Tecnológico de Monterrey

Intelligent Systems

SAT Scores (LR, MSE and GD)

Miguel Ángel Marines Olvera | A01705317

PURPOSE

The purpose of this project is to code in python and provide a real life example of the application of the hypothesis function (linear regression), mean square error function and the gradient descent function.

PROBLEM

This project helps predict the score that a student will get in the SAT based on the study hours and the score of the practice test. The SAT is the most important test that students take in the United States when they are finishing high school in order to get into collage.

This problem is base on real data, which is provided by College Board and Prep Institutions.

Study Hours, Practice Test and Test Relation:
https://collegedunia.com/exams/sat/sat-preparation-time

Data Sets:
https://www.kaggle.com/datasets/new-york-city/new-york-city-sat-results
https://nces.ed.gov/programs/digest/d17/tables/dt17_226.40.asp

DATA SET

This project uses an SAT Test dataset.

In order to make it easier to understand, this document only takes into account the first five cases (rows) of the data set.

First 5 Cases:

x_feature_inputs		y_result
Study Hours	**Practice Test**	**Test**
80	1360	1400
40	1250	1270
150	1580	1600
20	1180	1200

HYPOTHESIS FUNCTION

The hypothesis function (linear regression) is what will help the project predict the score that a student will get based on certain parameters (thetas) and the study hours and the practice test score (x_features_inputs).

Hypothesis Function: $y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \ldots + b$

When the code executes the hypothesis function, it only works with one row, which would correspond to the case of one student

x_feature_inputs		y_result
Study Hours	Practice Test	Test
80	1360	1400
40	1250	1270
150	1580	1600
20	1180	1200

The code has a cycle that runs the hypothesis function (linear regression), the number of times according to the number of rows (student cases) in the dataset.

x_feature_inputs		y_result
Study Hours	Practice Test	Test
80	1360	1400
40	1250	1270
150	1580	1600
20	1180	1200

For this Project:

Hypothesis Function:

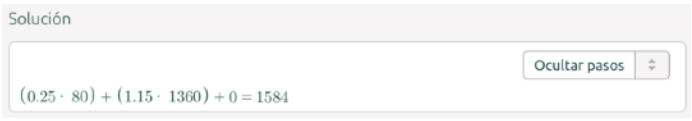
$y = \theta_1 x_1 + \theta_2 x_2 \ldots + b$

Thetas:

$\theta_1 = 0.25$
$\theta_2 = 1.15$
$b = 0$

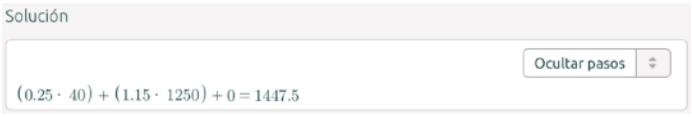
Student (Case) 1:

y = (0.25 * 80) + (1.15 * 1360) + 0
y = 1584



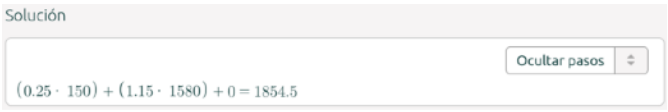
Student (Case) 2:

y = (0.25 * 40) + (1.15 * 1250) + 0
y = 1447.5



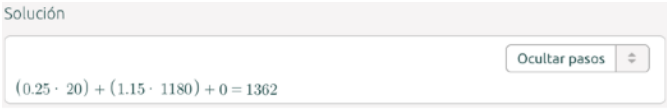
Student (Case) 3:

y = (0.25 * 150) + (1.15 * 1580) + 0
y = 1854.5



Student (Case) 4:

y = (0.25 * 20) + (1.15 * 1180) + 0
y = 1362



<u>Project Run on Terminal</u>:

```
Hypothesis Function:

Student (Case) 1 :  1583.9999999999998

Student (Case) 2 :  1447.5

Student (Case) 3 :  1854.4999999999998

Student (Case) 4 :  1362.0
```

MEAN SQUARE ERROR

The project uses the mean square error to know how concentrated the data is around the line of best fit, in other words, it tells us the mean of how far are the result predictions from the real results.

Mean Square Error: MSE = $1/n * \Sigma(X_1 - Y_2)^2$

y_result (Calculated)	r_result (Real)
Test	**Test**
1584	1400
1447.5	1270
1854.5	1600
1362	1200

For this Project:

Mean Square Error:

$MSE = 1/n * \Sigma(Y_1 - Y_2)^2$

$MSE = ((1584 - 1400)^2 + (1447.5 - 1270)^2 + (1854.5 - 1600)^2 + (1362 - 1200)^2) / 4$

$MSE = 39094.125$

Solución

Ocultar pasos

$$\frac{(1584-1400)^2+(1447.5-1270)^2+(1854.5-1600)^2+(1362-1200)^2}{4}=39094.125$$

Project Run on Terminal:

```
Y Computed and Y Real Results:

Y - Hypothesis Function Results:  1583.9999999999998 1447.5 1854.4999999999998 1362.0

Y - Real Results:  1400 1270 1600 1200

Mean Square Error:  39094.12499999995
```

GRADIENT DESCENT FUNCTION

The gradient descent function is used to correct the parameters (thetas) and obtained better predictions (y_results).

Gradient Descent Function: $\theta j = \theta j - \alpha/m \, \Sigma[(h\theta(Xi) - Y)Xi]$

x_feature_inputs		y_result
Study Hours	**Practice Test**	**Test**
80	1360	1400
40	1250	1270
150	1580	1600
20	1180	1200

Parameters:	θ_1	θ_2	**b**	**α**
Old	0.25	1.15	0	0.1
New	-1580.625	-26633.475	-19.45	0.1

θ_1

Solución

Mostrar pasos

$$0.25 - \frac{0.1}{4}\left(\left(\left(\left(0.25 \cdot\ 80\right) + \left(1.15 \cdot\ 1360\right) + 0\right) - 1400\right) \cdot\ 80\right) + \left(\left(\left(\left(0.25 \cdot\ 40\right) + \left(1.15 \cdot\ 1250\right)\right.\right.\right.$$

Pasos

$$0.25 - \frac{0.1}{4}\left(\left(\left(\left(0.25 \cdot\ 80\right) + \left(1.15 \cdot\ 1360\right) + 0\right) - 1400\right) \cdot\ 80\right) + \left(\left(\left(\left(0.25 \cdot\ 40\right) + \left(1.15 \cdot\ 1250\right.\right.\right.\right.$$

Quitar los parentesis: $(a) = a$

$$= 0.25 - \frac{0.1}{4}\left(\left(0.25 \cdot\ 80 + 1.15 \cdot\ 1360 + 0 - 1400\right) \cdot\ 80 + \left(0.25 \cdot\ 40 + 1.15 \cdot\ 1250 + 0 - 1270\right)\right.$$

Mostrar pasos

$$\frac{0.1}{4}\left(\left(0.25 \cdot\ 80 + 1.15 \cdot\ 1360 + 0 - 1400\right) \cdot\ 80 + \left(0.25 \cdot\ 40 + 1.15 \cdot\ 1250 + 0 -\right.\right.$$

$$= 0.25 - 1580.875$$

Restar: $0.25 - 1580.875 = -1580.625$

$$= -1580.625$$

θ_2

Solución

Mostrar pasos

$$1.15 - \frac{0.1}{4}\left(\left(\left(\left((0.25 \cdot\ 80) + (1.15 \cdot\ 1360) + 0\right) - 1400\right) \cdot\ 1360\right) + \left(\left(\left((0.25 \cdot\ 40) + (1.15 \cdot\ 125\right.\right.\right.$$

Pasos

$$1.15 - \frac{0.1}{4}\left(\left(\left(\left((0.25 \cdot\ 80) + (1.15 \cdot\ 1360) + 0\right) - 1400\right) \cdot\ 1360\right) + \left(\left(\left((0.25 \cdot\ 40) + (1.15 \cdot\ 12\right.\right.\right.$$

Quitar los parentesis: $(a) = a$

$$= 1.15 - \frac{0.1}{4}\left(\left(0.25 \cdot\ 80 + 1.15 \cdot\ 1360 + 0 - 1400\right) \cdot\ 1360 + \left(0.25 \cdot\ 40 + 1.15 \cdot\ 1250 + 0 - 127\right.\right.$$

Mostrar pasos

$$\frac{0.1}{4}\left(\left(0.25 \cdot\ 80 + 1.15 \cdot\ 1360 + 0 - 1400\right) \cdot\ 1360 + \left(0.25 \cdot\ 40 + 1.15 \cdot\ 1250 + 0\right.\right.$$

$$= 1.15 - 26634.625$$

Restar: $1.15 - 26634.625 = -26633.475$

$$= -26633.475$$

b

Solución

Mostrar pasos

$$0 - \frac{0.1}{4}\left(\left(\left(\left(0.25 \cdot\ 80\right) + \left(1.15 \cdot\ 1360\right) + 0\right) - 1400\right) \cdot\ 1\right) + \left(\left(\left(0.25 \cdot\ 40\right) + \left(1.15 \cdot\ 1250\right) + 0\right.\right.$$

Pasos

$$0 - \frac{0.1}{4}\left(\left(\left(\left(0.25 \cdot\ 80\right) + \left(1.15 \cdot\ 1360\right) + 0\right) - 1400\right) \cdot\ 1\right) + \left(\left(\left(0.25 \cdot\ 40\right) + \left(1.15 \cdot\ 1250\right) +\right.\right.$$

Quitar los parentesis: $(a) = a$

$$= 0 - \frac{0.1}{4}\left(\left(0.25 \cdot\ 80 + 1.15 \cdot\ 1360 + 0 - 1400\right) \cdot\ 1 + \left(0.25 \cdot\ 40 + 1.15 \cdot\ 1250 + 0 - 1270\right) \cdot\ 1 +\right.$$

Mostrar pasos

$$\frac{0.1}{4}\left(\left(0.25 \cdot\ 80 + 1.15 \cdot\ 1360 + 0 - 1400\right) \cdot\ 1 + \left(0.25 \cdot\ 40 + 1.15 \cdot\ 1250 + 0 -\right.\right.$$

$$= 0 - 19.45$$

Restar: $0 - 19.45 = -19.45$

$$= -19.45$$

Project Run on Terminal:

```
Gradient Descent Function:
Old Thetas:  [0.25, 1.15, 0]
New Thetas [-1580.6249999999986, -26633.474999999984, -19.449999999999999]
```