ULaval **IFT&GLO**

Mini-Projet #2 (sur 15%)

à rendre le 19 mars 2021 à 23h59mn

Professeur Brahim Chaib-draa *GLO-7050 : Apprentissage machine en pratique*

Les projets du cours sont inspirés des projets du cours COMP 551, donné par les collègues de McGill, un grand merci à eux pour nous avoir donné l'autorisation de les utiliser.

Préambule

- Important : Ce mini-projet constitue un travail pouvant se faire en individuel ou en équipe de deux au maximum. Vous pouvez discuter avec les autres étudiants qui suivent le cours, mais en aucun cas vous ne devez reprendre le code et l'écrit d'autrui ; il vous est demandé d'élaborer les vôtres.
- Vous devriez soumettre votre travail sur MonPortail, et également à une compétition Kaggle. Pour celle-ci vous devez vous inscrire pour la compétition Kaggle en utilisant le courriel auquel vous êtes associé sur MonPortail (c.-à-d. @ulaval.ca). Pour s'inscrire à la compétition Kaggle utilisez l'adresse suivante : `https://www.kaggle.com/t/3d52b21ba1634ae6acaa07603af60cde`. Pour ce MiniProject 2, vous devez tout d'abord enregistrer votre équipe sur MonPortail. Ensuite, formez une équipe sur la compétition Kaggle portant le même nom d'équipe que MonPortail (vous devez utiliser le nom de votre équipe MonPortail comme nom d'équipe sur Kaggle). Toutes les soumissions Kaggle doivent être associées à une équipe valide enregistrée sur MonPortail.
- Si vous "empruntez" des idées, méthodes, démarches ou autres, merci d'indiquer vos sources dans le rapport.
- Il vous est fortement suggéré d'argumenter et/ou justifier vos réponses.
- Après la date de remise, vous avez jusqu'à une semaine pour remettre votre travail avec une pénalité de 20%. Au delà le travail vaut 0.
- Vous êtes libres d'utiliser les librairies telles que Numpy pour Python. Toutefois à moins d'être explicitement autorisées, vous ne devez pas utiliser des implémentations pré-existantes des algorithmes demandés, vous devez les implémenter par vous même.
- Si vous avez des questions concernant le travail, merci de passer par le Forum, en posant clairement vos questions.

Énoncé

Dans ce mini-projet, vous devez développer des modèles pour l'analyse des sentiments à partir du texte provenant du site web Amazon (`https://www.amazon.fr/`), un commerce en ligne où les clients commentent et évaluent les produits qu'ils ont achetés. L'objectif de ce projet est de

développer un modèle de classification binaire (supervisée) qui peut prédire le sentiment exprimé dans un commentaire (positif ou négatif). Vous serez en compétition avec d'autres équipes pour obtenir le meilleur score possible. **Cependant, votre performance sur la compétition n'est qu'un aspect de votre note finale. Il vous ai également demandé de mettre en oeuvre quelques modèles d'apprentissage et de rapporter leurs performances dans un rapport écrit.**

La page web de la compétition Kaggle comporte un lien pour télécharger les données. Cela correspond à un problème de classification en 2 classes (positif, négatif). L'ensemble de données (le dataset) n'est pas équilibré (c'est-à-dire qu'il y a plus de commentaires positifs que de commentaires négatifs). Les données sont fournies dans un fichier CSV. Chaque entrée dans le fichier CSV d'entraînement (l'ensemble d'entraînement) contient un ID (pour identifier chaque commentaire), le texte du commentaire et le sentiment cible pour ce commentaire (positif ou négatif). Pour le fichier CSV de test, chaque ligne (ou entrée) contient un ID de commentaire et le texte de ce commentaire (bien entendu, on ne fournit pas le sentiment pour les données de test, c'est à votre modèle de les prédire). Vous pouvez consulter et télécharger les données via ce lien : `https://www.kaggle.com/c/analyse-de-sentiments-amazon-glo-7050-h21/data`

Pour pouvoir soumettre votre solution sur Kaggle, vous devez produire un fichier CSV de prédiction où chaque ligne contient l'ID du commentaire et le sentiment prédit pour ce commentaire. Les données étant non équilibrées, vous serez évalués selon le score F1 moyen (Mean F1-score) qui tient compte de la précision et le rappel de chaque classe. Un exemple de format approprié pour le fichier de soumission peut être consulté à l'adresse suivante : `https://www.kaggle.com/c/analyse-de-sentiments-amazon-glo-7050-h21/overview/evaluation`.

Tâches à réaliser

Vous pouvez expérimenter avec n'importe quel modèle de votre choix, et vous êtes libres d'utiliser n'importe quelle bibliothèque pour l'extraction des caractéristiques. Cependant, vous devez satisfaire les conditions suivantes :

- Vous devez implémenter un modèle Bernoulli Bayes Naif (BN) (le modèle Bernoulli BN vu en cours) à partir de zéro (c'est-à-dire sans utiliser de bibliothèques externes telles que SciKit learn). Vous êtes libre d'utiliser n'importe quelle technique de prétraitement de texte avec ce modèle. **Astuce 1 :** vous auriez peut-être besoin d'utiliser le lissage (ou correction) de Laplace avec votre modèle Bernoulli BN. **Astuce 2 :** vous pouvez choisir le vocabulaire de votre modèle (c'est-à-dire les mots que vous voulez inclure ou ignorer), mais vous devez justifier vos choix. (**Note** : les commentaires dans ce dataset sont écrits en français)
- Vous devez aussi utiliser au moins deux classificateurs différents de la bibliothèque d'apprentissage Scikit-learn (qui ne sont pas des Bernoulli Bayes Naif). Les options possibles sont les suivantes :
  - La régression logistique `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`
  - Les arbres de décision `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html`

  - Les machines à vecteurs de support `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`
- Vous devez développer un pipeline de validation de modèles (par exemple, en utilisant la validation croisée K-Fold ou un ensemble de validation maintenu de l'ensemble d'entrainement) et rapporter les performances des différents modèles mentionnés ci-dessus.
- IMPORTANT : Vous devez évaluer tous les modèles que vous avez choisis ainsi que le Bernoulli BN que vous avez implémenté (c'est-à-dire les modèles d'apprentissage de scikit-learn et le Bernoulli BN) en utilisant votre pipeline de validation (c'est-à-dire sans les soumettre sur Kaggle) et rendre compte des différentes performances dans votre rapport (Rappel, Précision, F1-score, etc.). Idéalement, vous ne devriez utiliser votre "meilleur" modèle que dans le cadre de la compétition Kaggle, puisque vous êtes limité à deux soumissions Kaggle par jour.

Déliverables

Vous devez soumettre deux fichiers à MonPortail (**en utilisant précisément les noms et types de fichiers suivants**) :

1. **code.zip** : Une collection de tous les fichiers de code .py, .ipynb et tout autre code de soutien, qui sont écrits avec la version Python 3.x. Vous devez inclure votre implémentation de Bernoulli BN et il faut que ça soit possible pour Amar de reproduire tous les résultats dans votre rapport ainsi que votre soumission au classement Kaggle en utilisant le code que vous avez fourni. Veuillez soumettre un LISEZ-MOI (ou README) détaillant les paquets (packages) Python que vous avez utilisés ainsi que les instructions pour reproduire vos résultats (optionnel : vous pouvez inclure un fichier "requirements.txt" pour mentionner les versions des packages que vous avez utilisés).
2. **rapport.pdf** : Votre rapport de projet (5 pages maximum) en format pdf (voir détails ci-dessous).

Rédaction du rapport

Vous devez soumettre un rapport de 5 pages maximum (à simple interligne, police de 10 pt ou plus ; des pages supplémentaires pour les références et les annexes peuvent être utilisées). Nous vous recommandons d'utiliser LaTeXpour rédiger votre rapport et d'utiliser bibtex pour les citations. Vous êtes libre de structurer le rapport comme bon vous semble ; vous trouverez ci-dessous des directives générales et des recommandations, mais il ne s'agit que d'une suggestion de structure vous n'êtes pas obligé de la suivre.

- **Résumé (100-250 mots)** : Récapitulatif du projet et vos principales conclusions.
- **Introduction (5+ phrases)** : Un résumé du projet, de l'ensemble des données et de vos conclusions les plus importantes. Ceci doit être similaire au résumé mais plus détaillé.
- **Travaux connexes (4+ phrases)** : Un résumé de la littérature relative au problème de l'analyse des sentiments.

— **Dataset (3+ phrases)** : Décrivez très brièvement le dataset et toute méthode de prétraitement commune à vos approches (par exemple, le tokenizing). Note : Vous n'avez pas besoin de vérifier explicitement que les données satisfont l'hypothèse de la i.i.d. (ou à toute autre hypothèse formelle de classification linéaire).
— A**pproche proposée (7+ phrases)** : Décrivez brièvement les différents modèles que vous avez mis en oeuvre ainsi que les caractéristiques (features) que vous avez conçues et choisies, en fournissant des citations si nécessaire. Si vous utilisez ou développez un modèle existant basé sur des travaux publiés, il est essentiel que vous le citiez et le reconnaissiez correctement. Discutez également le choix et l'implémentation de vos modèles. Incluez toute décision concernant la répartition train/validation, les stratégies de régularisation, toute astuce d'optimisation, le choix des hyperparamètres, etc. Il n'est pas nécessaire de fournir le détail des modèles que vous utilisez, mais vous devez fournir au moins quelques phrases concernant les fondements (et la motivation) pour chaque modèle.
— **Résultats (7+ phrases, éventuellement avec des figures ou des tableaux)** : Présentez les résultats sur les différents modèles que vous avez mis en oeuvre (précision, rappel, exactitude et score F1 sur l'ensemble de validation, temps d'exécution, ...etc.). Vous devez indiquer dans cette section le score sur l'ensemble de test de votre meilleur modèle (le score que vous avez eu sur Kaggle), mais la plupart de vos résultats doivent être sur votre ensemble de validation (ou de validation croisée).
— **Discussion et conclusion (3+ phrases)** : Récapitulez les principales conclusions du projet et les éventuelles futures pistes de recherche.
— **État des contributions (1 à 3 phrases)** : Indiquez la répartition de la charge de travail de chaque membre de l'équipe s'il y a lieu.

Évaluation

Le mini-projet est noté sur 100 points, et la répartition des points est la suivante :

— Compétition (50 points)
  — Les performances de votre modèle seront évaluées sur la compétition Kaggle. Votre note sera calculée sur la base du score obtenu par votre modèle sur un ensemble de test séparé Remarque : Lorsque vous faîtes une soumission, Kaggle vous calcule un score et vous donne la possibilité de l'afficher sur le tableau de classement (Leaderboard). Cependant, ce score est calculé sur une partie de l'ensemble de test. L'autre partie est gardée secrète par Kaggle et ne sera utilisée qu'à la fin de la compétition.
  — Le calcul de la note est une interpolation linéaire entre la performance d'un modèle aléatoire, d'une référence TA (score obtenu par Amar) et du deuxième meilleur score de la compétition. Les trois premières équipes de la compétition reçoivent toutes la note complète (50 points).
  — Ainsi, si par exemple, X indique votre score sur l'ensemble de test, R indique le score du modèle aléatoire, B indique le score de la deuxième meilleure équipe, et T indique le score de l'assistant (TA Baseline), votre score sera calculé comme suit :

$$\text{points} = 50 * \begin{cases} 0 & \text{if } X < R \\ \frac{X-R}{T-R} * 0.75 & \text{if } X > R \text{ and } X \leq T \\ \frac{X-T}{B-T} * 0.25 + 0.75 & \text{if } X > T \text{ and } X \leq B \\ 1 & \text{if } X > B \end{cases}$$

L'équation peut paraître compliquée, mais l'idée de base est la suivante :

- La référence R représente le score nécessaire pour obtenir plus de 0% sur la compétition (il faudrait avoir un meilleur score qu'un modèle aléatoire), la référence de l'assistant T représente le score nécessaire pour obtenir au moins 75% sur la compétition, et B le score de la 2ème meilleure équipe représente le score nécessaire pour obtenir 100%.
- Si votre score se situe entre R et T, alors votre note est une interpolation linéaire entre 0% et 75%.
- Si votre score se situe entre T et le score du 2ème meilleur groupe B, alors votre note est une interpolation linéaire entre 75% et 100% des points de la compétition.

— De plus, la première équipe reçoit un bonus de 5 points.

— Qualité de la rédaction et de la méthodologie proposée (50 points)
Votre rédaction sera jugée en fonction de sa qualité scientifique en rapport avec les questions suivantes (incluses mais non limitée à) :

- Avez-vous rapporté sur toutes les expériences et comparaisons requises ?
- La méthodologie que vous proposez est-elle techniquement valable ?
- Dans quelle mesure vos expériences sont-elles détaillées / rigoureuses / étendues ?
- Votre rapport décrit-il clairement la tâche sur laquelle vous avez travaillé, la configuration expérimentale, les résultats et les figures (par exemple, n'oubliez pas les axes et les légendes sur les figures, n'oubliez pas d'expliquer les chiffres dans le texte).
- Votre rapport est-il bien organisé et cohérent ?
- Votre rapport est-il clair et exempt d'erreurs grammaticales et de fautes de frappe ?
- Votre rapport comprend-il une discussion adéquate des travaux et citations connexes ?

Remarques

Vous êtes censé faire preuve d'initiative, de créativité, de rigueur scientifique et d'esprit critique, et de bonnes capacités de communication. Vous n'avez pas besoin de vous limiter aux exigences énumérées ci-dessus - n'hésitez pas à aller au-delà, et d'explorer davantage.

Vous pouvez discuter des méthodes et des questions techniques avec les membres d'autres équipes, mais vous ne devez en aucun cas partager le code ou les données. Toute équipe qui triche (par exemple en utilisant des données externes à la compétition ou en utilisant des ressources sans références appropriées) sur le code, les prédictions ou le rapport écrit recevra une note de 0 pour le mini projet 2.

Règles spécifiques à la compétition Kaggle

Important

1. Ne pas tricher ! Vous devez soumettre un code qui permet de reproduire le score de votre solution soumise sur Kaggle.
2. Les données de la compétition sont basées sur un ensemble de données publiques. Vous ne devez pas tenter de tricher en cherchant des informations sur l'ensemble de test. Les soumissions dont la précision et/ou les prédictions sont douteuses seront signalées et donneront lieu à la note 0.
3. Ne pas essayer de faire plus de soumissions. Vous recevrez la note 0 pour avoir créé intentionnellement un nouveau compte dans le but de faire plus de soumissions sur Kaggle.