Machine Learning

Efrei Paris
Big Data & Machine Learning
M1

2020/2021

Machine Learning

Salim NAHLE

Organization:

- You can work on any Python environment
- 2 **PDF /HTML report files** are expected. The first one is your work on the Power plant dataset and the second one is this mini-project on the Bike rental dataset.
- They shall contain the code (executed, explanations and necessary screenshots). You can simply print your notebooks into a PDF/HTML files.
- Please work in **pairs**! Each group (composed of 2 persons at most) shall submit one report. Do not forget to indicate your names in the report. The same pairs shall be maintained all the semester.
- The report shall be uploaded on the Moodle's page before **Tuesday 01/12/2020 at 11:45 pm.**
- Late reports are penalized (2 points/20 per day).

Abstract:

- The objective of this mini-project is to build a predictive model by implementing gradient descent for linear regression, compare your model to models obtained by Normal equation and scikit learn's. Then improve the model by doing feature engineering.
- An open data set is provided. The correct answers are given. Supervised learning algorithms are thus used.
- In the data set, the output is continuous, you shall build several regression models, tune them and compare them

Bike Rental Data Set from UCI Machine Learning Repository

1. Citations

Reconsider the Bike Rental data set and the provided notebook.

Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelber

2. Attributes on original data

- 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 [Web Link])
- 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 derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (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)

3. URL:

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

4. Consulting Project

You have been contacted to build a predictive model to help Bike Rental companies in predicting the hourly and daily demand on bikes.
For this reason, you have to start by building 3 models and comparing them:

- First, by using your implementation-from-scratch of linear regression with gradient descent
- Second, by using your implementation of the closed form solution (normal equation)
- Third, by using Scikit Learn library
- For each model display meanAbsoluteError and r2.
- Improve any of the models by tuning the different hyperparameters.
- Try to get some insights from the results you obtained:
    - Display, for instance, the average real demand versus the average predicted demand and the standard deviation of both by grouping your data by:
        - hour
        - season
        - other features that you think useful
- Add dummy variables/Do feature engineering to improve the accuracy of the selected model.
- Optional: Try other machine learning algorithms and compare.