COMS 4030A: Assignment 1

Tamlin Love (1438243)

1 Problem Description

At Wits, the university runs a bus service that transports students between campuses and residences. These busses run on a schedule and are typically scheduled so that each stop is visited by a bus every 30 minutes. Many students, including myself, make frequent and regular use of this service. In particular, I use catch the bus that goes between the AMIC deck on Main Campus and the stop at Wits Education Campus (WEC), and as such these will be the only two stops considered in this report.
However, despite the presence of a schedule, certain conditions may result in a bus arriving either late or, somewhat less frequently, early. As such, a student arriving at the bus stop may have to wait for longer than necessary (in the case of a late bus) or may miss the bus entirely (in the case of an early bus). If students could better predict the deviation in arrival time of a bus, they could better manage their time and could better avoid the above two situations.

2 Formulation

To solve the problem of predicting arrival time deviation posed above, the problem is formulated as a supervised learning problem, where each of the conditions that may affect the arrival time of a bus are quantified as features and the target variable to be predicted is the difference in time between the bus's actual arrival time and its scheduled time.
More formally, we seek a function f such that $y = f(\underline{x})$ where y represents the target variable and $\underline{x}$ represents a vector of features.

2.1 Target Variable

As stated above, the target variable $y \in \mathbb{R}$ is the difference in seconds between the bus's actual arrival time and its scheduled arrival time. Thus a late bus will be labeled with a positive number and an early bus with a negative number. A bus which is exactly on time would give a value $y = 0$. Because the domain of y is continuous, the problem is formulated as a regression problem, rather than a classification problem.

2.2 Features

We now consider the following features which will make up $\underline{x}$.

Scheduled Arrival Time - the time at which the bus is scheduled to arrive. This is considered as certain times of the day may be more conducive to lateness or earliness (e.g. busses may tend to be late during peak hour traffic times). This can be represented as a decimal number $t \in [0,24)$, where, for example, $t = 17.5$ denotes the 17:30 bus.

Bus Driver ID - this number identifies the bus driver. This is considered as certain drivers may have a pattern of earliness/lateness. This can be represented as $I \in \mathbb{Z}$, where each integer corresponds to a bus driver.

Week Day - the day of the week. This is considered as different days of the week may feature regular weekly traffic patterns that may affect bus arrival time. This can be represented as an integer $d \in [0,4]$, where 0 enumerates Monday, 1 enumerates Tuesday, 2 enumerates Wednesday, 3 enumerates Thursday and 4 enumerates Friday.

Stop - the stop which we are considering. This is considered as traffic may interrupt flow from stop A to stop B, but not necessarily from B to A. This is represented as a boolean $s \in [0,1]$, where 0 represents AMIC deck and 1 represents WEC.

Rain Status - the intensity of rain at the time of scheduled arrival. This is considered as rain can have a noticeable effect on traffic flow [1]. This can be represented as an integer $r \in [0,2]$, where 0 represents no rain, 1 represents light rain and 2 represents heavy rain.

Therefore, our feature vector $\underline{x}$ will be the vector $\underline{x} = [t, I, d, s, r]^T$

2.3 Data Labeling

When discussing applications of machine learning, it is important to state how one intends to gather and label data. This report proposes an electronic system, most likely in the form of a software application, which could allow bus stop coordinators to label the above variables with ease.
Of the above variables, t, I and r could be entered manually by the on-duty coordinator, as the scheduled time and bus driver ID is known beforehand and the rain status is easily observed. The other variables, d, s and the target y could be labelled automatically by the software system, with the week day and actual arrival time deviation being caulculated from the timestamp when the data is entered, and the stop being calculated from the GPS coordinates of the coordinator's device.

References

[1] Edward Chung, O. Ohtani, H. Warita, Masao Kuwahara, and H. Morita. Effect of rain on travel demand and traffic accidents. In *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.* IEEE, September 2005.