Washington DC Weather and Bike Share Usage – Evan Galloway

I started by looking over the built-in .arff datasets to see if anything looked interesting there. I soon realized that Weka can easily process .csv files, so I expanded my search, looking at various government websites, before coming across the NOAA website, which has many great datasets. However, when I saw the historical weather station data, I had an idea. I wanted to combine data from Washington DC weather stations with data from Capital BikeShare. I am using the latter dataset as part of a visualization project, but I have been curious about what other external factors might influence bikeshare usage, and how it might be possible to predict demand. Weather seemed like a possibility.

I downloaded a large dataset from DC and Maryland weather stations. Most appeared unreliable, so I decided to just choose Reagan National. I then downloaded a couple years’ worth of bikeshare data. I was able to count rides per day, which I then added as a column to the weather data. I also added a categorical column, which indicated if the volume of riders for that day was High or NotHigh. I chose to High to be greater than one standard deviation from the mean.

I loaded the data into Weka and first looked at the data with the categorical class as the target, and used the J48 tree building algorithm. After a couple iterations, it appeared that weather and precipitation were decent attribute for prediction. This is what I would expect.

I then ran a linear regression using the numerical rider class as the target. I was able to get the equation down to including just two attributes, again max daily temperature and precipitation, without really affecting the r value.

Max temperature and precipitation make sense. However, I think there are better predictors. For instance, when I looked at the distribution of rides over the course of the year, there is a clear seasonal influence. Perhaps season would be a better predictor, and temperature and precipitation are just proxies. However, I also know that DC is strongly affected by various events. For instance, the Cherry Blossom festival is a huge event, and, indeed, the bike share data shows a significant peak on those days. But at the time it occurs – in early spring – max temperature would be a poor predictor. Perhaps this is part of the reason for the weakness of the prediction.

(The days actually vary from year to year; in fact, prediction of the timing is pretty important for DC, as people schedule events and trips around that time, and must represent an interesting analytics problem.)

Here are my videos. I apologize for the poor quality and the watermark. For me, this ended up being much more the “ScreenCapture Challenge”, as I struggled to find a reliable tool for screen capture. (I’d be happy for any suggestions for future assignments.)

<https://youtu.be/8AViQMSbMIM>

<https://youtu.be/50r9OYWLmek>

List of sources investigated/used:

Weka built-in .arff datasets

Google Public Data Explorer - <http://www.google.com/publicdata/directory>

Data.gov

NOAA - <http://www.ncdc.noaa.gov/cdo-web/>

Capital Bikeshare - <https://www.capitalbikeshare.com/trip-history-data>