TASK 1 – BIKE SHARING (5%)

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automated. Through these systems, users are able to easily rent a bike from a particular station and to return the bike at another station. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues. In 2014, there were close to 1 million bikes deployed in bike sharing systems across the globe.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for research, as 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.

In this first task, you will carry out basic analysis on a dataset gathered from such a bike sharing system, specifically, the bike sharing system of Washington DC. A short video explaining this system can be found here, and you can find more general information about bike-sharing systems here.

Download the dataset here.

You will find two different sets, a training set, and a test set. To obtain comparable results among all participants, please make sure that you use the training set to develop your algorithm (see instructions below) and use the test set to evaluate your algorithm.

Your task here is to:

1. **Analyze the dataset to obtain helpful information about the data.** For instance, you should analyze the data on how seasonal trends, weekday/weekend patterns, etc., affect the number of rides taken within the system. Also, have a close look at outliers and try to explain why these outliers appear (you can use external information like, e.g. public holiday calenders, wikipedia events for a certain date, etc., to support your explanation).
2. **Based on your analysis, you should build a ML model that predicts the number of rides on a given input day.** Use the results of your analysis (or more advanced ML techniques) to determine important features, engineer new features with higher predictive power, etc. For instance, the weathersit feature of the dataset might not be the most accurate description of the weather on a day as it only provides four very coarse categories.
3. **Please submit a short written report that visualizes your most interesting findings from the analysis (those which have impacted your model design) and illustrates the performance of your model.** Your grade in this task is determined by the clarity of your analysis, how you link your analysis to your model, and finally the mean squared error (MSE) of your model for the test set (the lower the error, the more points you will obtain).

Note that in this task the primary goal is to see whether or not you are fit for the course. To pass this exercise, your analysis and model do not need to be overly complex, and in fact, simple models might be able to deliver very good predictions for this dataset, while additionally preventing overfitting. If you have trouble with this task, please consider deepening your knowledge in data science / machine learning before taking the course in the next edition.

This dataset was published in: Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013)