ITESO, Universidad Jesuita de Guadalajara

Maestría de Ciencia de Datos

Optimización Convexa

Tarea 3: Regresión lineal múltiple

Estudiante: Daniel Nuño
Profesor: Dr. Juan Diego Sanchez Torres
Fecha entrega: 21 de febrero de 2022

Introduction

In most of the real-life problems related to data analysis, when the objective is to explain or predict a given variable, several other explanatory variables are often used. It is pretty common to propose input-output models where the explanatory variables are the inputs and the variable to explain or predict is the output. For this modeling process, the first hypothesis is the output variable admits a representation as a linear combination of the input variables. Since in real-world data, the output does not belong to the input-set span, it is necessary to perform a sort of approximation method. As a solution, the approximate output or model's prediction is defined as an output's projection on the set spanned by the inputs. This straightforward reasoning leads to multiple linear regression, a case of linear regression with multiple input variables.

Problema 1

First, let start with a calculus refresher.

Problem 1: Warm Up

Solve the following calculus exercises:

1. For $f_1(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$, calculate the gradient $\frac{df_1}{dx} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \end{bmatrix}$. Similarly, for $f_2(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin(t)$ and $x_2 = \cos(t)$, calculate the gradient $\frac{df_2}{dt} = \begin{bmatrix} \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix}$.
2. For $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, and $\boldsymbol{x} \in \mathbb{R}^N$, calculate the gradient $\frac{d\boldsymbol{f}}{d\boldsymbol{x}}$.
3. Consider the function $h : \mathbb{R} \to \mathbb{R}, h(t) = (f \circ g)(t)$ with

$$f : \mathbb{R}^2 \to \mathbb{R}$$
$$g : \mathbb{R} \to \mathbb{R}^2$$
$$f(\boldsymbol{x}) = \exp\left(x_1 x_2^2\right)$$
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}$$

and compute the gradient of h with respect to t

4. Let us consider the linear model $\boldsymbol{Y} = \boldsymbol{X\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter vector, $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ are input features and $\boldsymbol{Y} \in \mathbb{R}^N$ are the corresponding observations. Also, let the functions $L(\boldsymbol{e}) = \|\boldsymbol{e}\|_2^2$ and $\boldsymbol{e}(\boldsymbol{\theta}) = \boldsymbol{Y} - \boldsymbol{X\theta}$. Calculate critical point of $L(\boldsymbol{e})$, that is the solution to $\frac{\partial L}{\partial \boldsymbol{\theta}} = 0$.

Daniel Nuño

HW 3.1: Regresion lineal múltiple

Problema 1: Calentamiento. Resuelve los siguientes ejercicios de Cálculo.

1 Para $f_1(X_1, X_2) = X_1^2 X_2 + X_1 X_2^3 \in \mathbb{R}$, calcula el gradiente.

$$\frac{\partial f_1}{\partial X_1} = 2X_1 X_2 + X_2^3$$

$$\frac{\partial f_1}{\partial X_2} = X^2 + 3X_1X_2^2$$

$$\nabla f(X_1,X_2) = \begin{bmatrix} \frac{\partial f_1}{\partial X_1} \\ \frac{\partial f_2}{\partial X_2} \end{bmatrix}$$

Para $f_2(X_1, X_2) = X_1^2 + 2X_2$, donde $X_1 = \sin(t)$, $X_2 = \cos(t)$ calcula el gradiente.

$$\frac{\partial f_2}{\partial t} = \begin{bmatrix} \frac{\partial f_2}{\partial X_1} & \frac{\partial f_2}{\partial X_2} \end{bmatrix} \begin{bmatrix} \frac{\partial X_1(t)}{\partial t} \\ \frac{\partial X_2(t)}{\partial t} \end{bmatrix}$$

$$\frac{\partial f_2}{\partial X_1} = 2X_1 \quad \frac{\partial X_1}{\partial t} = \cos(t)$$

$$\frac{\partial f_2}{\partial X_2} = 2 \quad \frac{\partial X_2}{\partial t} = -\sin(t)$$

$$\nabla \frac{\partial f_2}{\partial t} = [2X_1 \ \ 2] \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix}$$

2 Para $f(X) = AX$, $f(X) \in \mathbb{R}^M$, $A \in \mathbb{R}^{M \times N}$, and $X \in \mathbb{R}^N$, calcula el gradiente

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ A_{M1} & \cdots & & A_{MN} \end{bmatrix}_{M\times N} \qquad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$$

$$N \times 1 = M \times 1$$

$$AX = \begin{bmatrix} A_{11}X_1 + A_{12}X_2 + \dots + A_{1N}X_N \\ \vdots \\ A_{M1}X_1 + A_{M2}X_2 + \cdots + A_{MN}X_N \end{bmatrix} \begin{matrix} f_1(X_1, X_2, \dots, X_N) \\ \vdots \\ f_M(X_1, X_2, \dots, X_N) \end{matrix}$$

$$\nabla \frac{\partial F}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \frac{\partial f_2}{\partial x_1} & & & \frac{\partial f_2}{\partial x_N} \\ \vdots & & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \frac{\partial f_M}{\partial x_2} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix}$$

3. Considera la funcion $h: \mathbb{R} \to \mathbb{R}, h(t) = (f \circ g)(t)$ con $f: \mathbb{R}^2 \to \mathbb{R}$, $g: \mathbb{R} \to \mathbb{R}^2$

$$F(x) = \exp(x_1 x_2^2)$$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t \sin t \end{bmatrix}$$ Calcula el gradiente de h con respecto t

$$\frac{\partial f}{\partial x_1} = x_2^2 e^{x_1 x_2^2} \qquad \frac{\partial x_1}{\partial t} = \cos(t) - t\sin(t)$$

$$\frac{\partial f}{\partial x_2} = 2x_1 x_2 e^{x_1 x_2^2} \qquad \frac{\partial x_2}{\partial t} = \sin(t) + t\cos(t)$$

$$\frac{\partial f}{\partial t} = \begin{bmatrix} x_2^2 e^{x_1 x_2^2} & 2x_1 x_2 e^{x_1 x_2^2} \end{bmatrix} \begin{bmatrix} \cos(t) - t\sin(t) \\ \sin(t) - t\cos(t) \end{bmatrix}$$

$x_1 = t\cos(t)$

$x_2 = t\sin(t)$

4. Considera el modelo lineal $Y = X\theta$, sea $\theta \in \mathbb{R}^D$ es un vector parametros. $X \in \mathbb{R}^{N \times D}$ son entradas y $Y \in \mathbb{R}^N$ son las observaciones correspondientes. Tambien, sea las funciones

$L(e) = \|e\|_2^2$ ~~ay~~ $e(\theta) = y - X\theta$. Calcula el punto critico de $L(e)$, que es solución a $\frac{\partial L}{\partial \theta} = 0$

$$\frac{\partial L(e)}{\partial \theta} = 2X^T(y - X\theta) = 0$$

$$= 2X^T Y - 2X^T X\theta = 0$$

$$= 2X^T Y = 2X^T X\theta$$

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

Propiedad de la norma 2

$$\|e\|_2^2 = e^T e$$

$$\frac{d\|e\|_2^2}{de} = 2e$$

Problema 2

The following case considers the application of the ML and MAP approaches to the multiple regression.

Problem 2: Multiple Linear Regression

Consider a standard linear regression problem, in which for $i = 1, \ldots, n$ the mean of the conditional distribution of y_i is specified given a $k \times 1$ predictor vector $\mathbf{x}_i$ of the form $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta} + \varepsilon_i$ where $\boldsymbol{\theta}$ is a $k \times 1$ vector, and the ε_i are independent and identically normally distributed random variables $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

1. Show that the likelihood function for this problem is

$$L\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \sigma^2\right) \propto \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})\right)$$

   and find the parameter vector $\boldsymbol{\theta}$ that minimizes L.

2. Propose a prior normal distribution for $\boldsymbol{\theta}$, find the posterior distribution for $\boldsymbol{\theta}$.

3. Propose a prior Laplace distribution for $\boldsymbol{\theta}$, find the posterior distribution for $\boldsymbol{\theta}$.

4. (5 pt) Revisit this problem using the least-squares approach, exposing in a very detailed way the conditions in such the probabilistic formulation is equivalent to the Tikhonov (RIDGE) regularization and the LASSO regularization. Finally, explain with an example how the LASSO regularization can lead to a sparse solution for the regression problem.

Problema 2. Regresion lineal Multiple

Considero una regresion lineal estandar, en cual para $i = 1, 2, \ldots, n$ la media de la distribucion de probabilidad condicional de $y_i = x_i^T \theta + \epsilon_i$ sea θ un vector $k \times 1$, y la ϵ_i son independientes e identicamente distribuidos. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

1.- Muestra que la ~~probabilidad~~ funcion de maxima verosimilitud es

$$L(Y \mid X, \theta, \sigma^2) \propto (\sigma^2)^{-n/n} \exp\left(-\frac{1}{2\sigma^2}(Y - X\theta)^T (Y - X\theta)\right)$$

y encuentre un parametro que ~~minimice~~ maximize L.

$L(Y \mid \theta X, \sigma)$ puede ser escrito como $\propto \left(\frac{1}{\sqrt{\sigma}}\right) \exp\left[\frac{1}{2\sigma^2} \|Y - X\theta\|_2^2\right]$

$$\ln(L(Y \mid \theta X, \sigma)) = -\frac{1}{2}\ln(\sigma) - \frac{1}{2\sigma^2}\|Y - X\theta\|_2^2$$

$$\frac{\partial \ln(L(Y \mid X\theta, \sigma))}{\partial \theta} = \frac{1}{2\sigma^2} \cdot 2X^T(Y - X\theta) = 0$$

$$X^T Y = X^T X \theta$$

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

$$\frac{\partial^2 \ln(L)}{\partial \theta^2} \quad \frac{1}{\sigma^2}(X^T Y - X^T X \theta) = \frac{-X^T X}{\sigma^2}$$

2. Proponga una distribucion normal priori para θ, encuentre la distribucion posteriori para θ.

Una distribucion normal $\theta \sim \mathcal{N}(0, \sigma^2)$ donde $e = Y - X\theta$ y $\theta \sim \mathcal{N}(a, \frac{1}{\lambda^2})$ obtenemos

$$P(e) = \frac{1}{(2\pi)^{n/2} \cdot \sigma^n} \exp\left[-\frac{1}{2\sigma^2}\|e\|_2^2\right]$$

$$P(\theta) = \frac{\lambda^m}{(2\pi)^{m/2}} \exp\left[\frac{-\lambda^2}{2}\|\theta - A\|_2^2\right]$$

$$P(\theta \mid D) \propto \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[\frac{-1}{2\sigma^2}\|e\|_2^2\right] \cdot \frac{\lambda^m}{(2\pi)^{m/2}} \exp\left[\frac{-\lambda^2}{2}\|\theta - A\|_2^2\right]$$

3. proponga una distibucion de laplace priori para θ, encuentre la distribucion posteriori para θ.