Université Paris-Panthéon-Assas

Projet Machine Learning

L'objectif du projet est de prédire les décisions de la Cour de Cassation en utilisant les arguments (moyens) des avocats. La base de données contient 9525 décisions de la Cour de Cassation. Il y a deux variables :

— La variable *solution* : elle contient la décision de la Cour de Cassation
— La variable *moyens* : elle contient les moyens avancés par les avocats

Ces deux variables sont au format texte.

Présentation globale

Le projet a deux objectifs :

— Obtenir le meilleur taux de prédiction en utilisant la métrique du f1-score pondéré
— Implémenter les algorithmes vus en cours et comprendre comment ils fonctionnent

1 Critères de notation 1 : Implémenter des algorithmes (/6)

Vous devez comparer la performance des algorithmes suivants en utilisant les mêmes variables explicatives :

— Régression Logistique
— K plus proches voisins
— Random Forest
— Machine à vecteur de support
— Réseaux de neurones
— Un modèle non vu en cours

Pour chaque modèle, indiquez la précision, rappel, f1-score pondéré, temps nécessaire pour obtenir les résultats, principaux hyperparamètres utilisés.

2 Critères de notation 2 : Comprendre comment obtenir la meilleure performance

2.1 Choix des variables explicatives (/2)

Pour le modèle de votre choix parmi la liste de la Section 1, vous devez :

— Utiliser au moins quatre différents ensembles de variables explicatives (X). Ces changements doivent être substantiels, ex : faire varier la dimension des n-grams, le choix du nombre et de la fréquence des caractéristiques.
— Montrer comment l'erreur de test évolue avec ces choix

2.2 Choix des Hyperparamètres (/2)

Pour quatre modèles de votre choix parmi la liste de la Section 1, vous devez :

— Faire varier les hyperparamètres de votre choix
— Montrer comment l'erreur de test évolue avec les hyperparamètres

3 Critères de notation 3 : Choix et évaluation du modèle le plus performant (/4)

Pour le modèle obtenant le meilleur f1-score pondéré, vous devrez :
— Expliquer pourquoi vous avez choisi ce modèle (algorithme, hyperparamètre, variables explicatives)
— Indiquer la précision, le rappel, le f1-score pondéré et le taux de classification correct
— Comparer vos résultats avec un classificateur naïf (*DummyClassifier*)

4 Critères de notation 4 : Présentation et note de synthèse (/6)

Pour la présentation

1. Contenu
   — Une réponse à chacune des sections ci-dessus
   — Présentation de la démarche et des limites potentielles
   — Présentation rapide de la structure et du contenu du code (ainsi que les fonctions utilisées sur python pour répondre à chacune des sections)
2. Forme
   — Respect du temps imparti (10 minutes)
   — Présentation sans notes écrites

Pour la note de synthèse :
— Note de 2 pages
— Réponse aux critères de notation 1, 2.1, 2.2 et 3
— Résumé de la démarche

Important : vous devez m'envoyer votre code la veille de la présentation.

Date de la présentation : Mardi 13 décembre 2022