Projet de Machine Learning

Deadline: 15 Mars 2022

Création du jeu de données

Vous allez travailler sur le jeu de données du Defi IA. Vous choisissez une station sur laquelle vous allez faire la prédiction de la variable Y : cumul de pluie sur la journée. Pour cela, vous pouvez utiliser les observations de la veille sur cette station et sur les 5 stations les plus proches, ainsi que les prévisions des modèles de Météo France sur les 5 positions les plus proches et les données météorologiques disponibles (température, vitesse du vent etc.). La première étape sera de créer un dataset d'apprentissage contenant les données que vous allez utiliser pour construire vos modèles conformément à ce qui est indiqué ci-dessus et un dataset de test pour comparer les différents modèles optimisés. Pour constituer le jeu d'apprentissage et le jeu de test, vous prendrez les données des années 2016 et 2017 puis partagerez ces données de manière aléatoire en un échantillon d'apprentissage et un échantillon de test. L'échantillon d'apprentissage comportera 75% des données et l'échantillon de test 25% des données. Les deux fichiers, nommés "dataapp" et "datatest" doivent être déposés sous Moodle pour le 17 Janvier.

Questions posées

Analyse exploratoire des données

L'objectif dans un premier temps est d'explorer les différentes variables, étape préliminaire indispensable à l'analyse. Ci-dessous sont précisées quelques questions basiques. Vous pouvez compléter l'analyse selon vos propres idées.

1. Commencez par une analyse descriptive unidimensionnelle des données. Voyez-vous des anomalies? Les distributions sont-elles comparables entre le jeu d'apprentissage et le jeu de test ?
2. Poursuivez avec une analyse descriptive multidimensionnelle. Utilisez des techniques de visualisation : par exemple scatterplot, correlation plot, boxplot. Quelles variables vous semblent les plus corrélées avec la variable à prédire ? Analysez les interactions.
3. Réalisez une analyse en composantes principales des données quantitatives et interprétez les résultats.

Modélisation

Nous considérons maintenant le problème de la prédiction du point de vue de l'apprentissage automatique, c'est-à-dire en nous concentrant sur les performances du modèle. L'objectif est de déterminer les meilleures performances que nous pouvons attendre, et les modèles qui les atteignent. Voici quelques questions pour vous guider.

1. Tout d'abord, divisez les données en un échantillon d'apprentissage et un échantillon test. Vous utiliserez les données de l'année 2016, 2017 comme échantillon d'apprentissage. Les données de l'année 2018 formeront l'échantillon de test. Pourquoi cette étape est-elle nécessaire lorsque nous nous concentrons sur les performances des algorithmes ?
2. Comparez les performances d'un modèle linéaire avec/sans pénalisation, d'un SVM, d'un arbre optimal, d'une forêt aléatoire, du boosting, et de réseaux de neurones. Justifiez vos choix (par exemple le noyau pour le SVM), et ajustez soigneusement les paramètres (à l'aide d'un échantillon de validation ou par validation croisée). Interprétez les résultats et quantifiez l'amélioration éventuelle apportée par les modèles non linéaires.

3. Comparez les différents modèles optimisés sur votre échantillon test. Quels sont les modèles les plus performants ? Quel est le niveau de précision obtenu ?

4. Interprétation et retour sur l'analyse des données. Vos résultats sont-ils cohérents avec l'analyse préliminaire des données, par exemple en ce qui concerne l'importance des variables ?

Organisation et rapport à rendre

Vous réaliserez le projet par groupe de 3 ou 4 étudiant.e.s. **Deadline: 15 Mars 2022**. Comme livrable, vous rendrez un rapport en format pdf ne dépassant pas 30 pages. Il doit comprendre une introduction, une description succinte des algorithmes utilisés, une interprétation des résultats, une conclusion, etc. De plus, vous rendrez un notebook Jupyter, soit en R, soit en Python, ainsi qu'un fichier contenant le jeu de données utilisé. N'oubliez pas de commenter votre code. Le dépôt se fera sur Moodle : chaque groupe téléchargera un fichier zip contenant le rapport (format pdf), le jeu de données et le notebook Jupyter.

L'évaluation tiendra compte de la présentation du rapport et de la rédaction (clarté, argumentation, etc.), de la cohérence de l'étude, de la qualité de présentation du notebook, des interprétations des résultats (graphiques et autres).