Machine Learning Project Proposal 2014 Spring

Capital BikeShare Data Set Analysis and Prediction

Pavan Reddy, Shreya Ghosh, Nianzu Ma, Paul Landes

Thursday, March 20th, 2014

1 Introduction

Bike sharing systems have made renting bikes efficient and quick with memberships, multiple bike locations and easy rental and return process. Through these systems, a user is able to easily rent a bike from a particular station and return back at another station. Currently, there are over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles[7].

1.1 Objectives

The University of California at Irvine, data set has features like date, weather, weekdays, holidays and count of bikes rented. These features of the data make it apt for research such as finding patterns between different features and number of bikes rented.

In this project we will analyze operational data from the bike-sharing systems, which was originally generated from Capital Bike Share and derive bike activity patterns. The problem addressed in this project is finding how many bikes will be rented based on features given in the dataset. We will use concepts like regression, kernel function and classification algorithms to generate these patterns and then compare their accuracies.

1.2 Team

Pavan Reddy CS master student, has experience of machine learning algorithm application and have research in this domain.
Paul Landes CS master student, programming expert, has experience of AI and data visualization.
Shreya Ghosh CS Ph.D. candidate, has experience of data mining and NLP.
Nianzu Ma CS master student, has experience of data mining and NLP.

2 Approach

The Bike Sharing Dataset provides data about the counts of total rental bikes between years 2011 and 2012 for the Capital bikeshare system. The data includes information about the bike usage of both casual and registered users available in two resolutions on the time scale; hourly and daily count of rental bikes. The data also includes the corresponding weather, seasonal, holiday, time information. The data set is a collation of data from multiple sources–Capital Bike Sharing system data[1], Weather information[2] and holiday schedule[3].

We propose to use this data set to predict the counts of rental bikes based on:

- Seasonal holidays
- Weather: wind chill, wind speeds, humidity, precipitation
- Calendar: week day vs. weekend, time of the day and month of the year

Predictions will include not only the usage of rental bikes for casual users and registered users but also the usage of the rental bikes daily and hourly. These predictions can play an important role in traffic, environmental and health issues management. The problem will be treated as both a classification problem and regression problem.

Our approach will make use of supervised machine learning methods, specifically:

1. The first step is data pre-processing: sanitize the available data and handle any missing data.
2. Identify the best parameters for the regression function: after choosing the best parameters we will build regression models using linear regression analysis and support vector machines[6].
3. Post-processing to improve the results using various methods such as regularization to minimize over-fitting the model, which involves estimating the best regularization parameters.
4. Improve results further with ensemble learning methods such as bootstrap aggregating (bagging).

We will transform the bike rental data to discrete classes as the value of class attribute changes dramatically of each record. For example, using the discredited counts of total rental bikes ranges between 0 and 50 in conjunction with supervised learning classification algorithms such as decision tree, naive bayes and logistic regression[4]. Post-processing data improvements will be employed after best feature selection and the building of a classification model. The classification model will then be used to predict the number of rental bikes usage hourly and daily and compare the results we obtained with classification and regression methods.

Java and the Weka[1] technologies for the implementation of the project.

[1] http://www.cs.waikato.ac.nz/ml/weka/

3 Model Evaluation

Since the total record is more than 17,000, it is big enough for training a model. So sample a set of training example from hours.csv for training and the rest of testing. The data will be as 80% training set and 20% test set. To further test the robustness of the classifier, 10-fold cross-validation will be used for evaluation. Root Mean Squared Error and Mean Absolute Error would be used as measure for evaluation of regression model[5]. Accuracy, precision, recall and F-score will be used as the measurement for the evaluation of the classifier.[8]

References

[1] Capital bike share. http://capitalbikeshare.com/system-data/.

[2] Freemeteo: Weather forcasts for the entire planet. http://www.freemeteo.com/.

[3] Holiday schedules. http://dchr.dc.gov/page/holiday-schedule/.

[4] Logistic regression. http://en.wikipedia.org/wiki/Logistic_regression/.

[5] Regression model evaluation. http://www.saedsayad.com/model_evaluation_r.htm/.

[6] Support vector machine. http://en.wikipedia.org/wiki/Support_vector_machine/.

[7] Uci machine learning repository: Bike sharing. http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#/.

[8] Bing Liu. *Web data mining*. Springer, 2007.