EXAMEN FINAL: MATHEMATIQUES DU BIG DATA

BENSALEM Akram

1er choix de travaux: La Régression linéaire

Lien vers le github: https://github.com/Nina809/PSBX/blob/main/Regression.Rmd Par ZOUMANIGUI Nina

La régression linéaire est un concept très répandu et utilisé dans le cadre d'études statistiques et économétriques. Elle d'apréhender l'impact d'une variable explicative sur une variable endogène dans le cadre d'une régression linéaire simple et l'impact de plusieurs variables explicatives sur une variable endogène dans le cadre d'une régression linéaire multiple. Cela se modélise par une droite traversant un nuage de points, un modèle linéaire parfait serait une droite traversant l'intégralité des points. La régression linéaire renvoie à des notions de dispersion comme la variance et c'est graphiquement montré par la distance séparant les points et la droite.

C'est aux personnes elles-même de définir quelle(s) est/sont la ou les variables explicatives et quelle est la variable endogène. On peut voir dans ce travail l'import d'une dataframe afin d'illustrer sur R la régression linéaire. Dans le code on voit la construction d'un graphe (plot) avec lm qui est une fonction dont les acronymes signifient linear model et des poids ont été posés dans cette régression.

Je trouve ce travail très pertinent pour plusieurs raisons. Il est bien structuré car on a bien une introduction, un corps de texte, du code et une conclusion. Le choix du sujet est très intéressant car la connaissance des régressions linéaires est quasiment un requis pour des personnes travaillant avec de la data ou des statistiques. La lecture est simple et facile à comprendre même si on débute sur le sujet. L'utilisation d'une dataframe pour illustrer les propos est très pratique et rend le contenu plus visuel. Les cas d'application de ce document peuvent être variés.

2nd choix de travaux: ML for financial products

Lien vers le github: https://github.com/soukainaElGhaldy/PSB-X/blob/main/Mathematics/maths.pdf Par EL GHALDY Soukaina

Ce travail mené renvoie à plusieurs notions comme la finance, les mathématiques, de l'algorithmie, du machine learning mais également de réseaux neuronaux. Le marché financier repose sur des échanges (ventes et achats) d'actions à plusieurs échelles mais également sur des variables exogènes. Le but des acteurs dans le milieu de la finance est de réaliser des profits via l'achat/revente d'actions, pour prévoir les prix des actifs financiers les acteurs prennent en considération plusieurs facteurs qu'ils traduisent mathématiquement afin de prévoir au mieux les cours des actifs. Ils ont même créé des algorithmes de machine learning afin de spéculer beaucoup plus précisément.

Certaines techniques comme l'utilisation d'un taux d'accroissement permet d'apréhender le prix d'une obligation (actif financier qui est en réalité un actif couvrant un crédit) sur plusieurs années. Mathématiquement lorsque n, le nombre d'années, croît plus l'obligation prend de la valeur du fait d'une sommme plus importante des intérêts. Lorsque n est très grand, la valeur de l'obligation croît de façon exponentielle car en effet le développement limité de l'exponentiel est

$$\sum_{1}^{n} \frac{x^n}{n}$$

L'algorithme "Experts network" va déterminer les meilleures stratégies de création d'un portefeuille d'actions en prenant en considération mathématiquement l'hétérogénéité des acteurs mais également leur comportement que l'algorithme va classer grâce à du mapping. Finalement on a un schéma simplifié des réseaux neuronaux sur le marché financiers qui ressemble à un DAG.

Le plan du travail est bien structuré avec des explications plutôt simplifiées pour le lecteur. Le travail est très pertinent car il présente un côté fonctionnel et donc un exemple direct d'application de mathématiques et machine learning dans le monde du travail. Les formules mathématiques sont très explicites car on sait à quoi correspond chaque variable. Finalement on a quelques notions de réseaux neuronaux et commment cela peut représenter l'imbrication entre plusieurs acteurs.

3ème choix de travaux: Tree-based pipeline optimization

Lien vers le github: https://github.com/OlfaLmt/PSBX/blob/main/Thèses/Automating%20biomedical%20data%20science%20through%20tree-based%20pipeline%20optimization.pdf Par LAMTI Olfa

Ce travail traite de l'optimisation automatisée de pipelines à partir d'un arbre de décision. Il faut savoir que les pipelines sont des séries d'action réalisée afin de récupérer de la data dite brute, traiter ces données brutes afin de les exploiter par la suite. En l'occurence dans ce travail on nous présente le fonctionnement d'un algorithme nommé TPOT qui permet de réaliser des pipelines avec diverses configurations possibles, chacune répondant à une problématique spécifique.

Il faut savoir que pour que cet algorithme soit efficient il faut qu'une personne réalise une partie de traitement sur les données brutes d'où l'importance d'une bonne gouvernance de la data. La chaîne de transition de la data est schématiquement représenté par un graphe orienté acyclique. Les tâches réalisées par le TPOT sont complexes, le TPOT fait des choix notamment grâce à des "Genetic algorithms" qui sont des algorithmes évolutifs, c'est-à-dire qu'ils sont dotés d'un ensemble de fonctions et le choix de la combinaison des fonctions est fait par l'algorithme selon le contexte. De plus, l'algorithme prend en considération des choix passés afin d'élaborer le modèle, c'est un principe de machine learning intégré.

Le travail est bien structuré avec un schéma très limpide qui nous permet de visualiser au mieux le fonctionnement. Nous n'avons pas de formulation mathématique explicite mais nous avons tout de même des notions d'algorithmie et de machine learning. Le choix du travail est très pertinent car répond à des enjeux non seulement sur la gouvernance de la donnée mais également de l'élaboration de stratégies dans un éco-système devenu complexe avec la digitalisation, le TPOT répond à un manque de main d'oeuvre qualifié dans ce domaine.

4ème choix de travaux: Naïve Bayes Classifier

Lien vers le github: https://github.com/Jordyhsn/PSB_Hounsinou/blob/main/Naïve-Bayes.pdf Par HOUNSINOU Jordy

Ce travail traite de probabilités bayesiennes et l'élaboration d'un algorithme nommé Naive Bayes qui se base justement sur des probabilités bayesiennes. La formule de Bayes est très célébre et utilisée dans plusieurs domaines et notamment en médecine afin de connaître la probabilité d'un bon diagnostic ou la fiabilité de certains outils. L'algorithme reprend exactement le principe de la loi de Bayes et l'applique à plusieurs variables explicatives potentielles.

Ici, comme très souvent, l'exemple d'application est appliqué à la médecine et la santé. En premier lieu, on a un test de probabilité permettant de savoir la fiabilité du pcr dans le diagnostic sur le coronavirus. La formule de Bayes prend en compte des évémenements distincts, l'un des événements est le complémentaire de l'autre. L'algorithme effectue le même calcul mais pose au préalable une sorte de tableau croisé avec plusieurs variables explicatives et plusieurs diagnostics possibles. Ces variables explicatives sont indépendantes entre elles. Grâce à ce tableau croisé, les résultats de probabilité sont quasi immédiats.

Ce travail est très bien struturé avec une intro, un corps de texte, des exemples concrets et une conclusion. De plus les avantages et limites sont mis en lumière et argumentés. Le choix du sujet est très pertinent étant donné la crise sanitaire actuel, il est bon de savoir comment la fiabilité des tests sont calculés. La formule de

Bayes est très explicite et ses composantes bien décrites. Le focntionnement de l'algorithme est très limpide également.

5ème choix de travaux: Cryptographie et théories des nombres

Lien vers le github: https://github.com/WilliamRbc/PSBX/blob/main/CRYPTO/Cryptographie.pdf Par ARSIC Marko et ROBACHE William

Ce travail traite de la cryptographie et de ses principes. La cryptographie assure une certaine confidentialité de données stockées sur le web et des communications par réseau. On peut voir que le principe de cryptographie existe depuis très longtemps et que des nouvelles techniques de cryptographie ont vu le jour, s'inspirant souvent des moyens ancestraux. De plus on a une formalisation mathématique du chiffrement qui est une forme de cryptographie. Chaque forme de cryptographie a ses propriétés et un fonctionnement propre.

L'un des premiers ancêtres de la cryptographie est le codage de César, la méthode était simple, chaque lettre de l'alphabet correspond à un nombre et cela dans l'ordre. Il attribuait également les symboles à des nombres. Cela a été repris par Vigénère qui a complexifié en y ajoutant une clé dont la taille correspond au décalage entre les lettres. Dans ce travail on nous explique également le fonctionnement des cryptages à clé symétrique et ses limites. En effet une unique clé sert pour le déchiffrement et le chiffrement, autrement dit deux individus doivent posséder tous deux la même clé pour pouvoir communiquer entre eux contrairement au cryptage à sens asymétrique où la clé sert uniquement au déchiffrement. Les contraintes semblent moindres pour la clé asymétrique. La clé asymétrique a été formalisée mathématiquement par la fonction à sens unique avec comme illustration une bijection difficile à exprimer. C'est-à-dire qu'il sera soit difficile de trouver l'antécédent soit l'image d'une information. On a également l'explication du principe de chiffrement, si une certaine fonction correspond à une information cryptée, le seul moyen de connaître l'information cryptée est de passer par une seconde fonction qui est justement optimisée pour obtenir l'image de celle-ci.

Je trouve ce travail bien structuré avec beaucoup d'exemples concrets. On a quelques principes mathématiques mis en avant. Ce sujet est pertinent et d'actualité car avec les nouvelles technologies on s'inquiète toujours de la confidentialité de nos données et la cybersécurité en général pour les entreprises ayant des données sensibles. Les domaines d'application peuvent être multiples comme nous le rappele le petit bond dans le passé au début.

Auto-évalutation de mon travail en maths: Systèmes de conduite autonomes

Lien vers le github:https://github.com/AkramBensalemPSB/PSB/blob/main/Systemes%20de%20conduite%20autonomes%20maths.pdf Par BENSALEM Akram

Ici le sujet traite du fonctionnement de voitures autonomes et les algorithmes sous-jacents. On a donc des notions de topologie, de géométrie dans l'espace, d'intégration et d'algorithmies. On a plusieurs illustrations qui appuient les formules mathématiques et les variables sont explicites. Cependant je trouve que j'aurais pu développer davantage les formules mathématiques quitte à avoir certaines lignes de développement algébrique. J'ai aurais pu également introduire des notions pour le calcul d'intégrales comme le théorème de Fubini, cependant cela était volontaire car je me voulais concis dans l'explication mathématique et mettre l'accent sur le côté fonctionnel des systèmes autonomes. Dans l'ensemble le travail est bien structuré et le choix du sujet est pertinent car de plus en plus de voitures sont dotés de ces systèmes et qu'il serait bon pour la culture d'en connaître les rudiments.