Projet

Titre : Apprentissage machine pour mesurer la tendance d'évolution du Covid-19

Le but : tester des algorithmes simples et ceux du deep learning pour prédire l'évolution de la pandémie Covid-19 pour un pays ou une région.

Question : ayant des données de quelques pays pour une période antérieur peut-on avoir un modèle qui permettra de prédire une valeur de nombre de cas ou de décès, si on observe le même phénomène au début ?

Hypothèse : les performances des algorithmes du machine learning permettent de modéliser une tendance d'une manière efficace.

Le travail demandé est de vérifier cette hypothèse vrai ou fausse (sans obligation de résultats). Cela pourra nous amuser durant cette période si particulière ! en espérant que les résultats annonceront des tendances rassurantes !

Cependant, ici, nous simplifions les choses car le lien entre les données est complexe et joue un rôle important, par exemple la politique sanitaire, la culture des pays, les actions mises sur le terrain…

Les données : plusieurs site internet publient chaque jours les données mondiales de l'expansion du coronavirus par pays. A titre d'exemple l'université Johns Hopkins, Kagle… Nous optons pour les données de l'université de Johns Hopkins. Kagle prose des challenges intéressants pour ceux qui souhatent allez plus loin.

Afin de pouvoir utiliser les données de manière cohérence, il faut normaliser la variable temporelle, 1 jours premiers cas confirmé, puis 2eme jours, … En d'autres termes on doit être indépendant de la date du début et de fin de la série temporelle. On peut aussi normaliser le nombre de cas par rapport à la population d'un pays ou une région un nombre entre [0 1], si cela améliore les performances.

Les données sont obtenues du dépôt Github de l'université de Johns Hopkins https://github.com/CSSEGISandData/COVID-19. https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series
Chaque 24 heures il y a une mise à jour des données.

- Série temporelle 1 : date vs nombre de cas confirmés par pays. (time_series_covid19_confirmed_global.csv.html)
- Série temporelle 2 : date vs nombre de décès par jours (time_series_covid19_deaths_global.csv.html)
- Série temporelle 3 : date vs nombre de guéri par jours (time_series_covid19_recovered_global.csv.html)
- Variable de l'âge >65 par pays, à utiliser seulement avec la question 4.b. (https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS)

Le travail à réaliser :

1. Choix et analyse des données sur lesquelles vous aller travailler, choix des pays,…
2. Préparer les ensembles de données d'apprentissage pour chaque série temporelle. 2/3 de données pour l'entrainement et 1/3 pour le test.
3. Test d'un modèle de régression de votre choix pour chaque série temporelle, vous pouvez utiliser des bibliothèques python come scikit-learn, keras... En donnant un jour après le début de l'épidémie le modèle donnes une valeur prédite (nombres de cas...).
4. Test des réseaux récurrents LSTM (télécharger sur le web le code en python). Par exemple : https://adventuresinmachinelearning.com/keras-lstm-tutorial/. Le but est d'apprendre un modèle avec des séries temporelles, puis on donne au modèle une séquence de donnée (historique). Le modèle génère la prédiction pour les jours suivants.
    a. Variable monodimensionnelle, chaque série séparée.
    b. Variable multidimensionnelle, combinaison des séries. Utilisation de vecteur de paramètres en entrée [cas confirmé, nombre de décès, âge]. Avec cela nous souhaitons prédire le nombre de décès pour un pays. En d'autres termes, étant donné une série de vecteurs consécutifs des jours précédents que sera le nombre de décès ou des guéris dans les prochains jours ?
5. Question bonus ajout du paramètre de confinement.

Remarque : pour le test l'erreur pourra se calculer avec la différence entre la valeur réelle et la valeur prédite. On peut utiliser la norme(v1-v2) ou la valeur absolue abs(v1-v2) pour calculer l'erreur ainsi que l'erreur moyenne sur tout l'échantillon du test.

Rendu : (sur celene)

1. Un rapport d'environ une dizaine de pages, expliquant les techniques utilisées, les données, et les résultats.
2. Code et données