Bike sharing Data set Analysis

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, user is able to easily rent a bike from a particular position and return back to another position. Currently, there are about over 500 bike-sharing programs around the world which are 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.

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system in Washington, DC with the corresponding weather and seasonal information.

Content

Both **hour.csv** and **day.csv** have the following fields, except *hr* which is not available in day.csv

- **instant:** Record index
- **dteday:** Date
- **season:** Season (1:springer, 2:summer, 3:fall, 4:winter)
- **yr:** Year (0: 2011, 1:2012)
- **mnth:** Month (1 to 12)
- **hr:** Hour (0 to 23)
- **holiday:** weather day is holiday or not (extracted from Holiday Schedule)
- **weekday:** Day of the week
- **workingday:** If day is neither weekend nor holiday is 1, otherwise is 0.
- **weathersit:** (extracted from Freemeteo)
- 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 derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- **atemp:** Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
- **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

1. Examine your data

Data : https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Download the and examine your data :

- How do the temperatures change across the seasons? What are the mean and median temperatures?
- Is there a correlation between the temp/atemp/mean.temp.atemp and the total count of bike rentals?
- What are the mean temperature, humidity, windspeed and total rentals per months?
- Is temperature associated with bike rentals (registered vs. casual)?

In the following, we you build a predictive model ff the number of bike sharing by day (daily variable names __cnt_)

- Plot the *cn*t vs *dteday* and examine its patterns and irregularities
- Clean up any outliers or missing values if needed
- Smooth your time series and compare with the original
- ….

2. Now you will be using the smoothed version of cnt: choose the smoothing method and justify your choice.

    - Add the right frequency to your smoothed time series et justify your choices
    - What could you tell about this new time series in term of stationarity and seasonality? Justify your conclusions.

3. Could you model the smoothed time series using ARIMA model:
   - What are the candidate model
   - Choose your model and justify your choice

4. Forecasting with ARIMA Models
   I. Fit an ARIMA model on de-seasonal cnt (remove the season of cnt before fitting the model)
      - What are the candidate models? What is your best model? Justify your choices
      - What is your conclusion?

   II. Fit an ARIMA with Auto-ARIMA
      - Use **auto.arima()** function to fit an ARIMA model of de-seasonal cnt
      - Check residuals, What are your conclusions?

   III. Evaluate and iterate
      - If there are visible patterns or bias, plot ACF/PACF. Are any additional order parameters needed?
      - Refit model if needed. Compare model errors and fit criteria such as AIC or BIC.
      - Calculate forecast using the chosen model
      - plot both the original and the forecasted time series
      -

   IV. Forecasting

   Split the data into training and test times series (test starting at observation 700, use function **window**)
      - fit an Arima model, manually and with Auto-Arima on the training part
      - forecast the next 25 observation and plot the original ts and the forecasted one.
      - What do you observe?