Projets ADD

Polytech Lyon, MAM 3A

5 mai 2025

1 Objectif

Ce projet vise à vous permettre de mobiliser l'ensemble des compétences nécessaires à la réalisation d'une analyse de données. Vous serez amenés à travailler sur toutes les étapes du processus, de l'acquisition des données à leur prédiction, en passant par le nettoyage et la préparation, si nécessaire.

Ce projet pourra également servir de vitrine lors de votre recherche de stage en 4A ou 5A, en mettant en valeur votre capacité à mener un projet d'analyse de données de manière autonome et professionnelle.

2 Consignes

Le projet peut être réalisé individuellement ou en groupe de deux personnes. Chaque sujet est divisé en trois grandes parties : Prétraitement des données (TP 1,2 et 3), Exploration des données (TP 3 et 4) et Machine Learning (TP 4 et 5). Pour chaque sujet, vous trouverez une brève description du contexte ainsi que quelques propositions pour la partie 3 (Machine Learning).

Vous devez fournir (i) le code source du projet, ainsi que (ii) un rapport **sourcé** résumant votre travail sur ce projet. Votre rapport devra également détailler les améliorations possibles de votre approche. Veuillez noter que le code source et le rapport sont tous deux obligatoires. Il est possible de rendre les deux en même temps en utilisant un notebook.

Vous avez également la possibilité de proposer votre propre sujet en m'envoyant par mail `corentin.constanza@creatis.insa-lyon.fr` : le lien du dataset, les membres de votre groupe, ainsi que, si possible, l'objectif de votre analyse (i.e. la partie 3).

Sujet 1 : Analyse multidimensionnelle et classification des morceaux Spotify

Dans un contexte où la musique numérique occupe une place centrale dans notre quotidien, les plateformes de streaming telles que Spotify génèrent une quantité massive de données sur les morceaux, les artistes et les comportements d'écoute. L'analyse multidimensionnelle permet d'explorer ces données sous différents angles afin de mieux comprendre les caractéristiques musicales et les tendances d'écoute. Ce projet s'intéresse à l'application de techniques de classification sur les morceaux Spotify [1] à partir de leurs attributs audio (tels que la danseabilité, l'énergie, ou encore le tempo), dans le but d'identifier des regroupements cohérents et potentiellement révélateurs de genres ou de styles musicaux.

Idées pour la partie Machine Learning :

1. *Classification des morceaux par genre*
2. *Système de recommandation*
3. *Comparaison entre la classification et le clustering*

Tache Bonus : Lyrics Embedding

1. *Récupérez les paroles de chaque morceaux (API Spotify ou autres ressources)*
2. *Utilisez un LLM type GPT-3[2], BERT[3] ou autres pour convertir les paroles en une représentation vectorielle latente.*
3. *Utilisez le vecteur latent comme variable supplémentaire afin d'améliorer la performance de la classification.*

Références

[1] *spotify-tracks-dataset · Datasets at Hugging Face.*
URL : https://huggingface.co/datasets/maharshipandya/spotify-tracks-dataset.

[2] Tom B. Brown et al. « Language Models are Few-Shot Learners ».
In : *arXiv preprint arXiv :2005.14165* (2020). URL : https://arxiv.org/abs/2005.14165.

[3] Jacob Devlin et al.
« BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ».
In : *arXiv preprint arXiv :1810.04805* (2019). URL : https://arxiv.org/abs/1810.04805.

Sujet 2 : Étude du protéome dans le cancer du sein

Le cancer du sein constitue l'une des principales causes de mortalité chez les femmes à travers le monde. C'est une maladie hétérogène aux multiples sous-types biologiques, dont la classification précise est essentielle pour adapter les traitements et améliorer les pronostics. L'analyse du protéome, c'est-à-dire l'ensemble des protéines exprimées dans les cellules tumorales, offre une perspective plus directe sur les mécanismes fonctionnels de la maladie par rapport aux seules données génétiques. Ce projet s'appuie sur le jeu de données *iTRAQ proteome profiling* [1] portant sur 77 échantillons de cancers du sein, tel que présenté dans l'étude publiée dans Nature (2016) [2], afin d'explorer les profils d'expression protéique. L'objectif est de mieux comprendre le paysage protéomique de ces tumeurs et d'identifier, à l'aide de méthodes d'analyse de données et de classification, d'éventuels sous-types tumoraux fondés sur les signatures protéiques.

Outils : BioPython [3] pourrait vous être utiles pour certaines analyses liées aux protéines

Idées pour la partie Machine Learning :

1. *Clustering sur les données protéiques*
2. *Comparer différentes méthodes de clustering*
3. *Entraîner différents modèles d'apprentissage automatique (arbres de décision, SVM, réseaux de neurones) pour prédire les sous-types de cancer ou d'autres issues cliniques à partir des profils d'expression protéique.*

Références

[1] *Breast cancer proteomes*. Nov. 2019. URL : https://www.kaggle.com/datasets/piotrgrabo/breastcancerproteomes?resource=download.

[2] Philipp Mertins et al.
« Proteogenomics connects somatic mutations to signalling in breast cancer ».
In : *Nature* 534.7605 (2016), p. 55-62.

[3] *BioPython*. URL : https://biopython.org/.

Sujet 3 : Étude spectrale à partir du relevé SDSS

L'analyse spectroscopique est l'un des outils les plus puissants en astronomie moderne, permettant de déduire des propriétés physiques fondamentales des objets célestes, tels que les galaxies, à partir de la lumière qu'ils émettent. Les spectres galactiques contiennent des informations précieuses sur la composition chimique, la vitesse, la distance (via le redshift), la luminosité, ou encore l'activité stellaire d'une galaxie. Dans ce contexte, la base de données issue du Sloan Digital Sky Survey (SDSS) [1] constitue une ressource exceptionnelle : elle regroupe des millions de spectres optiques de galaxies et d'autres objets célestes, collectés de manière systématique depuis plus de deux décennies. Ce projet vise à exploiter les spectres galactiques extraits du SDSS pour prédire certaines quantités physiques clés — notamment le redshift et la magnitude apparente — à l'aide de techniques d'apprentissage automatique (random forest, SVM, etc...).

Astuces :

Traiter les spectres comme une ligne dans vos dataframe (i.e. chaque longueur d'onde est une variable).

Les données du SDSS sont publiques et accessibles via le `SkyServer`, où elles sont stockées dans une base de données relationnelle interrogeable avec le langage SQL. L'interface `CasJobs` [2] permet d'exécuter directement des requêtes SQL sur les serveurs de données et de télécharger les résultats dans plusieurs formats. Une alternative plus récente, `Betelgeuse`, est disponible sur `SciServer`, une plateforme généraliste de data science qui prolonge le projet SkyServer.

Attention : le volume total des données SDSS atteint **652 To** ! Il est donc recommandé de ne télécharger qu'un sous-ensemble gérable.

Les données photométriques se trouvent dans la table `PhotoObjAll`, et les données spectroscopiques dans `SpecObjAll`. Ces deux tables peuvent être jointes via la colonne `specObjID`, identifiant unique des objets ayant des données spectroscopiques.

Exemple de requête SQL pour joindre les deux tables :

```
SELECT <colonnes souhaitées>
FROM PhotoObjAll AS p
JOIN SpecObjAll AS s
ON p.specObjID = s.specObjID
```

Références

[1] *SkyServer SDSS*. URL : https://skyserver.sdss.org/.

[2] *CasJobs*. URL : https://skyserver.sdss.org/CasJobs/.

Sujet 4 : Analyse de la pollution de l'air en milieu urbain

La pollution de l'air représente une menace sérieuse pour les conditions environnementales durables au XXIe siècle, affectant des millions de personnes à travers le monde. Son importance dans la détermination de la santé et des conditions de vie en milieu urbain devrait seulement croître avec le temps. Divers facteurs allant des émissions artificielles aux phénomènes naturels sont connus pour être des agents causaux principaux ou des influenceurs derrière l'augmentation des niveaux de pollution de l'air. Les principaux polluants tels que les particules fines et le dioxyde d'azote proviennent des transports, de l'industrie et du chauffage.

Vous travaillerez sur le jeu de données DEAP [1]. Votre objectif sera à partir des données météorologiques, de trafics routiers et émission énergétique d'arriver à prédire le niveau de différents polluants.

Références

[1] Mayukh Bhattacharyya, Sayan Nag et Udita Ghosh.
« Deciphering Environmental Air Pollution with Large Scale City Data ». In : *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*.
International Joint Conferences on Artificial Intelligence Organization, 2022.
URL : https://github.com/mayukh18/DEAP.

Sujet 5 : Hiérarchisation des publications sur les réseaux sociaux

Dans le contexte actuel des réseaux sociaux, les utilisateurs sont souvent submergés par une multitude de publications dans leurs fils d'actualités, rendant difficile la découverte de contenu réellement pertinent. En utilisant des modèles d'apprentissage sur les interactions passées des utilisateurs avec des tweets pour prédire la pertinence des publications futures, dans le but d'améliorer l'expérience utilisateur et de proposer du contenu personnalisé.

Il est important de noter que de tels algorithmes soulèvent des enjeux sociétaux majeurs, notamment en ce qui concerne la privatisation des réseaux sociaux, la manipulation de l'information, la propagation de contenus biaisés ou la création de chambres d'écho.

Pour ce projet, vous travaillerez sur le Twitter Vaccination Dataset [1], un scrapping de tout les tweets portant sur la vaccination sur la période 2006 à novembre 2019. Il inclut le texte des tweets, leur date et heure de publication, ainsi que la localisation des utilisateurs (lorsqu'elle est fournie). Des informations sur les identifiants des utilisateurs, leurs abonnés et leurs amis ont également été collectées.

Références

[1] *Twitter vaccination dataset*. Avr. 2020. URL : https://www.kaggle.com/datasets/keplaxo/twitter-vaccination-dataset/data.

Sujet 6 : Étude des ilots de chaleur urbain dans la métropole de Toulouse

Les îlots de chaleur urbains (ICU) sont des zones dans les villes où la température est significativement plus élevée que dans les zones rurales environnantes, principalement en raison de l'urbanisation, de la densité de la population et des activités humaines. Dans la métropole de Toulouse, ces phénomènes sont de plus en plus marqués, particulièrement pendant les périodes estivales.

Pour étudier le phénomène la métropole à installer plus de 60 stations météo. Les données sont en libre accès sur le site de la métropoles [1].

Idée pour la partie 3 : Peut-on se passer de certaines stations en utilisant les mesures des stations avoisinantes pour prédire la mesure attendue ? Si oui, de combien de stations a-t-on besoin pour couvrir l'ensemble de la métropole ?

Tache Bonus : Les données sont mises à jour toutes les 15 minutes, faites un affichage en temps réel sur une carte.

Références

[1] *Explore — Open Data Toulouse Metropole - ICU.* URL : https://data.toulouse-metropole.fr/explore/?sort=modified&refine.keyword=ilot+de+chaleur.

Sujet 7 : Étude des caractéristiques du noyau cellulaire pour le diagnostic du cancer du sein

Le cancer du sein constitue l'une des principales causes de mortalité chez les femmes à travers le monde. Son diagnostic précoce joue un rôle crucial dans l'amélioration du pronostic et de la survie des patientes. Dans cette optique, l'analyse des caractéristiques morphologiques des noyaux cellulaires, obtenues à partir de biopsie à l'aiguille fine (BAF) de masses mammaires, représente une méthode prometteuse pour affiner les techniques de diagnostic assisté par ordinateur.

Ce projet s'inscrit dans cette démarche. Vous étudierez le *Breast Cancer Wisconsin Dataset* [1] comportant des mesures précises des noyaux cellulaires extraites d'images BAF. Ces caractéristiques comprennent notamment des descripteurs tels que la texture, la concavité, la symétrie ou encore la compacité des noyaux. Ces variables ont été sélectionnées via une recherche exhaustive parmi des combinaisons de caractéristiques pertinentes. [2]

L'objectif de cette étude est de mieux comprendre les attributs morphologiques associés aux tumeurs bénignes et malignes, et d'évaluer leur pertinence pour le diagnostic automatisé du cancer du sein.

Références

[1] *UCI Machine Learning Repository - Breast Cancer Wisconsin (Diagnostic)*. URL : https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

[2] Kristin P Bennett et Olvi L Mangasarian.
« Robust linear programming discrimination of two linearly inseparable sets ».
In : *Optimization methods and software* 1.1 (1992), p. 23-34.

Prédiction du décrochage et de la réussite académique des étudiants

Ce projet s'appuie sur un jeu de données [1] élaboré à partir de plusieurs bases de données indépendantes [2]. Chaque observation correspond à un étudiant inscrit dans un cursus de licence (agronomie, design, éducation, journalisme, technologies, etc.) et contient des données connues dès l'inscription, telles que le parcours académique antérieur, les caractéristiques démographiques et socio-économiques, ainsi que les performances académiques durant les deux premiers semestres.

L'objectif est de développer des modèles de classification capables de prédire l'issue du parcours académique selon trois catégories : abandon, poursuite ou obtention du diplôme. Ce problème est particulièrement complexe en raison du déséquilibre des classes, une majorité d'étudiants étant encore inscrits à la fin de la durée normale du cursus. Une attention particulière est donc portée à la sélection des variables pertinentes et à l'évaluation des performances des modèles.

Au cours de ce projet, vous prendrez soin d'analyser tous les biais potentiels du jeu de donnée et de vos algorithmes. Ces biais peuvent reproduire ou amplifier des inégalités sociales ou académiques présentes dans les données, en désavantageant inconsciemment certains groupes d'étudiants. Il est donc important de les renseigner afin de connaître les limites de vos analyses.

Références

[1] *UCI Machine Learning Repository - Predict Students' Dropout and Academic Success.* URL : https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success.

[2] Mónica V Martins et al.
« Early prediction of student's performance in higher education : A case study ».
In : *Trends and Applications in Information Systems and Technologies : Volume 1 9.* Springer. 2021, p. 166-175.

Sujet 9 : Analyse des préférences œnologiques à partir de données physico-chimiques

L'analyse des préférences œnologiques est un enjeu central pour les producteurs de vin, désireux d'adapter leurs produits aux attentes des consommateurs et d'améliorer la qualité perçue de leurs cuvées.

Ce projet s'appuie sur deux ensembles de données distincts [1], relatifs à des vins portugais rouges et blancs. Ces jeux de données, bien que dépourvus d'informations commerciales ou variétales pour des raisons de confidentialité, contiennent des mesures précises de propriétés physico-chimiques (comme l'acidité, la teneur en alcool, le pH, etc.) ainsi qu'une note qualitative globale attribuée à chaque vin. L'objectif est d'identifier les relations entre les caractéristiques mesurées en laboratoire et la qualité perçue du vin.

L'analyse peut être abordée sous l'angle de la régression ou de la classification bien que la distribution déséquilibrée des classes (la majorité des vins étant de qualité moyenne) constitue un défi supplémentaire. La détection d'anomalie peut aussi être étudiée.

Références

[1] *UCI Machine Learning Repository - Wine Quality.*
URL : https://archive.ics.uci.edu/dataset/186/wine+quality.

Sujet 10 : Reconnaissance d'activités humaines à l'aide de smartphones

La reconnaissance automatique d'activités humaines est devenue un champ de recherche prometteur notamment pour les domaines de la santé, du sport ou de l'assistance à la personne. En exploitant les signaux issus de l'accéléromètre et du gyroscope intégrés dans un téléphone mobile, il est possible d'identifier, en temps réel, les mouvements ou postures de l'utilisateur.

Ce projet s'appuie sur un ensemble de données collectées [1] auprès de 30 volontaires âgés de 19 à 48 ans, réalisant six activités quotidiennes (marche, montée et descente d'escaliers, position assise, debout et allongée), tout en portant un smartphone fixé à la taille. Les signaux inertiels ont été prétraités pour extraire des caractéristiques temporelles et fréquentielles sur des fenêtres temporelles glissantes, permettant de construire des modèles de classification fiables.

L'objectif de cette étude est de développer un système capable de reconnaître automatiquement l'activité en cours à partir des données des capteurs.

Références

[1] *UCI Machine Learning Repository - Human Activity Recognition Using Smartphones.* URL : https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones.