Performance of Multi-Layer Perceptron in Classification of Graduates' Salary

Raj Choudhary*

*School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127
Email: raj.choudhary2016@vitstudent.ac.in *

Abstract: **A multilayer perceptron (MLP) is a class of feedforward artificial neural network. An 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. It is based on the idea of support vector machines (SVM's). An upper bound on the Vapnik–Chervonenkis (VC) dimension is iteratively minimized over the interconnection matrix of the hidden layer and its bias vector. The output weights are determined according to the support vector method.**

Index Terms—**MLP, Artificial Neural Networks, Supervised Learning, Classification, Employability**

1. Introduction

Artificial neurons represent the building blocks of the multi-layer artificial neural networks that we are going to discuss in this paper. The basic concept behind artificial neural networks was built upon hypotheses and models of how the human brain works to solve complex problem tasks.

The figure (Fig. 1) explains the concept of an MLP consisting of three layers: one input layer, one **hidden layer**, and one output layer. The units in the hidden layer are fully connected to the input layer, and the output layer is fully connected to the hidden layer, respectively. If such a network has more than one hidden layer, we also call it a *deep* artificial neural network.
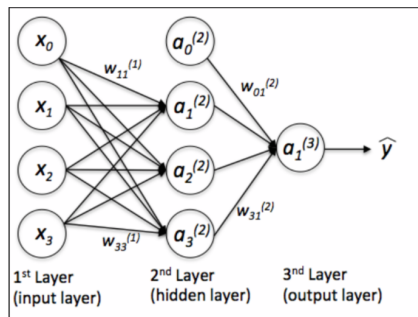

Figure 1 Multilayer Perceptron Model

Training a MLP model involves adjusting the parameters, or the weights and biases, of the model in order to minimize error. **Backpropagation** is used to make those weigh and bias adjustments relative to the error, and the error itself can be measured in a variety of ways, including by root mean squared error (RMSE). Neurons in MLPs calculate a smoothed variant of discontinuous function:

$$\vec{x} \mapsto f_{log}(w_0 + \langle \vec{w}, \vec{x} \rangle)$$

with flog is called **logistic function.**

2. Methodology

The Aspiring Minds' Employability Outcomes 2015 dataset is collected from the Aspiring Minds official website. The data is cleaned of missing values and anomalous data points. The attributes are made numeric and categorical for the computation of split points by the MLP. In each of the *forward passes*, the signal flow moves from the input layer through the hidden layers to the output layer, and the decision of the output layer is measured against the actual truth labels. On the other hand, in the *backward pass*, using backpropagation and the chain rule of calculus, partial derivatives of the error function w.r.t. the various weights and biases are back-propagated through the MLP which gives us a gradient, or a landscape of error, along which the parameters may be adjusted as they move the MLP one step closer to the error minimum.

The performance of the multi-layer perceptron model is evaluated on the Aspiring Minds' Employability Outcomes 2015 dataset. It aims to predict the salary of an engineering graduate fresh out of college given the candidates' academic background, standardized test performance and personality test scores.

3. Dataset – AMEO 2015

For every engineer, AMEO dataset provides anonymized bio data information along with their respective skill scores and employment outcome information. Specifically, the following information is available for every engineer: AMEO 2015 has gained traction since its public release. Aspiring Minds annually publishes the National Employability Report, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset.

4. Analysis and Result

Here, on learning the MLP model based on various parameters and activation function we have observed that the *hyperbolic-tan (tanh)* function

when chosen as an activation function of the model, it performed well in classifying the data points to their correct salary levels compared to the predictions done by model using *identity* as an activation function. We also observed that the accuracy of classification depends on the number of iterations and most importantly the number of layers in the MLP too. On layer, the output prediction varies with a huge extent. For instance, MLP model with 4 hidden layers each having 4, 3, 2 and 1 neurons respectively and identity as an activation function, then on performing 5000 iterations, the model was able to classify the data points in two of the salary level (1 and 4) correctly with an accuracy score of 33.8% approx. which is good, but the problem here was the data points in the salary level 2 and 3 are highly misclassified. So, on changing the activation function to 'tanh' and changing the number of neurons in each layer to 7, 6, 6 and 4 respectively, this model was able to classify data points of all the 4 salary levels with an accuracy of 30% on performing the same number of iterations which is far more better than the previous model since we are not misclassifying all points belonging to any of the 4 categories of salary.

Hidden layer	Number of iterations	Activation function	Accuracy
(4,3,2,1)	1000	Identity	27.75%
(4,3,2,1)	5000	Identity	25.5%
(7,6,6,4)	6500	Tanh	31.87%
(6,5,5,3)	5000	Tanh	31.87%

From the above table we are able to infer that as we increase the dimension of the hidden layers and the number of iterations, we get better accuracy for the AMEO dataset.

5. Conclusion

Hence, we have discovered some interesting insights into a typical engineering graduate and his/her job prospects. One of the curious observations is the large overlap of data points, as we are still getting an maximum accuracy of at most 35% using such a complex ANN model like MLP, which will prove to be a challenge during the model development phase. The analytics can also suggest graduate's methods of improving their job prospects.

6. Reference

[1]. V. Aggarwal, S. Srikant, and H. Nisar, "Ameo 2015: A dataset com- prising amcat test scores, biodata details and employment outcomes of job seekers," in AMEO 2015: A Dataset Comprising AMCAT Test Scores, Biodata Details and Employment Outcomes of Job Seekers. ACM, 2016.