Optimization

Before jumping into optimization lets discuss error and cost function first.

1. Error/loss Function:

It is for single training examples.

error = $Y_t - Y_p$

2. Cost Function :

It is an average loss over the entire dataset.

$$\text{cost function} = \sum_{i=0}^{n} (Y_t - Y_p)$$

back to the optimization:

Optimization is the process where we train the model iteratively that results in a maximum a
function evaluation. It is one of the most important phenomena in Machine Learning to get be

Example :

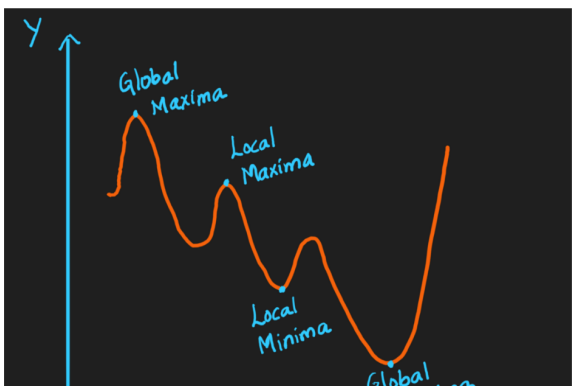
let $w = \frac{(a-5)^2}{(a+2)^2}$

optimize the value of a so function value will be maximum and minimum.

1. Initialize the value of a.
2. calculate the value of func.
3. repeat the step and 2 until we get max/min

MAXIMA AND MINIMA

Maxima is the largest and Minima is the smallest value of a function within a
We represent them as below:



- Global Maxima and Minima: It is the maximum value and minimum value res
  entire domain of the function.
- Local Maxima and Minima: It is the maximum value and minimum value resp
  function within a given range.

Types of Optimization

GRADIENT DESCENT

Let's say you are playing a game where the players are at the top of a mountain,
of the mountain. Additionally, they are blindfolded. So, what approach do you think

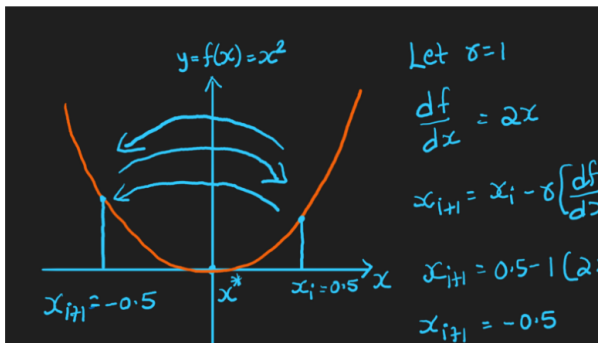
Gradient descent is an iterative optimization algorithm for finding the local minimu

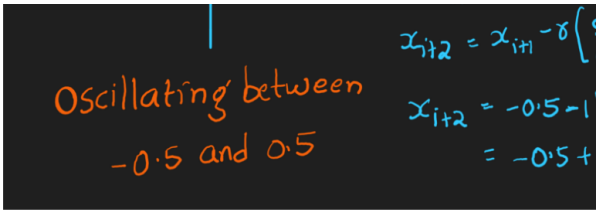
1. Compute the gradient (slope), the first order derivative of the function at th
2. Make a step (move) in the direction opposite to the gradient, opposite directi

Objective: Calculate X*- local minimum of the function $Y=X^2$.

1. Pick an initial point X_0 at random
2. Calculate $X_1 = X_0 - r[df/dx]$ at X_0. r is Learning Rate. Let us take r=1. Here,
3. Calculate $X_2 = X_1 - r[df/dx]$ at X_1.
4. Calculate for all the points: $X_1, X_2, X_3, \ldots\ldots, X_{i-1}, X_i$
5. General formula for calculating local minima: $X_i = (X_{i-1}) - r[df/dx]$ at X_{i-1}
6. When $(X_i - X_{i-1})$ is small, i.e., when X_{i-1}, X_i converge, we stop the iteration and

Learning Rate/steps is a hyperparameter or tuning parameter tha
iteration while moving towards minima in the function. For example
taken as r=0.01 in the next step. Likewise, it can be reduced ex
used more effectively in deep learning.





The disadvantage of Gradient Descent:

- When n(number of data points) is large, the time it takes for k iterations to co
  becomes very large.
- This problem is solved with Stochastic Gradient Descent.

STOCHASTIC GRADIENT DESCENT(SGD/ MB-Sc

- In SGD, we do not use all the data points but a sample of it to calculate the
- Stochastic basically means Probabilistic
- So we select points randomly from the population.

GD

- All the data points consider for each iteration.

- loss = $\sum_{i=1}^{n} (Y - Y_p)^2$

- Take to much space complexity for taking all data points

SGD

- Each data points consider for each iteration

- loss = $(Y - Y_p)^2$

- take too much time complexity to consider each data points

- S

-

Step 1

Let m = 0 and c = 0. Let L be our learning rate. It could be a small value like 0.01 for good accuracy

Step 2

Calculate the partial derivative of the Cost function with respect to m. Let partial derivative of the Cost function with respect to m be Dm (With little change in m how much Cost function changes).

$$D_m = \frac{\partial(Cost\ Function)}{\partial m} = \frac{\partial}{\partial m}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial m}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial m}\left(\sum_{i=0}^{n}(y_i^2 + m^2x_i^2 + c^2 + 2mx_ic - 2y_imx_i - 2y_ic)\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - (mx_i + c))$$

$$= \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - y_{i\,pred})$$

Similarly, let's find the partial derivative with respect to c. Let partial derivative of the Cost function to c be Dc (With little change in c how much Cost function changes).

$$D_c = \frac{\partial(Cost\ Function)}{\partial c} = \frac{\partial}{\partial c}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial c}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial c}\left(\sum_{i=0}^{n}(y_i^2 + m^2x_i^2 + c^2 + 2mx_ic - 2y_imx_i - 2y_ic)\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))$$

$$\frac{-2}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})$$

Step 3

Now update the current values of m and c using the following equation:

$$m = m - LD_m$$

$$c = c - LDc$$

Step 4

We will repeat this process until our Cost function is very small (ideally 0).

The Boston Housing Dataset

Notebook Data Logs Comments (13)

355 Co

Detail Compact Column 1 of 1 columns

A 0.00632 18.00 2.310 0 0.5380 6.5750 65.20 4.0900 1 296.0 15.30 396.90 4.98 24.00

- CRIM per capita crime rate by town - ZN proportion of residential land zoned for lots over 25,000 sq.ft. - INDUS proportion of non-retail business acres per town - CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) - NOX nitric oxides concentration (parts per 10 million) - RM average number of rooms per dwelling - AGE proportion of owner-occupied units built prior to 1940 - DIS weighted distances to five Boston employment centres - RAD index of accessibility to radial highways - TAX full-value property-tax rate per $10,000 - PTRATIO pupil-teacher ratio by town - B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town - LSTAT % lower status of the population - MEDV Median value of owner-occupied homes in $1000's

505
unique values

Valid	505	100%
Mismatched	0	0%
Missing	0	0%

unique values	Unique	505
	Most Common	0.02731 0.0 0%

Error / Cost F

Absolute Error :

$$L = |Y - Y_p|$$

Mean Absolute Error :

$$MAE = \frac{\sum_{i=1}^{N} (Y_i - Y_{pi})}{N}$$

Square Error :

$$L = (Y - Y_p)^2$$

Mean Square Error :

$$MSE = \frac{\sum_{i=1}^{N} (Y_i - Y_{pi})^2}{N}$$

Root Mean Square Error :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - Y_{pi})^2}{N}}$$

R- Squared :

How much the model is working well in comparison with a dump mean model.

$$SS_{res} = \sum_{i=1}^{N} (Y_i - Y_{mi})^2$$

-- Y_m - Mean of the data.
Y - True value

$$SS_{total} = \sum_{i=1}^{N} (Y_i - Y_{pi})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

$(0 =< R^2 <= 1)$

let's discuss working of linear Regression :

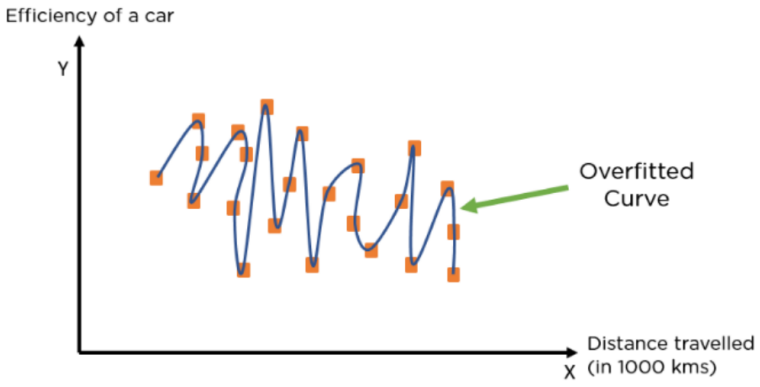
let we have MSE :

$$MSE = \frac{\sum_{i=1}^{N} (y_i - y_{pi})^2}{N}$$

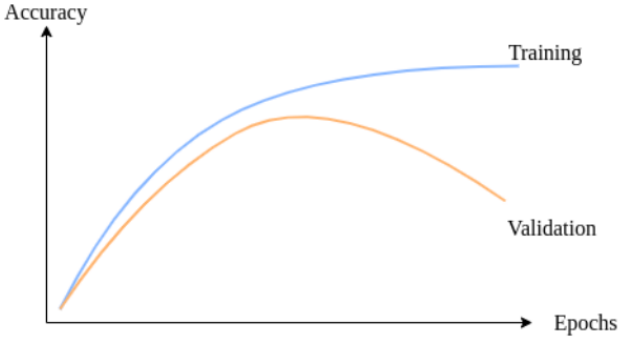
Overfitting and Underfitting

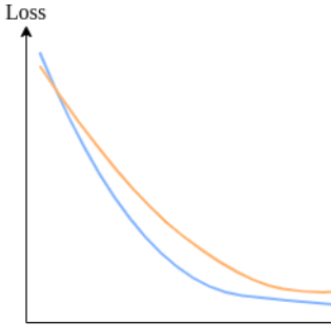
Overfitting :

When a model performs very well for training data but has poor performance with
it is known as overfitting. In this case, the machine learning model learns the deta
data such that it negatively affects the performance of the model on test data



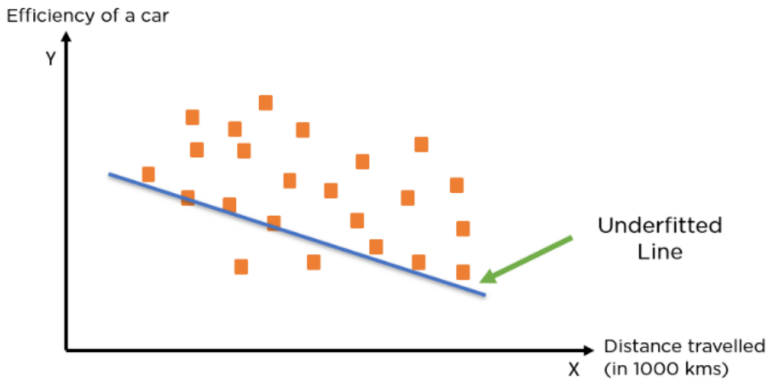
We need to closely watch model loss and accuracy to decide how the model is fitt



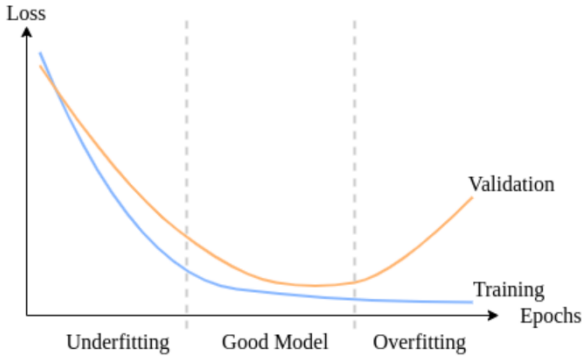


Underfitting :

When a model has not learned the patterns in the training data well and is unable
new data, it is known as underfitting. An underfit model has poor performance on
in unreliable predictions.



Conditions



Let's summarize what we've discussed so far in a comparison table:

Overfitting	Underfitting
Model is too complex	Model is not complex enough
Accurate for training set	Not accurate for training set
Not accurate for validation set	Not accurate for validation se
Need to reduce complexity	Need to increase complexity
Reduce number of features	Increase number of features
Apply regularization	Reduce regularization
Reduce training	Increase training
Add training examples	Add training examples

Column Standardization and normalizat

Standardization :

For any given data matrix we process each column separately to make the range of values under particul

let a, b, c ,d _____
after std : a1, b1, c1, d1, _____ (Mean = 0, std dev. = 1)

$$A_i' = \frac{ai - a(mean)}{std}$$

std = $\sqrt{var}$

Var = $1/n(\sum_{i=0}^{n}(xi - mean)^2)$

Normalization :

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

$$A_i' = \frac{ai - a(min)}{a(max) - a(min)}$$

Values will be between [0,1].

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for s
2.	It is used when features are of different scales.	It is used when we want to ensure zero m
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normaliza

Summary

1. Linear Regression
2. Optimisation
   - Gradient descent -- Maxima/Minima/Learning rate
   - STOCHASTIC GRADIENT DESCENT(SGD)
   - MB SGD
3. Loss/ cost functions for Regression
   - Absolute Error
   - MAE

- Square Error
- MSE
- RMSE
- R-squared

4. Working of linear regression using optimization

5. Overfitting and underfitting

6. Feature scaling

7. Boston House price prediction ipynb project

MCQ

1. If Linear regression model perfectly first i.e., train error is zero, then _____
a) Test error is also always zero
b) Test error is non zero
c) Couldn't comment on Test error
d) Test error is equal to Train error

2. Which of the following metrics can be used for evaluating regression models?

i) R Squared
ii) Adjusted R Squared
iii) F Statistics
iv) RMSE / MSE / MAE

a) ii and iv

b) i and ii
c) ii, iii and iv
d) i, ii, iii and iv

3. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

a) by 1
b) no change
c) by intercept
d) by its slope

4. In the mathematical Equation of Linear Regression $Y = \beta1 + \beta2X + \epsilon$, ($\beta1$, $\beta2$) refers to _____

a) (X-intercept, Slope)
b) (Slope, X-Intercept)
c) (Y-Intercept, Slope)
d) (slope, Y-Intercept)

5. True- False: Overfitting is more likely when you have huge amount of data to train?

A) TRUE
B) FALSE