

# Advanced Computational Econometric: Machine Learning Final Exam

August 19, 2019

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

The bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc., can affect rental behaviors. The dataset `bike_sharing_day.csv` is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA. It contains aggregated daily data with the corresponding weather and seasonal information.

The data file contains 16 columns with the following variables:

- `instant`: record index
- `dteday`: date
- `season`: season (1: spring, 2: summer, 3: fall, 4: winter)
- `yr`: year (0: 2011, 1:2012)
- `mnth`: month (1 to 12)
- `holiday`: weather day is holiday or not (extracted from <http://dchr.dc.gov/page/holiday-schedule>)
- `weekday`: day of the week
- `workingday`: if day is neither weekend nor holiday is 1, otherwise is 0.

- **weathersit:**
  - 1: Clear, Few clouds, Partly cloudy, 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:** Normalized temperature in Celsius. The values are divided to 41 (max)
- **atemp:** Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- **hum:** Normalized humidity. The values are divided to 100 (max)
- **windspeed:** Normalized wind speed. The values are 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

The task is to predict bike rental count (variable **cnt**) based on the environmental and seasonal settings, using at least four machine learning methods studied in the class.

You have to give two files:

1. Commented R code with name `<your_name>.R`. (Insert comment lines in your code to explain each step of the analysis).
2. A Word or pdf file with name `<your_name>.doc` or `<your_name>.pdf` describing your main results and conclusions.

Send both files to `tdenoeux@utc.fr`.