Regression Based Prediction Model for Bike Sharing Resource Management System

HASHAN PERERA, SUDAM KALPAGE, THILINI DESHIKA

Department of Computer Engineering

University of Peradeniya

1 Introduction

Up to date bike sharing systems are able to automate the system and replace the traditional bike rentaling systems, From user registration, subscription and all the activities are connected. This system enables an efficient service for the clients while reducing a larger amount of carbon footprint from the world. Since there are about over 500 bike sharing programs around the world. This also results in an important role in traffic, nature and for the health of the living beings.

2 Problem Specification

Although having a fully automated system may not resolve all the requirements. This system is capable of solving most of the requirements of the client phase. But in order for more financial and sustainable optimization the data collected during the system operates should help to develop the system more efficiently. Currently, one of the main problems and challenges in the management of Bike Sharing Systems is to assure that users will be able to find available bicycles and parking slots in a station independently of the time. However, users' behaviour causes bicycles to be asymmetrically distributed. Therefore a Rebalancing System is needed to maintain the adequate number of bikes at each station, in order to satisfy the demand. Rebalancing is a costly operation in terms of logistics and its inefficiency represents the main cause of dissatisfaction within customers. The design of algorithms that help and advise operators to redistribute bicycles accurately and efficiently through the cities is a very important step in terms of sustainability of these systems.

3 Input

Bike-sharing rental process is highly correlated to the environmental and seasonal attributes. Such as; weather conditions, precipitation, day of week, season, hour of the day, etc. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systemsTherefore the dataset about a 2-year usage log of a bike sharing system namely Capital Bike Sharing (CBS) at Washington, D.C., USA which is publicly available at http://capitalbikeshare.com/system-data . The corresponding weather and seasonal information has been extracted from http://www.freemeteo.com and aggregated. There are two datasets containing daily and hourly details. In the daily dataset there are 16 attributes and 731 data records(instances). In the daily dataset there are 17 attributes and 17379 instances. There are some missing weather data for some hours. Both hour.csv and day.csv have the following fields commonly except hr:hour which is only available in hour.csv

- instant: record index
- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month
- hr : hour (0 to 23)

- holiday : weather day is holiday or not
- weekday : day of the week
- workingday : if the day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit : (1:Clear, 2:Misty, 3:Light rain/snow, 4:Heavy rain/snow)
- temp : Normalized temperature in Celsius. (divided to 41 -max)
- atemp: Normalized feeling temperature in Celsius. (divided to 50 max)
- hum: Normalized humidity. (divided to 100 max)
- windspeed: Normalized wind speed. (divided to 67 max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

4 Output

Model
Basically the model predicts the number of total rental bikes using regression algorithms in this project based on previous data.

Output
The output of the model is the count of total rental bikes for newly given instances. That count is predicted according to the given dataset according to environmental and seasonal settings according to hourly and daily .

5 ML Technique

To perform the regression task of predicting hourly and daily bike rental count, following algorithms will be considered as the initial approach.

Multiple Linear Regression attempts to fit a straight hyperplane to the dataset that is closest to all data points. It is most suitable when there are linear relationships between the variables in the dataset. This is quick to compute and can be updated easily with new data. Regularization techniques can be used to prevent overfitting.

Decision trees learn how to best split the dataset into separate branches, allowing it to learn non-linear relationships. Random Forests (RF) and Gradient Boosted Trees (GBT) are two algorithms that build many individual trees, pooling their predictions. As they use a collection of results to make a final decision, they are referred to as “Ensemble techniques”. Here, it is fast to train as there is single decision tree and also robust to noise and missing values.

K-Nearest Neighbors (KNN) makes a prediction for a new observation by searching for the most similar training observations and pooling their values. This algorithm is a simple yet powerful algorithm where no training involved.

6 Dataset

The detailed description and the dataset itself can be found under the following URL:
https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset