

Day-11: Linear regression

- Linear regression is a Supervised machine learning algorithm used to model the relationship between one or more independent variable and a continuous dependent variable by fitting a straight line that minimizes the prediction error
- The model assumes the form

$$y = mx + c$$

where :

y = target / predict

m = Slope (How steep the line is)

x = input feature

c = intercept (Starting point)

Slope

- It tells, how much y changes when x increases by 1

bias

The value of y when $x = 0$

- In linear regression, x is the input & y is the output, the model assumes y changes at a constant rate with respect to x and tries to find a straight line that predicts y from x with minimal error compared to observed data and the best regression line is the one that minimizes total square error, not the line that pass through most of the data points

Error:

- The model never be perfect, so we measure how wrong is it for each data

- Actual = y

• predict = \hat{y} ($\hat{y} = mx + b$)

- we define error as

$$\text{error} = \text{prediction} - \text{actual}$$

$$= \hat{y} - y$$

- we square the error that is $(\hat{y} - y)^2$ because

- Remove sign
- penalizes large error
- differentiable and it enable optimization

cost function:-

- This function is used to turn best line into a measurable objective
- for multiple data points

$$J(m, b) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

$$J(m, b) = \text{MSE}$$

where

J = cost (How bad the model is)

n = No. of Samples

- the smaller J the better model

code

```
from sklearn.metrics import mean_squared_error as MSE
```

```
mse = MSE(y_Actual, y_Prediction)
```

- How to find m, b
 - we know that, linear regression follows $y = mx + b$ form
 - To find m, b we use gradient descent
- Gradient descent**
- The objective of gradient descent is to minimize the cost function by finding a parameter value (m & b) that produces lowest prediction error
 - The gradients provides both direction and magnitude of the steepest change in cost function, allowing us to move parameters towards lower cost value, minimizing the cost corresponding to improve model's prediction
- Cost function**
- $$J(m, b) = \begin{bmatrix} \frac{\partial J}{\partial m} \\ \frac{\partial J}{\partial b} \end{bmatrix}$$
- (gradient is a vector of partial derivative)
- $$m = m - \alpha \frac{\partial J}{\partial m}$$
- $$b = b - \alpha \frac{\partial J}{\partial b}$$
- where,
- α = learning rate
- learning rate control size.
- Large α
 - Fast movement
 - Risk of overshooting
 - possible divergence
 - Small α
 - Stable
 - Slow convergence
 - Stability is prioritized over speed

Model training-training and testing.

- During training

- model see the data
- They adjust the parameters to reduce cost on the data
- The model is incentivized to perform well on those exact points

- if the model works based only past data or unable to predict the future data, the model is useless so that we split the data into

80:20

70:30

- Dataset splitted into two parts

- Training set used to compute gradients
update m, b
minimize cost
- Testing set used only for training

- when the model gives high train & test error it is called underfit

- when the model give low train error & high error, it is called overfit

- If test set leaks into training

- performance metrics lie
- overfitting goes undetected
- production failure occurs

e) How to solve linear regression with one dependent variable and one independent variable

i. Assume the dataset

$$x = [1, 2, 3, 4, 5], y = [2, 3, 4, 5, 6]$$

try to find the best line using

1. General method
2. Gradient descents

A) i - General method

we know that

x is independent variable

y is dependent variable

1) compute mean

$$x \text{ mean } = \bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$y \text{ mean } = \bar{y} = \frac{2+3+4+5+6}{5} = \frac{20}{5} = 4$$

2) compute slope (m)

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
1	2	-2	-2	4	4
2	3	-1	-1	1	1
3	4	0	0	0	0
4	5	+1	1	0	1
5	6	+2	2	4	4

$$(x_i - \bar{x})(y_i - \bar{y}) = 4 + 1 + 4 = 9$$

$$(x_i - \bar{x})^2 = 4 + 1 + 1 + 4 = 10$$

$$m = \frac{9}{10} = 0.9$$

③ compute intercept "b"

$$y = mx + c$$

$$c = y - mx$$

$$c = 4 - (0.9)(3)$$

$$c = 4 - 2.7$$

$$c = 1.3$$

Linear regression

$$m = 0.9, c = 1.3, y = 4, x = 3$$

④ prediction

x	actual y	predicted \hat{y}
1	2	$\hat{y} = 0.9 \times (1) + 1.3 = 2.2$
2	3	$\hat{y} = 0.9 \times (2) + 1.3 = 3.1$
3	5	$\hat{y} = 0.9 \times (5) + 1.3 = 5.8$
4	4	$0.9 \times (4) + 1.3 = 4.9$
5	6	$0.9 \times (6) + 1.3 = 5.9$