Polytechnique Montréal

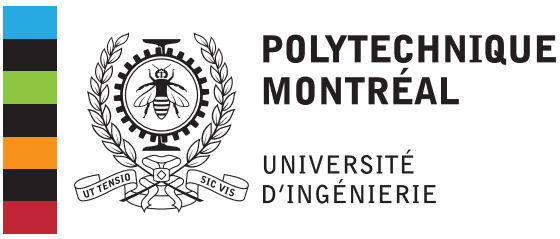
Travail Pratique 4

Fouille de données

CIV8760 : Gestion des données en transport

Chargé de TP :
Guillaume Neven

Automne 2024



1 Introduction

Ce travail pratique a pour objectif de vous familiariser avec les techniques de classification et d'analyse de données appliquées aux données. Vous utiliserez des données réelles provenant de Toronto pour effectuer diverses tâches de classification et d'analyse.

2 Données

Pour la tache demandée, vous allez devoir utiliser les données de la ville de Toronto. Les données ont été récoltées dans la cadre de du papier de Loder *et al.* (2019), comparant le trafic dans plusieurs ville à travers le monde. Voici comment les données sont organisées :

Colonne	Description
time	Timestamp des données collectées
detid	Identifiant du détecteur
flow	Débit de circulation (en veh/h)
occ	Taux d'occupation (fraction du temps pendant lequel un véhicule est détecté)
speed	Vitesse moyenne des véhicules (en km/h)

Table 1: Description des données de circulation

Figure 1: Carte de la position des detecteurs

3 Segmentation

Cette première tâche va se concentrer sur la classification non supervisée des données. Le diagramme fondamental, qui représente le flux de véhicules en fonction du ratio d'occupation du détecteur, permet de caractériser l'état de la circulation sur la route sur laquelle le détecteur est installé. Des explications plus précises sont disponibles sur les diapositives du cours.

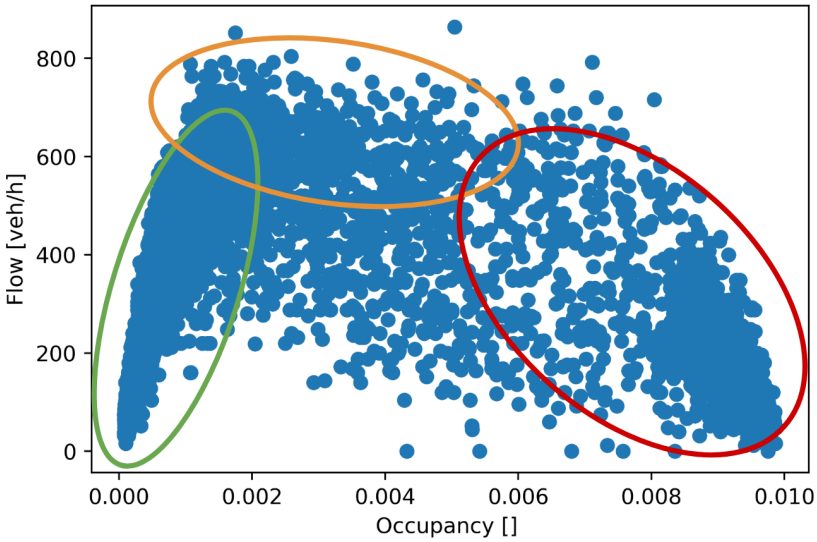


Figure 2: Diagramme fondamental, avec les différents états de circulation: libre en vert, à capacité en orange et congestionnées en rouge, le reste étant des états de transitions.

Pour la première tâche, sélectionnez un détecteur sur une route avec de la congestion (c'est-à-dire des points dans la zone rouge). En utilisant la méthode des k-moyennes en utilisant le flux, l'occupation et la vitesse comme variables, regroupez les états de circulation. Notez qu'il est possible de grouper des zones si elles représentent selon vous un niveau de service similaire. Par exemple, il est possible que deux zones représentent la vitesse libre, une pour peu de flux et une pour beaucoup de flux, il est possible de les grouper une fois la classification effectuée en un seul groupe. Utiliser le nombre de groupes que vous jugez nécessaire. Décrivez les groupes résultats en fonction de leurs attributs (donner un nom aux groupes) et créez au moins un graphique présentant les différents états selon plusieurs attributs

4 Arbre décisionnel

Dans cette deuxième section, vous devez utiliser un arbre de décision en utilisant les classes obtenues dans la première tâche. Retenez 15% des données, cela sera votre jeu de donnée test. Entrainez votre modèle sur le jeu d'entrainement (les 85% restant).

Entrainez un arbre sur le jeu d'entrainement, créez le graphe de l'arbre de décision (ne pas utiliser plus de 3-4 niveaux, sinon le graphique sera difficile à interpréter). Interprétez les règles et leur niveau de confiance, et commentez sur la logique de l'arbre, auriez-vous utilisé des règles similaires ?

Maintenant, prédisez sur le jeu test. Est-ce que les résultats sont toujours satisfaisants ?

5 Classification

Maintenant, vous allez créer un classificateur général, qui pourra classer l'état de trafic de n'importe quel détecteur. Pour ce faire, répétez la tâche 1 avec 5-6 détecteurs (réappliquer une méthode de k-moyennes par détecteurs), essayer de prendre des détecteurs avec des états de trafic différents (des diagrammes fondamentaux différents). Il est possible que certains détecteurs aient beaucoup de points en phase de transition, il sera plus difficile pour vous d'établir les classes, n'hésitez pas à ne sélectionner que des diagrammes clairs.

Entrainez un modèle global (il n'est plus nécessaire de bloquer le nombre de niveaux), encore une fois, retenez 15% des données. Créez un graphe des prédictions pour les jeux tests (un graphe par détecteurs). Êtes-vous satisfait des résultats ?

Maintenant, prédisez les états de trafic pour les données d'un détecteur pas encore utilisé jamais vu par la modèle. Êtes-vous satisfait du résultat ? Si les performances sont mauvaises, commentez pourquoi. Pouvez-vous essayer de fournir une solution, sans l'implémenter pour autant.

6 Soumission du rapport

Ce TP se fait individuellement. Veuillez soumettre le rapport au format **jupyter notebook** d'ici au 5 décembre à 23h59 sur Moodle. Répondre aux questions directement dans le notebook. Un exemple des modèles utilisés est disponible dans les ressources du cours.

Assurez-vous que le rapport est exempt d'erreurs grammaticales et comprend des graphiques précis et des explications claires. Des points seront retirés pour les erreurs de rédaction et les inexactitudes dans l'analyse des données. Si un modèle de langage est utilisé (par exemple, ChatGPT), vous devez divulguer son utilisation précise.

References

Loder, A., L. Ambühl, M. Menendez and K. W. Axhausen (2019) Understanding traffic capacity of urban networks, *Scientific Reports*, **9** (1) 16283, ISSN 2045-2322.