Machine Learning Progress Report 2014 Spring

Capital BikeShare Data Set Analysis and Prediction

Pavan Reddy, Shreya Ghosh, Nianzu Ma, Paul Landes

Tuesday, April 8th, 2014

1 Detailed Objectives

The weather condition in our data set has a lot of impact on the usage of bicycling. Specifically, usage can be affected by colder weather, precipitation, and excessive heat. Our research will analyze the effect of weather, working day, holidays and other factors of bike rentals. We are currently building statistical models for this and estimating the number of both kinds of users (registered and casual users).

2 Progress

Our approach includes using the bike rental data to build a multiple linear regression model for both daily and hourly data. We built multiple regression models based on different feature selections. We performed basic data analysis to help us choose the features for the regression model. The weather in Washington DC includes all variations and has an impact on the number of bike rentals for example the weather situation such as rain has a direct consequence on the total number of bike rentals. The impact of seasons also affects the bike rentals. We also see good correlation of the of features such as temperature and "feels like" temperature with the total number of bike rentals. Based on such data analysis we built multiple linear regression model, which takes the general form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where p is the number of variables. Examples of values of x are weather (i.e. rain), temperature, number of rentals, etc. The β values are the learned parameters.

We built similar regression models using different feature selections for the daily dataset.

We performed 3-fold cross validation to evaluate our model. We also measured the residuals to estimate the error in the model.

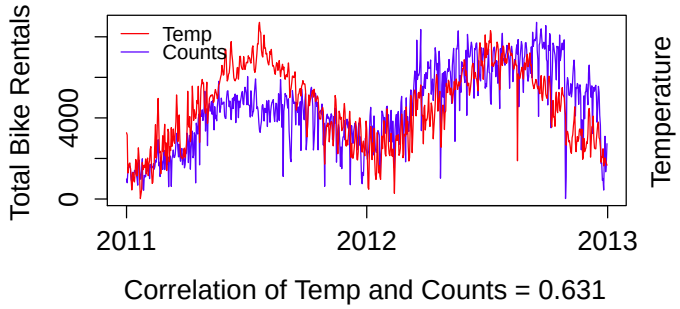

Figure 1: Temperature Bike Counts Correlation

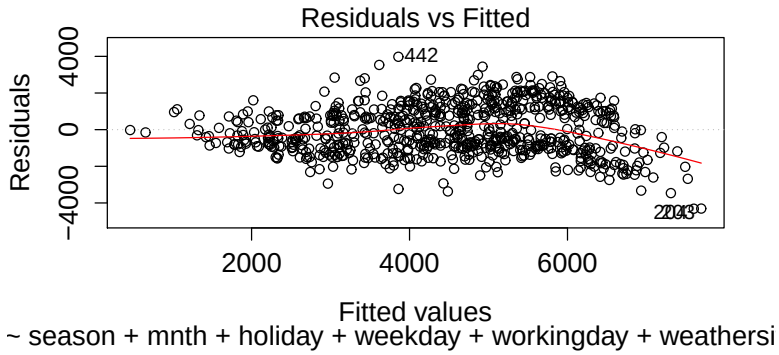

Figure 2: Residuals vs. Fitted Values

We built similar regression models for the hourly dataset but we obtained partial results, we are still working on improving our model based on different feature selections.

We used R[1] to build the regression models and to generate the plots and we used Java for basic data processing.

Action items executed:

- Multiple linear regression for hourly and daily bike rental data.
- Different feature selections for regression model of daily data. (Plot)
- Partial results for hourly bike rental data regression model
- 3-fold cross validation for bike rentals for daily data regression model
- Evaluation based on residuals (Residuals vs fitted plot)
- Implementaion details - R, java for data processing

3 Remaining Work

We are currently trying to improve the regression models by using the following methods - Experiment and discover better feature selection Addition of dummy variables e.g. replacing the 'weather situation' variable by 4 boolean variables 'clear', 'rain', 'snow', 'cloudy' Experiment with squares of significant variables in the regression. Regularization

We are also going to identify the impact of weather conditions, time of day, day of week, holidays etc. on bike rentals by registered users and casual users.

We are going to process the data to discretize the total number of bike rentals and employ classification techniques. We plan to use classification methods such as Naive Bayes classifier and Logistic Regression and compare the results of regression and classification.

3.1 Summary

- Discretize bike rental data
- Use classification models and compare results with regression
- Experiment and improve regression / classification models based on various feature selections.
- Process data - add dummy variables (weather situation variable replaced by rainy, fog, snow, sunny variables)
- Employ regularization
- Impact on registered users and casual users.

References

[1] The r project for statistical computing. http://www.r-project.org/.