

Estimation en temps réel des retards dans un réseau de bus à l'aide de données historiques

Nikita Marchant¹

¹Université Libre de Bruxelles, Département d'Informatique
nimarcha@ulb.ac.be

Abstract

Je n'ai pas encore écrit d'abstract, voici donc du latin pour remplir. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean purus tellus, fringilla id orci nec, molestie dignissim neque. Sed luctus vitae ante non condimentum. Pellentesque sollicitudin mi aliquam, euismod est at, cursus erat. Nulla porttitor enim tellus, sed aliquam nisl facilisis quis. Proin at odio convallis leo mattis vehicula. Nam ac suscipit lectus. Nullam tincidunt lobortis tincidunt. Praesent sodales leo ut quam mollis sodales. Aliquam non mauris ex. Donec sed elementum elit, pellentesque dapibus arcu. Cras convallis venenatis bibendum.

Introduction

Le but de ce projet est de prédire, en temps réel, l'heure d'arrivée aux arrêts d'un véhicule de transport en commun. Dans le cadre de ce travail, la prédiction sera effectuée pour des lignes de bus, tram et métro de la STIB¹.

Un réseau de transport en commun étant très difficile à modéliser à cause de sa complexité et par ce qu'il est très influencé par des événements stochastiques, l'approche qui sera utilisée ici sera non pas de modéliser le réseau pour prédire son état futur mais plutôt d'extrapoler les trajets des ses véhicules grâce à des données historiques récoltées au préalable.

Métriques de performance

Pour pouvoir comparer plusieurs méthodes, il est important d'avoir des métriques bien définies mesurant la performance de celles-ci.

La littérature utilise souvent le RMSE²: l'erreur quadratique moyenne.

Cependant, cette métrique souffre d'un problème : qu'un bus soit annoncé une minute à l'avance ou une minute en retard compte de la même manière alors que dans un cas l'usager attend son bus une minute de plus et que dans l'autre il le rate.

J'ajouterai ici une métrique qui résoud ce problème.

¹Société de transports en commun à Bruxelles (Belgique)

²Root Mean Squared Error

Algorithmes des prédiction

Plusieurs algorithmes seront implémentés et leur performances seront comparées.

Méthodes naïves

Deux méthodes assez naïves seront implémentées pour servir de base pour comparer les méthodes proposées par la suite. La première, la plus simple, est de prédire tout le temps la même durée de trajet entre deux arrêts en utilisant simplement la durée spécifiée dans les horaires statiques.

La seconde, celle utilisée par la STIB, est de prédire le temps de trajet entre deux arrêts comme étant la moyenne des temps de trajets des trois derniers véhicules de la ligne étant passés sur ce tronçon.

k plus proches voisins

L'algorithme des k plus proches voisins³ est une méthode d'intelligence artificielle qui peut être utilisée aussi bien pour de la classification que pour de la régression. L'idée de cette méthode est de trouver les k trajets les plus similaires à la cible et de les utiliser pour extrapoler le temps de trajet futur.

Pour cela, on projette chaque trajet dans un espace à n dimensions avec $n+1$ étant le nombre d'arrêts déjà effectués par le véhicule dont on cherche à prédire le temps de trajet (ce véhicule sera appelé α).

Les trajets sont donc représentés par le vecteur colonne $T_\alpha = (d_{1,\beta}, d_{2,\beta}, \dots, d_{n,\beta})^T$ avec $d_{i,\beta}$ étant la durée du trajet entre l'arrêt i et $i+1$ pour le véhicule β . La similarité $s_{\alpha,\beta}$ entre deux trajets α et β est définie comme la distance Euclidienne entre deux vecteurs :

$$s_{\alpha,\beta} = \sqrt{\sum_{j=0}^n (d_{j,\alpha} - d_{j,\beta})^2} \quad (1)$$

La prédiction du temps de trajet $\hat{d}_{n+1,\alpha}$ est donnée par la moyenne des temps de trajet pour ces arrêts des k plus proches voisins de α dans l'espace.

³ k -nearest neighbors ou encore k -NN. (Hastie et al. (2009))

$$\hat{d}_{n+1,\alpha} = \frac{1}{k} \sum_{j=0}^k d_{n+1,k} \quad (2)$$

Pondération : Il serait peut-être appréciable que les performances récentes d'un véhicule influencent plus la prédiction que ses performances plus anciennes. Pour cela, on peut introduire la notion d'une distance pondérée dans laquelle certaines dimensions subiraient un homothétie pour en augmenter ou diminuer le poids.

J'ajouterai ici la définition de cette distance

Prise en compte d'autres critères que le temps de trajet : Il pourrait être utile de prendre en compte d'autres critères que le temps de trajet pour déterminer la similitude entre deux trajets. Le jour de la semaine, la météo ou même l'heure sont des facteurs influençant le temps de trajet. L'heure étant une variable continue, nous pouvons simplement l'ajouter au vecteur T_α et utiliser ce vecteur de $n + 1$ éléments pour la distance euclidienne. Pour ce qui est du jour de la semaine ou de la météo, ce sont des variables discrètes pour lesquelles il faudra définir une nouvelle fonction de distance.

Il faut encore que je fasse des recherches sur les k-NN avec variables discrètes si je m'aventure dans cette direction là.

References

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.