**Introduction to Machine Learning**

**And its Python Implementation**

Author: Li Hong Zephan

Version: 1.0

**Table of Content**

[1 Introduction to Machine Learning 3](#_Toc47261399)

[1.1 What is NOT Machine Learning 3](#_Toc47261400)

[1.2 What is Machine Learning 7](#_Toc47261401)

[2 An Introduction to Gradient Descent 8](#_Toc47261402)

[2.1 Notation 8](#_Toc47261403)

[2.2 Standard Gradient Descent 11](#_Toc47261404)

[2.3 Stochastic Gradient Descent 13](#_Toc47261405)

[2.4 Mathematical Alternative of Gradient Descent 14](#_Toc47261406)

[2.5 Python Implementation of Gradient Descent 24](#_Toc47261407)

[2.6 The influence of the Learning Rate on the Convergence Rate 27](#_Toc47261408)

# 1 Introduction to Machine Learning

## 1.1 What is NOT Machine Learning

Instead of defining what is Machine Learning by statements, I would like to help the readers understand what is NOT machine learning by a simple example.

Suppose we built a robot which has a speaker and a camera, and the camera can read a 8 x 8 pixels monitor, each pixel can either be turned on or off (Let’s use “1” to represent ON, and “0” to represent OFF). If we represent these pixels by a Two Dimensional Array, we’ll have something like this:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 |  |  |  | 1 |  |  |
|  | 1 |  | 1 |  | 1 |  |  |
|  |  | 1 |  |  | 1 |  |  |
|  |  |  |  |  |  |  | 1 |
|  |  |  |  |  | 1 |  |  |
| 1 |  | 1 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  | 1 |  |  | 1 |  |  |

Table 1.1 – 8 x 8 pixels monitor

(For better looking, we omitted all zeros in the above table, i.e. empty space “ ” means zero “0” in the array)

In computer programs such as Python or JavaScript, we usually say that the above array has 8 rows (i from 0 to 7) and 8 columns (j from 0 to 7)

Now suppose, we program the robot so that, if the number of “1” in the upper part of the table, is more than or equal to the number of “1” in the lower part of the table, the robot will say “Up” (or say “1”). On the contrary, the robot will say “Down” (or say “0”).

For a computer programmer that does not understand Machine Learning, he can write a simple program, that counts the number of “1” in rows 0 to 3 (then assign this value to, say intUp), and then counts the number of “1” in rows 4 to 7(then assign this value to intDown), and compare the values of intUp and intDown to determine whether the robot should say “Up” or “Down”.

The Python Implementation of this program is:

1. # In this NON-Machine Learning Example, we try to use a simple program
2. # to determine whether the no. of 1 in the upper part of the array
3. # is more than the no. of 1 in the lower part of the array
5. # Define Array
6. aryT = []
7. aryT.append([1,1,0,0,0,1,0,0])
8. aryT.append([0,1,0,1,0,1,0,0])
9. aryT.append([0,0,1,0,0,1,0,0])
10. aryT.append([0,0,0,0,0,0,0,1])
11. aryT.append([0,0,0,0,0,1,0,0])
12. aryT.append([1,0,1,0,0,0,0,0])
13. aryT.append([0,0,0,0,0,0,0,0])
14. aryT.append([0,0,1,0,0,1,0,0])
16. # No. of Rows and Columns
17. intM = len(aryT)
18. intN = len(aryT[0])
20. # Init No. of Up and Down
21. intUp = 0
22. intDown = 0
24. # Count No. of Up and Down in Array
25. **for** i **in** range(intM):
26. **for** j **in** range(intN):
27. **if** (i <= 3):
28. **if** (aryT[i][j] == 1):
29. intUp += 1
30. **else**:
31. **if** (aryT[i][j] == 1):
32. intDown += 1

35. # Print result
36. **print**('intUp: ' + str(intUp))
37. **print**('intDown: ' + str(intDown))
39. **if** (intUp >= intDown):
40. **print**('The result is: 1')
41. **else**:
42. **print**('The result is: 0')

Program 1.1 - NON Machine Learning classification program ml0101.py

This is the output of the program:

1. intUp: 9
2. intDown: 5
3. The result **is**: 1

Output 1.1

As expected, the numbers of “1” in the upper part and lower part of the array are printed and the result is “1”, i.e. the array has more “1”s in the upper part of the array.

Now, if we want the robot to recognize the digits from “0” to “9”, for examples:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | 1 | 1 |  |  |  |
|  |  | 1 |  | 1 |  |  |  |
|  |  |  |  | 1 |  |  |  |
|  |  |  |  | 1 |  |  |  |
|  |  |  |  | 1 |  |  |  |
|  |  |  |  | 1 |  |  |  |
|  |  |  |  | 1 |  |  |  |
|  |  | 1 | 1 | 1 | 1 | 1 |  |

Or other “similar figures” to be recognized as “1” and:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1 | 1 | 1 | 1 |  |  |
|  | 1 |  |  |  |  | 1 |  |
|  |  |  |  |  |  | 1 |  |
|  |  |  |  |  |  | 1 |  |
|  |  | 1 | 1 | 1 | 1 |  |  |
|  | 1 |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  |
|  | 1 | 1 | 1 | 1 | 1 | 1 |  |

Or other “similar figures” to be recognized as “2”, and so on for the “3” to “9”.

How can we modify the Python program so that the robot can “act like a human being”, and tell us what is the digit shown in 8x8 pixels monitor, if there are so many combinations to represent each digit?

In this scenario, we cannot simply explicitly tell the program how is each digit represented, we need to use some machine learning algorithm(s) for this program, and we’ll explain those algorithms in our later chapters.

## 1.2 What is Machine Learning

So what is Machine Learning? According to Wikipedia, “Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.”

In our next chapter, we’ll learn our first Machine Learning Algorithm – Gradient Descent.

# 2 An Introduction to Gradient Descent

## 2.1 Notation

m = # training examples

x = “input” variables / features

y = “output” variables / “target” variables (Figures inputted by human)

(x, y) = a training example.

The jth training example = (x(j), y(j))

Hypothesis hϴ(x), or simply h(x) = ϴ0 + ϴ1 x1 + ϴ2 x2 + …. + ϴnxn (for an x with n features)

We shall explain the meaning of Hypothesis very soon.

And Defining x0 = 1, =>

Now, let ϴ and x be matrices with 1 column and n row (i.e. consider ϴ and x be vectors represented by matrices), we have:

ϴ = and x =

ϴT =

Recall some matrix operation techniques, suppose a matrix A with i rows and j columns, another matrix B with j rows and k columns, and the multiplication of the two matrices C = AB, will be a matrix of i rows and k columns.

Now ϴT is just matrix A, with i = 1 and j = n, x is just matrix B with j = n and k = 1, so the multiplication of ϴTx is a matrix with 1 row and 1 column, and the value of the single element of this matrix product is:

ϴ0 x0 + ϴ1 x1 + ϴ2 x2 + …. + ϴnxn

Thus we have:

We call the ϴs “parameters”.

So what is a Hypothesis? (i.e. What is hϴ(x) ?) A hypothesis is an equation that we “guess”, how y is related to x, in the above example, we “guess” that y is directly proportional to the sum of the x features, with a Real Number Parameter ϴi before each xi . In later chapters, we’ll also handle hypothesis with higher order of xi .

Then, for m training examples, we try to find all ϴs, in order to minimize:

The is just for the simplification of mathematical operations later on.

Now, let:

By putting in different set of ϴs, we try to minimize J(ϴ), in other words, we want to find some optimistic ϴs, such that the value hϴ(x) that we guess, can be as close as possible to the actual value y, for all m training examples. Here we call J(ϴ) the Loss Function.

So how can we find those ϴs? We’ll introduce the Gradient Descent!

## 2.2 Standard Gradient Descent

To start, we randomly pick a value for ϴi, and iterates this ϴi with the following equation:

The := denote a computer operation that a new value of ϴi will be replaced by some calculations based on the old value of the previous ϴi. And the alpha “α” is called the learning rate, which is often set by human beings based on experiences on different machine learning algorithm, we’ll talk about how to set this learning rate later.

So, this computer operation means that, we try to find an optimal ϴi such that the J(ϴ) is minimum with respect to ϴi, in other words, we try to minimize the differences between the “calculated value by the hypothesis h(x)” and the “actual value y” provided by the trainer (i.e. figures provided by the human being), so we have, for 1 training example (x,y):

=

= (By Chain Rule)

since the only term inside hϴ(x) related to ϴi is ϴi xi, the derivative of other terms are zero, we have:

=

Thus,

ϴi := ϴi – … (Eq. 5) for 1 set of training example (x,y) only

(Note: in the above equation, the “i” stands for the ith feature of x)

for m sets of training examples, we have

(Note: Again, the “i” stands for the ith feature of x, and “j” stands for a particular training example)

The above algorithm to find each ϴi by iteration is called “**Batch Gradient Descent**” or just “**Standard Gradient Descent**”, it is obvious that, for EACH iteration, you need to sum all the (h(x) – y). xi in order to get a “better” ϴi to minimize J(ϴ). Consider if there are so many training examples, say, 10000 training examples, it’s very time consuming for the “Batch Gradient Descent” to find all ϴs such that J(ϴ) is minimized.

To handle a large number of training examples, we would like to introduce the “**Stochastic Gradient Descent**”.

## 2.3 Stochastic Gradient Descent

Unlike “Batch Gradient Descent”, “Stochastic Gradient Descent” randomly pick a few subset of the training examples, to approximately guesses the next ϴi, and repeat the process until the ϴi converges (“converges” means the ϴi does not change much, or the J(ϴ) does not change much, while the iterations continue).

Say, for example, if we only pick a random of 100 examples from a total of 100,000 training examples for each iteration, the total time consumed to optimize ϴs may be a few hundred times faster than the “Batch Gradient Descent”.

It is important to note that, picking fewer random examples in each iteration means faster iterations speed, but lower convergence rate. (i.e. Picking too little examples, the calculations may not converge, which means that you can’t get the minimum J(ϴ) after changing ϴs)

The above algorithms use iterations to obtain ϴs in order to minimize J(ϴ). In some particular cases, e.g. least square equation such as Eq. 3, we can obtain the ϴs by mathematical calculations instead of computer iterations.

## 2.4 Mathematical Alternative of Gradient Descent

To do so, we’ll need to introduce a few more notations:

∇ϴ J is a vector represented by a “n+1 rows, 1 column, Real Number Matrix”, which means the derivative of J with respect to ϴ, i.e. :

∇ϴ J= ϵ ℝ … (Eq. 7)

The notation “ϵ ℝ”, in case of font problem you can’t see it, it means “belong to Real Number”

Then, we rewrite ϴ as:

ϴ := ϴ - α ∇ϴ J … (Eq. 8)

We’ll now introduce more notations before we continue:

suppose A ϵ ℝ m×n, in other words, A is a matrix of real numbers with m rows and n columns, and f(A) is a function of A, then:

∇A f(A) = … (Eq. 9)

which is also a matrix of real numbers with m rows and n columns.

Now suppose B is a Square Matrix, i.e. B ϵ ℝ n×n

The trace of a square matrix is defined as:

Let’s recall some properties of trace of matrix, assume A, B and C are all square matrices with n rows and columns:

tr(AB) = tr(BA) … (Eq. 11)

We’ll try to verify Eq. 11 with a simple example by setting n = 2:

A = … (Eq. 12)

B = … (Eq. 13)

AB = … (Eq. 14)

tr(AB) = a11b11+a12b21+a21b12+a22b22 … (Eq. 15)

BA = … (Eq. 16)

tr(BA) = a11b11+a21b12+a12b21+a22b22 … (Eq. 17)

by reordering the 2nd and the 3rd terms, we have

tr(BA) = a11b11+a12b21+a21b12+a22b22 = tr(AB)

By similar examples, we can also verify that:

tr(ABC) = tr(CAB) = tr(BCA) … (Eq. 18)

Now, suppose, if we can express f(A) as tr(AB), i.e.:

f(A) = tr(AB) … (Eq. 19)

It turns out that:

∇A f(A) = BT … (Eq. 20)

Let’s verify Eq. 20 by using a particular example of A and B stated in Eq. 12 and Eq. 13:

In Eq. 9, and replace f(A) with tr(AB) in Eq. 15, we have:

∇A f(A) =

= = BT

And it is also easy to verify that:

tr(A) = tr(AT) … (Eq. 21)

since the Transpose of A does not change the diagonal elements of square matrix A.

Let us introduce 2 more equations here without the actual proving:

(1), if a ϵ ℝ, we have:

tr(a) = a … (Eq. 22)

(2),

∇A tr(ABATC) = CAB + CTABT … (Eq. 23)

With the help of the above equations: Eq. 11, Eq. 18, Eq. 20, Eq. 21, Eq. 22 and Eq. 23, we are going to learn how we can find ϴs in order to minimize J(ϴ) without using computer iterations.

Let X be a matrix which contains all inputs of each training example (m training examples with n features)

X = … (Eq. 24)

=

(Note: )

Then Xϴ will be the hypothesis for each training example

Xϴ = ϴ

=

=

= … (Eq. 25)

Define y as a matrix which contains outputs of m training examples

= … (Eq. 26)

Then Xϴ - y will be a matrix which contains the difference between hypothesis and actual outputs of m training examples

Xϴ - y = … (Eq. 27)

=

=

=

Multiply by one half, we have:

To minimize J, we have to set the derivative of J() closest to 0, which means the slope of J is closest to 0.

= … (Eq. 30 Step 1)

= … (Eq. 30 Step 2)

= … (Eq. 30 Step 3)

For Eq. 30 Step 1, expand the :

Note that X is a m x n matrix while is a n x 1 matrix, when is transposed, is a n x m matrix while is a 1 x n matrix.

Therefore, but

For Eq.30 Step 2, we know that is just a real number (See Eq. 28), so we can add a trace to it, where a trace of a real number is just itself (See Eq. 22).

For Eq.30 Step 3, ignore the one half and divide it into three parts:

First part:

For the first part, apply Eq. 18 to and move to the front, then we have

Second part: 1st

For the second part, we know that is a real number, where and it is a 1 x m matrix, while y is a m x 1 matrix. The transpose of a real number is just itself. Therefore, we can transpose it, then we have .

Third part: 2nd

For the third part, for later calculation, we do not have to change it.

For the , since we cannot differentiate it with respect to , it will be equal to 0 and we can just ignore it.

Add an identity matrix *I*, which will be ignored, to , then apply Eq. 23 to it, we have:

= … (Eq. 31)

Where refers to A, I refers to B, refers to , and refers to C.

Note: the identity matrix I is a m x m square matrix. Here is an example of identity matrix with 3 rows and columns:

I3 =

By applying Eq. 19, we have:

= … (Eq. 32)

Where refers to B, refers to A, and refers to

Apply Eq. 31 and Eq. 32 to Eq. 30, we have:

=

=

Then set to 0, we have:

… (Eq. 33 – Normal Equation)

Note: Eq. 33 is also called the Normal Equation.

By using the normal equation, we can solve for without the use of the iterative Gradient Descent algorithm.

## 2.5 Python Implementation of Gradient Descent

The following is a simple python program to illustrate how we can find all the ϴs in order to minimize the J(ϴ) as described in our last chapter. We shall explain this program briefly after the source codes.

1. # In this Machine Learning Example, we try to use Gradient Descent
2. # to estimate the weight of a person y, with given height x of a person
4. # Define No. of Features of x
5. intN = 1
7. # Define Training Examples
8. aryT = []
9. aryT.append({'x': [20], 'y': 2})
10. aryT.append({'x': [40], 'y': 4})
11. aryT.append({'x': [60], 'y': 6})
12. aryT.append({'x': [120], 'y': 12})
13. aryT.append({'x': [160], 'y': 16})
14. aryT.append({'x': [162], 'y': 16.2})
15. aryT.append({'x': [171], 'y': 17.1})
17. # No. of Training Examples
18. intM = len(aryT)
20. # Learning Rate
21. intAlpha = 0.00001
23. # Init Array Parameters
24. aryTheta = []
25. **for** i **in** range(intN+1):
26. **if** (i == 0):
27. # Init. Theta0 with 0
28. aryTheta.append(0)
29. **else**:
30. # Init. Theta1 with 0
31. aryTheta.append(0)

34. # Hypothesis H, t stands for theta
35. # H(t) = t0 x0 + t1 x1
37. intMaxTrainTimes = 100

40. # Init. Temp vars.
41. intSum = 0
42. intTemp = 0

45. # Iterates for intMaxTrainTimes
46. **for** t **in** range(intMaxTrainTimes):
47. # Iterate for each feature
48. **for** i **in** range(0, intN+1):
49. # Init Summation
50. intSum = 0
52. **for** j **in** range(intM):
53. intTemp = (aryTheta[0] + aryTheta[1] \* aryT[j]['x'][0]) - aryT[j]['y']
54. **if** i == 1:
55. intTemp = intTemp \* aryT[j]['x'][0]
57. **print**('i: 0 | t: ' + str(t) + ' | j: ' + str(j) + ' | intTemp: ' + str(intTemp))
58. intSum += intTemp
59. #End for
61. **print**('intSum: ' + str(intSum))
63. # Calculate New Theta(0)
64. aryTheta[i] = aryTheta[i] - intAlpha \* intSum
66. #print('i: 0 | aryTheta[0]: ' + str(aryTheta[0]))
67. **print**('aryTheta[0]: ' + str(aryTheta[0]))
68. **print**('aryTheta[1]: ' + str(aryTheta[1]))


72. # Print all Theta(s)
73. **print**('aryTheta[0]: ' + str(aryTheta[0]))
74. **print**('aryTheta[1]: ' + str(aryTheta[1]))

Explanations:

1. The intN is the no. of features in an input x of a training example (x, y).
2. The intM is the total no. of training examples.
3. The array aryT contains all the training examples, aryT[0] means the first training example.
4. The “x” of aryT contains all the feature inputs of a training example, e.g. if intN = 2, aryT[3][’x’][1] means the x24 , i.e. The value of the second feature of x, of the 4th training example.
5. The “y” of aryT contains the actual value provided by human being, e.g. aryT[2][’y’] means the y3, i.e. The actual value of the 3rd training example.
6. intAlpha is the Learning Rate.
7. aryTheta is an array that stores all the values of ϴs.
8. intMaxTrainTimes is the no. of times that we want to iterates in order to get the final values of aryTheta.
9. The main objective of this python program is to simulate Eq. 6, so as to find all ϴs.

## 2.6 The influence of the Learning Rate on the Convergence Rate

It is obvious that, we can use the figures provided by the training examples to plot a graph of a straight line with a slope of 0.1 and an intercept on (0,0). After running this program using python3, we can find that:

aryTheta[0] = 0.00072392

aryTheta[1] = 0.099995

which is very close to the calculated values of 0 and 0.1 for these ϴs.

So how about we change the value of the Learning Rate? Will the calculations still converge with larger learning rate? Here is the table that contains the figures of the ϴs for different learning rate:

|  |  |  |
| --- | --- | --- |
| Learning Rate | ϴ0 | ϴ1 |
| 0.000010 | 0.00072392 | 0.099995 |
| 0.000011 | 0.00072381 | 0.099995 |
| 0.000012 | 0.00072369 | 0.099995 |
| 0.000013 | 0.00072357 | 0.099995 |
| 0.000014 | 0.00072345 | 0.099995 |
| 0.000015 | 0.00072333 | 0.099995 |
| 0.000016 | 0.00072322 | 0.099995 |
| 0.000017 | 0.00072310 | 0.099995 |
| 0.000018 | 0.00072298 | 0.099995 |
| 0.000019 | 0.00072268 | 0.099969 |
| 0.000020 | 0.00541728 | -0.746647 |
| 0.000021 | -76.78181 | -10587.30267 |

So, it is quite obvious that, the calculations start to diverge, when the learning rate is >= 0.000019 for these particular training examples.

i.e. For the calculations to converge, we need to set the learning rate <= 0.000018 for these particular training examples.

To simplify the selection of the Learning Rate, we can choose a initial learning rate of 0.001 divided by (10 to the power of the order of the maximum x + 1), for example, in our training set, the maximum value of x is 171, so we divide 0.001 by 1000 and set the learning rate to 0.000001. And since the maximum learning rate allowed is 0.000018, by using 0.000001(18 times smaller that the maximum learning rate allowed), we may get the convergence fast and safe enough.

In our later chapters, we’ll try to understand how we can find the optimal value(s) of the Learning Rate using other algorithms.