Artificial Neural Networks

Benchmark Problem Report

Antonio Peters

September 13, 2016

1 Introduction

For this report, all data sets are obtained from `https://archive.ics.uci.edu/ml/datasets.html`. The code run for this report is written in the MATLAB programming language. All code can be accessed in their respective folders in the same directory as this document.

2 Mushrooms

The mushroom classification set seeks to classify mushrooms into two groups, poisonous and edible. It does this by looking at 22 input parameters: cap shape, cap surface, cap color, bruises, odour, gill attachment, gill spacing, gill size, gill color, stalk shape, stalk root, stalk surface above ring, stalk surface below ring, stalk color above ring, stalk color below ring, veil-type, veil color, ring number, ring type, spore print color, population and habitat.

We begin by importing the data and normalising it so it can be used as input into the net. The data is presented as a series of characters with 23 characters to a row, each separated by a comma, this represents the 22 parameters and the corresponding result. The data is first separated into a $22xn$ matrix for the input and a vector of n length for the output. Each element of the input is then converted to its corresponding ASCII value and 97 is subtracted so that $a = 0, b = 1, ..., z = 25$ for each element. In the output vector, there are only two possible answers, p and e for poisonous and edible, these are translated to 0 and 1 respectively. It has been stated by the supplier of the data that the dataset is incomplete, as such there are elements of the input data which are labelled as ?, these then correspond to a negative value in the transformed input matrix, therefore any rows containing negative values are removed and their corresponding output value is also removed from the output vector as this could mislead our network.

The data is then split into training and testing sets in a ratio of 3 : 1. A 3-layer feed-forward neural network will be used to train on the training data. A triple nested for-loop is generated to iterate through the number of neurons for each of the net's layers from 1 to 20. From this, the best net is selected and stored, the criteria for the best network is based resulting in the highest R^2 and Correlation Coefficient values determined by simulating the network on the remaining test data. Once a best network has been found, its error is determined and plotted as shown in Figure 1.

Figure 1: Mushroom Best NN Test Error