COMS 4030A: Assignment 1

Tamlin Love
1438243

February 27, 2019

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 offset 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 offset 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.

2.1 Features

In this report we consider the following features.

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.

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 often has a big effect on traffic flow. 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.

Temperature - the temperature at the stop. This is considered as we hypothesise that ambient temperature may have an effect on traffic. This is represented as **TODO: Look at Occupancy thing and see how they do it**
Thus, we