

MULTILAYER PERCEPTRON

*Solution of linearly non-separable problems*

Frankfurt University of Applied Sciences

Presented by,

NARAYAN NARVEKAR

ABOBARIN OPEYEMI

*Under the guidance of Prof. Damir Dobric*

CONTENTS

1. Introduction

* What is a neural network?

1. Perceptron
2. Multilayer perceptron

* Architecture
* Stages

1. Project Details

* Link to repository
* Class MLPerceptronAlgorithm
* Class BackPropagationNetwork

1. Training process flowchart
2. Validation process flowchart
3. Training datasets
4. Conclusion
5. References

INTRODUCTION

**What is a neural network?**

Neural networks are one of the most beautiful programming paradigms ever invented. In the conventional approach to programming, the computer is told what to do, breaking big problems up into smaller ones, with the tasks being precisely defined that the computer can easily perform. In a neural network we don't tell the computer how to solve our problem. Instead, it learns from observational data, figuring out its own solution to the problems presented to it.

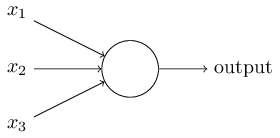
Learning from data automatically sounds promising. However, until 2006 there was no defined strategy to train neural networks to surpass more traditional approaches, except for a few specialized problems. What changed in 2006 was the discovery of techniques for learning in so-called deep neural networks. These techniques are now known as deep learning. They've been developed further, and today deep neural networks and deep learning have managed to achieve outstanding performance on many important problems in computer vision, natural language processing and speech recognition.

In the most general form, a neural network is a machine designed to model the way in which our brain can perform a particular task of interest. The heart of a neural network is the learning algorithm, whose function is to modify the weights of the neural network in an optimized manner to achieve the desired objective. How this is achieved would be explained in greater detail in the following chapters.

PERCEPTRON

**What is a Perceptron?**

A Perceptron, sometimes also referred to as an artificial neuron, takes several inputs and produces a binary output:



In the above example, the perceptron has three inputs, x1, x2 and x3. In general it could have more or fewer inputs. Key parameters of a perceptron are the ‘***weights’***.  Weights, usually denoted as w1, w2, w3 and so on, are real numbers which express the importance of the respective inputs to the output. The neuron's output, 0 or 1, is determined by whether the weighted sum ∑jwjxj  is less than or greater than some **‘threshold value’**. Similarly, the threshold is a real number which is a parameter of the neuron. This can be represented in more precise algebraic terms as follows:

**Output = { 0 if ∑jwjxj <= threshold**

**{ 1 if ∑jwjxj > threshold**

That's all there is to how a simple perceptron works.

The condition ∑jwjxj > threshold could become cumbersome, and following notational changes can be made to simplify it. The first change is to write ∑jwjxj as a dot product, w.x = ∑jwjxj, where ‘w’ and ‘x’ are vectors whose components are the weights and inputs, respectively. The other change is to move the threshold to the LHS of the inequality, and replace it by what's known as the perceptron's ‘***bias’***, b = −threshold. Using the bias instead of the threshold, the perceptron rule can be rewritten:  
 **Output = { 0 if w.x + b <= 0**

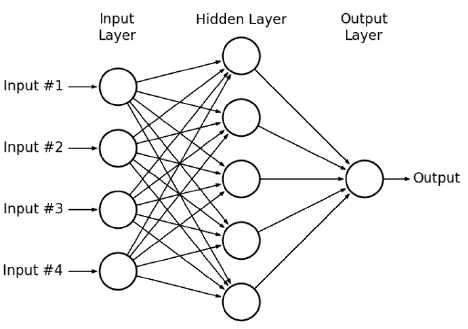
**{ 1 if w.x + b > 0**

The bias can be thought of as a measure of how easy it is to get the perceptron to output a 1. In other words, the bias is a measure of how easy it is to get the perceptron to fire. For a perceptron with a really big bias, it's extremely easy for the perceptron to output a 1. But if the bias is very negative, then it's difficult for the perceptron to output a 1.

MULTILAYER PERCEPTRON

**Architecture:**

A multilayer perceptron (MLP) is a class of feed forward artificial neural network.  A MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear [activation function](https://en.wikipedia.org/wiki/Activation_function). An example of a simple MLP neural network is shown below:



The leftmost layer in this network is called the input layer, and the neurons within the layer are called input neurons*.* The rightmost or output layer contains the output neurons, or, as in this case, a single output neuron. The middle layer is called a hidden layer, since the neurons in this layer are neither inputs nor outputs. The term "hidden" perhaps sounds a little mysterious, but it really means nothing more than "not an input or an output". The above network has only one hidden layer, but there could be networks with multiple hidden layers.

Every neuron in a particular layer is connected via a specific weight to every neuron in the following layer. In the above network, each one of the four neurons in the input layer would be connected to each of the five neurons in the hidden layers by a unique weight assigned to each connection. Likewise, each of the five neurons in the hidden layer would be connected to the single neuron in the output layer by unique weights. In addition to the weights, every neuron in the hidden and output layers would be assigned a bias, which along with all the weights, contributes to something known as the weighted sum that is calculated at each node at these layers.

**Stages:**

Following are the important stages that we encounter in a MLP neural network:

1. **Input a set of training examples:**

Training the MLP neural network is the most important step and in this context, we would need precisely defined ***training examples***, also known as ***training datasets***. These training datasets consist of a set of inputs and also the actual outputs that correspond to these inputs. These datasets are then fed to the MLP neural network, which is thus trained by updating the weights and biases at each neuron such that the error at the output is minimized.

1. **Feedforward path:**

For each training example, set the corresponding input activation ax,1. In general, we can represent the activations at each neuron in the network as follows:

ax,l, where ‘x’ refers to the (x+1)th neuron(since we start numbering the neurons from zero and not one) and ‘l’ refers to the number of the layer that we are referring to.

For eg., a1,1 is the input activation at the 2nd neuron of the 1st layer(input layer) and a3,2 is the input activation at the 4th neuron of the 2nd layer(1st hidden layer).

If L is the total number of layers in the network, then for each l = 2, 3..., L, the weighted sum is computed at each neuron as follows:

**zx,l = wlax,l + bl**

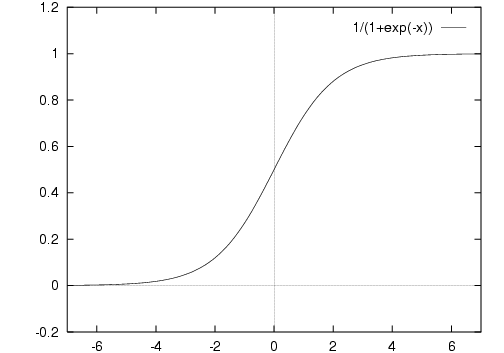
where, wl is the weight matrix which represents all the weights connecting the neurons in the lth layer to the neurons in the (l-1)th layer, ax,l is the vector of activations input to all the neurons in the lth layer and bl is the vector of biases applied to all the neurons in the lth layer.

This weighted sum is then used to calculate the output at each neuron by passing it through the ***activation function, σ***.

The two activation functions that are supported by the project are both sigmoid functions and are described by:

**σ(x) = tanh(x) and σ(x) = (1 + e-x)-1**

The first is a [hyperbolic tangent](https://en.wikipedia.org/wiki/Hyperbolic_tangent) that ranges from -1 to 1, while the other is the [logistic function](https://en.wikipedia.org/wiki/Logistic_function), which is similar in shape but ranges from 0 to 1. Consider the second function. Here is the shape of the function:

******

One big difference between perceptrons and sigmoid neurons is that sigmoid neurons don't just output 0 or 1. They can output any real number between 0 and 1, so values such as 0.23, 0.45 are legitimate outputs. This can be useful if we want to use the output value to represent the average intensity of the pixels in an image input to a neural network. But sometimes it can be a nuisance. Suppose we want the output from the network to indicate either "the input image is a 5" or "the input image is not a 5". Obviously, it'd be easiest to do this if the output was a 0 or a 1, as in a perceptron. But in practice we can set up a convention to deal with this, for example, by deciding to interpret any output of at least 0.5 as indicating a "5", and any output less than 0.5 as indicating "not a 5".

The outputs thus calculated serve as input activations to the neurons in the following layer. At the end of the feedforward path stage, we have thus calculated outputs at each of the neurons in the network, terminating with the output layer neurons.

1. **Output layer error:**

At the end of the feedforward path when the output is calculated at the output neuron, this value represents the “calculated output” of the network. The goal of the MLP learning algorithm is to calculate the error between the calculated output and the actual output of the different training datasets and minimize this error by adjusting the weights and biases in the network. The error at the output neuron is called the cost function. In this project we have used the quadratic cost and this is given by:

**C = (1/2n) \* ∑ | y(x) – aL(x) |2**

where n is the total number of training examples, y(x) is the actual output which is a part of the training dataset and aL(x) is the calculated output of the MLP network.

The output error for each neuron in the output layer is given by the equation:

**δx,L = ∇aCx ⊙ σ|(zx,L)**

where ∇aCxis vector which represents the gradient of the quadratic cost at each output neuron with respect to its corresponding input activation, σ|(zx,L)is the vector which represents the derivative of the sigmoid function applied to the weighted inputs calculated at the output layer. The symbol “⊙” represents the Hadamard product.

The Hadamard product between the above two vectors gives the error at the output layer, δx,L.

1. **Backpropagation:**

The error that is calculated at the output layer is then distributed backwards towards all the neurons in the network. In other words, the error is propagated backwards from the output layer to all the hidden layer neurons, starting from the last hidden layer and ending at the first hidden layer. Hence the name backpropagation.

For each l = L-1, L-2,…, 2, we need to compute the error at each neuron. This error is given by,

**δx,l = ((wl+1)T δx,l+1 ) ⊙ σ|(zx,l)**

where (wl+1)T is the matrix of weights at the (l+1)th layer, δx,l+1 is the vector of errors at the (l+1)th layer, σ|(zx,l)is the vector which represents the derivative of the sigmoid function applied to the weighted inputs calculated at the lth layer.

Once the errors have been determined and back propagated to each neuron in the hidden layers, the next step is to calculate the rate of change of the cost function with respect to all the weights and biases in the network.

The rate of change of cost function with respect to the biases is given by,

**∂C/∂blj = δlj**

where blj represents the bias of the (j+1)th neuron of the lth layer and δlj is the error calculated at the (j+1)th neuron of the lth layer.

This means that the error δlj is exactly equal to the rate of change ∂C/∂blj.

Similarly, the rate of change of cost function with respect to the weights is given by,

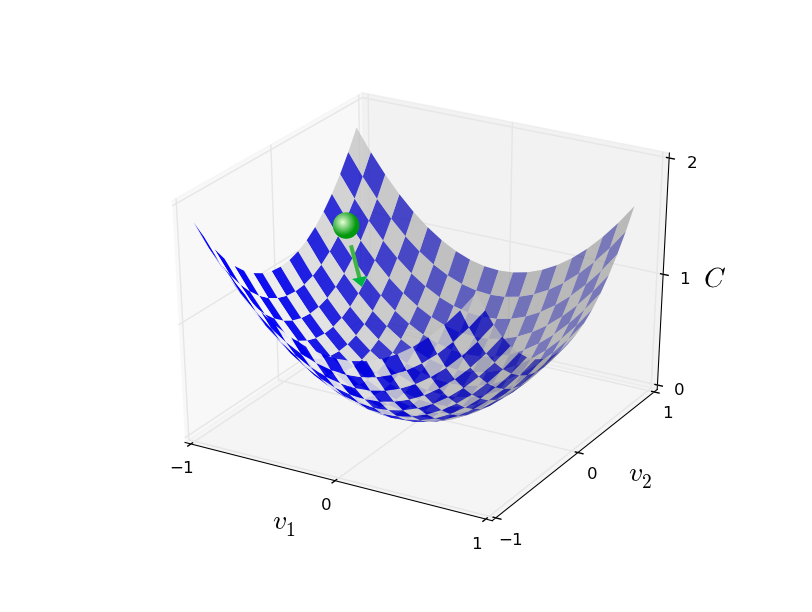
**∂C/∂wljk = akl-1 δlj**

where wljk represents the weight connecting the (j+1)th neuron in the lth layer to the (k+1)th neuron in the (l-1)th layer, akl-1 is the calculated output at the (k+1)th neuron of the (l-1)th layer and δlj is the error calculated at the (j+1)th neuron of the lth layer.

1. **Gradient descent:**

The main purpose of the MLP learning algorithm is to find a set of weights and biases that makes the cost as small as possible. This can be achieved by a process known as gradient descent. The working principle of gradient descent is explained below:

The quadratic cost is a function of weights and biases. Consider the below figure which shows an example of the quadratic cost, C, taking on a set of values for different values of weights, v1 and biases, v2.



We will succeed when the cost function has reached its minimum value. If suppose the cost function has a current value which is represented by the ball in the figure, then the slope of the tangent at that particular point will give us a direction to move towards. We then make steps down the cost function in the direction with the steepest descent. The size of each step is determined by the learning rate α.

After each step is taken, we have a new set of weights and biases. Continuing this procedure of taking steps in the direction of the steepest descent, finally we end up at a point where the cost function is minimum. The weights and biases which are thus calculated when the cost function is minimum are said to be the optimal weights and biases of the network. Based on this principle, we can come up with the following equations:

For l = L, L-1, L-2,…, 2, the biases are updated according to the following rule,

**blj -> blj – (α δlj)**

Similarly, the weights are updated according to the following rule,

**wljk -> wljk – (α akl-1 δlj)**

These updates are carried out until all the training examples have been exhausted and all the defined number of iterations have been executed.

PROJECT DETAILS

**Link to Repository**

Algorithm: *SE-2017-2018\SE\Projects\Narayan Narvekar\Prototyping\ Task.Jan\MLPerceptron\MLPerceptron\MLPerceptron*

Test Project: *SE-2017-2018\SE\Projects\Narayan Narvekar\Prototyping\ Task.Jan\MLPerceptron\MLPerceptron\MLPUnitTests*

CSV files: *SE-2017-2018\SE\Projects\Narayan Narvekar\Prototyping\ Task.Jan\MLPerceptron\MLPerceptron\MLPUnitTests\TestFiles*

**Class MLPerceptronAlgorithm**

This class contains the MLP learning algorithm. A brief description of the important methods contained in this class is given below:

1. ***IScore Run(double[][] featureValues, IContext ctx):***

This method accepts the training examples as input and performs the training of the MLP neural network.

The weights and biases are initialized to a random value between 0 and 1.

For each training example, the feedforward path is implemented by calculating the outputs at each neuron.

For each training example, a BackPropagationNetwork object is instantiated, and the methods to calculate the errors at each node, and cost function changes with weights and biases are executed. The new weights and biases are updated in the methods defined in BackPropagationNetwork class.

1. ***double[] Predict(double[][] data, IContext ctx)***

This method accepts the test data as input and determines the output for each sample of test data.

The return value of the function is the calculated output. The return value is then compared in the unit test method with the correct outputs, and thus used to determine whether the unit test has passed or failed.

**Class BackPropagationNetwork**

This class contains methods that implement the back propagation part of the MLP learning algorithm. A brief description of the important methods contained in this class is given below:

1. ***void ErrorOutputLayer(double[] calculatedop, int[] hiddenlayerneurons, double[] weightedip, double[] featureValues)***

This method calculates the error at the output layer as defined by the equation,

**δx,L = ∇aCx ⊙ σ|(zx,L)**

1. ***void ErrorHiddenLayers(double[][] calculatedop, double[][,] weights, int[] hiddenlayerneurons, double[][] weightedip, double[] featureValues)***

This method calculates the error that is back propagated at each neuron present in the hidden layers as defined by the equation,

**δx,l = ((wl+1)T δx,l+1 )⊙ σ|(zx,l)**

1. ***void CostFunctionChangeWithWeights(double[][,] currentweights, double[][] calculatedop, int[] hiddenlayerneurons, double learningrate, double[] featurevalues, out double[][,] newweights)***

This method calculates the rate of change of cost function with respect to all the weights that are a part of the network as defined by the equation,

**∂C/∂wljk = akl-1 δlj**

This method also calculates the newly update weights as defined by the equation,

**wljk -> wljk – (α akl-1 δlj)**

1. ***void CostFunctionChangeWithBiases(double[][] currentbiases, int[] hiddenlayerneurons, double learningrate, out double[][] newbiases)***

This method calculates the rate of change of cost function with respect to all the biases that are a part of the network as defined by the equation,

**∂C/∂blj = δlj**

This method also calculates the newly update biases as defined by the equation,

**blj -> blj – (α δlj)**

TRAINING PROCESS FLOWCHART



Using gradient descent update the new weights and biases based on results from back propagation

Execute the Back Propagation path on the same training example

Execute the feedforward path on a training example

Initialize the weights and biases with random values between 0 and 1

Call the Run method by passing the training examples as input parameters

No Yes No Yes

Repeat the training procedure on all training examples

Stop training the network

All iterations completed?

Loop to next training example

All training examples exhausted?

VALIDATION PROCESS FLOWCHART

Invoke the predict method from the unit test case. Pass the test data as parameter

All test data exhausted?

Loop to next test data

Compare the calculated result with the correct output

Calculate the output for a test data sample and update it in the result array

No

Yes

Return the result array

Calculated result == correct output?

Yes No

Unit test failed

Unit test passed

TRAINING DATASETS

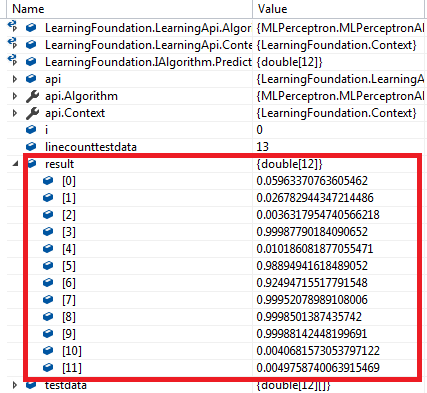
**Unit test 1:** The first unit test consists of the following training data pattern. Blue dots have the label of 0 and red dots have the label of 1.

The MLP learning algorithm was run on the above training data set, with Learning rate = 0.1, number of iterations = 10000, hidden layer neurons = (6) and sigmoid activation function.

The test data is specified in the below table:

|  |  |  |
| --- | --- | --- |
| X | Y | Output |
| 0.3 | 0.01 | 0 |
| 0.09 | 0.09 | 0 |
| 0.11 | 0.98 | 0 |
| 0.65 | 0.25 | 1 |
| 0.4 | 0.82 | 0 |
| 0.25 | 0.45 | 1 |
| 0.26 | 0.38 | 1 |
| 0.38 | 0.6 | 1 |
| 0.48 | 0.45 | 1 |
| 0.88 | 0.15 | 1 |
| 0.5 | 0.88 | 0 |
| 0.6 | 0.81 | 0 |

The attached snapshot indicates the results calculated on the test data by the MLP learning algorithm:



Applying the rule that a value greater than or equal to 0.5 represents a 1 and less than 0.5 represents a zero and comparing the calculated results with the actual output from the test data table, we see that the MLP learning algorithm has correctly classified all the test data.

**Unit test 2:** The second unit test consists of the “Iris” dataset which classifies iris plants into three species. It includes 100 samples distributed between three species, with some properties about each flower. The columns in this dataset are:

* SepalLengthCm
* SepalWidthCm
* PetalLengthm
* PetalWidthCm
* Species

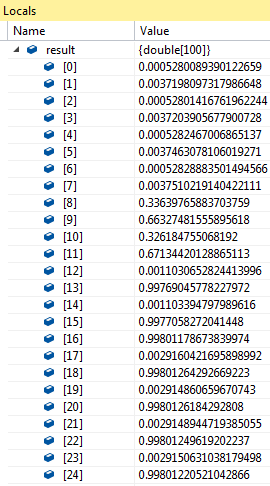
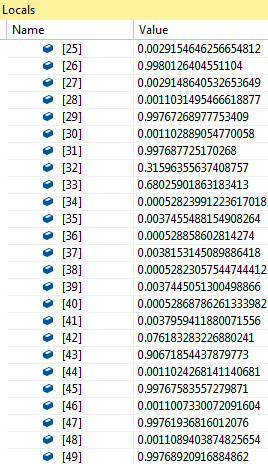
The data is represented graphically as follows:

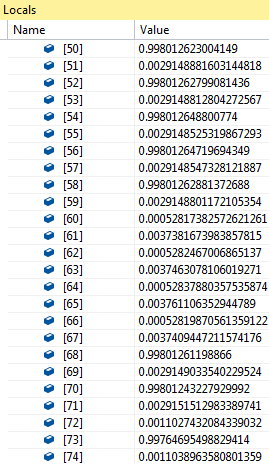
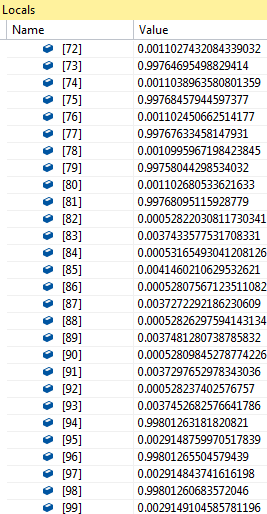
The MLP learning algorithm was run on the above training data set, with Learning rate = 0.1, number of iterations = 10000 and hidden layer neurons = (3, 4, 3) and hyperbolic tangent activation function for the hidden layers and sigmoid activation function for the output layer.

The test data is specified in the below table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SepalLength | SepalWidth | PetalLength | PetalWidth | Output |
| 5.2 | 4.1 | 1.5 | 0.1 | setosa |
| 5.5 | 4.2 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 5 | 3.2 | 1.2 | 0.2 | setosa |
| 5.4 | 3 | 4.5 | 1.5 | versicolor |
| 6 | 3.4 | 4.5 | 1.6 | versicolor |
| 6.7 | 3.1 | 4.7 | 1.5 | versicolor |
| 6.3 | 2.3 | 4.4 | 1.3 | versicolor |
| 6.1 | 2.6 | 5.6 | 1.4 | virginica |
| 7.7 | 3 | 6.1 | 2.3 | virginica |
| 6.3 | 3.4 | 5.6 | 2.4 | virginica |
| 6.4 | 3.1 | 5.5 | 1.8 | Virginica |
| 6 | 3 | 4.8 | 1.8 | Virginica |
| 6.9 | 3.1 | 5.4 | 2.1 | Virginica |
| 5.6 | 3 | 4.1 | 1.3 | Versicolor |
| 5.5 | 2.5 | 4 | 1.3 | Versicolor |
| 5.5 | 2.6 | 4.4 | 1.2 | Versicolor |
| 4.4 | 3.2 | 1.3 | 0.2 | Setosa |
| 5 | 3.5 | 1.6 | 0.6 | Setosa |
| 5.1 | 3.8 | 1.9 | 0.4 | Setosa |
| 4.8 | 3 | 1.4 | 0.3 | setosa |
| 6.1 | 3 | 4.6 | 1.4 | versicolor |
| 5.8 | 2.6 | 4 | 1.2 | versicolor |
| 5 | 2.3 | 3.3 | 1 | versicolor |
| 5.6 | 2.7 | 4.2 | 1.3 | versicolor |
| 6.8 | 3.2 | 5.9 | 2.3 | virginica |
| 6.7 | 3.3 | 5.7 | 2.5 | virginica |
| 6.7 | 3 | 5.2 | 2.3 | virginica |
| 6.3 | 2.5 | 5 | 1.9 | virginica |
| 6.5 | 3 | 5.2 | 2 | virginica |
| 5.5 | 3.5 | 1.3 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 4.4 | 3 | 1.3 | 0.2 | setosa |
| 5.1 | 3.4 | 1.5 | 0.2 | setosa |
| 6.2 | 3.4 | 5.4 | 2.3 | virginica |
| 5.9 | 3 | 5.1 | 1.8 | virginica |
| 5.7 | 3 | 4.2 | 1.2 | versicolor |
| 5.7 | 2.9 | 4.2 | 1.3 | versicolor |
| 6.2 | 2.9 | 4.3 | 1.3 | versicolor |
| 5.1 | 2.5 | 3 | 1.1 | versicolor |
| 5.7 | 2.8 | 4.1 | 1.3 | versicolor |
| 5 | 3.5 | 1.3 | 0.3 | setosa |
| 4.5 | 2.3 | 1.3 | 0.3 | setosa |
| 5.1 | 3.8 | 1.6 | 0.2 | setosa |
| 4.6 | 3.2 | 1.4 | 0.2 | setosa |
| 5.3 | 3.7 | 1.5 | 0.2 | setosa |
| 5 | 3.3 | 1.4 | 0.2 | setosa |
| 6.7 | 3.1 | 5.6 | 2.4 | virginica |
| 6.9 | 3.1 | 5.1 | 2.3 | virginica |
| 5.8 | 2.7 | 5.1 | 1.9 | virginica |

The attached snapshots indicate the results calculated on the test data by the MLP learning algorithm:

**Unit test 3:** This test case uses the German Dataset as the training dataset. This dataset classifies people described by a set of attributes as good or bad credit risks. The following table describes the original attributes in the German dataset:

|  |  |  |
| --- | --- | --- |
| Number | Description | Class |
| Attribute 1 | Status of existing checking account | Qualitative |
| Attribute 2 | Duration in month | Numerical |
| Attribute 3 | Credit History | Qualitative |
| Attribute 4 | Purpose | Qualitative |
| Attribute 5 | Credit amount | Numerical |
| Attribute 6 | Savings account/bonds | Qualitative |
| Attribute 7 | Present employment since | Qualitative |
| Attribute 8 | Installment rate in percentage of disposable income | Numerical |
| Attribute 9 | Personal status and sex | Qualitative |
| Attribute 10 | Other debtors/ guarantors | Qualitative |
| Attribute 11 | Present residence since | Numerical |
| Attribute 12 | Property | Qualitative |
| Attribute 13 | Age in years | Numerical |
| Attribute 14 | Other installment plans | Qualitative |
| Attribute 15 | Housing | Qualitative |
| Attribute 16 | Number of existing credits at this bank | Numerical |
| Attribute 17 | Job | Qualitative |
| Attribute 18 | Number of people being liable to provide maintenance for | Numerical |
| Attribute 19 | Telephone | Qualitative |
| Attribute 20 | Foreign worker | Qualitative |

Some of these attributes are numerical and, but others are qualitative and cannot be computed in the training of neural networks. Thus, a numerical version of the dataset is used in this unit test. It transforms all qualitative variables to numeric and adds four more attributes.

The dataset contains 1000 instances, with 700 approved application cases and 300 rejected ones. 800 of these are a part of the training dataset and 200 are a part of the testing dataset. Each dataset group contains the same ratio of approved and rejected applications, i.e., 7:3.

Following table gives the results of the trained MLP network for different number of hidden neurons:

|  |  |  |
| --- | --- | --- |
| hidden layer neurons | Number of credit approvals classified correctly from 140 samples | Number of credit rejections classified correctly from 60 samples |
| 6 | 130 | 23 |
| 7 | 125 | 21 |
| 8 | 126 | 20 |
| 9 | 131 | 24 |
| 10 | 130 | 22 |
| 11 | 130 | 20 |
| 12 | 130 | 17 |
| 13 | 133 | 21 |
| 14 | 129 | 19 |
| 15 | 133 | 16 |
| 16 | 129 | 19 |
| 17 | 133 | 19 |
| 18 | 132 | 17 |
| 19 | 135 | 18 |
| 20 | 130 | 19 |
| 21 | 133 | 16 |
| 22 | 133 | 14 |
| 23 | 132 | 14 |
| 24 | 133 | 14 |
| 25 | 130 | 14 |

From the above table, the observation is that the model with 9 hidden neurons gives the best results, with a correct result percentage of 77.5%.

The number of hidden neurons can have a great impact on the network performance. Too few hidden neurons cannot solve complex credit rating problems, while too many of them may result in low efficiency and accuracy.

CONCLUSION

The multilayer perception performs a good job at solving linearly non separable problems. Some of the key factors which affect the performance of an MLP network are listed down below:

**Activation function:**

The Sigmoid and HyperbolicTan functions are the activation functions that are being supported by the MLP network.

The main reason for choosing Sigmoid is because it exists between 0 and 1 and hence it is especially used for models where we have to predict the probability of the output. However, if the input to the Sigmoid function is too large, then it gives rise to the problem of “vanishing gradients”. The network refuses to learn or is drastically slow.

In case of the HyperbolicTan function, the characteristics are similar to sigmoid. The gradient however is stronger for HyperbolicTan than sigmoid (derivatives are steeper). Deciding between sigmoid or tanh will depend on the requirement of gradient strength. Like Sigmoid, HyperbolicTan function also has the vanishing gradient problem.

To solve the problem of vanishing gradients, the Rectified Linear Unit (ReLU) activation function is used. ReLUs are much simple computationally. ReLUs only saturate when the input is less than 0. And even this saturation can be eliminated by using leaky ReLUs. For very deep networks, saturation hampers learning, and so ReLUs provide a nice workaround.

**Gradient Descent vs. Stochastic Gradient Descent:**

In both gradient descent and stochastic gradient descent, we update the weights and biases in an iterative manner to minimize an error function.

While in gradient descent, we run through all the samples in our training dataset to do a single update for a parameter in a particular iteration, in stochastic gradient descent we use only one or a subset of training samples from the training dataset to perform the update of the parameter in a particular iteration.

Stochastic gradient descent often converges much faster compared to gradient descent but the error function is not as well minimized as in the case of gradient descent. In most cases, the close approximation that we get in stochastic gradient descent for the parameter values are enough because they reach the optimal values and keep oscillating there.

Another advantage of stochastic gradient descent is that the randomness or noise introduced by it allows escaping from local minima to reach a better minimum.

**Learning rate:**

Choosing of learning rate helps us in determining how fast we can reach the minima. If we choose larger value, then we might overshoot the minima and smaller values might take long time for convergence.

In general, we need to find a learning rate that is low enough for the network to converge to something useful, but high enough such that we don’t have to spend years training it.

**Epoch:**

One Epoch (iteration) is when the entire training dataset is passed forward and backward through the neural network only once.

It doesn’t make sense to pass the entire dataset through a neural network only once. We need to pass the full dataset multiple times to the same neural network. Gradient descent is an iterative process and updating the weights and biases with one epoch is not enough. There is no hard and fast rule on the right number of epochs for a network. The answer is different for different datasets and is generally dependant on how diverse the data is.

**Number of hidden layers:**

One of the common rules of thumb is that *“the optimal size of the hidden layer is usually between the size of the input and size of the output layers”.* Also, the situations in which the performance improves with a second (or third, etc.) hidden layer are very few. One hidden layer is sufficient for majority of the problems.

In general, we can get decent performance by setting the hidden layer configuration using the below two rules:

* Number of hidden layers equals one
* Number of neurons in that layer is the mean of the neurons in the input and output layers.

REFERENCES

1. <https://en.wikipedia.org/wiki/Multilayer_perceptron>
2. Neural Networks and Deep Learning – Michael Nielsen
3. Neural Networks, A Comprehensive Foundation – Simon Haykins
4. Investigation and improvement of multi-layer perceptron neural networks for credit scoring – Zongyuan Zhao, Shuxiang Xu, Byeong Ho Kang, Mir Md Jahangir Kabir, Yunling Liu, Rainer Wasinger
5. [Neural network approach to Iris dataset](https://www.kaggle.com/louisong97/neural-network-approach-to-iris-dataset) – Louis Ong
6. NET & C# In-Browser tutorial – Microsoft