Homework 2

Climate Change

There have been many studies documenting that the average global temperature has been increasing over the last century. The consequences of a continued rise in global temperature will be dire. Rising sea levels and an increased frequency of extreme weather events will affect billions of people.

In this problem, you will attempt to study the relationship between average global temperature and several other factors. The file climate_change_1.csv contains climate data from May 1983 to December 2008. The available variables include:

- *Year*: the observation year.
- *Month*: the observation month.
- *Temp*: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.
- *CO2*, *N2O*, *CH4*, *CFC.11*, *CFC.12*: atmospheric concentrations of carbon dioxide (CO_2), nitrous oxide (N_2O), methane (CH_4), trichlorofluoromethane (CCl_3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl_2F_2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division.
- CO2, N2O and CH4 are expressed in ppmv (parts per million by volume -- i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere)
- CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).
- *Aerosols*: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.
- *TSI*: the total solar irradiance (TSI) in W/m^2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.
- *MEI*: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

Problem 1 —Creating Your First Model

We are interested in how changes in these variables affect future temperatures, as well as how well these variables explain temperature changes so far. To do this, first read the dataset climate_change_1.csv into Python or Matlab.

Then, split the data into a training set, consisting of all the observations up to and including 2006, and a testing set consisting of the remaining years. A training set refers to the data that will be used to build the model, and a testing set refers to the data we will use to test our predictive ability.

After seeing the problem, your classmate Alice immediately argues that we can apply a linear regression model. Though being a little doubtful, you decide to have a try. To solve the linear regression problem, you recall the linear regression has a closed form solution:

$$\boldsymbol{\theta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

1. **Implement a function *closed_form_1* that computes this closed form solution given the features X, labels Y (using Python or Matlab).**

2. **Write down the mathematical formula for the linear model and evaluate the model R^2 on the training set and the testing set.**

3. **Which variables are significant in the model?**

4. **Write down the necessary conditions for using the closed form solution. And you can apply it to the dataset *climate_change_2.csv*, explain the solution is unreasonable.**

Problem 2 — Regularization

Regularization is a method to boost robustness of model, including L$_1$ regularization and L$_2$ regularization.

1. **Please write down the loss function for linear model with L$_1$ regularization, L$_2$ regularization, respectively.**

2. **The closed form solution for linear model with L$_2$ regularization:**

$$\boldsymbol{\theta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

where I is the identity matrix. Write a function *closed_form_2* that computes this closed form solution given the features X, labels Y and the regularization parameter λ (using Python or Matlab).

3. **Compare the two solutions in problem 1 and problem 2 and explain the reason why linear model with L_2 regularization is robust. (using *climate_change_1.csv*)**

4. **You can change the regularization parameter λ to get different solutions for this problem. Suppose we set λ = 10, 1, 0.1, 0.01, 0.001, and please evaluate the model R^2 on the training set and the testing set. Finally, please decide the best regularization parameter λ. (Note that: As a qualified data analyst, you must know how to choose model parameters, please learn about cross validation methods.)**

Problem 3 — Feature Selection

1. **From Problem 1, you can know which variables are significant, therefore you can use less variables to train model. For example, remove highly correlated and redundant features. You can propose a workflow to select feature.**

2. **Train a better model than the model in Problem 2.**

Problem 4 — Gradient Descent

Gradient descent algorithm is an iterative process that takes us to the minimum of a function. Please write down the iterative expression for updating the solution of linear model and implement it using Python or Matlab in *gradientDescent* function.