43923 MILLET Yann

Prédiction de feux de forêt grâce au Machine Learning

J'ai le projet d'exercer dans le domaine de l'intelligence artificielle, ce qui a orienté ma recherche de sujet vers une thématique applicable dans ce domaine. La sécurité passe notamment par la prévention de problèmes, auxquels le machine learning s'applique bien.
Dans les 3 sous-thèmes explicités et attrayant aux enjeux sociétaux on retrouve la sécurité et l'environnement. L'anticipation de feux de forêt s'inscrit ainsi bien dans le thème de l'année puisqu'elle concerne la sécurité d'un environnement et des individus y évoluant.

Positionnement thématique (ETAPE 1)

INFORMATIQUE (Informatique pratique), MATHEMATIQUES (Mathématiques Appliquées), INFORMATIQUE (Informatique Théorique).

Mots-clés (ETAPE 1)

Mots-Clés (en français)	**Mots-Clés** (en anglais)
Apprentissage Automatique	*Machine Learning*
Machine à Vecteurs de Support	*Support Vectors Machine*
feux de forêt	*forest fires*
Régression	*Regression*
Optimisation	*Optimization*

Bibliographie commentée

Dans le cadre de ce TIPE, j'ai séparé mon travail en *trois parties*, la **recherche de données et leur analyse**, la **compréhension de l'algorithme** à utiliser et enfin la mise en **application pratique**.

La *première* de ces parties a d'abord eu lieu sur le site *UCI Machine Learning Repository* (référence bibliographique 7), où j'ai cherché des jeux de données en rapport avec l'environnement, si possible de manière à pouvoir les relier avec leur sécurité pour coïncider au mieux avec le thème de l'année et mes préoccuppations sociétales actuelles. Pour leur analyse, seul un environnement de développement python a été nécessaire, ainsi que les bibliothèques bien connues *numpy*, *matplotlib* et *pandas*. La principale difficulté pour **analyser ces données** a été de bien se les représenter, de trouver quelle caractéristique représenter en fonction de quelles autres. Mon principal obstacle a été que l'interprétation de certaines caractéristiques nécessite des connaissances en météorologie dont je ne disposais pas. L'analyse s'est donc restreinte à l'étude des **tailles des données** (leur échelle) et la **représentation** des fréquences d'apparation de certaines valeurs, ce qui a cependant été suffisant pour trouver une manière presque optimale d'entraîner l'algorithme et de régler ses paramètres.

La *deuxième* et la *troisième* partie ont nécessité, pour la compréhension de l'algorithme **SVR non**

linéaire qui sera utilisé dans ce TIPE, l'appui des ouvrages *The Elements of statistical learning* (référence bibliographique 1), *Le Machine Learning avec python* (référence 2), *Analyse Numérique et Optimisation* (référence 4) ainsi que le cours de Paul Paisitkriangkrai de l'Université d'Adelaide *Linear Regression and Support Vector Regression* (référence 6), l'étude *A Data Mining approach to predict forest fires using meteorological data* (référence 3) et le site de scikit-learn (référence 5) qui est la bibliothèque python dont provient l'algorithme utilisé pour l'application pratique.
Pour aborder l'algorithme je me suis d'abord aidé du livre *Le Machine Learning avec python*, qui permet de bien débuter dans le domaine du Machine Learning et l'implémentation en langage python. Je me suis ensuite intéressé aux **aspects mathématiques de l'algorithme** choisi (SVM pour la regression, ou SVR), dont j'ai compris le fonctionnement grâce aux références 1, 4, 5 et 6. La faible utilisation actuelle de ce type d'algorithme, et par conséquent le peu d'informations à son sujet ont rendu nécessaire le croisement de ces documents pour éviter au maximum toute erreur de compréhension. Les références 1, 5 et 6 donnent les **formules et équations qui régissent le fonctionnement de l'algorithme**, tandis que *Analyse Numérique et Optimisation* présente les concepts mathématiques qui permettent de **résoudre lesdites équations**.
Enfin pour la *troisième* et dernière partie de ce TIPE, il a fallu mettre en pratique ce qui a été présenté précédemment. Pour cela, *Le Machine Learning avec python* et dans une plus forte mesure la *documentation de scikit-learn* ont été très utiles pour comprendre comment utiliser les algorithmes développés par la communauté scikit-learn. En effet, la référence bibliographique n°2 donne de bonnes bases pour entraîner ces algorithmes dans un premier temps, mais comme souvent en informatique la documentation de la bibliothèque utilisée se révèle d'un grand secours pour une utilisation plus fine et optimale de ses objets. Ainsi la référence bibliographique n°5 m'a permis d'**ajuster** au mieux les **hyper-paramètres** du SVR non linéaire afin de tendre vers les meilleurs résultats possibles.

Pour finir, l'étude de P. Cortez et A. Morais qui constitue la référence bibliographique n°3 était jointe au set de données que j'ai utilisé pour l'application pratique. Elle m'a été utile pour ajuster les données et comparer quelques paramètres communs à leur algorithme et au mien, mais également pour choisir certaines parties de l'algorithme, car leurs choix avaient obtenu des résultats convenables. Enfin, cette étude m'a aussi permis de mettre en perspective mes résultats, en comparant les ajustements que j'ai fait (et pas fait) avec les leurs.

Problématique retenue

Comment **anticiper les feux de forêt** par **apprentissage automatique** (Machine Learning) et statistique afin d'améliorer la réponse humaine devant ces catastrophes?

Objectifs du TIPE

Les objectifs de ce TIPE sont de montrer l'utilité de l'apprentissage automatique dans l'**aide à la décision**, de présenter une manière de procéder grâce à un **algorithme de regression**, ainsi que d'expliquer son **fonctionnement** et son **réglage**. Enfin, ce TIPE vise également à donner un **exemple d'application** d'un algorithme de machine learning à la préservation de l'environnement, à travers la problématique des feux de forêts.

Références bibliographiques (ETAPE 1)

[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman : The Elements of statistical learning : *Springer, Series in Statistics, https://link.springer.com/book/10.1007%2F978-0-387-84858-7*

[2] Andreas C. Müller, Sarah Guido : Le Machine Learning avec python : *O'Reilly, Jupyter, https://livre.fnac.com/a11113496/Sarah-Guido-Le-machine-learning-avec-Python*

[3] Paulo Cortez, Anibal Morais : A Data Mining approach to predict forest fires using meteorological data : *http://www.dsi.uminho.pt/~pcortez/fires.pdf*

[4] Grégoire Allaire : Analyse Numérique et Optimisation : *https://www.editions.polytechnique.fr/?afficherfiche=78*

[5] Communauté scikit-learn : Documentation bibliothèque open-source scikit-learn : *https://scikit-learn.org/stable/index.html*

[6] Paul Paisitkriangkrai : Linear Regression and Support Vector Regression : *The University of Adelaide, https://cs.adelaide.edu.au/~chhshen/teaching/ML_SVR.pdf*

[7] Paulo Cortez, Anibal Morais : UCI Machine Learning Repository : *https://archive.ics.uci.edu/ml/datasets/Forest+Fires*

DOT

[1] *Septembre 2020 - Octobre 2020 : Recherche du jeu de données, analyse de ses caractéristiques et détermination en conséquence de l'algorithme qui sera utilisé. Tentative puis renoncement de codage de l'algorithme lui-même.*

[2] *Novembre 2020 - Février 2021 : Compréhension du fonctionnement mathématique de l'algorithme et détermination des transformations nécessaires à sa bonne utilisation.*

[3] *Décembre 2020 : Première implémentation de l'algorithme SVR issu de la bibliothèque scikit-learn, listage des améliorations à y apporter (hyperparamètres du modèle, paramètres de séparations et d'apprentissage).*

[4] *Janvier 2021 : Ecriture des fonctions annexes à l'algorithme (calcul de score, transformation des données textuelles et données numériques).*

[5] *Février 2021 : Début de l'implémentation pratique du processus d'apprentissage statistique (extraction complète des données, transformations, entraînements, tests).*

[6] *Mars 2021 : Création des graphiques et histogrammes pour la représentation.*