SAE S5 Maths : prédictions d'évaluations à partir d'avis textuels

Présentation du sujet

L'objectif de cette SAE est de mettre en place des modèles pour réaliser des tâches de clustering, régression et classification, sur des données textuelles.
Vous travaillerez en binôme.

Vous devez choisir un jeu de données parmi ceux disponibles sur le site suivant : `https://amazon-reviews-2023.github.io`. Vous pouvez également choisir un jeu de données sur un autre site (par exemple, parmi ceux du site kaggle).

Quel que soit le jeu de données choisi, les contraintes suivantes doivent être respectées :

— il ne faut pas que 2 binômes différents travaillent sur le même jeu de données, ni sur un des jeux de données déjà utilisés par les étudiants des autres groupes de TD ;
— il faut au moins qu'une variable caractéristique corresponde à des avis d'utilisateur, rédigés en langage naturel ;
— la variable cible correspondra à une évaluation (une note) qui sera souvent un nombre (mais pas nécessairement). Il faudra que cette évaluation ne soit pas binaire, mais concerne au moins 3 valeurs différentes. C'est cette variable cible qui sera prédite ;
— les jeux de données peuvent être volumineux (de l'ordre du giga ou plus pour certains). Vos traitements doivent pouvoir tourner sur les machines de l'IUT. Cela signifie que vous pouvez tout-à-fait récupérer des données volumineuses hors de l'IUT mais il faudra en choisir un extrait pour que les traitements puissent s'effectuer dans des temps de calcul raisonnables.

Vous utiliserez exclusivement `scikit-learn`, comme bibliothèque pour réaliser de l'apprentissage automatique, ainsi que `spacy` pour réaliser les traitements linguistiques. Une classe de `scikit-learn` vous sera utile pour représenter le texte, qui est en langage naturel, sous une forme exploitable par les systèmes de classification : `CountVectorizer`. Cette classe permet de représenter un texte sous la forme d'une matrice de nombres. Vous pouvez également regarder la classe `TfidfVectorizer`.

Travail à réaliser

Le travail à réaliser est composé de plusieurs étapes (seule la dernière étape n'est pas obligatoire). Comme vous travaillerez eb binôme, vous pourrez vous répartir le travail, une fois les deux premières étapes réalisées ensemble.

(a) **Choix du jeu de données**
Après avoir constitué votre binôme, choisissez le jeu de données sur lequel vous souhaitez travailler, indiquez-le dans le tableur dont le lien se trouve sur madoc puis demandez la validation de votre jeu de données à votre enseignant.

(b) **Pré-traitement des données**
En général, les données ne peuvent pas être utilisées directement. Il faudra mettre en place des pré-traitements (suppression des éventuelles lignes en doublon, traitement des cellules vides, ...). Il faudra également séparer vos données en un ensemble d'apprentissage, un ensemble de validation et un ensemble de test. Ces 3

sous-ensembles seront utilisés tout au long des expérimentations effectuées pour répondre aux questions suivantes. Vous pouvez également considérer, dans un premier temps, un sous-ensemble de votre ensemble d'apprentissage, pour mettre au point chacune de vos approches.

(c) **Clustering des avis**
Vous utiliserez l'algorithme des k-moyennes pour trouver le nombre de clusters optimal pour partitionner vos données. Pour ce faire, vous devrez :
— réaliser plusieurs expérimentations, avec des pré-traitements différents sur le texte des avis ;
— optimiser les hyperparamètres du modèle entraîné, pour chaque expérimentation.
Le but est de déterminer, à l'aide de plusieurs expérimentations différentes, la configuration permettant d'obtenir les meilleurs résultats, sur votre ensemble de validation, en utilisant les métriques vues en TD de Modélisations mathématiques. Vous appliquerez enfin le modèle de clustering choisi, sur votre ensemble de test.

(d) **Classification binaire des avis (deux classes à prédire)**
Il y aura deux classes à prédire : elles correspondront à *avis favorable* et *avis défavorable*. Vous utiliserez un des algorithmes de classification qui a été vu en TD de Modélisations mathématiques. Vous devrez :
— réaliser plusieurs expérimentations, avec des pré-traitements différents sur le texte des avis ;
— optimiser les hyperparamètres du classifieur, pour chaque expérimentation.
Le but est également de déterminer, à l'aide de plusieurs expérimentations différentes, la configuration permettant d'obtenir les meilleurs résultats, sur votre ensemble de validation, puis de calculer les résultats obtenus sur votre ensemble de test. Faites attention à la prise en compte du déséquilibre que vous pourrez avoir entre vos deux classes, en termes de nombre d'avis dans chacune des classes.

(e) **Classification multiclasse des avis (plus de deux classes à prédire)**
Le but est de réaliser les mêmes expérimentations qu'à la question précédente mais en utilisant uniquement les pré-traitements du texte ayant permis d'obtenir les meilleurs résultats lors de la classication binaire. Vous devrez maintenant prédire plus de deux classes possibles. Vous pouvez reprendre le nombre de clusters optimal, que vous avez trouvé précédemment, pour fixer le nombre de classes considérées ici. Vous indiquerez les notes auxquelles correspondent chacune de vos classes.

(f) **Régression pour la prédiction des notes des avis**
Le but est maintenant de prédire les notes des avis, en utilisant la régression linéaire. Dans un premier temps, vous utiliserez uniquement les avis textuels pour effectuer votre régression linéaire, sur votre ensemble d'apprentissage. Dans un second temps, vous pourrez essayer d'ajouter les autres informations disponibles sur les avis, pour voir si ces informations peuvent améliorer les résultats de la régression.

(g) **Bonus (pour aller plus loin)**
Vous pouvez utiliser d'autre méthodes de clustering ou de classification, en plus de celles vues en cours, mais elles doivent pouvoir être exécutées sur les ordinateurs de l'IUT. En revanche, vous devrez avoir compris le fonctionnement de ces nouvelles méthodes et être capable d'en expliquer le principe (dans le rapport et lors de la soutenance).

Rendu du travail

Le rendu sera réalisé sur madoc, au plus tard le **vendredi 10 janvier 2025, à 23h55**.

Vous devez rendre le travail dans une archive au format zip qui sera nommée avec le nom de chaque étudiant(e) du binôme. L'archive comprendra :

— le code (notebooks) de vos expérimentations (qui doit fonctionner à l'IUT), dans un répertoire *code* ;
— les données utilisées, dans un répertoire *data* (si elles sont trop volumineuses pour être stockées sur madoc, vous indiquerez un lien vers un cloud sur lequel elles seront stockées) ;
— un rapport de votre travail, en pdf, dans un répertoire *report*.

Voici les différentes parties attendues pour le rapport (comprenant entre 5 et 15 pages) :

— une introduction ;
— une présentation du jeu de données choisi ;
— des explications sur les pré-traitements réalisés sur les données ;
— une présentation des différentes expérimentations (les caractéristiques choisies, comment vous avez optimisé les hyperparamètres, un tableau récapitulatif des résultats globaux obtenus...) ;
— une partie bonus, si elle a été réalisée ;
— une conclusion.

Soutenance finale

La soutenance de votre travail aura lieu en semaine 3. Le temps prévu pour chaque soutenance vous sera indiqué ultérieurement.

Chaque binôme présentera le travail réalisé. Chaque membre du binôme devra pouvoir répondre aux questions. Le jury sera vigilant à ce que tout ce qui a été codé ou écrit ait été compris par les deux étudiants du binôme.