**HOMEWORK 1**

1. Load the "boston house prices" dataset from the boston\_house\_prices.csv file using the pd.read\_csv function (you don't need index\_columnhere). You can find a description of this dataset in the boston\_house\_prices.txt file.

This is a regression dataset with "MEDV" the median house value in a block in thousand dollars the target. How many features are there and how many samples?

Split the data into a training and a test set for learning. Optionally you can plot MEDV vs any of the features using the plot method of the dataframe (using kind="scatter").

1. Load the iris dataset from the sklearn.datasets module using the load\_iris function.

Split it into training and test set using train\_test\_split. Then train and evaluate the following models on the iris dataset:

sklearn.neighbors.KNeighborsClassifier sklearn.linear\_model.LogisticRegression

How do these perform on the training set vs the test set? Which one is the best on the training set, which one is the best on the test set?

1. Load the dataset bike\_day\_raw.csv, which has the regression target cnt. This dataset is hourly bike rentals in the citybike platform. The cnt column is the number of rentals, which we want to predict from date and weather data.

Split the data into a training and a test set using train\_test\_split.

Use the LinearRegression class to learn a regression model on this data.

You can evaluate with the score method, which provides the or using the mean\_squared\_error function from sklearn.metrics (Challenge: You can also write it yourself in numpy).

1. Load the diabetes dataset using sklearn.datasets.load\_diabetes. Apply LinearRegression, Ridge and Lasso and visualize the coefficients. Try polynomial features.