étape: dépassez les points de référence (21 Oct)

Vous pouvez voir deux scores de référence sur le classement. Le premier score correspond à un classificateur qui attribue des étiquettes aléatoires à chaque document. Le deuxième score de référence correspond à un classificateur de régression logistique de base. Pour la première étape, vous devrez battre le classificateur de régression logistique de référence sur le classement public.

Remarque importante: Pour battre la ligne de base, vous n'êtes PAS autorisé à utiliser une bibliothèque d'apprentissage automatique, par exemple `scikit-learn`. Vous devez implémenter votre solution à partir de zéro en utilisant uniquement NumPy et les fonctionnalités de base de Python.

4 Deuxième étape: Compete (9 Nov)

Vous avez jusqu'au 9 novembre 23h59 pour obtenir les meilleures performances possibles sur la tâche. Dans cette phase, vous êtes libre de mettre en œuvre la méthode qui vous semble la plus efficace. Le classement Kaggle comporte un composant public et privé pour empêcher les participants de "overfitting" au classement. Le classement public affiche votre score calculé sur 30% de l'ensemble de test, tandis que le classement privé est basé sur votre score sur les 70% restants de l'ensemble de test. Vous ne pouvez voir le classement public que pendant la compétition. Les points de cette phase seront attribués en fonction de votre classement dans le classement privé qui sera publié à la fin de la compétition.

Remarque importante: Vous devez soumettre deux solutions distinctes, une pour la première phase (dépasser la ligne de base) et une pour la deuxième phase (votre modèle le plus performant). Vous devez nommer vos fichiers de soumission pour faire la distinction entre les deux. Pour votre soumission de code sur Gradescope, vous devez également séparer les deux solutions. Un seul membre de l'équipe doit soumettre le code à Gradescope.

5 Troisième étape: Soumettre le code et le rapport (12 Nov)

Vous devez rédiger un rapport qui détaille votre pipeline d'apprentissage automatique, y compris le prétraitement, les algorithmes, l'optimisation et l'apprentissage, le réglage des hyperparamètres et la procédure de validation. Vous devez également fournir et comparer les résultats d'autres méthodes que vous avez mises en œuvre avant d'atteindre le modèle le plus performant. Le rapport doit contenir les éléments suivants. Vous perdrez des points si vous ne suivez pas ces directives.

- Titre du projet
- Votre nom et celui de votre coéquipier (pour IFT3395)
- Introduction : décrivez brièvement le problème et résumez votre approche et vos résultats.
- Conception des fonctionnalités : Décrivez et justifiez vos méthodes de prétraitement et d'extraction de fonctionnalités.
- Algorithmes : Donnez un aperçu des algorithmes d'apprentissage utilisés sans entrer trop dans les détails.

- Méthodologie: inclure toutes les décisions concernant la répartition formation/validation, la stratégie de régularisation, les astuces d'optimisation, la définition des hyperparamètres, etc.

- Résultats: présenter une analyse détaillée de vos résultats, y compris des graphiques et des tableaux, le cas échéant. Cette analyse doit être plus large que les seuls résultats de Kaggle: inclure une brève comparaison de différentes valeurs pour les hyperparamètres importants de votre algorithme le plus performant et comparer également les performances de cette méthode avec au moins deux autres méthodes que vous avez mises en œuvre.

- Discussion: discuter des avantages/inconvénients de votre approche et suggérer des idées d'amélioration.

- Références (très importantes si vous utilisez des idées et des méthodes que vous avez trouvées dans un article 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.

Le texte principal du rapport ne doit pas dépasser 6 pages. Les références et l'annexe peuvent dépasser les 6 pages.

Vous devez soumettre votre code (premier et deuxième jalons) et votre rapport (troisième jalon) sur Gradescope avant le **12 novembre 23:59**. Une seule soumission de rapport est requise par équipe (pour les étudiants de premier cycle).

Instructions de soumission

- Vous devez avoir des fichiers .py/notebooks séparés pour les premier et deuxième jalons. Le code doit être bien documenté. Si vous n'utilisez pas de notebooks Jupyter, vous devez inclure un fichier README contenant des instructions sur la façon d'exécuter le code. Vous devrez soumettre un fichier zip contenant votre code et les fichiers associés à Gradescope.

- Le fichier de prédiction contenant vos prédictions sur l'ensemble de tests ne doit être soumis qu'à Kaggle.

- Le rapport au format PDF (rédigé selon la présentation générale décrite précédemment) doit être soumis à Gradescope.

- (POUR LES ÉTUDIANTS DE PREMIER CYCLE) Une seule soumission par équipe est requise sur Gradescope.

6 Critères d'évaluation

1. Vous recevrez un nombre minimum de points si vous battez la ligne de base de la régression logistique dans le classement public de Kaggle (à condition que vous respectiez les instructions susmentionnées).

2. Vous serez noté en fonction de la qualité et de la solidité technique de votre rapport final.

3. Vous recevrez des points **bonus** en fonction de votre classement final dans le classement privé de Kaggle à la fin de la compétition.

7 Dates limites

- La date limite pour s'inscrire sur Kaggle et remplir le Google form est le **13 octobre, à 23h59** (pour les étudiants de premier cycle qui souhaitent participer en équipes de deux) .
- La date limite pour battre la ligne de base est le **21 octobre, à 23h59**.
- Le compétition Kaggle se terminera le **9 novembre, à 23h59**.
- Vous devez télécharger votre rapport et votre code sur Gradescope avant le **12 novembre 23h59**.