SY19
TP5: Splines and modèles additifs généralisés

Exercise 1

We consider again the `Boston` dataset from package `MASS`. We wish to predict variable `medv` (median value of owner-occupied homes in \$1000s) as a function of `lstat` (lower status of the population in percent).

1. Estimate the expected value of `medv` as a function of `lstat` using order-p polynomial regression. Represent graphically the data and the estimated regression function for different values of p. Which values of p seem visually suitable ?
2. Determine the optimal value of p by cross-validation.
3. Same questions using natural splines. This time, the coefficient to be determined in the number of degrees of freedom (parameter `df` in function `ns`).
4. Same questions using smoothing splines (function `smooth.spline`). Find the optimal value of coefficient `df` using the leave-one-out, then let this coefficient vary around its optimal value and estimate the cross-validation error using the same folds as in the two previous questions.
5. Represent the regression functions estimated by the three methods on the same graph and compare their cross-validation errors.

Exercise 2

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 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 training set is composed of the data for 2011.

The file `bike_sharing_day.csv` contains 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)
— `season` : weather day is holiday or not
— `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)
— `cnt` : count of total rental bikes

Construct a GAM to explain variable `cnt` from variables `season`, `season`, `workingday`, `weathersit`, `atemp`, `hum` and `windspeed`. Plot and interpret the results.