

867002105_HW1

Charles Dotson

September 3, 2022

Contents

Problem 1	2
Problem 2	3
Problem 3	4
Problem 4	5
Problem 5	7

Problem 1

Let the training examples be $x^{(1)} = (1, 0.5, 3)^T$, $x^{(2)} = (1, 0.2, -1)^T$, $y^{(1)} = 5$, $y^{(2)} = 2$. Write the MSE loss function $J(\theta)$ with the parameters θ and the above training data.

We will start by introducing the our cost function $J(\theta)$ where:

$$h(x; \theta) = \langle \theta^T x \rangle$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (h(x^{(i)}; \theta) - y^{(i)})^2$$

Introducing our vectors:

$$J(\theta) = \frac{1}{2} \left[\frac{1}{2} (\theta^T x^{(1)} - 5)^2 + \frac{1}{2} (\theta^T x^{(2)} - 2)^2 \right]$$

$$= \frac{1}{4} \left[\left(\theta_0 + \frac{1}{2} \theta_1 + 3\theta_2 - 5 \right)^2 + \left(\theta_0 + \frac{1}{5} \theta_1 - \theta_2 - 2 \right)^2 \right]$$

$$J(\theta) = \frac{1}{2} \theta_0^2 + \frac{7}{20} \theta_0 \theta_1 + \theta_0 \theta_2 + \frac{29}{400} \theta_1^2 + \frac{13}{20} \theta_1 \theta_2 + \frac{5}{2} \theta_2^2 - \frac{7}{20} \theta_0 - \frac{29}{20} \theta_1 - \frac{13}{2} \theta_2 + \frac{29}{40}$$

Problem 2

Derive the gradient (a 3 dimensional column vector) of $J(\theta)$ with respect to θ . Need to derive the partial derivatives of $J(\theta)$ with respect to each element of θ explicitly. Then write the gradient vector in the form of a matrix-vector product plus a constant vector. We called this the “vectorized” gradient that will be useful for gradient descent.

$$\frac{\partial J(\theta)}{\partial \theta_i} = \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \frac{\partial J(\theta)}{\partial \theta_1} \\ \frac{\partial J(\theta)}{\partial \theta_2} \end{bmatrix}$$

Where,

$$J(\theta) = \frac{1}{2}\theta_0^2 + \frac{7}{20}\theta_0\theta_1 + \theta_0\theta_2 + \frac{29}{400}\theta_1^2 + \frac{13}{20}\theta_1\theta_2 + \frac{5}{2}\theta_2^2 - \frac{7}{20}\theta_0 - \frac{29}{20}\theta_1 - \frac{13}{2}\theta_2 + \frac{29}{40}$$

Thus our columns vector is

$$\frac{\partial J(\theta)}{\partial \theta_i} = \begin{bmatrix} \theta_0 + \frac{7}{20}\theta_1 + \theta_2 - \frac{7}{20} \\ \frac{7}{20}\theta_0 + \frac{29}{200}\theta_1 + \frac{13}{20}\theta_2 - \frac{29}{20} \\ \theta_0 + \frac{13}{20}\theta_1 + 5\theta_2 - \frac{13}{2} \end{bmatrix}$$

Now we can see that our Gradient Vector = $A\theta - b$ where,

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & \frac{7}{20} & 1 \\ \frac{7}{20} & \frac{29}{200} & \frac{13}{20} \\ 1 & \frac{13}{20} & 5 \end{bmatrix}$$

$$b = \begin{bmatrix} \frac{7}{20} \\ \frac{29}{20} \\ \frac{13}{2} \end{bmatrix}$$

Problem 3

Use matrix calculus to derive the gradient of $J(\theta)$ with respect to θ . The gradient must use the design matrix with the i -th rows being $(x^{(i)})$, $i = 1, 2$, and the target value vector $y = [y^{(1)}, y^{(2)}]$. Please don't derive it based on partial derivatives as in the above question.

We start by writing our $J(\theta)$ in vector notation:

$$J(\theta) = \frac{1}{2m}(X\theta - y)^T(X\theta - y)$$

where

$$X = \begin{bmatrix} 1 & \frac{1}{2} & 3 \\ 1 & \frac{1}{5} & -1 \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \quad y = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

We also know that our residuals, $r = X\theta - y$, thus:

$$r = X\theta - y = \begin{bmatrix} \theta_0 + \frac{1}{2}\theta_1 + 3\theta_2 - 5 \\ \theta_0 + \frac{1}{5}\theta_1 - \theta_2 - 2 \end{bmatrix}$$

Taking the gradient of $J(\theta)$ w.r.t θ

$$\frac{\partial}{\partial \theta} J(\theta) = \frac{\partial}{\partial \theta} \frac{1}{2m} (X\theta - y)^T (X\theta - y)$$

$$= \frac{1}{4} (2X^T r)$$

$$= \frac{1}{2} (X^T r)$$

$$= \frac{1}{2} \begin{bmatrix} \theta_0 + \frac{1}{2}\theta_1 + 3\theta_2 - 5 \\ \theta_0 + \frac{1}{5}\theta_1 - \theta_2 - 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ \frac{1}{2} & \frac{1}{5} \\ 3 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} \theta_0 + \frac{7}{20}\theta_1 + \theta_2 - \frac{7}{20} \\ \frac{7}{20}\theta_0 + \frac{29}{200}\theta_1 + \frac{13}{20}\theta_2 - \frac{29}{20} \\ \theta_0 + \frac{13}{20}\theta_1 + 5\theta_2 - \frac{13}{2} \end{bmatrix}$$

Problem 4

Use matrix calculus to find the gradient of the gradient $\frac{\partial J(\theta)}{\partial \theta}$. This is the so-called “second-order derivatives” or the Hessian matrix H . Show that this matrix is positive-semidefinite (PSD), defined as for any vector $u \in \mathbb{R}^3$, $u^T H u \geq 0$. Note: the result is a 3-by-3 matrix.

Recalling our gradient from Problem 3:

$$\frac{\partial J(\theta)}{\partial \theta} = 1/2(X^T r)$$

where,

$$r = X\theta - y$$

Thus,

$$\frac{\partial J(\theta)}{\partial \theta} = 1/2(X^T X\theta - y)$$

To find our Hessian matrix we must take the second order derivation of $J(\theta)$ w.r.t θ ,

$$H = \frac{\partial^2 J(\theta)}{\partial \theta^2} = 1/2(X^T X)$$

Plugging in our matrix X

$$H = \frac{\partial^2 J(\theta)}{\partial \theta^2} = 1/2(X^T X)$$
$$= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ \frac{1}{2} & \frac{1}{5} \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & 3 \\ 1 & \frac{1}{5} & -1 \end{bmatrix}$$

We arrive at

$$H = \begin{bmatrix} 1 & \frac{7}{20} & 1 \\ \frac{7}{20} & \frac{29}{200} & \frac{13}{20} \\ 1 & \frac{13}{20} & 5 \end{bmatrix}$$

And we can clearly see that the matrix H is symmetric. Thus to prove PSD we shall show all eigen values are ≥ 0

```
[42]: '''importing numpy'''
import numpy as np
import pandas as pd
from IPython.display import Markdown as md
'''creating the matrix'''
H = np.array([
    [1, 7/20, 1],
    [7/20, 29/200, 13/20],
    [1, 13/20, 5]
])
'''Eigen-Value Calculations'''
lambdas = pd.DataFrame([round(np.linalg.eigh(H)[0][i],4) for i in range(3)],
    index=['$\lambda_{}$'.format(i) for i in range(3)], columns=['Value'])

'''Printing the Value'''
md(lambdas.T.to_markdown())
```

[42]:

	λ_0	λ_1	λ_2
Value	0	0.8108	5.3342

We have proven H is PSD since a square symmetrical matrix is PSD iff all $\lambda \geq 0$

Problem 5

(Graduate only) Stochastic gradient descent samples one training example to calculate the gradient vector. What needs to be changed in the gradient found in Question 3 to find a stochastic gradient? Then find the two gradients with $(x^{(1)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$, respectively.

Question 3 finds the gradient descent using all of the training data. This is what needs to be changed. Using stochastic gradient descent will only use a randomly selected batch of training data to train the algorithm. For our purposes we only have two entries for training data and thus the stochastic gradient descents will use either or to mimic the method in question.

Recall our cost function

$$h(x; \theta) = \langle \theta^T x \rangle$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (h(x^{(i)}; \theta) - y^{(i)})^2$$

Only using $(x^{(1)}, y^{(1)})$, we arrive at:

$$J(\theta) = \frac{1}{4}\theta_0^2 + \frac{1}{4}\theta_0\theta_1 + \frac{3}{2}\theta_0\theta_2 + \frac{1}{16}\theta_1^2 + \frac{3}{4}\theta_1\theta_2 + \frac{9}{4}\theta_2^2 - \frac{5}{2}\theta_0 - \frac{5}{4}\theta_1 - \frac{15}{2}\theta_2 + \frac{25}{4}$$

Calculating the partial derivatives with respect to θ_i vector we get:

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{2}\theta_0 + \frac{1}{4}\theta_1 + \frac{3}{2}\theta_2 - \frac{5}{2}$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{1}{4}\theta_0 + \frac{1}{8}\theta_1 + \frac{3}{4}\theta_2 - \frac{5}{4}$$

$$\frac{\partial J(\theta)}{\partial \theta_2} = \frac{3}{2}\theta_0 + \frac{3}{4}\theta_1 + \frac{9}{2}\theta_2 - \frac{15}{2}$$

Thus our columns vector is

$$\frac{\partial J(\theta)}{\partial \theta_i} = \begin{bmatrix} \frac{1}{2}\theta_0 + \frac{1}{4}\theta_1 + \frac{3}{2}\theta_2 - \frac{5}{2} \\ \frac{1}{4}\theta_0 + \frac{1}{8}\theta_1 + \frac{3}{4}\theta_2 - \frac{5}{4} \\ \frac{3}{2}\theta_0 + \frac{3}{4}\theta_1 + \frac{9}{2}\theta_2 - \frac{15}{2} \end{bmatrix}$$

Recall our cost function

$$h(x; \theta) = \langle \theta^T x \rangle$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (h(x^{(i)}; \theta) - y^{(i)})^2$$

Only using $(x^{(2)}, y^{(2)})$, we arrive at:

$$J(\theta) = \frac{1}{4}\theta_0^2 + \frac{1}{10}\theta_0\theta_1 - \frac{1}{2}\theta_0\theta_2 + \frac{1}{100}\theta_1^2 - \frac{1}{10}\theta_1\theta_2 + \frac{1}{4}\theta_2^2 - \theta_0 - \frac{1}{5}\theta_1 + \theta_2 + 1$$

Calculating the partial derivatives with respect to θ_i vector we get:

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{2}\theta_0 + \frac{1}{10}\theta_1 + \frac{1}{2}\theta_2 - 1$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{1}{10}\theta_0 + \frac{1}{50}\theta_1 - \frac{1}{10}\theta_2 - \frac{1}{5}$$

$$\frac{\partial J(\theta)}{\partial \theta_2} = -\frac{1}{2}\theta_0 - \frac{1}{10}\theta_1 + \frac{1}{2}\theta_2 + 1$$

Thus our columns vector is

$$\frac{\partial J(\theta)}{\partial \theta_i} = \begin{bmatrix} \frac{1}{2}\theta_0 + \frac{1}{10}\theta_1 + \frac{1}{2}\theta_2 - 1 \\ \frac{1}{10}\theta_0 + \frac{1}{50}\theta_1 - \frac{1}{10}\theta_2 - \frac{1}{5} \\ -\frac{1}{2}\theta_0 - \frac{1}{10}\theta_1 + \frac{1}{2}\theta_2 + 1 \end{bmatrix}$$