Job de veille

Kawther ELTARR

December 14, 2020

1 Outils

- Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique, qui vise à simplifier la gestion des paquets et de déploiement. [1]
  Anaconda propose un outil de gestion de packages appelé Conda. Ce dernier permettra de mettre à jour et installer facilement les librairies dont on aura besoin pour nos développements.[2]

- Jupyter est une application web utilisée pour programmer dans plus de 40 langages de programmation, dont Python, Julia, Ruby, R, ou encore Scala2. Jupyter est une évolution du projet IPython. Jupyter permet de réaliser des calepins ou notebooks, c'est-à-dire des programmes contenant à la fois du texte en markdown et du code en Julia, Python, R... Ces calepins sont utilisés en science des données pour explorer et analyser des données.[3]

2 Librairies

- Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. Pandas est un logiciel libre sous licence BSD2. [4]

- Seaborn est une bibliothèque de visualisation de données Python basée sur matplotlib et s'intègre étroitement aux structures de données des pandas. Elle fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.[5]
  La liste des tracés pris en charge comprend des tracés de distribution univariés et bivariés, des tracés de régression et un certain nombre de méthodes pour tracer des variables catégorielles.[6]

- Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques5. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy6. Matplotlib est distribuée librement et gratuitement sous une licence de style BSD4.[7]

- Plotly est une société d'informatique technique basée à Montréal, au Québec, qui développe des outils d'analyse et de visualisation de données en ligne. Plotly fournit des outils de graphiques, d'analyse et de statistiques en ligne pour les individus et la collaboration, ainsi que des bibliothèques de graphiques scientifiques pour Python, R, MATLAB, Perl, Julia, Arduino et REST. [8]

- Scikit-Learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs. Elle propose dans son framework de nombreuses bibliothèques d'algorithmes à implémenter clé en main, à disposition des data scientists.

  Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy. [9]

- StatsModels est un package Python qui permet aux utilisateurs d'explorer des données, d'estimer des modèles statistiques et d'effectuer des tests statistiques . Une liste complète de statistiques descriptives, de tests statistiques, de fonctions de traçage et de statistiques de résultats est disponible pour différents types de données et chaque estimateur. Il complète le module de statistiques de SciPy. [10]

3 Machine Learning

- Classification (non supervisée) : trouver dans un ensemble d'objets des groupes homogènes (classes) et bien distincts les uns des autres.

- Classement (supervisée) : à partir d'exemples objets répartis en groupes, construire un algorithme, qui détermine le groupe adapté pour le nouvel objet.

- L'analyse de régression est principalement utilisée à deux fins conceptuellement distinctes. Premièrement, l'analyse de régression est largement utilisée pour la prédiction et la prévision, où son utilisation recoupe substantiellement le domaine de l'apprentissage automatique. Deuxièmement, dans certaines situations, l'analyse de régression peut être utilisée pour déduire des relations causales entre les variables indépendantes et dépendantes. Il est important de noter que les régressions en elles-mêmes ne révèlent que les relations entre une variable dépendante et un ensemble de variables indépendantes dans un ensemble de données fixe [11].

- Réduction de dimensions désigne ainsi toute méthode permettant de projeter des données issues d'un espace de grande dimension dans un espace de plus petite dimension. Cette opération est cruciale en apprentissage automatique pour lutter contre ce qu'on appelle le fléau des grandes dimensions (le fait que les grandes dimensions altèrent l'efficacité des méthodes).

- Deep vs Machine Learning : Le Machine learning et le Deep learning font partie de l'intelligence artificielle. Ces approches ont toutes deux pour résultat de donner aux ordinateurs la capacité de prendre des décisions intelligentes. Cependant, le Deep learning

est une sous-catégorie du Machine learning, car il s'appuie sur un apprentissage sans surveillance. [14]

4 Maths

- Fonction de coût mesure l'erreur entre le modèle et les valeurs de y du Dataset
- La descente de gradient est un algorithme d' optimisation itérative du premier ordre pour trouver un minimum local d'une fonction différentiable. L'idée est de faire des pas répétés dans la direction opposée du gradient (ou gradient approximatif) de la fonction au point courant, car c'est la direction de la descente la plus raide. [16]
- Équation normale : Dans un plan affine euclidien, l'équation d'une droite affine $ax + by + c = 0$ est dite normale si $a^2 + b^2 = 1$.

5 Bibliographie

[1] «**Anaconda (Python distribution)**», n.d., Wikipedia

[2] Y. Benzaki. «**Installer un environnement Python pour Machine Learning avec Anaconda**», 20 septembre 2017.
Lien : https ://mrmint.fr/installer-environnement-python-machine-learning-anaconda

[3] «**Jupyter**», n.d., Wikipedia

[4] «**Pandas**», n.d., Wikipedia

[5] Seaborn. «**An introduction to seaborn**»
Lien : https ://seaborn.pydata.org/introduction.html

[6] Riptutorial. «**Seaborn**»
Lien : https ://riptutorial.com/fr/python/example/8598/seaborn

[7] «**Matplotlib**», n.d., Wikipedia

[8] «**Plotly**», n.d., Wikipedia

[9] «**Scikit-learn**», n.d., Wikipedia

[10] «**Statsmodels**», n.d., Wikipedia

[11] «**Regression analysis**», n.d., Wikipedia

[12] X. Dupre. «**Régressions linéaires et autres variations**».
Lien : http ://www.xavierdupre.fr/app/mlstatpy/helpsphinx/c_ml/index_reglin.html

[13] Data Analytics Post. «**Réduction de dimensionnalité**».
Lien : https ://dataanalyticspost.com/Lexique/reduction-de-dimensionnalite/

[14] IONOS. «**Quelles sont les différences entre le Deep learning et le Machine learning ?**».
Lien : https ://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/deep-learning-vs-machine-learning/

[15] Lemagit. «**Machine Learning vs Deep Learning ? La même différence qu'entre un ULM et un Airbus A380**».
https ://www.lemagit.fr/conseil/Machine-Learning-vs-Deep-Learning-un-avion-a-helices-et-un-avion-a-reaction

[16] «**Gradient_descent**», n.d., Wikipedia