Projet TP - Machine Learning

Licence Flex

24 avril 2024

L'objectif de ce projet est de mener une étude de cas en utilisant des outils de machine learning. Un sujet de classification ou de régression vous sera remis en début de la première séance de projet afin de faire les meilleures prédictions. Merci de respecter les sujets qui vous seront donnés :

- sur le jeu de données 'ozone.csv', prédire la concentration d'ozone dans l'air. Les algorithmes que vous essayerez sont :
  - k-plus proches voisins
  - régression avec pénalité lasso
  - régression avec pénalité ridge

  Métrique d'évaluation : RMSE
- Classification : sur le jeu de données 'titanic.csv', prédire si un individu survit ou non. Les algorithmes que vous essayerez sont :
  - k-plus proches voisins
  - SVM
  - régression logistique

  Métrique d'évaluation : Accuracy

Dans l'ensemble de ces sujets, vous suivrez un processus similaire et proche de ce que nous avons fait en TP. Assurez-vous de nettoyer les données, d'utiliser convenablement les algorithmes, et d'utiliser les hyperparamètres optimaux pour les algorithmes qui vous ont été attribués. Il sera sûrement nécessaire de faire du feature engineering afin d'améliorer vos modèles.

Il sera important de soigner votre présentation des résultats. Pour les étudiants travaillant sur un algorithme de classification, entraîner un modèle simplifié en 2 dimensions par ACP puis afficher ses zones de décision peut être pertinent. Par ailleurs, il est possible d'utiliser les outils d'évaluation vus en TP (matrice de confusion, courbe ROC si c'est possible...). Pour les étudiants travaillant en régression, vous pourrez comparer les résultats obtenus avec votre algorithme à ceux donnés par une régression linéaire effectuée en parallèle sur les données.

Questions pratiques :

- 3 étudiants par groupe
- Par groupe un jeu de données
- Jour de la présentation : 21 Mai 2024 sur votre créneau de TP
- Format de la présentation : 10 minutes de présentation + 5 minutes de questions.
- Support de présentation : diapositives

Pour la présentation voici une idée de plan que vous pourriez suivre :

- Présentation de la question, des données (dont étude statistique)
- Nettoyage des données
- Feature engineering
- Présentation d'un des algorithmes (au choix)
- Recherche des hyperparamètres optimaux
- Présentation des résultats.
- Conclusion : regard critique sur les résultats et méthodes employées

Sujet sur la régression

Le jeu de données que vous utiliserez se trouve dans le fichier 'ozone.csv'. L'objectif, sur ces données, est d'améliorer la prévision déterministe (MOCAGE), calculée par les services de MétéoFrance, de la concentration d'ozone dans certaines stations de prélèvement. Les données ont été extraites et mises en forme par le service concerné de Météo France. Elles sont décrites par les variables suivantes :

- **JOUR** Le type de jour ; férié (1) ou pas (0) ;
- **O3obs** La concentration d'ozone effectivement observée le lendemain à 17h locales correspondant souvent au maximum de pollution observée ;
- **MOCAGE** Prévision de cette pollution obtenue par un modèle déterministe de mécanique des fluides (équation de Navier et Stockes) ;
- **TEMPE** Température prévue par MétéoFrance pour le lendemain 17h ;
- **RMH2O** Rapport d'humidité ;
- **NO2** Concentration en dioxyde d'azote ;
- **NO** Concentration en monoxyde d'azote ;
- **STATION** Lieu de l'observation : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache et Plan de Cuques ;
- **VentMOD** Force du vent ;
- **VentANG** Orientation du vent.

Le but est de prédire la concentration d'ozone.

Sujet sur la classification

Le jeu de donénes que vous utiliserez se trouve dans le fichier 'titanic.csv'. L'objectif, sur ces données, est de prédire la mort des passagers du titanic.

Les données sont décrites par les variables suivantes :

- **Survived** Le passager a survécu au naufrage ; survécu (1) ou mort (0) ;
- **Pclass** Classe du ticket du passager ; 1ère, 2ème ou 3ème ;
- **Name** Nom du passager ;
- **Sex** Sexe du passager ; male ou female ;
- **Age** Age du passager ;
- **SibSp** Nombre de frères/soeurs/époux/épouse à bord du bateau ;
- **Parch** Nombre de parents ou enfants à bord du bateau ;
- **Ticket** NUméro de ticket du passager ;
- **Fare** Prix du ticket ;
- **Cabin** Numéro de la cabine attribuée ;
- **Embarked** Port d'embarcation ; C = Cherbourg, Q = Queenstown, S = Southampton ;