Travail Pratique 2

8PRO408 - Outils de programmation pour la science des données

Professeur Julien Maitre, Ph.D.
Automne 2022

1. Description Générale

L'objectif général de ce travail pratique 2 (TP2) est de vous familiariser avec le traitement des données, plus spécifiquement des signaux (***time series***), qui peut avoir lieu avant l'étape d'exploitation des algorithmes d'apprentissage machine pour l'extraction des connaissances ou la classification par exemple. Ainsi, ce TP2 vous permettra de manipuler les données temporelles pour extraire des caractéristiques et ainsi créer des ensembles de données. Pour cela, vous utiliserez le langage de programmation Python et de ses bibliothèques pour la science des données (ex. : NumPy, Pandas, SciPy, …).

De plus, dans ce deuxième travail, vous aurez à produire un rapport scientifique qui décrit les données exploitées et les étapes d'extraction des caractéristiques que vous aurez appliqué pour la création d'un ensemble de données. Le nombre de pages pour le rapport est limité à 10.

Vous aurez besoin des cours 2 à 5 et le cours 8 du cours 8PRO408 afin de réaliser ce TP2. Aussi, vous devrez chercher les informations nécessaires par vous-même (sur internet) pour coder votre travail.

2. Formalités

Le travail est à réaliser en groupe de **six (il peut y avoir des exceptions)** et se composera d'un rapport scientifique et des scripts Python que vous aurez réalisé.

La date limite pour le rendu de votre travail est le **29 novembre 2022 à 8h00**. Après cette date, il y aura une **pénalité de 10% par jour de retard**.

Vous déposerez votre travail sous le format d'un dossier `.zip` sur Moodle (du cours 8PRO408) dans la sous-section « **TP2** » de la section « **Remise des Travaux Pratiques** ». Le nom du fichier prendra la forme de **TP2.zip**. **Une seule personne par groupe déposera les travaux**. N'oubliez pas d'inclure votre code permanent sur la page de garde du rapport scientifique. **Le nom respect des règles citées ci-dessus entraînera une pénalité de 5%**.

3. Ce qui est attendu

Le rapport du TP2 devra comprendre :

- la description de votre ensemble de données
    - Par exemple :
        - dans quel(s) objectif(s) les données ont été collectées (s'il y a lieu);
        - comment ont été collectées les données ? (s'il y a lieu);
        - quelles sont les variables?;
        - le nombre d'instances ou de fichiers;
        - le nombre de classes;
- la description des étapes pour la création de l'ensemble de données
    - Vous devrez :
        - définir une fenêtre temporelle glissante et un ratio de chevauchement;
        - extraire des caractéristiques (ceux citées dans le cours du Chapitre 6);
        - stocker les caractéristiques dans un ***DataFrames*** de Pandas;
        - donner un nom à chaque colonne des caractéristiques extraites;
        - enregistrer le ***DataFrames*** dans le format de fichier `.pickle`.
- la présentation des résultats obtenus dans la réduction du nombre de colonnes (caractéristiques) de votre ensemble de données
    - Vous devrez :
        - importer votre ***DataFrames*** enregistré précédemment;
        - appliquer la réduction de dimensionnalité;
        - citer (avec identification) les caractéristiques restantes;
        - utiliser des outils de visualisation pour interpréter vos résultats;
        - recommencer ce travail depuis le point deux en définissant une nouvelle largeur de fenêtre temporelle glissante et un nouveau ratio de chevauchement pour créer un deuxième ensemble de données;
        - comparer les deux résultats de la réduction pour les deux ensembles de données ;

    Explication : sur Moodle, vous avez des fichiers de code Python qui sont fournis. Ceux-ci utilisent une librairie appelée ***`scikit-learn`*** *et exploite un algorithme de réduction de dimensionnalité. Avant de l'utiliser pour traiter les points ci-dessus, j'aimerai que vous fassiez une recherche sur internet au sujet de l'algorithme. Vous devrez comprendre l'objectif (à quoi sert et comment il le fait – dans les grandes lignes) de cet algorithme. Ainsi, dans le rapport du TP2, vous devrez décrire cet algorithme (en 1 page maximum). Cela vous permettra de comprendre les résultats et d'interpréter ceux-ci. Cette partie du travail est très représentative de votre quotidien en tant que* ***data scientist****.*

- une conclusion
    - résumer les informations essentielles de votre extraction de caractéristiques.
- une conclusion générale
    - résumer ce que vous avez apprécié, appris, moins apprécié par rapport aux données, les outils utilisés ou encore le travail demandé dans ce TP2.

Ainsi, la programmation avec Python doit être implémentée pour atteindre les objectifs du rapport.

J'attends également que vous programmiez ***au minimum 2 nouvelles caractéristiques que nous n'avons pas vu en cours***.

Je veux également que votre code soit constitué d'un fichier `main` et d'autres fichiers ou vous créerez des fonctions ou des classes.

4. Précisions

En ce qui concerne l'ensemble de données (***dataset***), ***la restriction est qu'il doit concerner des données temporelle*** (***time series***) ***pour un problème de classification***. Vous allez chercher sur le Web un ensemble de données dans un domaine qui vous intéresse pour plus de « fun » (ex. : bio-informatique, marketing, commerce, …). Voici un échantillon de liens Web qui donnent accès à des ensembles de données :

- https://www.data.gov/
- https://www.reddit.com/r/datasets/
- https://www.reddit.com/r/data/
- https://registry.opendata.aws/
- https://rs.io/100-interesting-data-sets-for-statistics/
- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/datasets.php
- https://datasetsearch.research.google.com/
- etc.

Annexe A

Grilles des barèmes

Bien que le TP2 ne représente que 20% de la notes finale, celui-ci est noté sur 100 points.

Rapport Scientifique **Points notés**	**Barème**
Structure (Plan) de votre rapport	5/100
Description de votre ensemble de données	10/100
Présentation de l'extraction des caractéristiques	10/100
Présentation de l'algorithme de réduction de dimensionnalité	10/100
Présentation des résultats de la réduction	10/100
Interprétations	10/100
Comparaison	5/100
Total	**60/100**

Scripts Python **Points notés**	**Barème**
Structure de votre code (*fonctions, classes, fichiers …*)	10/100
Atteintes des objectifs par rapport au ***DataFrame***	15/100
Atteintes des objectifs par rapport à l'affichage/identification des caractéristiques restantes après la réduction	10/100
Niveau des commentaires	5/100
Total	**40/100**

En ce qui concerne la qualité du Français, ce sont des points de pénalité pouvant aller jusqu'à 10/100.