Faculty of Mathematics and Informatics

Semiparametric Regression

Project Proposal

Bike Rentals

Authors:
Diana Gogołowicz
Michał Makowski

Index number:
258512
258544

May 24, 2017

Introduction

In our proposal we would like to describe the data we choose for **Semiparametric regression** final project. Course is conducted by Prof. J. Harężlak. Our goal is to explore the data and build a model which will help us answer several questions which we will propose in the latter part of this proposal.

Data Set

The data we choose came from Kaggle.com, they are available for over 2 years right now. Data concerns Washington bike sharing system, which is the mean of renting bicycles where the process of rental and bike return is automated via a network of kiosk locations throughout a city.
We are going to analyse hourly bike rental data combined with weather data observed during years 2011 -2012.
Our data consists od 10886 observations of 14 variables:

- datetime - hourly date + timestamp
- season - categorical variable (1 - spring, 2 - summer, 3 - fall, 4 - winter)
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather - categorical variable with 4 levels:
  - 1 - Clear, Few clouds, Partly cloudy
  - 2 - Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3 - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4 - Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp - temperature in Celsius
- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

Goal

We want to build model to predict bike rental demand for each hour and anwser questions:

- which weather parameters encourage users and which discourage.
- how relationships between number of rentals and predictors differs between holiday and working day
- how relationships between number of rentals and predictors differs between registered and unregistered users.

Methods

We plan to use General Additive Mixed Model to explore association between predictors and response. We want to take into account interactions between categorical variables.
Set is divided into training set is comprised of the first 19 days of each month and the test set is the 20th to the end of the month. This will allow to test accuracy of our model.