Softmax and Sigmoid

Thursday, June 3, 2021 11:45 PM

The **softmax function**, also known as **softargmax** or **normalized exponential function** is a generalization of the logistic function to multiple dimensions. It is used in multinomial logistic regression and is often used as the last **activation function** of a neural network to normalize the output of a network to a probability distribution over predicted output classes, based on Luce's choice axiom.

The Sigmoid Activation Function

The Sigmoid Activation Function is a mathematical function with a recognizable "S" shaped curve. It is used for the logistic regression and basic neural network implementation. If we want to have a classifier to solve a problem with more than one right answer, the Sigmoid Function is the right choice. We should apply this function to each element of the raw output independently. The return value of Sigmoid Function is mostly in the range of values between 0 and 1 or -1 and 1.

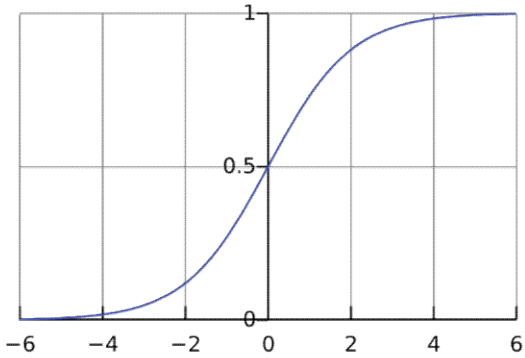
There is a wide range of these functions. Apart from Logistic, there is also a Hyperbolic Tangent Function that has been used in Artificial Neurons. Apart from this, it has also been used as a Cumulative Distribution Function. It is straightforward and reduces the time required for implementation. On the other hand, there is a significant drawback due to derivative having a short-range, which leads to significant information loss.

This is how the Sigmoid Function looks like:

$$f(x) = sigmoid(x) = \frac{1}{1 + e^{-x}}$$

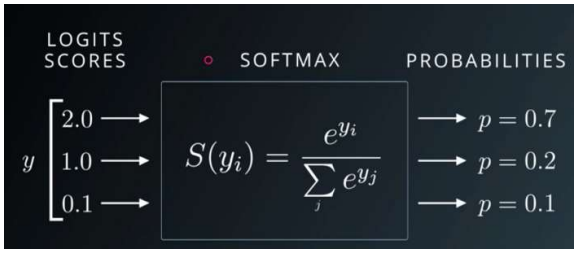
If there are more layers in our Neural Network, the more data is compressed and lost per layer, and this amplifies and causes significant data loss overall.

This is how the sigmoid function looks like:



The Softmax Activation Function

The Softmax Activation Function, also know as *SoftArgMax* or *Normalized Exponential Function* is a fascinating activation function that takes vectors of real numbers as inputs, and normalizes them into a probability distribution proportional to the exponentials of the input numbers. Before applying, some input data could be negative or greater than 1. Also, they might not sum up to 1. After applying Softmax, each element will be in the range of 0 to 1, and the elements will add up to 1. This way, they can be interpreted as a probability distribution. For more clarification, the larger the input number, the larger the probabilities will be.



Softmax is often used in:

- *Artificial and Convolutional Neural Networks* — Idea is to map the non-normalized output of data to the probability distribution for output classes. It is used in the final layers of neural network-based classifiers. They are trained under either the log-loss or cross-entropy regime. This way, the result is a non-linear variant of multinomial logistic regression (Softmax Regression).
- Other Multiclass Classification Methods such as *Multiclass Linear Discriminant Analysis, Naive Bayes Classifiers*, etc.
- *Reinforcement Learning* — Softmax function can be used to convert values into action probabilities.

Softmax is used for multi-classification in the Logistic Regression model, whereas Sigmoid is used for binary classification in the Logistic Regression model.

This is how the Softmax function looks like this:

$$\text{softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \text{ for j} = 1,\ldots,\text{K}$$

This is similar to the Sigmoid function. The difference is that, in the denominator, we sum together all of the values. To explain this further, when calculating the value of Softmax on a single raw output, we can't just look at one element alone, but instead, we have to take into account all the output data.

This is main reason why the Softmax is cool. It makes sure that the sum of all our output probabilities is equal to one.

For example, if we're classifying numbers and applying a Softmax to our raw outputs, for the Artificial Network to increase the probability that a particular output example is classified as "5", some other probabilities for other numbers (0, 1, 2, 3, 4, 6, 7, 8 and/or 9) needs to decrease.

Summary

*Characteristics of a **Sigmoid Activation Function***

- Used for Binary Classification in the Logistic Regression model
- The probabilities sum does not need to be 1
- Used as an Activation Function while building a Neural Network

*Characteristics of a **Softmax Activation Function***

- Used for Multi-classification in the Logistics Regression model
- The probabilities sum will be 1
- Used in the different layers of Neural Networks