Data Memo

Data Science II (STAT 301-2)

Derex Wangmang

January 31, 2021

Contents

Overview

The dataset contains concerning rental bikes in Seoul, South Korea. This dataset comes from the University of California at Irvine's Machine Learning Repository, found here. Collection of the data requires a simple download of the csv file.

The dataset contains 8760 observations and 14 characteristics. Most are numerical, but some are nominal in nature. Some of the characteristics are listed below.

- Date
- Hour
- Temperature
- Humidity
- Holiday
- Windspeed
- Rented Bike Count

There are no missing values within this dataset.

Research Questions

My project will be focused on a predictive analysis. Some potential questions include:

- Based on some or all of the factors, is a given day a holiday?
- Based on some or all of the factors, what is the temperature?
- Based on some or all of the factors, how many bikes will be rented?

My approach will depend upon the specific question I pursue. To answer the first question, I will take a classification-based approach with a binary output variable: that day is a holiday or not a holiday. The second and third questions require a regression-based approach, as the temperature and number of rental bikes are numerical.

Of these, predicting the bike is the most similar to a real world scenario. Thus, I will most likely focus on answering that question. I suspect the variables that will be helpful in modeling the response include day, hour, temperature, and amount of rainfall.

Difficulties

Since the data collection mechanism is through a website download, I do not expect it to be difficult. After scanning through the data, the dataset appears to be tidy and well-formatted.

However, these observations may not be independent. While it can be plausibly inferred that the observations between each day are independent, the observations within a day may not be. That is, the rented bike count within day 1 will not affect the count on day 2. However, the rented bike count in hour 1 may affect the rented bike count in the next hour, as the same people may be on the bikes for multiple hours at a time, potentially leading to dependence between observations. To combat this, I may space out the observations over time. Rather than using the entire dataset with hours 0 through 23, I may retrieve every 4th hour.

Proposed Timeline

I expect to have my dataset loaded into R by February 13. I expect to start my analysis that same weekend and complete the first draft by the early submission deadline, March 1.

Citations

Seoul Bike Sharing Demand Data Set. (2020). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand]. Irvine, CA: University of California, School of Information and Computer Science.