LINEAR REGRESSION

Scientific Computing II
Fundación Universitaria Konrad Lorenz
February 17, 2024

You are expected to communicate accurately the solution of the exercises in this assignment. Full score will only be given to correct and completely justified answers. Miraculous, obtuse and unnecessarily complex solutions will receive partial or null score. You can ask any question of this assignment during class or through email: julian.jimenezc@konradlorenz.edu.co.

The deadline to submit this assigment is **February 24, 2024, 8.15am.**

This assignment aims to guide you to understand and apply the "first machine learning model", Linear Regression.

Note: All one-dimensional vectors are assumed to be column vectors.

Suppose you have m data vectors of the form $(\boldsymbol{x_i}, y_i)$, where $\boldsymbol{x_i} \in \mathbb{R}^k$ is a vector of descriptive variables that determine the dependent variable y_i. We aim to fit a linear model that best adjusts the data. For this, we define the matrix $\boldsymbol{X} \in \mathbb{R}^{m\times(k+1)}$ to have the vectors of descriptive variables $\boldsymbol{x_i}$ as rows, and appending a column of ones to account for the bias term. The linear model is then

$$y(\boldsymbol{x}) = \boldsymbol{c}^T\boldsymbol{x}, \tag{1}$$

with $\boldsymbol{c} \in \mathbb{R}^{k+1}$ being the vector of parameters of the model. To get the best choice of parameters $\boldsymbol{c}$ for our data, we must minimize the difference between the predicted and expected values. For this, we define the cost function

$$F : \mathbb{R}^{k+1} \to \mathbb{R}_{\geq 0},\ \boldsymbol{c} \mapsto (\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y}), \tag{2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)$ is the vector of descriptive variables.

1. (0/5) Show that the cost function can be written as

$$F(\boldsymbol{c}) = \boldsymbol{c}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y}^T\boldsymbol{y}.$$

2. (0/10) Show that the gradient of F is

$$\nabla_{\boldsymbol{c}} F = 2\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} - \boldsymbol{X}^T\boldsymbol{y}\right), \tag{3}$$

so that any equilibrium point should satisfy

$$\boldsymbol{c} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

3. (0/10) Show that F is a convex function by finding its Hessian matrix and proving that it is positive-(semi)definite. *Hint: Calculate the Hessian matrix* $\boldsymbol{H}$ *and show that* $\boldsymbol{\xi}^T\boldsymbol{H}\boldsymbol{\xi} \geq 0$ *for any* $\boldsymbol{\xi} \in \mathbb{R}^{k+1}$. Conclude that the solution of (3) is a global minimum.

4. (0/20) Here you can find the data related to the cost in million dollars of certain terrains as a function of its area (in acres). Estimate the best linear model that fits this data. Plot the linear model and the data.

Note: The data consists of 3 columns, the first being the column of ones (bias), the second one corresponds to the descriptive variable and the third one corresponds to the dependent variable.

5. (0/5) Use the model of the previous item to estimate the price of a land with $\{25, 30, 35, 40\}$ acres.