Polytechnique Montréal

Lab Work 4

Data mining

CIV8760E : Transport Data Management

Teaching Assistant :
Guillaume Neven

Fall 2024

1 Introduction

This practical work aims to familiarize you with classification and data analysis techniques applied to real-world data. You will use actual data from Toronto to perform various classification and analysis tasks.

2 Data

For the assigned task, you will use the Toronto city data. The data was collected as part of the study by Loder *et al.* (2019), which compares traffic across multiple cities worldwide. The data is organized as described in Table 1.

Column	**Description**
time	Timestamp of the collected data
detid	Detector identifier
flow	Traffic flow (in vehicles/hour)
occ	Occupancy rate (fraction of time a vehicle is detected)
speed	Average vehicle speed (in km/h)

Table 1: Description of traffic data

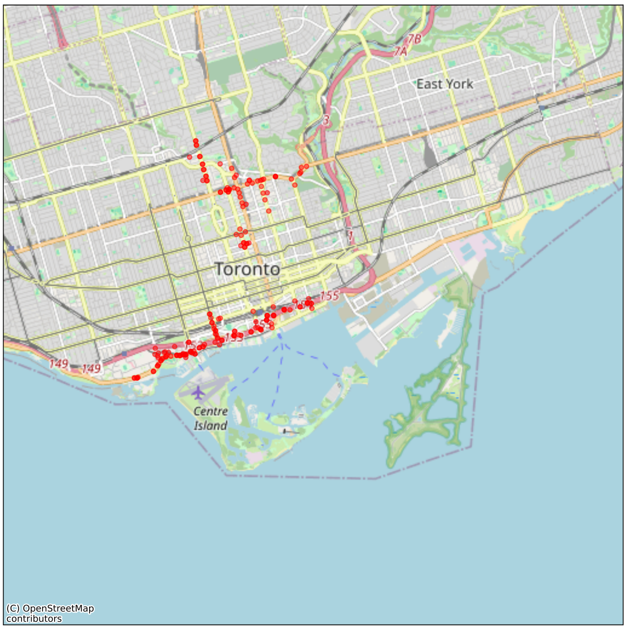

Figure 1: Map of detector locations

3 Clustering

The first task focuses on the unsupervised classification (clustering) of the data. The fundamental diagram, which represents vehicle flow as a function of detector occupancy rate, characterizes traffic conditions on the road where the detector is installed. More detailed explanations are available in the course slides.

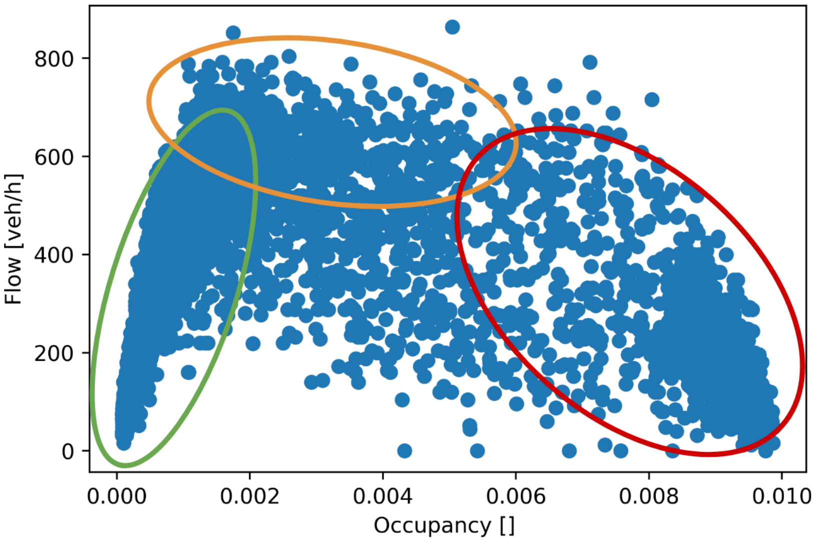

Figure 2: Fundamental diagram with traffic states: free-flow (green), capacity (orange), congested (red), and transition states.

For the first task, select a detector on a congested road (i.e., with points in the red zone). Using the k-means method and the variables flow, occupancy, and speed, cluster the traffic states. You may group clusters if they represent similar levels of service. For example, two clusters representing free-flow speed (one for low flow and another for high flow) can be combined into a single cluster after classification. Use as many clusters as you deem necessary. Describe the resulting groups based on their attributes (assign names to the groups) and create at least one graph showing the different states using multiple attributes.

4 Decision Tree

In this second section, use a decision tree with the classes obtained from the first task. Set aside 15% of the data as a test set. Train your model on the training set (the remaining 85%).

Train a decision tree on the training set and create a graph of the decision tree (limit to 3-4 levels to ensure clarity). Interpret the rules and their confidence levels, commenting on the tree's logic. Would you have used similar rules?

Next, make predictions on the test set. Are the results satisfactory?

5 Classification

Now, create a general classifier capable of classifying the traffic state for any detector. To do this, repeat Task 1 with 5-6 detectors, applying k-means clustering to each detector. Try to select detectors with different traffic states (different fundamental diagrams). Some detectors may have many transition points, making it harder to establish classes — feel free to select clear diagrams only.

Train a global tree model (you no longer need to limit the number of levels). Again, set aside 15% of the data. Create graphs of predictions for the test sets (one graph per detector). Are you satisfied with the results?

Now predict traffic states for data from a detector not previously seen by the model. Are the results satisfactory? If performance is poor, comment why. Propose a solution, even if you do not implement it.

6 Report Submission

This practical work is to be completed individually. Submit your report in **Jupyter Notebook** format by December 5 at 11:59 PM on Moodle. Answer the questions directly in the notebook. An example of the models used is available in the course resources.

Ensure the report is free of grammatical errors and includes precise graphs and clear explanations. Points will be deducted for writing errors and inaccuracies in data analysis. If a language model (e.g., ChatGPT) is used, disclose its precise usage.

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.