Projet de Machine learning: Prédiciton de la validation des titres de transport en île de France

Enzo DAMION et Armelle COUTAUX

December 13, 2022

Nous nous sommes interessés à l'évolution du nombre de validation de titres de transport en île de France.

1 Les données

Choix de la base de données

En tant qu'étudiant à Nancy, nous empruntons de manière occasionnelle le métro parisien en validant des tickets de métros à l'unité. Cependant, il existe bien plus de type de titre de transport (Imagine R, Navigo, etc). Il est donc légitime de se demander s'il existe un profil type d'utilisateur pour chaque type de titre de transport et si oui, comment le modéliser.

Pour le choix des données, nous avons donc opté pour l'historique des données de validation sur le réseau ferré disponible sur :
https://data.iledefrance-mobilites.fr/explore/dataset/histo-validations-reseau-ferre/table/
Il s'agit ici de séries temporelles répertoriant pour chaque année depuis 2015, le nombre de validation par arrêt et par catégorie de titre avec un jour comme granularité temporelle.
A la vue de ces données, nous essayerons donc de répondre à la question suivante : Peut-on prédire l'affluence dans les transports une semaine à l'avance ? Si oui, le modèle de prédiction est-il commun aux différents types de titre de transport ? Globalement, quel est le meilleur modèle de prédiction ?

Pour cela, nous effectuerons dans un premier temps une rapide analyse de données avant de les exploiter sur différents modèles de réseaux de neurones.

Analyse des données

Tout d'abord, si on s'intéresse au nombre de validation par jour de manière indépendante du type de titre de transport (voir
`Distribution du nombre de validation normalisé par jour.jpg` et
`Distribution du nombre de validation par jour.jpg`), on observe que la distribution des valeurs varie fortement pour l'année 2020 (et 2021 de manière moins significative). En effet, alors que les caractéristiques de distribution classiques sont quasiment identiques pour les années 2018 et 2019, on a un écart de 40% pour la médiane et pour le troisième quartile de l'année 2020. De manière, plus générale, on remarque que l'année 2020 comptabilise presque 2 fois de validations de titre de transport. Cette anomalie s'explique facilement par l'arrivée du covid-19 en France et les confinements qu'il a provoqué. Les utilisateurs du réseau ferré ne pouvant pas sortir de chez eux pendant plusieurs mois, la distribution des validation a été perturbé sans pour autant que le pic de validation sur une journée le soit (hors période de confinement). On évitera donc d'utiliser cette année comme référence.

En traçant, l'évolution du nombre de titres de transport validés par jour en 2020 (voir `2020.jpg`), on observe effectivement l'impact des deux confinement du 17 mars au 11 mai 2020 et du 30 octobre au 15 décembre 2020. On notera également un motif hebdomadaire avec un double pic le weekend suivi d'où l'interêt de prédire à j+7.

On peut également remarquer que la majorité des validations sont effectuées avec un titre de transport de type `Navigo` (voir
`Distribution par type de titre de transport.png`. Si l'on souhaite comparer des modèles sur différents types de titre de transport, il conviendra donc de normaliser les données au préalable.

Traitement des données

Nous avons décidé d'effectuer les prédictions à l'échelle de l'année pour un titre de transport donné. Nous avons ainsi sommé les données par titre de transport pour chaque jour, nous avons donc 365 lignes par années (non-bissextile).

2 Analyse du problème et choix du modèle

Pour avons essayer plusieurs modèles que nous comparerons dans la section résultat :

1. Un modèle `RNN` implémente le rnn de la libraire pytorch puis un simple perceptron à 10 couches cachée (`mlp`).

2. Un modèle `CNN` réalise les prédictions sans récurrence à l'aide d'une simple convolution en dimension 1 (`cnn` avec un kernel de 3 et un stride de 1) suivi d'un perceptron à une couche cachée de 10 neurones (`mlp`).

3. Un modèle `LSTM` et un modèle `GRU` implémentée avec la même architecture que le rnn.

3 Résultats

Nous avons étudié la convergence des quatres réseaux décrits précédemment. Nous réalisons un apprentissage supervisé avec une prédiction à 7 jours du nombre de validation sur toutes les stations de métro de l'île de France. La base d'apprentissage est constituée des données journalières de validation de l'année 2018 et nous testons la validité de notre modèle sur les données de 2019.

Les quatre modèles utilisent le critère de *mean square error* (`MSE`) comme erreur du modèle et le score R2 pour évaluer les performances du modèle.

On observe sur la Figure 1 `Figure1_RNN.png` que le rnn converge rapidement et sans surapprentissage pour un learning rate de 0,0001. Pour un learning rate de 0,001, on observe sur la `Figure2_RNN_overfitting.png` que du surapprentissage se produit à partir de 3500 époques. Au dela de ce point, le modèle effectue des saut et n'est plus pertinent.

En comparant avec les résultats obtenus avec le simple CNN, observables sur la `Figure3_CNN.png`, la convergence est rapide pour un learning rate de 0,001. Cependant, comme remarqué dans le précédent rendu, nous ne comparons pas exactement les mêmes données puisque nous évaluons [T(t), ..., T(t+2)] au lieu de [T(0),.., T(t)]. De plus, même en augmentant le nombre d'époques à 5000, nous n'observons pas de surapprentissage, ce qui est surprenant.

Le LSTM présente un comportement semblable à celui du rnn : on observe une convergence plus lente mais sans surapprentissage pour un taux d'apprentissage de 0,0001, visible sur la figure 4 : `Figure4_LSTM.png`. Pour un taux d'apprentissage de 0,001, on observe du surapprentissage dès l'époque 350 sur la Figure 5 : `Figure5_LSTM_overfitting.png`, contre l'époque 3500 pour le rnn.

Pour le réseaux GRU, on observe du surapprentissage à partir de l'époque 6000, même avec un learning rate de 0.0001 `Figure6_GRU.png`.

Parmi les 4 modèles testés, le RNN semble être le mieux qualifié avec un R2score allant jusqu'à ?