TD 3 : Data Science avec Python

Objectifs du TD

L'objectif de ce TD est de vous permettre d'appliquer ce que vous avez appris dans les cours sur *Data Camp*, en utilisant les données disponibles sur le covid-19 puis une base de chiffres manuscrits. Les données sur le covid-19 sont mises à jour quotidiennement sur `GitHub`, par John Hopkins University [1]. Vous travaillerez sur les données du 23 mars 2020 matin (allant ainsi jusqu'au 22/03/2020 inclus). Une partie de ces données est disponible sur madoc (uniquement les données que vous aurez à utiliser). Les données sur les chiffres manuscrits sont directement utilisables via `scikit-learn`.

Le travail sera réalisé en binôme. Vous devrez tout d'abord constituer un binôme et indiquer celui-ci sur madoc. En fonction de votre numéro de binôme, vous travaillerez sur les données des pays ou régions indiqués, concernant les données sur le covid-19.

Travail à rendre

Vous aurez à rendre le code Python de votre travail ainsi qu'un rapport de 4 à 10 pages présentant :

— principalement les figures demandées (pensez à numéroter vos figures et à leur donner un titre) et commentées un minimum (au moins une ou deux lignes) ;

— les réponses aux questions posées si la réponse est numérique ou n'est pas sous forme de figure.

1. Extraction des données correspondant aux pays/régions choisis

Utilisez le module `pandas` pour ouvrir les fichiers `csv` donnés et pour y sélectionner les pays et régions qui ont été attribués à votre numéro de binôme.

(a) Calculez et affichez le nombre de cas confirmés, pour chacun des pays/régions choisis (fichier *time_series_covid19_confirmed_global.csv*).

(b) Calculez et affichez le nombre de décès, pour chacun des pays/régions choisis (fichier *time_series_covid19_deaths_global.csv*).

2. Visualisation des données

Utilisez le module `matplotlib` pour créer des courbes, histogrammes et barres. Choisissez la visualisation qui vous paraît la plus adaptée pour visualiser chacune des informations suivantes demandées.

(a) Pour chacun des pays/régions choisis, créez une visualisation de l'évolution du nombre de nouveaux cas de contamination et du nombre de nouveaux décès journaliers.

(b) Pour chacun des pays/régions choisis, créez une visualisation de l'évolution du nombre de cas de contamination et du nombre de décès cumulés.

(c) Pour l'ensemble des pays, créez une visualisation de l'évolution du nombre de cas de contamination cumulés.

1. Données accessibles ici : `https://github.com/CSSEGISandData/COVID-19`

(d) Pour l'ensemble des pays, créez une visualisation de l'évolution du nombre de décès cumulés.

Vous pouvez également proposer d'autres visualisations des données qui vous paraissent pertinentes.

3. Classification supervisée par k plus proches voisins

Pour cette dernière partie, vous allez faire de la classification supervisée de chiffres, en utilisant les données du `Digit Dataset`, disponibles via `scikit-learn`. Vous pouvez regarder le tutoriel se trouvant sur le site de `scikit-learn` [2] et l'adapter pour utiliser un classifieur de type k plus proches voisins, comme vous l'avez vu dans le cours sur *Data Camp*. Afin de fixer la valeur k, vous pouvez utiliser une recherche de type `GridSearch` ou `RandomizedSearch` et utiliser des visualisations de type courbe ROC, pour choisir la meilleure valeur. Quels sont les résultats obtenus sur votre ensemble de test, en utilisant 40% des données dans l'ensemble de test ? Quels sont les résultats obtenus, en utilisant de la validation croisée ?

Si vous le souhaitez, vous pouvez tester d'autres types de classifieurs, en utilisant les mêmes ensembles d'apprentissage (*train*) et de test pour chacun des classifieurs. Vous pourrez alors comparer les résultats obtenus par chacun des classifieurs, sur l'ensemble de test.

2. `https://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html`