Evaluating Multi-Layer Perceptron Regression Versus Non-Neural Network Regression

Conor Farrell   
College Electronic Engineering  
University of GalwayGalway, Ireland  
c.farrell42@universityofgalway.ie

*Abstract*—

# Introduction

# Background

## The Dataset

The dataset comes from a 1998 paper in Cement and Concrete Research. It is made up of 1000 datapoints with 8 measurements for each datapoint: Cement(kg/m3), Fly ash (kg/m3), Blast furnace slag(kg/m3), water(kg/m3), Superplasticizer(kg/m3), Coarse aggregate(kg/m3), Fine aggregate(kg/m3), Age of testing(days). The components will need to be normalised as some components have a mean of 3.5kg/m3 (Superplasticizer) and others a mean of 943.5kg/m3 (Coarse Aggregate).

The original paper mentions that a portion of the dataset was removed from the original testing due to large aggregates ( >20mm) and other reasons, this cut the original dataset down to ~700 samples, the provided dataset does not distinguish between the original dataset and the simplified dataset. The evaluation will have to continue with a note that some of the data could be non-standard. Additionally, the Superplasticizer is not the same material for each sample and could have more of an impact than suggested by the mass.

## Multi-Layer Perceptron Regression (MLP)

Multi-Layer Perceptron regression is a neural network for continuous output prediction. The architecture is made up of an input layer, a series of hidden layers and an output layer. Each layer is made up of neurons, where a neuron takes the inputs, processes it through a series of weights and activation functions. Each neuron is connected to every neuron in adjacent layers.

There are multiple activation functions which can affect the quality of the regression model:

D

### Rectified Linear Unit (ReLU)

Is a computationally efficient function that outputs the input directly for any positive input. Represented as:

It’s a simple function which can make training quick but can suffer issues such as dying ReLU where neurons can become permanently inactive.

### Gaussian Error Linear Unit (GeLU)

GeLU is a probabilistic activation function, weighting the inputs by the probability of the input occurring. Represented as:

Where P(X ≤ x) is the probability of x occurring.

GeLU has a similar shape as ReLU except gated based on probability rather than sign.

### Leaky ReLU

Leaky ReLU is a method to solve the dying ReLU problem. Instead of 0 for negative values the value is multiplied by a constant α (~0.01). Represented as:

Leaky ReLU keeps the performance benefits or ReLU whilst fixing some of the limitations.

### Hyberbolic Tangent (Tanh)

Is a sigmoidal activation function, going from -1 to +1 and zero-centred. Represented as:

Tanh suffers from vanishing gradients for the most extreme values.

## Linear Regression

Linear Regression is a technique where a prediction is made by multiplying each factor by a scalar and adding the output. Represented by:

Where βn is a scalar coefficient and Xn is an input variable and ε is the error between the predicted value and the actual value. The coefficients are usually iteratively determined by using the method of least squares. Linear regression is an intuitive model where determinations are easy to determine based on the inputs however suffers when modelling non-linear relationships.

## Decision Tree Regression

Decision Tree Regression is a tree-based regression technique that recursively partitions the feature space to learn the relationship between the features and the target. The dataset is split using the feature that provides the best criterion such as lowest mean squared error. This creates two datasets which will be similarly split, this process continues until a stopping criterion is met e.g. maximum depth. The final nodes on the tree are known as leaf nodes, the prediction is the mean or median value of the training samples in that region.

## Metrics

A range of metrics are needed to evaluate the quality of a model.

### R2 Predicted vs Actual

A graph of the predicted value vs the actual value of a perfect model will be a straight line at 45° from the origin with an R2 of 1.0. The further the model is from this line the lower the R2 value will be and the greater the variance is.

### Mean Absolute Error (MAE)

Average of the absolute difference between the predicted value and the actual value. The MAE will be a value in the same units as the output data i.e. Concrete predictions are 20GPa off on average.

### Root Mean Squared Error (RMSE)

Is the square root of the average of the squared differences between the predicted and actual values. The values are the same units as the original output, but the data is more sensitive to outliers, 68% of values will be in the range of the reported value.

### Mean Absolute Percentage Error (MAPE)

Is a similar metric to MAE, taking the average percentage difference between the actual value and the predicted value. It’s scale independent, which makes it useful if the values can go over multiple orders of magnitude.

# Methodology

The dataset was pre-processed before running the models. An 80% training set and a 20% test set were used as the dataset of ~1000 datapoints allowed for sufficient testing. In the case of the MLP models the dataset was further split into a training set and validation set, which was used to verify that the model was stabilising and not overfitting the training set. This allowed for variable Epoch amounts.

## Non-Neural Network Regression

A linear regression and decision tree regression were created, and the metrics were used to determine the baseline minimum performance a successful MLP regression must meet.

[Complete after testing done]

## MLP Testing

The MLP quality was determined for different activation functions and different architectures for the hidden layers. The best architecture for each activation function was used to report the quality and compare with the Non-neural network functions.

The activation functions being compared are:

* ReLU
* Leaky ReLU
* GeLU
* Tanh

The architectures being tested are:

* IN-64-OUT
* IN-64-32-OUT
* IN-32-32-OUT
* IN-128-64-32-OUT
* IN-256-128-64-32-OUT

The determination for the best algorithm will be based on all 4 metrics mentioned in the background section with the RMSE being the most important of the 4 metrics for the purpose of the testing.

## Quality of Model

The quality of the models will then be determined based on what improvement is reached compared with the non-Neural Network based models.

* Normalization
* Evaluate the non-NN regressions as the baselines
* Determine Best Architecture for activation function
* Evaluate the Key Characteristics of the models

# results

# conclusions