Multivariable Calculus and its Applications in Artificial intelligence

Chong How, NUS High School

July 2020

1 Abstract

Mathematics plays an important role in modern science and technology. From speed of reaction and thermodynamics to electric circuits and computer network, the usage of math is found almost everywhere. In particular, math pushes the progress of science and technology, making advancements possible, while new problems arise because of unlocks in new scientific domains, this often cultivate the development of new Mathematical tools to analyse them. We will look into how Multivariable Calculus is related to Artificial Intelligence, such as Neural Network and Regression.

2 Introduction

"Mathematics is the queen of the sciences" - Carl Friedrich Gauss

Artificial Intelligence (AI) is closely linked to Mathematics. Why? Some keywords associated with AI are "machine learning" "algorithms" "neural network" and "graph theory". These domains are heavy in maths. Some of the related mathematical tools are Linear Algebra, Multivariable Calculus and Probability. We shall discuss Multivariable Calculus.

3 Multivariable Calculus

3.1 Coordinates and Functions

In secondary school, the 2-D Cartesian coordinate system (or xy coordinate system) is introduced. A function $y = f(x)$ is a function in x. For example, the function $y = x^2 - 1$ is a quadratic function that cuts through the x-axis at $x = 1$ and $x = -1$, and the y-axis at $y = -1$. In the Euclidean 3-D space, there is another axis — the z-axis which is orthogonal to xy axis and its positive direction determined by the right hand rule.

A function $z = f(x, y)$ is a function in x, y. The value of z depends on x and y. For example, the function $z = f(x, y) = x^2 + y^2$ is a elliptical paraboloid passing through (0,0,0).

3.2 Differentiation

Suppose $y = f(x)$, the derivative, $\frac{dy}{dx}$ is defined to be $\lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$. For example, $\frac{d}{dx}(2x^3) = 6x^2$

In 3-D coordinate, suppose $z = f(x, y)$, then the partial derivative $f_x(x, y)$ is defined to be $\lim_{h\to 0} \frac{f(x+h, y) - f(x, y)}{h} = \frac{\partial z}{\partial x}$. And the partial derivative $f_y(x, y)$ is defined to be $\lim_{h\to 0} \frac{f(x, y+h) - f(x, y)}{h} = \frac{\partial z}{\partial y}$.

What does this mean? Construct a plane $y = a$ that cuts through the surface of $z = f(x, y)$. $f_x(x, y)$ measures the rate of change of f on the plane in the direction of positive x-axis. $f_y(x, y)$ is interpreted the same way.

Example 1: Let $f(x, y) = x^2 + y^2$. Then $f_x(x, y) = 2x$, $f_y(x, y) = 2y$
Example 2: Let $f(x, y) = x^2y^2$. Then $f_x(x, y) = 2xy^2$,$f_y(x, y) = 2yx^2$

3.3 Gradient Vector and Directional Derivatives

Let f be a function of x and y. Then the gradient of f is the vector function

$$\nabla f(x, y) = \langle f_x(x, y), f_y(x, y)\rangle$$

and the directional derivative of f at (x_0, y_0) in the direction of a unit vector $\mathbf{u} = \langle a, b\rangle$ is

$$D_{\mathbf{u}} f(x_0, y_0) = \lim_{h\to 0} \frac{f(x_0 + ha, y_0 + hb) - f(x_0, y_0)}{h}$$

if this limit exists.

Example: Let $f(x, y) = x + y$, then $\nabla f(x, y) = \langle 1, 1\rangle$.

A useful property of $\nabla f(x_0, y_0)$ is that it provides the direction of the greatest rate of change of f at the point (x_0, y_0), which is $|\nabla f(x_0, y_0)|$. This let us to think: What if $D_{\mathbf{u}} f(x_0, y_0) = 0$ at some point (x_0, y_0) of f? What if we travel in the direction of $\nabla f(x_0, y_0)$?

4 Artificial Intelligence

Multivariable calculus has applications in AI development. These applications include:

Gradient Descent
Artificial Neural Networks
Maximising a expectation-maximization algorithm
Optimization problems
Finding maximal margin in support vector algorithm, etc

We will discuss the first two.

4.1 Linear Regression and Multiple linear Regression

Regression is a method in Statistics to model the relationship between a dependent variable and one or more independent variables. In Machine Learning, Regression Algorithm is a method to train the model. Simple Linear Regression predicts the dependent variable based on one independent variable. The aim of applying Linear Regression is to find a function $Y = mX + c$ and fit predicted values and actual values as close as possible. The difference between predicted and actual values are quantified by using a cost (sometimes called loss) function, which could be based on mean squared error algorithm.

Example: Suppose Y values $(\bar{y_1}, \bar{y_2}, \bar{y_3}) = (4.4, 4.8, 4.9)$ are predicted from a set of X values (x_1, x_2, x_3) through the relationship $Y = mX + c$, while the actual Y values are $(y_1, y_2, y_3) = (5.4, 5.8, 5.9)$, then the mean squared error is $\frac{1}{3}\sum_{i=1}^{3}(y_i - \bar{y_i})^2 = 1$.

What above Multiple Linear Regression? Multiple Linear Regression predicts the dependent variable based on two or more independent variable. The relationship is often given by $Y = m_1X_1 + m_2X_2 + ... + m_nX_n + c$.

4.2 Gradient Descent

Gradient descent is a iterative optimization algorithm to find the minimum value of a function. In this context, it is the cost function.

Imagine a man is going down a mountain. There is haze contributed by neighbourhood countries. He does not know how to get down. He tries to move to another point that is lower. When the slope of the mountain is steep, he takes a huge step, while if the slope is gentle, he take smaller steps. The next point is determined by previous point and once he reach the bottom, he stops this process, where hopefully he reached the bottom.

Now, Suppose we have a bunch of n data points. Define a cost function

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_i})^2 \tag{1}$$

Since $\bar{y_i} = mx_i + c$ We rewrite (1) as

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i + c))^2 \tag{2}$$

We wish to minimise E, i.e we need to find suitable values of m and c such that E is minimised.
To do so, we evaluate

$$\frac{\partial E}{\partial m} = \frac{-2}{n}\sum_{i=1}^{n} x_i(y_i - \bar{y_i}) \tag{3}$$

and

$$\frac{\partial E}{\partial c} = \frac{-2}{n}\sum_{i=1}^{n}(y_i - \bar{y_i}) \tag{4}$$

Can't we just solve $\frac{\partial E}{\partial m} = 0$ and $\frac{\partial E}{\partial c} = 0$? This method does work if the function is not complicated. It is not feasible for complicated functions, and unfortunately most error functions are complicated, and typically we also have a large data set, which makes solving $\frac{\partial E}{\partial m} = 0$ and $\frac{\partial E}{\partial c} = 0$ extremely difficult.

Learning rate, L, of model controls how much we modify our model every time. If the learning rate is too large, it cause the model to converge too quickly, while if the learning rate is too small, it may cause long training process.

To find the local minimum, we start by letting learning rate $L = 0.0001$, which controls how much m and c changes with each iteration. Let $m_1 = c_1 = 0$.
We plug in values of our data points and current m, c values into $\frac{\partial E}{\partial m}$ and $\frac{\partial E}{\partial c}$. Then, we update our m and c value, given by the recurrence relation:

$$m_{n+1} = m_n - L \times \frac{\partial E}{\partial m}$$

$$c_{n+1} = c_n - L \times \frac{\partial E}{\partial c}$$

As more iterations are being ran, finally we have $m = \lim_{n\to\infty} m_n$ and $c = \lim_{n\to\infty} c_n$. Hopefully, we have reached our desired linear relation $Y = mX + c$ that fits the actual value with predicted value optimally. Why? This is because using this method, we could have either reach a global minimum or local minimum of the cost function, which we would not know.

Often, the dependent variable that we wish to predict has relationship with

more than one variable. The idea is the same as Simple Linear Regression. For Gradient Descent of multiple variables, we will have to treat each variable separately by making all of the other variables constant and then find the partial derivative of the function.

Gradient Descent is also used in Neural networks.

4.3 Artificial Neural Networks and Machine Learning

As the name suggests, Artificial Neural Networks are inspired by the brain. They are sometimes also called models.

Imagine we have many neurons (or nodes). These neurons are organised in layers. Each neuron holds a number and a bias value. The layers of neurons are connected with neurons in other layers, forming a neural network.

A neural network accepts one or multiple inputs, processes it and give one or multiple outputs.

A neuron of one layer interacts with neuron of the next layer through weighted connections (a real valued number) between two neurons. Neuron in the next layer receives values of the previous layer neurons, each multiplied by their connection weight. The total sum of all those products plus that neuron's bias value is then put into an activation function, which then returns a value and assigns to that neuron. Information is passed through the entire network. The key is to decide on appropriate weights and bias values, which can be determined via machine learning. How?

Suppose we have an artificial neural network. We define a cost function

$$C = C(W)$$

where $W = (w_1, w_2, w_3, ..., w_n)$ is a matrix storing each weight/bias values. For each i, w_i is the value of one weight or bias in the network. C is our "error" and should be as low as possible. Clearly, we want to minimise the cost function, and this can be done by computing partial derivative with respect to each of the weights and bias in the network, which initially have any arbitrary initialised values that are subjected to change. We then modify the weight and bias values accordingly using the recurrence relation $W_{n+1} = W_n - \nabla C(W_n)$. By iterating this process, it is likely we can get to a point $C_{min} = \lim_{n\to\infty} C(W_n)$, At which C is minimised and the artificial neural network model is trained to give us our desired result.

This method that we have just discussed is an old variant of modern Machine Learning. It builds the foundation of more advanced concepts in Machine Learning. We will not be discussing that.

5 Concluding Remarks

It is truly remarkable how far technology has progressed over the last four decades. Artificial Intelligence especially, no doubt holds great potential in areas such as research and new development. According to a new report from the World Economic Forum (WEF), the growth of artificial intelligence could create 58 million net new jobs by the year 2022. Artificial Intelligence (AI) will define the next phase of the world's landscape, transforming our economy and society. Prevalence of AI is likely to rise in the next few decade. And it will dominate our lives, bringing immense benefits and also uncertainties into the future

Everything about AI is fascinating, yet without Mathematics, the study and development of AI will be impossible. This is yet another instance that Mathematics is inseparable from our modern world. I would like to end with the famous quote by Galileo Galilei: "In order to understand the universe you must know the language in which it is written and that language is mathematics."