ST 563 Final Project - Bike Sharing Data

Team #1

Peter Lung, David Shaw, Yijia Cai

11/4/2021

Introduction

For our project, we are analyzing the bike sharing dataset from the UCI Machine Learning database. The data is the number of daily bike rentals (registered plus "casual") over a two year period. Predictors for this model include variables pertaining to seasonality and weather.

The following is the description of the data from UCI's ML database:

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

The response variable from this dataset is *cnt* which is the total daily count of bike renters. Two other response variables are present in the dataset: *registered* and *casual*. The response we have chosen is *cnt*, which is the sum of the other two.

Our interest is twofold:

(1) Find the best function of the variable set for predicting the response

(2) Evaluate several candidate modeling types for variable selection and data reduction

The predictor variables all pertain to either weather conditions or seasonal effects, some of which can be highly collinear. That makes this data a great set for testing variable selection methods including forward, backward, best subsets, lasso and ridge. It also makes this data a great candidate for testing data reduction methods such as principal components analysis.

We are also interested in testing which variables have nonlinear effects and modeling them with splines. In all tests, we will evaluate the models with cross validation and test each model's predictive power with a consistent holdout sample.

Methods

This section will detail the methodologies and evaluations used for each type of model tested.

Exploratory Data Analysis

Lasso Regression

Variable Selection Methods

Splines to Capture Nonlinear Effects

Model Selection

Ridge Regression

Variable Selection Methods

Splines to Capture Nonlinear Effects

Model Selection

Principal Components Analysis

PCA and Component Selection

Splines to Capture Nonlinear Effects of the Components

Model Selection

Conclusions

Performance on the Holdout Dataset

Commentary on Model Performance

Appendix

References