Machine Learning Algorithms

**Supervised Learning**

Solve two things (Regression (output feature continuous) and Classification)

# Regression

## 1)Linear Regression

### Agenda

**→What problem we are solving**

* Aim is to find best fit line with minimal error
* use for Regression Problem

**→Geometric intuition**

**→Mathematical Intuition**

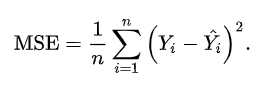
* y` = mx + c
* Y` is Predicted Output
* x is independent variable
* m is slope of coefficient
* c intercept
* Height (KG) (Y-axis) (Dependent Variable)
* Weight (cm) (X-axis) (Independent Variable)

**→Regression Analysis:**

* Dependent: (y) want to predict is also known as target variable.
* Independent: (x) used to predict is also known as predictor.
* Underfitting: algorithm does not perform well even with training dataset and test dataset
* overfitting: well in training dataset but not well with test dataset.
* Train data > model > Hypothesis get Weight data and predict Height data.
* model plot a best fit line after modelling.
* Learning Rate is always choose small

**→Mean Squared Error,Mean Absolute Error and RMSE**

* [Mean squared error (MSE)](https://statisticsbyjim.com/regression/mean-squared-error-mse/) measures the amount of error in statistical models. It assesses the average squared difference between the observed and [predicted values](https://statisticsbyjim.com/glossary/fitted-values/). When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).



* Advantage of MSE

Differentiable, it has one [local minima and one global minima](https://vitalflux.com/local-global-maxima-minima-explained-examples/)

* Disadvantage of MSE

Not Robust, it changes its unit

* [**Mean Absolute Error (MAE)**](https://deepchecks.com/glossary/mean-absolute-error/#:~:text=Mean%20Absolute%20Error%20(MAE)%20is,effectiveness%20of%20a%20regression%20model.) is a measure of the average size of the mistakes in a collection of predictions, without taking their direction into account. It is measured as the average absolute difference between the predicted values and the actual values and is used to assess the effectiveness of a [regression model](https://deepchecks.com/glossary/regression/).
* The MAE loss function formula
* MAE = (1/n) Σ(i=1 to n) |y\_i – ŷ\_i|
* Advantage

it is Robust of outlier, it will be in same unit

* DisAdvantage

conversion takes more time, Optimization is a complex process

* [Root mean square error (RMSE)](https://statisticsbyjim.com/regression/root-mean-square-error-rmse/) measures the average difference between a statistical model’s [predicted values](https://statisticsbyjim.com/glossary/fitted-values/) and the actual values. Mathematically, it is the standard deviation of the [residuals](https://statisticsbyjim.com/glossary/residuals/). Residuals represent the distance between the [regression](https://statisticsbyjim.com/glossary/regression-analysis/) line and the data points.

### Notes

1)[Mellon College Ch 9 Linear Regression](https://drive.google.com/file/d/1q8pPK_ylJ8YqQpHsUNtQCIV0jqdAKH1e/view?usp=sharing)

2)<https://drive.google.com/file/d/1G3mc6mPk2GYAKdOYttcZBWx2DAlesYQZ/view?usp=sharing>

3)<https://docs.google.com/presentation/d/1AZ5X0Zu2Xa54lvNDX8pKPG3jEeM-DCo1/edit?usp=sharing&ouid=101794246168461768671&rtpof=true&sd=true>

4)[Notebook](https://drive.google.com/file/d/1S8gHUCd9NKRiZAOkBanukAUirLPHwt-J/view?usp=sharing)

5)[Dataset for above notebook](https://drive.google.com/file/d/1RM_oLsQM-svL6qHA9bdwFY4y-dJ3MUs-/view?usp=sharing)

6)[Krish Naik Video Lecture](https://drive.google.com/file/d/1C32Zzu2uVi5fZ29fy6SmQuPSVk7sdDpB/view?usp=sharing)

7)[Notebook (California Dataset)](https://colab.research.google.com/drive/1sM08SGu2YgkmAO0PO7r72RTPSKOKkOk0?usp=sharing)

**Multiple Linear Regression:**

🡪used to explain the relationship between one continuous dependent variable and two or more independent variables.

Y = m1x1+ m2x2 + mnxn + C

## 2)Ridge and Lasso Regression

[Ridge and Lasso Regression](https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/#:~:text=Ridge%20and%20Lasso%20Regression%20are,hyperparameter%20tuning%20and%20cross%2Dvalidation.) are regularisation techniques used to prevent overfitting in linear regression models by adding a penalty term to the loss function. In Python, scikit-learn provides easy-to-use functions for implementing Ridge and Lasso regression with hyperparameter tuning and cross-validation.

### Agenda

**→What Problem we are solving**

* Regression Problem

**→Key Points**

* Over Fitting

Train Accuracy 90% (low bias)

Test Accuracy 70% (High Variance)

* Underfitting

Train Accuracy 60% (High bias)

Test Accuracy 62% (High variance)

* Generalised Model having low bias (90%) and low variance (89%) (Target)
* We Solve overfitting problems by using this algorithm.
* Ridge Regression (Overfitting is Removed)
* Lasso Regression (Prevent Overfitting and help in Feature Selection )

### Notes

1)[Krish Naik Video Lecture Note](https://drive.google.com/file/d/1XxsHgWrKyblzISP1A8wuhOxoEaiifg13/view?usp=sharing)

2)<https://drive.google.com/file/d/1xhu3_zSKZlp3ZsAw-Jped1Ka-2gqXr5q/view?usp=sharing>

3)<https://drive.google.com/file/d/1pwJzMlF94PiEqzBEK42viQAyiQJoVvue/view?usp=sharing>

4)[Notebook](https://colab.research.google.com/drive/1nAXy9M-V0ZvcvbAjMBm991wqPEzXvc2i?authuser=3#scrollTo=mzu3BcJYSTNv)

# Classification

## 1)Logistic Regression

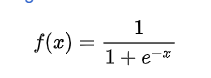
### Agenda

**→Problem Statement**

* Categorical dependent variables using given set of independent variables
* used to solve classification problem
* only two distinct values 0,1 True and False
* create a best fit line

**→Mathematical Intuition**

* (Sigmoid Function): mathematical function used to map predicted values to probabilities.



* it maps any value into another value within range of 0 and 1.
* the logistic regression equation can be obtained from the linear regression
* y/1-y; 0 for y=0 and infinity for y=1

range of +ive infinity to -ive infinity:

log(y/1-y) = bo + b1x1 + b2x2 + ... .... + bnxn

🡪threshold value = 0.5

Note: if predicted value is greater than 0.5 such as 0.7,0.8,0.9 then the values changes to 1

if predicted value is lesser than 0.5 such as 0.4,0.3,0.2 then the values changes to 0.

🡪We can check accuracy of trained model by using:

From sklearn.matrix import confusion\_matrix

### Notes

1)<https://drive.google.com/file/d/1Z_AJO7ybaJMI0-uAKcm6FFChbmirhk0G/view?usp=sharing>

2)<https://drive.google.com/file/d/1LayqIZuQ5YLja8NzRbA-SbXlUxieEPqU/view?usp=sharing>

3)<https://docs.google.com/presentation/d/1ROboC4GDAocGWt2Q1LCsRgZEXUa2bgVo/edit?usp=sharing&ouid=101794246168461768671&rtpof=true&sd=true>

4)[Krish Naik Video Notes](https://drive.google.com/file/d/1GzjbK828_81Mcnv1z7pHGH1ymEylVPh0/view?usp=sharing)

5)[Notebook](https://drive.google.com/file/d/11Jzd3m9RXW-R4FDvsbVEy_oy-tgJC4Vt/view?usp=sharing)

## 2)K Nearest Neighbour (KNN)

### Agenda

**→Problem Statement**

* Solve Classification Problem

**→Key Points**

* Simplest Machine Learning Algorithm based on supervised learning techniques.
* find value of k
* find k nearest data point
* Assumes similarity between new data and available data and put the new data into category that is most similar to available category.
* KNN both for Regression or Classification but mostly use for Classification.
* Non parametric algorithm which means it does not make any assumption on underlying data.
* Lazy learner algorithm because it does not quickly learn from training set data.
* find Euclidean distance: A(X1,Y1),B(X2,Y2):
* Euclidian distance from A1 to B2 =
* Limitations:

Huge dataset creates problems, Sensitive to Outliers and missing values.

### Notes

1)<https://drive.google.com/file/d/11O1MbkLmSv8Jvsq1kwgZbsGbFS7gd0WC/view?usp=sharing>

2)<https://drive.google.com/file/d/10tDM0tne4h5IUwfGc6tA-AoNFXS2SaiV/view?usp=sharing>

3)<https://drive.google.com/file/d/1_QYeitK4M9CYFWh_WFhB7y_dZFsb_3Wr/view?usp=sharing>

## 3)Naive Baye’s

### Key Points

* Probabilistic algorithm
* Supervised Machine Learning Algorithm
* Solve Classification Problem Statement by using Baye’s Theorem
* Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes.
* Applications:

Spam filter, document classification and sentiment analysis.

* Independent Events

Rolling a dice {1,2,3,4,5,6} (sample space)

pr(1)=⅙ ,pr(2)=⅙

Tossing a coin {head,tail}

* Dependent Events

Bag of Marble

red 3 marble and 2 green marble

pr(red)=⅗ ,

remaing marble

2 green and 2 red

pr(green given red)=2/4

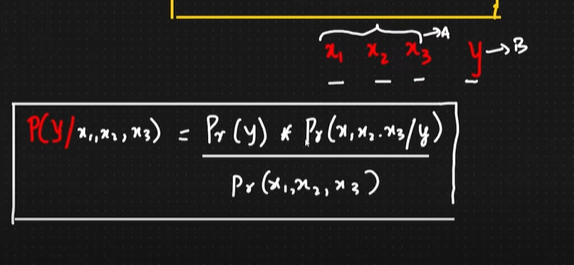
one event impact to another event

pr(red and green) = p(red) \* p(green/red)

pr(A and B) = pr(B and A)

pr(A)\*Pr(B/A) = pr(B) \* pr(A/B)

pr(B/A)=(pr(B)\*pr(A/B))/pr(A) (**Baye’s Theorem**)



y → Dependent Variable

x1,x2,x3 → Independent Variable

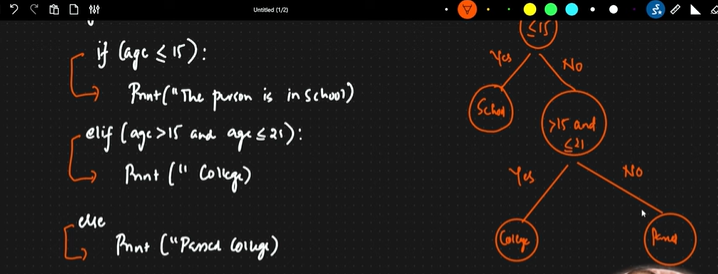
### Notes

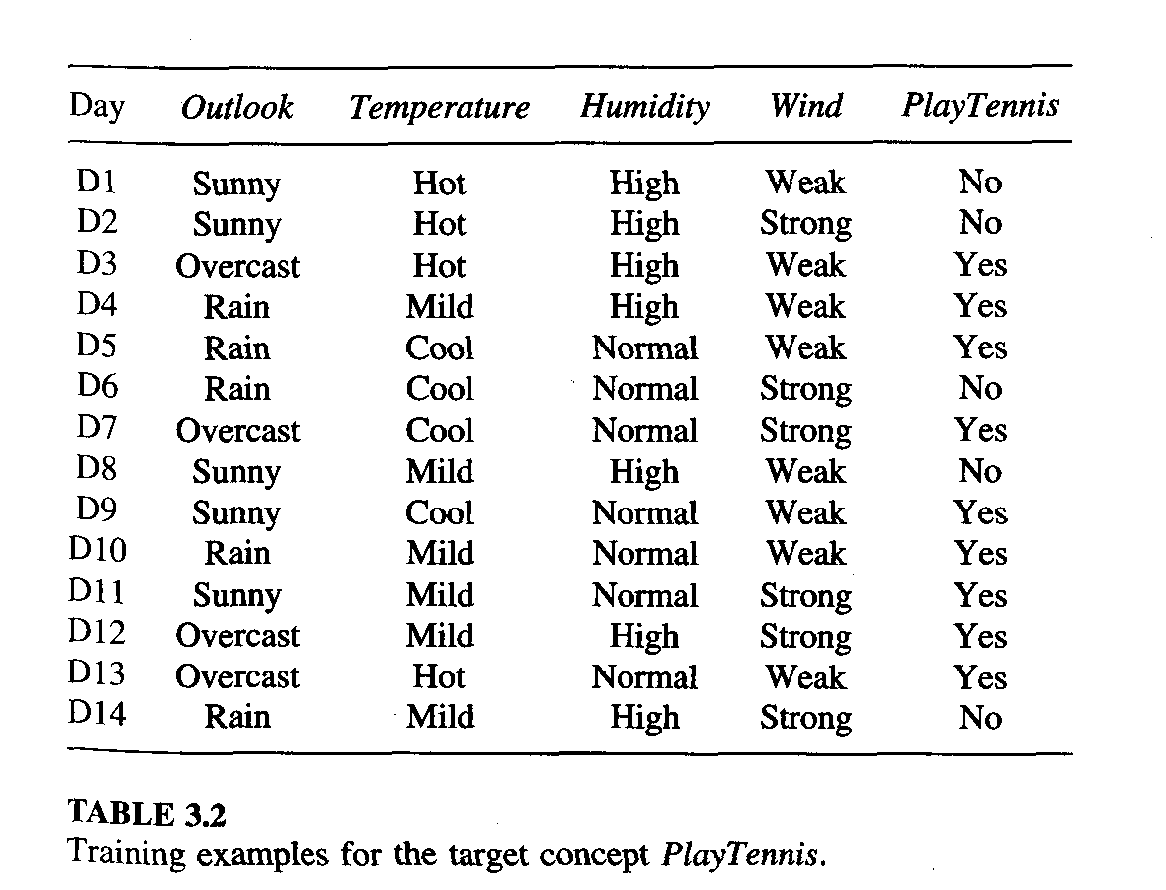
1. [Krish Naik Lecture](https://drive.google.com/file/d/1TJIuhPKvqZnIySGui99xwOJRSkYjWIvP/view?usp=drive_link)
2. [IBM BLOG](https://www.ibm.com/topics/naive-bayes#:~:text=Na%C3%AFve%20Bayes%20is%20part%20of,important%20to%20differentiate%20between%20classes.)
3. [PPT](https://docs.google.com/presentation/d/15UJpQS4zh9sALa4-cxg41tBuEExfVeuB/edit?usp=drive_link&ouid=101794246168461768671&rtpof=true&sd=true)
4. [PDF](https://drive.google.com/file/d/1QcaCz8ni-gO2pKgeb7bzy2VcjWBL6Nyp/view?usp=drive_link)

## 4)Decision Tree

### Key Points

* Decision Tree Use both for regression and classification problem.
* Two Techniques (ID3 and CART)
* Entropy and Gini Impurity(Purity Split)
* Information Gain (Feature Decision tree split)
* Multiple if else conditions apply



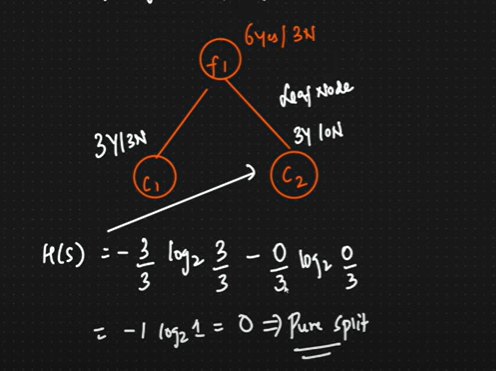


* Solve Example Yourself
* Purity? (Entropy and Gini Impurity) , Information Gain? (How the feature selcted)
* **Entropy Ranges (0 to 1)**

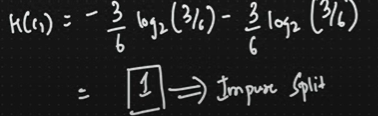
if a node have zero No(binary classification Problem) so this node is leaf node.

H(p) = −plg p − (1 − p) lg(1 − p)’

pure split

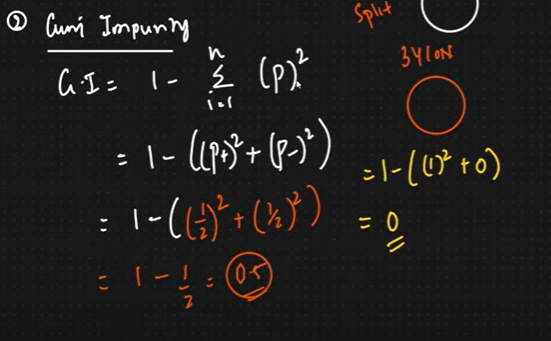


impure split

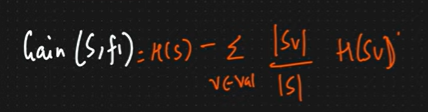


* **Gini Impurity Ranges (0 to 0.5)**

The Gini Impurity formula is: 1 – (p₁)² – (p₂)²,



* **Information Gain**



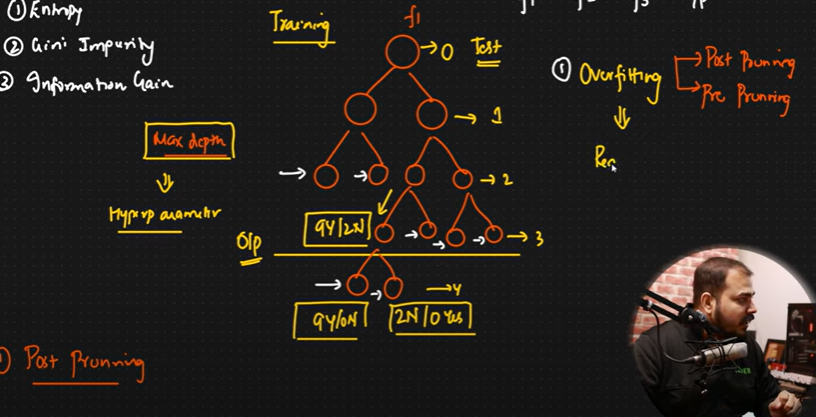
H(s) → Entropy of root node

H(c1) → leaf node entropy

if feature have high information gain than another feature having less information gain, so we choose high information gain feature.

Overfitting (Post Prunning and pre prunning)

* **Post Prunninig (apply on Small Dataset)**



* **Pre Prunning (Apply on Larger Dataset)**

When we construct tree use pre prunning technique choose feature that gives high accuracy.

* **k**

### Notes

1. Decision Tree Krish Naik (<https://drive.google.com/file/d/1UomP48jyDEBMOYZar6TPByX51nLAPn8z/view?usp=drive_link>)
2. post and Pre Prunning (<https://drive.google.com/file/d/1-x7tKrs-z9wM4QFpjTSWgX5HwcNc3FSe/view?usp=drive_link>)
3. Notebooks

<https://drive.google.com/file/d/108vP1PlfGoxntsLh3c-wsqHj9l2C0VqZ/view?usp=drive_link>

<https://drive.google.com/file/d/1FH43wYx5DADhIhpspeJo0hLlq8i3nmva/view?usp=drive_link>

<https://drive.google.com/file/d/1ZGjN5ypSwqja3MIQBXXlQcjLlVQiFjR4/view?usp=drive_link>

1. <https://drive.google.com/file/d/1OhYTXD0Sd8SLwDOL8D9rYLwcphevzk9v/view?usp=drive_link>
2. <https://drive.google.com/file/d/1j94L7p6v6zbIL_jepeSmu4m-_BbezBeT/view?usp=drive_link>
3. PPT (<https://docs.google.com/presentation/d/1ff2Ft0zlsrvfgANeTOYdNjxiEgtpq6Ul/edit?usp=drive_link&ouid=101794246168461768671&rtpof=true&sd=true>)

# Optimization Algorithm

## 1)Gradient Descent

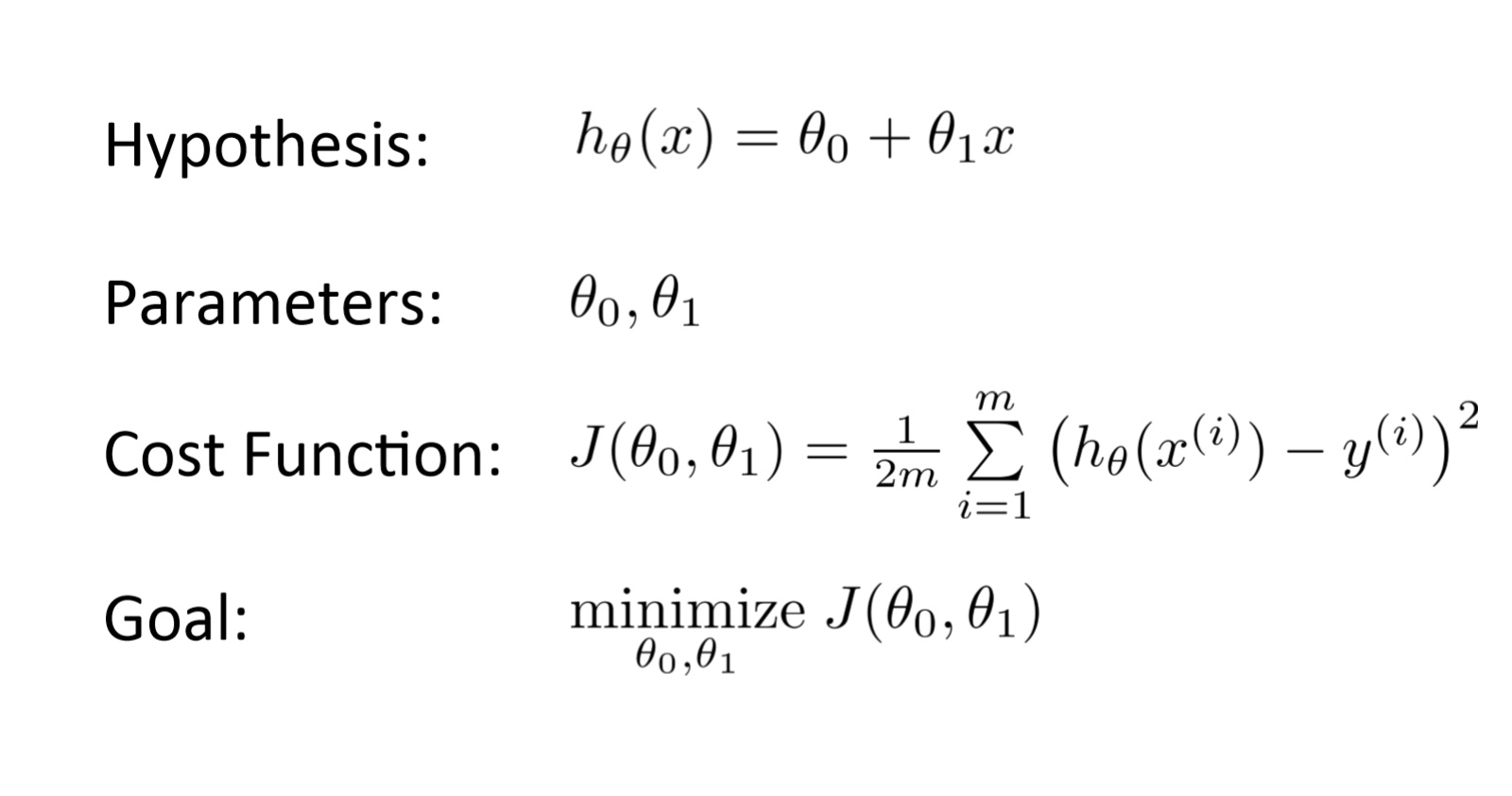
### Key Points

* Cost Function

Measures performance of the model for given data.

Quantify error between predicted value and actual value

Goal to reduce the cost function.



* Gradient Descent

Gradient descent is an optimization algorithm used in machine learning to minimize the cost function by iteratively adjusting parameters in the direction of the negative gradient, aiming to find the optimal set of parameters.

The cost function represents the discrepancy between the predicted output of the model and the actual output. The goal of gradient descent is to find the set of parameters that minimizes this discrepancy and improves the model’s performance.The Learning rate and hyperparameter determine step size in each iteration.

* Gradient Descent apply many machine learning algorithm Linear Regression,Logistic Regression, neural network and support vector machines.
* Example of Gradient Descent

Let’s say you are playing a game where the players are at the top of a mountain, and they are asked to reach the lowest point of the mountain. Additionally, they are blindfolded. So, what approach do you think would make you reach the lake?

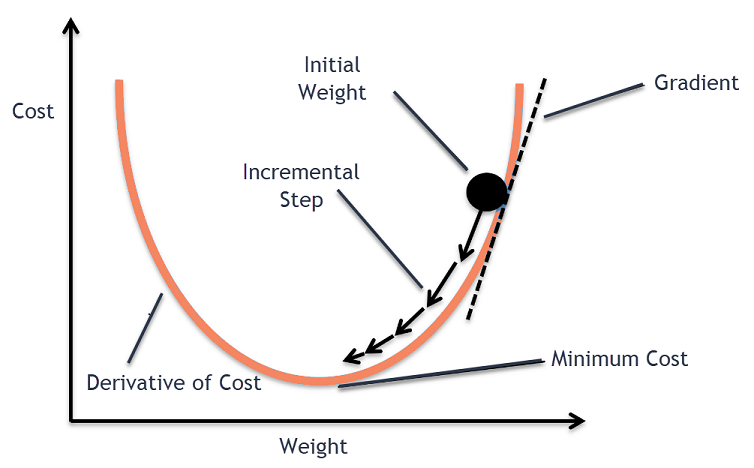
Take a moment to think about this before you read on.

The best way is to observe the ground and find where the land descends. From that position, take a step in the descending direction and iterate this process until we reach the lowest point.



To find the local minimum of a function using gradient descent, we must take steps proportional to the negative of the gradient (move away from the gradient) of the function at the current point. If we take steps proportional to the positive of the gradient (moving towards the gradient), we will approach a local maximum of the function, and the procedure is called Gradient Ascent.

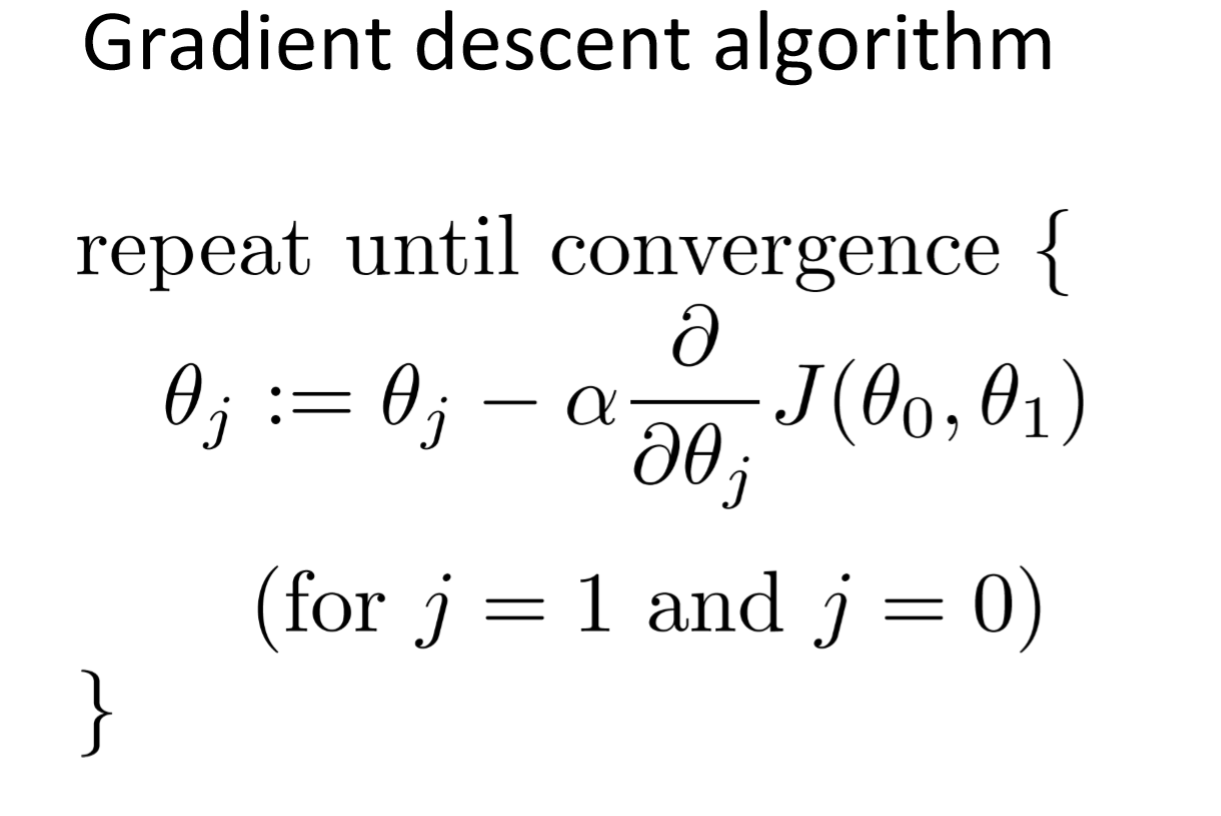
Gradient descent was originally proposed by CAUCHY in 1847. It is also known as steepest descent.



it performs two steps iteratively.

1-Compute the Gradient (the first order derivative of a function)

2-Make a step in the direction opposite to the gradient (opposite direction of slope increase from the current point aplatimes gradient at that point)



alpha is learning rate.

How Does Gradient Descent Work?

1. Gradient descent is an optimization algorithm used to minimize the cost function of a model.
2. The cost function measures how well the model fits the training data and is defined based on the difference between the predicted and actual values.
3. The gradient of the cost function is the derivative with respect to the model’s parameters and points in the direction of the steepest ascent.
4. The algorithm starts with an initial set of parameters and updates them in small steps to minimize the cost function.
5. In each iteration of the algorithm, the gradient of the cost function with respect to each parameter is computed.
6. The gradient tells us the direction of the steepest ascent, and by moving in the opposite direction, we can find the direction of the steepest descent.
7. The size of the step is controlled by the learning rate, which determines how quickly the algorithm moves towards the minimum.
8. The process is repeated until the cost function converges to a minimum, indicating that the model has reached the optimal set of parameters.
9. There are different variations of gradient descent, including batch gradient descent, stochastic gradient descent, and mini-batch gradient descent, each with its own advantages and limitations.
10. Efficient implementation of gradient descent is essential for achieving good performance in machine learning tasks. The choice of the learning rate and the number of iterations can significantly impact the performance of the algorithm.

* Types of Gradient Descent

Batch,mini batch,stochastic gradient descent

### Notes

1. Blog([Gradient Descent Algorithm in Machine Learning - Analytics Vidhya](https://www.analyticsvidhya.com/blog/2020/10/how-does-the-gradient-descent-algorithm-work-in-machine-learning/))
2. <https://drive.google.com/drive/folders/1h7xdnUFcACY8k0d-mS6yIgQYIILJSJwm?usp=drive_link>
3. <https://drive.google.com/file/d/1XhWVAXCHVTElA8VVUC-P3MZ7dm6r4XAu/view?usp=drive_link>