Ynov
2023/2024

Projet Maths-Datascience

Enseignant : Nabil El Malki

11 février 2024

Contexte général

Les algorithmes d'apprentissage automatique, sous-partie de l'intelligence artificielle, sont utilisés dans une grande variété d'applications, telles que la médecine, la chimie, la reconnaissance de la parole écrite et parlée, le filtrage du courrier électronique, la vision par ordinateur, où il est difficile ou irréalisable de développer des algorithmes conventionnels pour effectuer les tâches nécessaires. Les algorithmes d'apprentissage automatique ne reçoivent pas explicitement des instructions pour effectuer ces tâches mais ils apprennent automatiquement ces tâches à partir des données. Cette capacité d'apprentissage automatique a été possible par l'introduction dans ces algorithmes des mathématiques tels que l'algèbre linéaire et les statistiques.

L'objectif des datascientists est de concevoir des modèles d'apprentissage automatique et d'en optimiser la performance et l'efficacité. Parmi les défis majeurs auxquels ils sont confrontés figure le prétraitement des données, un processus qui influence directement la qualité du modèle élaboré. La complexité de ce défi réside dans l'absence d'une approche universelle pour le prétraitement des données, étant donné qu'il n'existe pas une seule méthode correcte. Les techniques employées dépendent étroitement du problème à résoudre et du type de données traitées.

Dans le contexte du monde réel les données présentent des valeurs manquantes. Ces dernières peuvent résulter de divers facteurs, notamment des erreurs humaines lors du traitement des données, des erreurs de mesure dues au dysfonctionnement des machines, des questionnaires non complétés entièrement, et la fusion de données sans relation apparente. L'exploitation de ces données peu fiables peut avoir des répercussions sur les résultats, conduisant fréquemment à des conclusions erronées. Le prétraitement des données se révèle ainsi essentiel pour obtenir des résultats fiables.

L'une des méthodes simples de traitement des valeurs manquantes dans un jeu de données structuré au format tabulaire consiste à éliminer les colonnes et/ou les lignes qui contiennent des valeurs manquantes. Cette approche peut être appropriée lorsque le nombre de valeurs manquantes est négligeable, et la suppression de ces données n'introduit pas de biais significatif. Une autre stratégie, appelée imputation, implique le remplacement des valeurs manquantes par d'autres valeurs pouvant être prédites ou estimées.

Parmi les approches d'imputation, on trouve les suivantes :

— Remplacement des valeurs manquantes par une valeur fixe, la moyenne, la valeur maximale ou la valeur minimale.
— Utilisation de techniques d'interpolation.
— Recours à des modèles statistiques ou d'apprentissage automatique pour remplacer les valeurs manquantes.

Pour plus d'informations sur le sujet, consultez les travaux de [1, 4, 3].

Contexte particulier

En tant qu'ingénieur data au sein du département de recherche et développement d'une entreprise de services informatiques, vous êtes fraîchement affecté à un nouveau pôle axé sur les données et l'intelligence artificielle. L'objectif est de développer des modèles d'apprentissage automatique

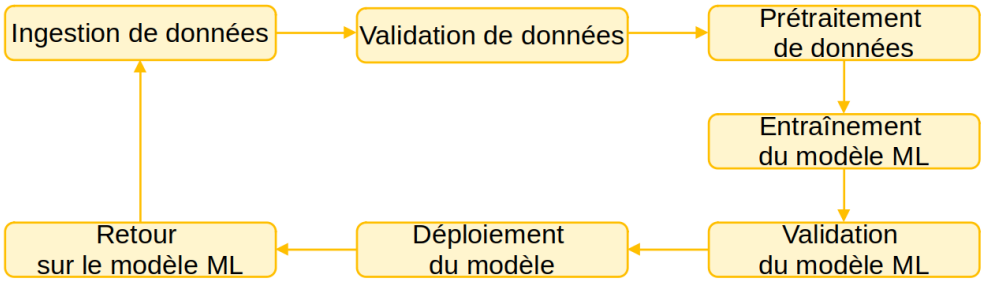

FIGURE 1 – Pipeline ML

pour diverses applications afin d'améliorer les chances de remporter des projets clients. Cependant, les résultats des modèles jusqu'à présent ont été mitigés.

Suite à des réunions entre les ingénieurs data et les ingénieurs en intelligence artificielle, des soupçons ont été émis concernant l'approche de traitement des valeurs nulles actuellement adoptée, qui consiste à les supprimer. En tant que personne ayant des compétences approfondies en mathématiques, il vous a été confié la responsabilité de mener une étude comparative, à la fois théorique et expérimentale, afin d'évaluer différentes stratégies de prétraitement de valeurs nulles. L'objectif est de fournir des éclaircissements à l'équipe data sur les choix les plus adaptés en matière de prétraitement en fonction de chaque contexte spécifique.

Tâches

Dans le cadre de ce projet, il est demandé d'étudier les différents types de valeurs manquantes ainsi que le fonctionnement de différentes solutions* [1] traitant les valeurs manquantes en utilisant la littérature scientifique (voir bibliographe ci-dessous ; d'autres références sont possibles) et en menant des expérimentations répondant à un certain de nombre de questions ci-dessous. L'attention sera portée sur les solutions suivantes (voir références de la bibliographie) :

- suppression
- imputation
  - remplacement par une valeur fixe (moyenne, valeur maximale, valeur minimale, constate arbitraire)
  - interpolation cubique ou linéaire ;
  - k-plus proches voisins ;
  - algorithme d'espérance-maximisation.

Au moins deux solutions d'imputation sont à analyser en plus de celle de la suppression. Vous avez le choix de choisir d'autres solutions d'imputation à la place de celles proposées ci-dessus. . Les caractéristiques à examiner incluent :

- **Hypothèses :**
  - **Tolérance aux variables corrélées :** Est-ce que la présence de variables corrélées, notamment entre la variable avec des données manquantes à estimer et une autre variable

1. Parfois le terme stratégie est employé.

(complète ou également avec des données manquantes à estimer ou pas), affecte l'efficacité de la stratégie de gestion de valeurs manquantes ? [2]

- **Distribution des variables :** Les variables doivent-elles suivre toutes une distribution probabiliste spécifique ? Si un ensemble de données comporte exclusivement des variables suivant une distribution gaussienne (ou toute autre distribution spécifique), cela impacte-t-il négativement l'efficacité de la méthode ? La question se pose également lorsque les variables suivent différentes distributions.

— **Scalabilité :**
  — La stratégie est-elle conçue pour le traitement efficace de volumes de données importants.
  — Dans le cas contraire, proposer les possibilités d'adaptation de la stratégie aux données massives (via des procédés de parallélisation ou autres technologies) . expérimentations ne sont pas nécessaires. * [3]

— **Sensibilité aux valeurs extrêmes ou aberrantes :** Évaluer l'effet de variations dans le nombre de valeurs aberrantes sur l'efficacité de la stratégie. A titre d'exemple, les valeurs aberrantes peuvent correspondre, pour la la colonne poids (pour humains), à des valeurs supérieures à 200 kg.

— **Sensibilité au taux de valeurs manquantes :** Examiner comment différents taux de valeurs manquantes affectent l'efficacité de la méthode.

— **Sensibilité à la grande dimensionnalité :** La méthode est-elle appropriée pour un petit nombre de variables ou peut-elle gérer efficacement un grand nombre de variables ? L'efficacité en fonction du nombre de colonnes est à évaluer.

Ensuite une étude comparative doit être menée en vue de mettre en évidence les disparités entre les différentes stratégies, les avantages et inconvénients de chacune, tout en fournissant des recommandations sur leur utilisation adaptée en fonction du contexte (caractéristiques des données, par exemple distribution probabilité, taux de valeurs manquantes...).

Pour chacune des solutions choisies, vous devez préciser le type de valeurs manquantes traité. Le jeu de données sujet aux expérimentations doit contenir ce type de valeurs manquantes. Consultez les travaux des auteurs Jager et al. [2] pour plus d'informations sur la génération de jeux de données.

Les données exploitées durant les expéimentations peuvent être soit réelles telles que celles fournies par scikit-learn [4] ou openml [5] soit simulées à l'aide de générateurs tels que ceux fournis par scikit-learn [6] ou numpy [7]. Si les jeux de données ne présentent pas naturellement de valeurs manquantes, il est demandé de les simuler de manière artificielle.

Il existe deux approches principales pour évaluer une méthode de traitement des valeurs manquantes. Premièrement, dans le contexte d'un projet d'apprentissage automatique (ML), on peut sélectionner une tâche spécifique de ML (telle que la régression ou la classification) et mesurer l'efficacité de la méthode de traitement des valeurs manquantes en fonction de la performance sur cette tâche ML. Dans cette approche, on part du principe que le traitement des valeurs manquantes est effectué dans le but exclusif de faciliter la réalisation de la tâche ML choisie. Alternativement, on peut évaluer la méthode de gestion des valeurs manquantes de manière indépendante de toute tâche ML spécifique, en utilisant des mesures telles que l'Erreur Quadratique Moyenne (RMSE) ou l'Erreur Absolue Moyenne (MAE) pour juger de son efficacité.

2. Pour aborder cette question, il pourrait être utile d'utiliser un ensemble de données comprenant au moins deux variables fortement corrélées. Cette analyse n'est à mener que pour les stratégies traitant plusieurs variables (Pour information les variables sont aussi appelées colonnes, caractéristiques ou dimensions).
3. La deuxième question sur la scalabilité est optionnelle.
4. https ://scikit-learn.org/stable/datasets/toy$_{d}ataset.html$
5. https ://www.openml.org/search ?type=datasort=runsstatus=active
6. https ://scikit-learn.org/stable/datasets/sample$_{g}enerators.html$
7. https ://numpy.org/doc/stable/reference/random/

Rendu du rapport

Chaque groupe doit fournir un rapport consignant les études demandées ainsi que les résultats expérimentaux.

Date de rendu :

Une première version du rapport est à rendre au plus tard le 23 février 2024 23h59. La dernière version est à rendre au plus tard le 16 Mars 2024 à 23h59.

Bibliographie

[1] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big Data*, 8(1) :1–37, 2021.

[2] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. *Frontiers in big Data*, 4 :693674, 2021.

[3] Ardalan Mirzaei, Stephen R. Carter, Asad E. Patanwala, and Carl R. Schneider. Missing data in surveys : Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 18(2) :2308–2316, 2022.

[4] Therese D Pigott. A review of methods for missing data. *Educational research and evaluation*, 7(4) :353–383, 2001.