Machine Learning

Lab 02: NumPy and Pandas

The goal of this lab exercise is to work with NumPy and Pandas to do some preliminary data analysis that is typically required before implementing any machine learning algorithms.

Task 1.

In this task you will perform a basic analysis of a bike sharing dataset available at https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. Load the CSV file "day.csv" using Pandas.

A. Compare the average number of casual rentals and registered rentals depending on if it is a holiday or not. What can you observe?
B. The temperature values are already normalised for classification. What is the minimum and maximum temperature in the data set in Celsius?
C. Usually there are more registered than casual renters. On which days in the data set is this not the case?
D. Plot the temperatures against the number of casual and registered rentals. What can you observe?
E. Create a 3d plot, which plots a point with coordinates "temp", "hum", and "windspeed" for each day. Now divide the dataset into two equal sized subsets, representing days with few casual rentals and days with a lot of casual rentals (Hint: use np.median()). Colour the dots in your plot corresponding to busy days red and the dots corresponding to non-busy days green.

Task 2.

In this task you will analyse the Titanic passenger dataset, which you can download from Canvas. Load the CSV file "titanic.csv" using Pandas.

A. How many passengers were on the titanic, and what percentage survived?
B. Determine the survival rate for male and female passengers. What can you observe?
C. What is the average fare paid by survivors compared to non-survivors?
D. Create a file "titanic_short.csv" containing only the name and age of all surviving passengers, who boarded the Titanic in Queenstown and whose age has been recorded in the data (this were only 8 passengers).