Applications of Mathematics in Machine Learning

Virender and Abhishek Pratap Singh
Department of Mathematics, Ramjas College, University of Delhi, Delhi. INDIA
Email :Virender57@yahoo.com; Abhishek.pr.singh03@gmail.com

Abstract

In this introductory report, use of mathematics in machine learning is studied and explained. Machine learning is an emerging area and now a day, it is one of essential components of technology in our day to day life. Through this study, an attempt is made to make an undergraduate student and a layman to understand mathematics involved in machine learning.

1. Introduction

Machine learning is basically a study of computer algorithms. It improves itself through experience and by the use of large data involved in experience. Artificial intelligence is the broad are involving machine learning. To develop an algorithm, a sample data known as 'training data' is required and a model is built based on the training data. After formation of the model, machine learning algorithm comes into consideration in order to make decisions or predictions. These algorithms are programmed to perform explicitly without human being. A wide variety of applications in medicine, email filtering, computer vision and similar fields uses the important concept of machine learning.

Though an American IBMer Arthur Samuel coined the term 'machine learning' in 1959, but it was not given proper attention. Arthur Samuel was a pioneer in the field of computer science and artificial intelligence. Nilsson wrote a book Learning Machine in 1960s which was referred as a book for research on the topic of machine learning. It was mainly used for pattern classifications till end of 1970s. Using strategies of teaching, a report was submitted in 1981 of neural networks. The report exhibited that a total of 40 characters i.e. 26 letters, 10 digits and 4 special symbols were learnt and recognized from a computer terminal by neural network. With this the research field of machine learning got a boost and it started expanding. Since then a lot have been done in machine learning.

In modern times, machine learning has two main objectives – to streamline the data related to models and to come up with expected future outcome based on these models. A typical example of algorithm specific to streamlining and classifying the data is used in classification of the cancerous moles. An example of other objective is the algorithms used for trading in stocks where the algorithms inform of potential future predictions of the stocks.

Machine learning is one of the parts of artificial intelligence (AI) and machine learning algorithms explore and make use of the

historical data to derive or predict new outcome values.

Mathematics in Machine Learning (M.L.)

The main concepts of mathematics used in M.L. are linear regression and the gradient descent. Linear regression is used to best fit to measure the relationship or interdependence between two quantities under consideration and the concepts of gradient descent are applied to optimize strategies for linear regression. One simple example of house price related to house area may be considered to understand the term 'relationship'. It is to note here that these techniques are not limited to linear regression. The other forms may be considered and used as and when required based on specific applications. Historical data of the model is collected and then used in machine algorithms to find the relationships between different variables. It is done with the help of regression and optimization. All these information are used by algorithms of machine learning to come up with an expected outcome. In this report, some mathematics aspects of this topic are discussed.

Before moving further, let us define these terms.

Linear Regression. It is an algorithm of machine learning based on supervised learning.

Linear regression performs a regression task and models target predictions based on independent variables. Finding out the relationship between variables and forecasting are the main objective of regression. Therefore, linear regression is used in almost all such requirements.

An example is considered to understand it.

Example 1. Consider the data of price of the house based on the area of the house given by

Area of the house (m^2)	Price of the house ($)
2700	555000
3000	566000
3200	615000
3500	670000

Our aim is to find the price of the house for a given area using the technique of linear regression.

To fit line best to the data, the best fitting line is the line around which data may be plotted closely. Mathematically, we consider the equation of fitting line as

$$y = mx + c,$$

where m is the gradient slope and c is a constant. Here x indicates the area of the house and y indicates the price of the house depending on the area.

If the actual price is y_{a_i} and the approximated price by the fitting line is y_i then the error in i^{th} price after approximation is $\left(y_{a_i} - y_i\right)^2$. Note the square is considered to avoid any sign in error.

Thus the total error in approximation in all observation is $\sum_i \left(y_{a_i} - y_i\right)^2$.

So, as per best fitting of data, we consider the fitting line to be best fitting line if it minimizes the above error. We leave the computation of equation of line as we can do using the simple calculus. Thus the equation of best fitting line is

$$y = 149.706x + 137412.$$

Plotting the best fitting line (in blue colour, Figure 1) together with some other lines in Figure 2, one may observe that the data is spread closest to blue line i.e. the best fitting line.

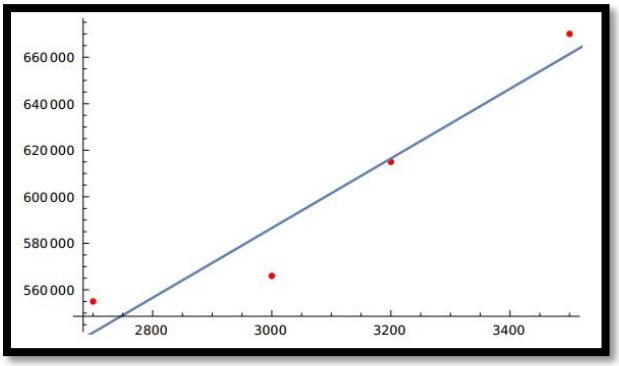

Figure 1

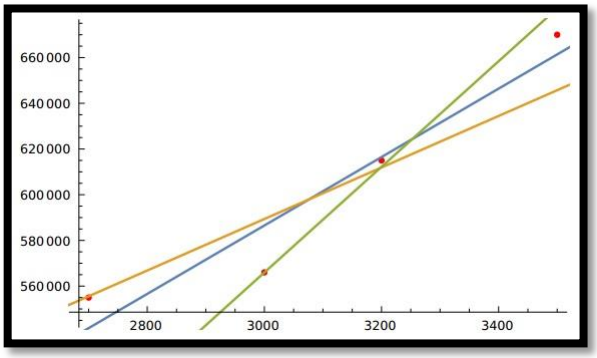

Figure 2

Now the line gives the relation between the price and area of the house. One can predict the price of the house for a given area using the relationship. For instance, $y = 601500.6$ for $x = 3100$ i.e. the price of the house is approximately \$ 601500.6 if the area of the house is 3100 m^2.

Since the fitting curves are not limited to one variable cases, let us take one more example to consider multivariable data.

Example 2. Let the data of the price of the house is given below

Area of the house (m^2)	No. of Bedrooms in the house	Age of the construction (Years)	Price of the house (\$)
2700	3	20	555000
3000	4	15	566000
3200	3	12	615000
3500	4	30	670000

To find the fitting line, we take equation of line as

$$y = m_1x_1 + m_2x_2 + m_3x_3 + c,$$

where m_1, m_2, m_3 are gradient slopes for area, bedroom, age, respectively and c is a constant. The variables x_1, x_2 and x_3 represent area, number of bedrooms and age of the house, respectively. The variable y denotes the price of the house depending on x_1, x_2 and x_3.

So the equation of line can be written as

$$y = m_0x_0 + m_1x_1 + m_2x_2 + m_3x_3 = \sum_{i=0}^{3} m_ix_i,$$

where $m_0 = 1$ and $x_0 = b$ are fixed.

To find the equation, one needs to find m_is. Here are two options in which the equation can be handled to compute m_is. Let us first consider in terms of matrices as we use matrices in linear algebra. The equation is represented in matrix form as

$$y = [x_0 \quad x_1 \quad x_2 \quad x_3]\begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{bmatrix} = Xm.$$

Clearly, $X = [x_0 \quad x_1 \quad x_2 \quad x_3]$ and $m = \begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{bmatrix}$.

Usually vector m is referred as weights. Thus

$$m = X^{-1}y.$$

In case, the above equation does not have any solution, it implies that y is not an element of the column space of X. In such cases, the projection of y onto the column space of X will be the closest vector to y which is also an element of the column space of X.

Thus multiplying the equation with the transpose of X i.e. X^T, it gives

$X^T y = X^T Xm$ or $(X^T X)^{-1} X^T y = m$.

Then m is given by

$$m = (X^T X)^{-1} X^T y.$$

This solution is computed using linear algebra. Another approach to compute the solution is using calculus. Let us understand that approach as well.

Let us assume that for any m, $f(m)$ gives the square error the m will generate on considered linear regression model. Obviously, the error is the variation or difference in computed value Xm and the actual value y, and to square the error, one may use the transpose just like one uses as norm. Then

$$f(m) = (Xm - y)^T (Xm - y).$$

Now if the requirement is to make the error minimum at a point, the point should be one of the critical points i.e. the gradient of f must be zero at that point.

Finding gradient of f,

$$\nabla f = \left[\frac{\partial f}{\partial m}\right]^T.$$

Now

$$\frac{\partial f}{\partial m} = \frac{\partial f}{\partial (Xm - y)} \frac{\partial (Xm - y)}{\partial m}$$

$$= 2(Xm - y)^T X.$$

Therefore

$$\nabla f = (2(Xm - y)^T X)^T = 2X^T (Xm - y).$$

$\nabla f = 0$ gives $2X^T (Xm - y) = 0$. Thus $X^T Xm - X^T y = 0$ and hence $m = (X^T X)^{-1} X^T y$. We get the same critical point. Further it is to investigate that it is a point of minima. To conclude, we will establish the convexity/concavity of the function f using the Hessian matrix.

Hessian matrix for f is given by

$$H = \nabla^2 f = \left(\frac{\partial(\nabla f)}{\partial m}\right)^T.$$

As $\frac{\partial(\nabla f)}{\partial m} = \frac{\partial(2X^T(XM-y))}{\partial m} = \frac{\partial(2X^TXm-2X^Ty)}{\partial m} = 2X^TX$, so $H = (2X^TX)^T = 2X^TX$.

Multiplying with a real-valued vector z and z^T, we get

$$z^THz = 2(Xz)^T(Xz) = 2\|Xz\|^2 \geq 0.$$

Then H is a semi-definite positive matrix and hence f is convex everywhere.

This gives that the critical point must be a point of minima.

To compute m, some conditions are required. Clearly X^TX is to be invertible and this requires X to have full column rank or in other words, all its columns are to be linearly independent. To satisfy this, the data must have number of rows to be greater than or equal to number of columns. Also this requirement assures $Xz \neq 0$ assuming $z \neq 0$ and hence H is positive definite which concludes that f is strictly convex.

In absence of the condition, the solution is computed using stochastic gradient descent, which is an iterative technique. This is beyond the scope of discussion in the report.

Using the techniques of linear regression, the relations between different variables are computed in machine learning algorithms as explained in above examples. It helps to predict new outcomes related to the model.

Sometimes, in M.L., cost function is considered.

Cost function. Cost function is the function that returns the error between the predicted outcomes compared with actual outcomes.

In linear regression models, this quantity is the minimum of the root mean squared errors of considered model and the mean squared error (M.S.E.) is given as $\frac{1}{n}\sum_{i=1}^{n}(y_i - y)^2; y = mx + c$, i.e. mean of the squares of errors. Next, we discuss the use of mathematical technique in optimization.

Gradient Descent. "It is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function."

As steepest descent is along the direction opposite to gradient at the considered point, we start with an iteration and step repeatedly in the direction opposite to the gradient/approximated gradient of concerned function at the iteration point.

Graphically, it may look like the below graphs

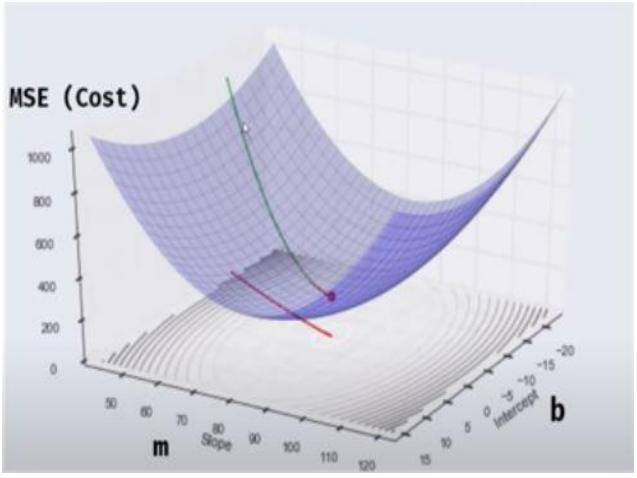

Figure 3

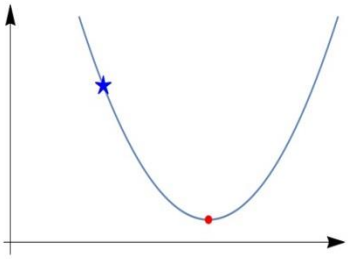

Figure 4

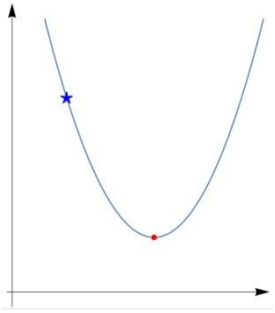

Figure 5

The aim is to arrive at the point of minima using iterative process. It is done using the measurement of descent and then proceeding towards the point.

Opposite to two different axes m and b, it may be visualized as given in Figure 4 and Figure 5.

One must take care of fixed step size in iterations as sometimes a bigger step size in iterations may miss the actual point of minima and it may cross over to other side as indicated in the below graph (Figure 6). Sufficiently small step size may avoid this scenario.

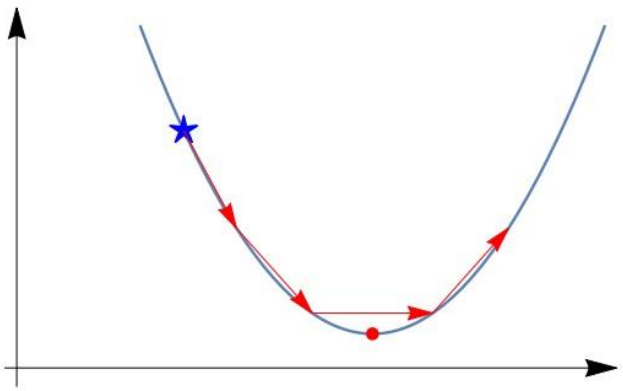

Figure 6

The following steps are usually performed in iterations in gradient descent technique

Steps:

1. Draw the line (tangent) from the point.
2. Find the slope.
3. Identify the change required in taking the derivatives (partial) of the function w.r.t. weights.
4. The change is multiplied with a variable called learning rate, denoted by α. Usually α is given the value 0.01.
5. Subtract the change value from earlier m value to get a new m value and again repeat the steps.

Let us try to understand the calculations in gradient descent which are indicated graphically in Figure 7.

Cost function involves M.S.E., so we start with M.S.E. given by

$$M.S.E. = \frac{1}{n}\sum_{i=1}^{n}(y_i - y)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (mx_i + b)\right)^2.$$

Take the partial derivatives of M.S.E. with respect to m and b,

$$\frac{\partial(M.S.E.)}{\partial m} = \frac{2}{n}\sum_{i=1}^{n}(-x_i(y_i - (mx_i + b))); \frac{\partial(M.S.E.)}{\partial b} = \frac{2}{n}\sum_{i=1}^{n}(y_i - (mx_i + b)).$$

Calculating the gradient of error with respect to m and b, we get

$$m = m - \alpha \frac{\partial(S.M.E.)}{\partial m}; b = b - \alpha \frac{\partial(S.M.E.)}{\partial b}.$$

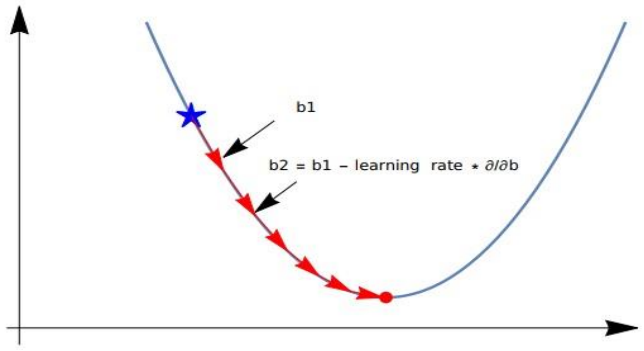

Figure 7

The learning rate α indicates how fast one can move down the slope. A bigger learning rate means bigger step size and a smaller learning rate means smaller step size. A bigger step size may miss the least square point and a smaller step size may take too long to optimize the model resulting in wastage of computational power. Hence a suitable size fits into the requirement to perform optimally.

Acknowledgement

Authors thank the anonymous reviewer(s) for valuable comments to improve the manuscript. Second author is financially supported by DST vide Application Reference Number DST/INSPIRE/02/2018/010561.

References

1. Deisenroth, M.P., Faisal, A.A. and Cheng, S.O., *Mathematics for Machine Learning*, Cambridge University Press, 2020.
2. Kohavi, R. and Provost, F., Glossary of terms, *Machine Learning, 30 (2 & 3), pp. 271–274, 1998*.
3. Nilsson, N., *Learning Machines*, McGraw Hill, 1965.
4. Samuel, Arthur (1959), Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal of Research and Development. 3(3), 210–229*.