Compétition Kaggle IFT3395/6390

Instructions pour les étudiants de l'UdeM

September 30, 2019

1 Présentation

Pour ce projet, vous allez participer à une compétition Kaggle de classification de texte. Vous devez construire un modèle qui pourra automatiquement classer des messages textes dans une liste donnée de sujets. L'ensemble de données que avons préparé contient des messages du extraits du forum Reddit. Nous avons sélectionné 20 subreddits: ce sont nos 20 sujets (classes) dans lesquels les messages seront classés. Nous avons échantillonné 3.500 messages pour chacun de ces 20 subreddits pour former un ensemble d'entrainement, et 1.500 messages pour l'ensemble de test. Vous devez implémenter et entrainer un certain nombre de classifieurs et vous allez être évalué sur la précision atteinte par votre modèle sur l'ensemble de test.

La compétition et les données sont disponibles ici:
`https://www.kaggle.com/c/ift3395-ift6390-reddit-comments`.

2 Création des équipes Kaggle

Chaque équipe sera composée de **2 étudiants gradués (IFT6390)** ou **3 étudiants au baccalauréat (IFT3395)**. 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/79d35aaec8f642a58ef289a52b67bfc7`
- 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/VvL28nGP1MMYrCnC7` avec les membres de votre équipe avant le **11 Oct à 23:59**. Toute équipe qui n'aura pas été enregistrée ou en retard ne sera pas évaluée.

Important: 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 (18 Oct)

Pour cette première partie, vous allez devoir implémenter un classifieur de Bayes naïf avec traits caractéristiques "sacs de mots" (bag of words), de manière à atteindre les points de références grisés dans le leaderboard de Kaggle. Ces références sont:

- un classifieur aléatoire qui choisit une classe au hasard pour chaque exemple de test
- un classifieur de Bayes naïf qui utilise des sacs de mots
- un classifieur de Bayes naïf avec lissage de Laplace

Pour participer à la compétition, vous devez fournir une liste de prédictions pour les éléments de l'ensemble de test dans l'onglet "Submit predictions". Vous pouvez effectuer 2 évaluations par jour pendant toute la durée de la compétition, donc nous vous suggérons d'effectuer rapidement vos premières évaluations, de manière à avoir suffisament de temps pour faire plusieurs évaluations et avoir une idée de la performance de votre modèle.

Vous devez déposer le code de votre classifieur de Bayes naïf sur Gradescope, avec un document Readme qui explique rapidement son organisation, avant le **18 Oct à 23:59**. Une partie de votre note pour cette compétition sera attribuée en fonction de votre meilleure évaluation test sur Kaggle en date du **Oct 18th 23:59**. Pour chacune des références que vous dépassez, vous améliorez votre note.

... mais la compétition n'est pas terminée ! Vous allez maintenant avoir l'occasion d'améliorer votre modèle, et de viser la meilleure performance possible, pendant la deuxième étape de la compétition.

4 Deuxième étape (6 Nov)

Vous avez jusqu'au **6 Nov à 23:59** pour obtenir la meilleure performance possible sur la même tâche de classification. Dans cette partie, vous pouvez implémenter les méthodes de votre choix, tout en respectant la consigne suivante.

Pour cette étape vous devez essayer au moins **2 autres modèles** en plus de Bayes naïf, et comparer leurs performances. Nous vons encourageons à implémenter les techniques vues en cours, et à effectuer des recherches pour trouver d'autres manières de résoudre ce problème. Voici une liste de suggestions:

- Machines à vecteurs de support (SVM) en utilisant des noyaux pour chaines de caractères
- Forêt d'arbres décisionnels
- Régression logistique avec traits caractéristiques créés à la main
- tout autre algorithme de votre choix...

Le but est d'obtenir la meilleure performance possible en évaluant votre prédictions de test sur Kaggle. Votre meilleure performance sur Kaggle compte comme un critère d'évaluation de ce projet (voir ci-dessous). Si un modèle testé n'est pas très performant, vous pouvez quand même l'ajouter à votre rapport, et expliquer pourquoi vous pensez qu'il n'est pas adapté pour cette tâche de classification. Ces discussions sont importantes et seront utilisées lors de l'évaluation de vos rapports.

5 Troisième étape: Rapport (8 Nov)

Vous devez rendre un rapport qui décrit le prétraitement, la méthode de validation, les algorithmes, les techniques d'optimisation, et qui donne les résultats lors de la comparaison de vos différents modèles. Le rapport doit contenir les éléments suivants. Vous allez perdre des points si vous ne suivez pas ces directives.

- Titre du projet
- Nom de l'équipe sur Kaggle, et liste des coéquipiers, avec le nom et code étudiant pour chacun des coéquipiers.
- Introduction: Décrivez brièvement la tâche et résumez votre approche et résultats
- Construction de traits caractéristiques: Décrivez et justifiez vos méthodes de prétraitement, et la manière dont vous avez choisis ou construis votre traits caractéristiques.
- Algorithmes: Faites une brève présentation des algorithmes que vous avez utilisé, sans donner trop de détails sauf si vous le jugez utile pour la compréhension.

- Méthodologie: Expliquez vos choix de division entrainement/validation, distribution pour Bayes naïf stratégies de régularisation, astuces d'optimisation, choix d'hyperparamètres, etc.

- Résultats: Presentez une analyse détaillée de vos résultats, en incluant des graphiques et tableaux lorsque nécessaire. Cette analyse sera plus large qu'une simple présentation de vos résultats sur Kaggle: incluez une discussion des hyperparamètres les plus importants, et toutes les méthodes (au moins 3) que vous aurez implémenté.

- Discussion: Donnez les points forts et points faibles des méthodes que vous avez testé, ainsi que la méthodologie. Donnez des suggestions d'amélioration de ces méthodes.

- Liste des contributions: Décrivez brièvement les contributions de chacun des membres de l'équipe à chaque partie du projet (par exemple définition du problème, développement de la méthodologie, programmation de la solution, analyse des performances, écriture du rapport, etc). À la fin de la liste de contributions, ajoutez la phrase suivante: "Nous certifions que nous sommes les auteurs des travaux présentés dans ce rapport".

- Références (important si vous avez utilisé des idées ou méthodes que vous avez trouvé dans des articles scientifiques ou en ligne; c'est une question d'intégrité académique).

- Appendice (optionnel). Vous pouvez y inclure des résultats ou détails supplémentaires

Le texte de votre rapport ne doit pas dépasser 6 pages. Les références et appendices peuvent utiliser des pages supplémentaires Your must submit your report and your code on Gradescope before **Nov 8th 23:59**.

Instructions pour le dépot

- Vous devez déposer le code développé pendant ce projet. Ce code doit être documenté, et inclure un document Readme qui contient des instructions sur comment exécuter le code.

- Les prédictions test doivent être évaluées en ligne en utilisant la fonction "Submit prediction" de Kaggle.

- Le rapport au format pdf (écrit en suivant les directives ci-dessus) et le code doivent être déposés sur Gradescope.

6 Critères d'évaluation

Les notes seront calculées en fonction des 3 étapes du projet:

1. Vous obtiendrez des points pour chacune des références grisées sur Kaggle que vous battez. Vous obtiendrez ces points uniquement si votre implémentation de Bayes naïf aura été déposée sur Gradescope avant la date limite pour l'étape 1.

2. Vous obtiendrez des points en fonction de la performance finale à la fin de la compétition (étape 2), donnée par votre classement sur le leaderboard privé sur Kaggle.

3. Vous obtiendrez des points en fonction de la qualité et de pertinence technique de votre rapport final

7 Dates limites

- La date limite pour former les équipes est **11 Octobre à 23:59**
- La date limite de la première étape (battre les références et déposer le code sur Gradescope) is **18 Octobre à 23:59**
- La compétition Kaggle se termine le **6 Novembre à 23:59**
- Vous devez déposer votre rapport sur Gradescope avant le **8 Novembre à 23:59**