Problem Definition and Characterization

Problem: “Bicycle Share Demand Prediction”

Stations where bicycles are docked and taken from are a common way of implementing a bicycle sharing system. As bicycle sharing becomes one of the most popular ways to commute, it plays a significant role in the public transportation system. However, the demand varies greatly between location, weekday and other factors, which leads to imbalances and congestions in the system. This results in customer dissatisfaction and unreliability of the system, jeopardizing the central role of bicycle in reaching emission-neutral transportation and providing convenience.

To properly implement dynamic solutions, such as adaptive dynamic pricing and terminal extensions, the demand needs to be reliably predicted. Since the demand fluctuates based on various aspects, we decided to train a machine learning model to investigate the relations between these aspects and the demand. With a sufficiently accurate model, bicycle sharing companies can adopt adjustments to provide more reliably service.

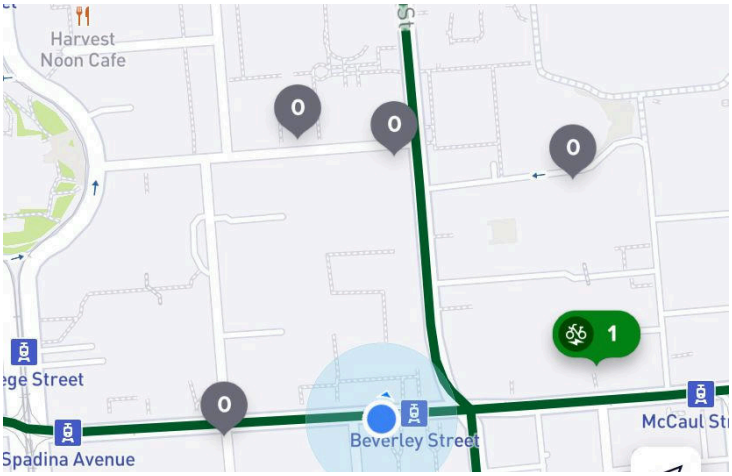

Figure 1: A screenshot of Bike Share Toronto app showing empty terminals

Problem Characterization

Our goal is to predict the demand of bicycle sharing using a machine learning approach given historical data and additional features, such as the day of the week and the daytime.

We consider a bicycle sharing system consisting of N_S stations with fixed capacities Cap_s, i.e. a maximum of Cap_s bicycles can be docked at station $s \in \mathbb{N}$. We denote the current number of bicycles at station s by $B_s(t)$ We distinguish two types of demands:

- $\mathrm{OutDem}_s(t)$: the number of bicycles that people take out the dock at (t, s) *if there are enough bicycles present.*
- $\mathrm{InDem}_s(t)$: the number of bicycles that people dock at (t, s) *if there is be enough space present.*

In order for the number of bicycles to be conserved, a pair of suitable demand functions has to satisfy

$$\sum_{t,s} \mathrm{InDem}_s(t) - \mathrm{OutDem}_s(t) = 0,$$

where the sum runs over all times and states.

These functions encode the psychological, social and logistical aspects of bicycle sharing demand. They define the behavior of the system (i.e. the number of bicycles at station s) via the difference equation

$$B_s(t+1) = B_s(t) + \min(\mathrm{InDem}_s(t), \mathrm{Cap}_s(t)) - \min(\mathrm{OutDem}_s(t), B_s(t)).$$

In order to formulate a proper machine learning problem, we now approximate the discrete quantities by real numbers, i.e. we introduce

- the hourly rates of In-Demand $\mathrm{ID}_s(t) \in \mathbb{R}$ and out demand $\mathrm{OD}_s(t) \in \mathbb{R}$
- the actual hourly rates of bicycles docked $\mathrm{In} \in \mathbb{R}$ and bicycles taken out $\mathrm{Out} \in \mathbb{R}$.

Our given given data consists of a list of all rides including start, end stations and start and and times in minutes. The discrete behavior previously stated now translates to the partially continuous loss

$$\mathrm{loss}(t) = \begin{cases} (\mathrm{ID}(t) - \mathrm{In})^2 + (\mathrm{OD}(t) - \mathrm{Out})^2 \text{ if } 0 < B(t) < \mathrm{Cap}(t) \\ (\mathrm{OD} - \mathrm{Out})^2 + \max(0, \mathrm{ID} - \mathrm{In})^2 \text{ if } B(t) = \mathrm{Cap}(t) \\ (\mathrm{ID} - \mathrm{In})^2 + \max(0, \mathrm{OD} - \mathrm{Out})^2 \text{ if } B(t) = 0 \end{cases},$$

i.e. the demands should normally be identically to the rates of the bicycles taken in or out, only if the station is either full or empty, the demands may be higher, but not lower, than the input or output bicycles.

Aspects of the Machine Learning Approach

- Controllable Parameters: These include the sharpness of the hourly rate approximation, which depends on the window. Choosing a large window to compute the rates leads to loss of precision in the time domain, while too small window might not capture the continuity of the demand well. Other control parameters of the historic data collection include the accuracy of time resolution which in the given dataset is minutes.
- Signals (i.e. input features): To predict the future hourly demand, we input
  - the historic input and output rates before the prediction time
  - the daytime
  - the day of the week.
- Error States (i.e. failure modes of the model):
  - Possible failure modes include cases if the future prediction are not accurate enough to be useful, the predictions become unstable or the in-demand does not match the out demand.
- Noise Factors:
  - Capacity changes of the stations during the month, which appears due to e.g. repositioning of stationing.These changes are not included in the published data.
  - Population density changes, irregular road and facility closures which are not present in the training data but have a significant influence on bicycle sharing behavior.

Possible Additional Modeling Approaches

In order to leverage the spatial relational structure of the bicycle stations for e.g. a Graph Neural Network approach, we introduce the weighted bicycle stations graph. This is a weighted undirected graph (V, E, w), with vertices $V = \{s \mid s \text{ is a bicycle station}\}$. To define a suitable connectivity structure, we draw an edge between two stations if a sufficient number of rides take place between the nodes, i.e.

$$\{s_1, s_2\} \in E :\Leftrightarrow |\{\text{ride between } s_1, s_2\}| > N_{\min}$$

with $N_{\min}$ suitably chosen depending on the data such that the graph is sufficiently sparse. Finally, we define the weights to be the number of rides between two stations, i.e. $w(s_1, s_2) := |\{\text{ride between } s_1, s_2\}|$