Capstone Proposal

Machine Learning Engineer Nanodegree

Christian Wolff

September 23, 2017

Proposal

Domain Background

In recent years, bike-share programs in cities around the world have grown in popularity. Wikipedia[1] currently lists 330 bike-sharing systems across the world. Their ultimate goal is to create a more sustainable transportation landscape that provides new mobility options for short trips and to improve connectivity to other modes of transportation.

As bike-sharing programs grow in popularity, cities begin to realize additional benefits which should mean plenty of related opportunities for years to come. But there are also challenges to overcome. One of the challenges are the running costs of said programmes, mostly from maintainance of bikes and stations, redistribution of bikes by vans during the day and the necessary IT infrastructure.

In this project, we will take a closer look at Hubway[2], Boston's regional bike sharing system. As of September 2017, it gives members access to more than 1,700 bikes at 170 stations. The station network is designed to offer a one-way transportation option and an alternative to public transport like bus or train.

A Hubway station

Problem Statement

Bike-sharing services can only be successful when they meet the demand for mobility. In this regard, two important factors for customer satisfaction are

- high availability of bikes at all stations,
- and the possibility to return a bike to any station.

But cities are one of the most complex systems known to us, and far from being homogenous or in equilibrium. Thus, some bike-sharing stations are frequented more often then others, with changing distributions over the day and week. This means, that bikes have to be continously redistributed to meet the demand.

Hubway tackles this problem on a real-time basis, as we can read on their website:

> Hubway employs 4-5 rebalancing vans, each with a payload of 20-25 bikes, used to redistribute bicycles between 6am and 10pm, 7 days a week. Bicycles are redistributed dynamically in response to real-time data.

[1] https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems#Cities

[2] https://www.thehubway.com

From this statement follows the question:

Can Hubway predict the demand for bikes and plan the necessary rebalancing ahead of time?

I assume that proper planning of rebalancing will decrease running costs and improve customer satisfaction. The goal of this project is the implementation of a workflow to predict the daily demand of bikes at the stations, so that rebalancing can be planned ahead of time, instead of relying on real-time rebalancing.

A Hubway rebalancing van

Datasets and Inputs

From the very beginning of their bike-sharing program, Hubway has collected usage data. In 2012, they ran a public challenge to visualize and analyze the data, and the results where published on Hubways' offical challenge website[3]. An expanded dataset, ranging from 2011 to 2013, is still available and will be used in this project as the main data source. It contains fine-grained data of single trips, e.g. start and end timestamp of each trip, together with the respective station.

Field definitions in Hubway trip data	
seq_id	Unique record ID
hubway_id	Trip ID
status	Trip status. ”closed” indicates a trip has terminated.
duration	Time of trip in seconds.
start_date	Start timestamp of trip with date and time, in EST.
strt_statn	ID of start station.
end_date	End timestamp of trip with date and time, in EST.
end_statn	Station ID of end station.
bike_nr	ID of bicycle used.
subsc_type	Subscription type. ”Registered” is a user with membership. ”Casual” is a user without membership.
zip_code	Zipcode of user (only available for registered users).
birth_date	Birth year of the user
gender	Gender of the user

I expect the trip dataset to be correlated with the following three factors, which will be used to augment it:

- Local weather conditions, mostly temperature and precipitation. This includes effects of the calendar seasons. Historic weather summaries for the US are available from the National Oceanic and Atmospheric Administration[4].

- We can expect a baseline of regular commuters during work days, and spikes of casual usage during weekends and public holidays. Information about public holidays in Boston can be aquired from the Massachusetts state government[5].

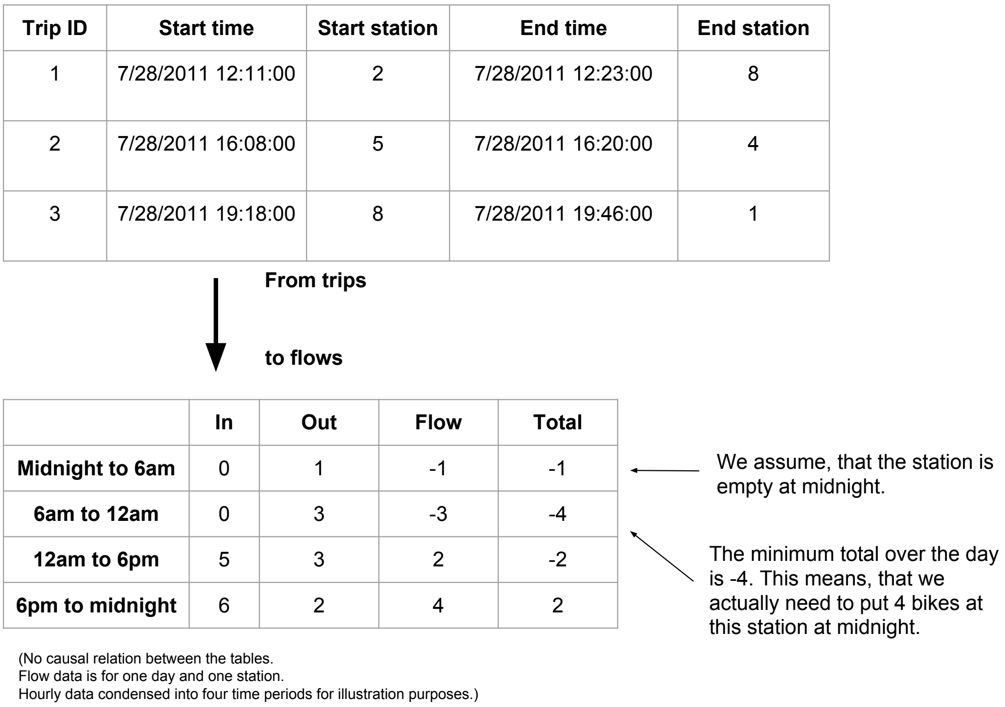
- Local events, like festivals and marathons increase the demand for mobility. These can be taken into account by using the methodology of event labeling, as proposed by Fanaee-T and Gama[6].

[3] http://hubwaydatachallenge.org
[4] https://www.ncdc.noaa.gov/cdo-web/
[5] https://www.mass.gov/
[6] https://doi.org/10.1007/s13748-013-0040-3

Solution Statement

The maximum demand at a station for a given day can be evaluated as follows:

1. From the trips dataset, compute the flow of bikes per hour. The flow at a station is defined as the number of bikes that are returned to the station, minus the number of bikes rented from the station, during a given hour.
2. Derive the total number of bikes for each hour from the flow. Assume that the station has no bikes at exactly midnight. The total nmber of bikes at a given hour is then the rolling total of the flow until this hour.
3. It is clear that the total number of bikes can be negative under the assumption that the station is empty at midnight. By taking the minimum of the running total of the day, we can evaluate how many bikes are actually needed at the start of the day, so that the station is never depleted.



Trip ID	Start time	Start station	End time	End station
1	7/28/2011 12:11:00	2	7/28/2011 12:23:00	8
2	7/28/2011 16:08:00	5	7/28/2011 16:20:00	4
3	7/28/2011 19:18:00	8	7/28/2011 19:46:00	1

	In	Out	Flow	Total
Midnight to 6am	0	1	-1	-1
6am to 12am	0	3	-3	-4
12am to 6pm	5	3	2	-2
6pm to midnight	6	2	4	2

We assume, that the station is empty at midnight.

The minimum total over the day is -4. This means, that we actually need to put 4 bikes at this station at midnight.

(No causal relation between the tables.
Flow data is for one day and one station.
Hourly data condensed into four time periods for illustration purposes.)

Transforming trips into flows

The maximum demand for each station and day can then be augmented with daily weather summaries and information about public holidays and weekends. The resulting dataset can be used to train different regression models or do unsupervised clustering for more data exploration.

So far I assumed that the rebalancing is done once each day at midnight. Because stations can also 'overflow', rebalancing is probably needed several times a day. I will try to extend the project to a point where it is possible to plan a good balancing of bikes, such that each station is close to half it's maximum capacity at any given time during the day.

Benchmark Model

As far as I know, Hubway has not yet implemented a predictive system for the given problem. This means that there is no benchmark model we can directly compare to.

One benchmark we could use instead is cost efficiency, by comparing the current costs for real-time rebalancing to the expected cost from the proposed machine learning approach. These costs can be approximated by the distance the rebalancing vans have to travel each day. For Hubway's current real-time approach I will try to derive these rebalancing tours from the real trip data, and for the ML approach from the flow forecasts. The calculated total distance for both approaches can then be compared and evaluated.

Evaluation Metrics

The trips dataset encompasses the time from 2011/07/28 to 2013/11/30. For the training and evaluation of the predictive model, the dataset will be split at a certain date into two sets. The set with the older data will be used for training the model, the younger set for testing. This ensures that the model can predict future developments from the past.

The target value of the model will be the demand for bikes at a station for given day. The root of the mean squared error over all stations and dates of the test set will be used to quantify the success of the model.

If the rebalancing tours can be derived from the data, I will compare Hubways' current real-time reaction approach with the results of the predictive approach by comparing the total distances of the rebalancing tours.

Project Design

Programming language and libraries

- Python 3.
- scikit-learn. Machine Learning library for Python.
- XGBoost. Open Source Gradient Boosting implementation for Python.

Project steps

1. Requisition of all necessary data.
2. Data exploration (tabular and visual).
3. Data cleansing and transformation.
4. Calculation of flow values.
5. Augmentation of flow data with weather data and public holidays.
6. Model training and optimisation of model parameters. Different regression models will be evaluated, with a focus on ensemble models.
7. Evaluation and comparison of real-time and predictive rebalancing.