2 Programming (50 points)

2.1. (Polynomial regression, 20 points) In this exercise, we will try to fit a non-linear function g with polynomial regression on the feasible space $\mathbf{X} = [0, 11]$:

$$\text{Unknown} \quad g(x) = ?$$

$$\text{Construct} \quad f(x) = \sum_{i=0}^{n} \alpha_i x^i \quad \Longleftrightarrow \quad f(x) = w^T x', \quad x' = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^n \end{bmatrix}, \quad s.t. \quad \forall x \in \mathbf{X}, \quad f(x) \approx g(x)$$

Where n is the polynomial degree of freedom and is manually chosen.
Follow the instructions given in the jupyter notebook. At the end of the exercise, you will retrieve an estimation of the desired function and make some comment on this method.

2.2. (Linear regression, 30 points) The CSV or XLS file contains a dataset for regression. There are 7750 samples with 25 features (described in the doc file). This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea.

This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. Hindcast validation was conducted for the period from 2015 to 2017.

You need to delete the first two attributes (station and date), and use attributes 3-23 to predict attributes 24 and 25. Randomly split the data into two parts, one contains 80% of the samples and the other contains 20% of the samples. Use the first part as training data and train a linear regression model and make prediction on the second part. Report the training error and testing error in terms of RMSE.

Repeat the splitting, training, and testing for 10 times. Use a loop and print the RMSEs in each trial.

Note that you need to write the codes of learning the parameters by yourself. Do not use the classification or regression packages of Sklearn. You can use their tools to shuffle the data randomly for splitting.