Project 5 – Neural Networks

*Joseph Park - x172917 Date: 4/10/20*

Objective

The assignment is an illustration of classification based on the neural networks model using particle collision data from a particle accelerator to identify particles.

Neural Networks model

Step 1: Download the data set from Kaggle, look for missing data and look at summary, size and shape.

Step 2: Since the data set is too large to process, randomly sample. (And, since the neuralnet function takes way too long with the 50,000 that I have been using up to this point, I am cutting that down to 5,000.

Step 3: Partition data into training vs testing and target vs classified data.

Step 4: Train data

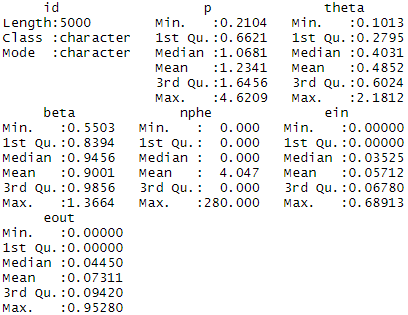
Step 5: Classify data with the one node.

Step 6: Repeat Step 3-5 but with normalized and again standardized (since ANN’s tend to work better that way and it is had to say from the histograms and summaries below which will work better).

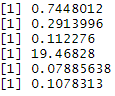
Step 7: Keep which of the above three options work the best and create a loop to optimize classification metric of choice, which is accuracy (for particles classification), with respect to number of hidden nodes (for performance of course as opposed to just using the default or some specific number of nodes).

Step 8: Plot accuracy of the index (of the loop iteration which is the number of trials) and write confusion matrix for the index value with max accuracy.

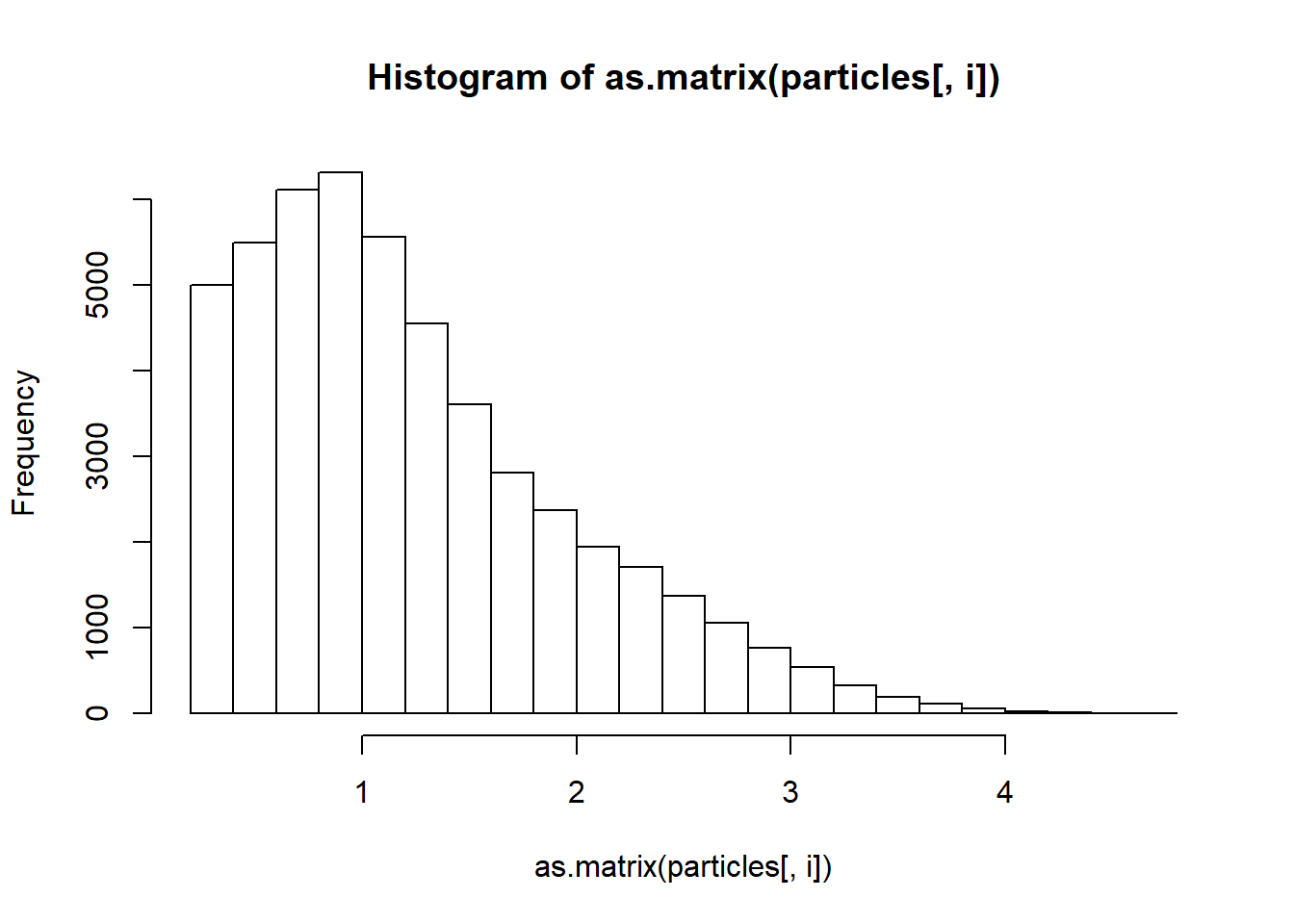
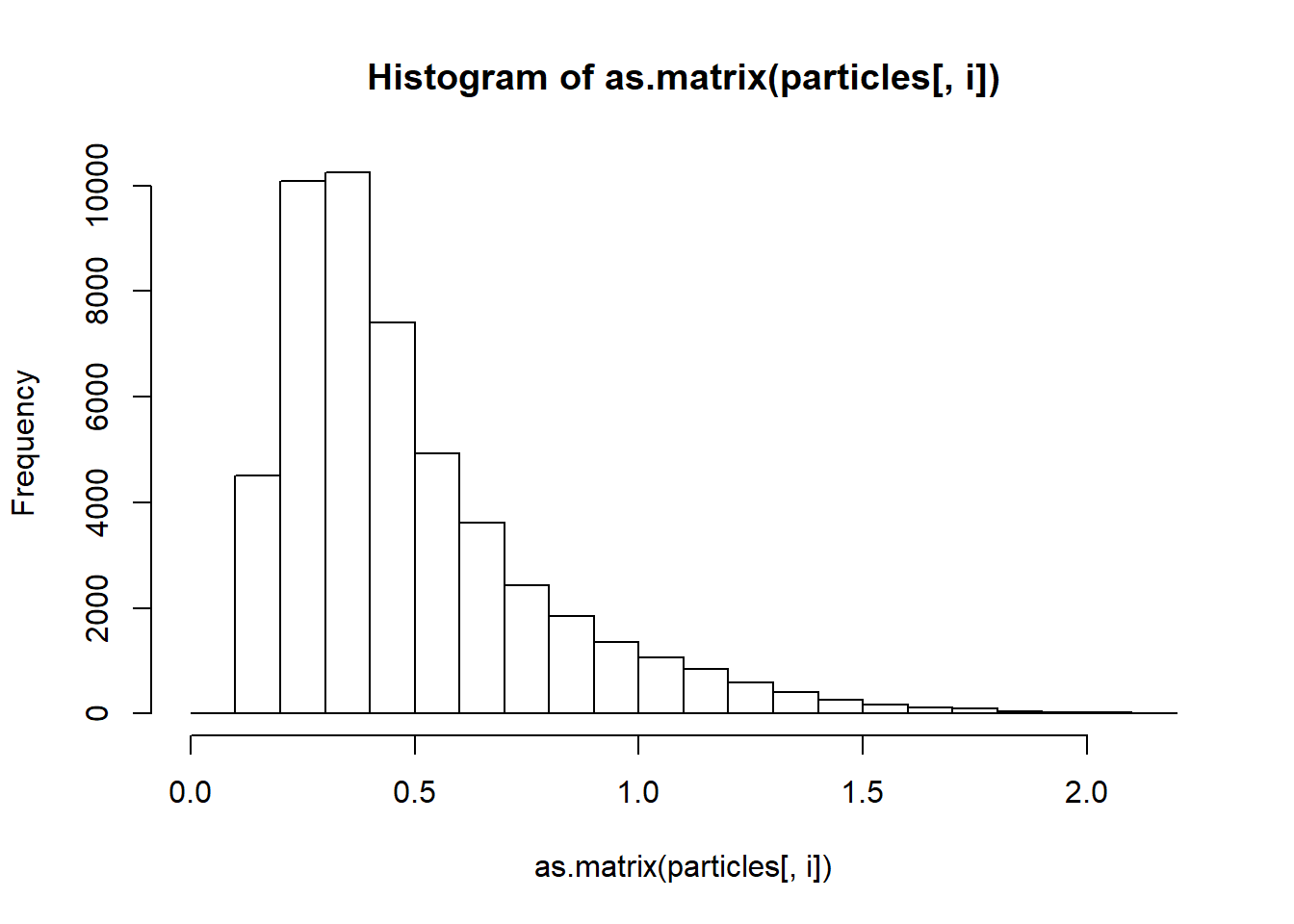
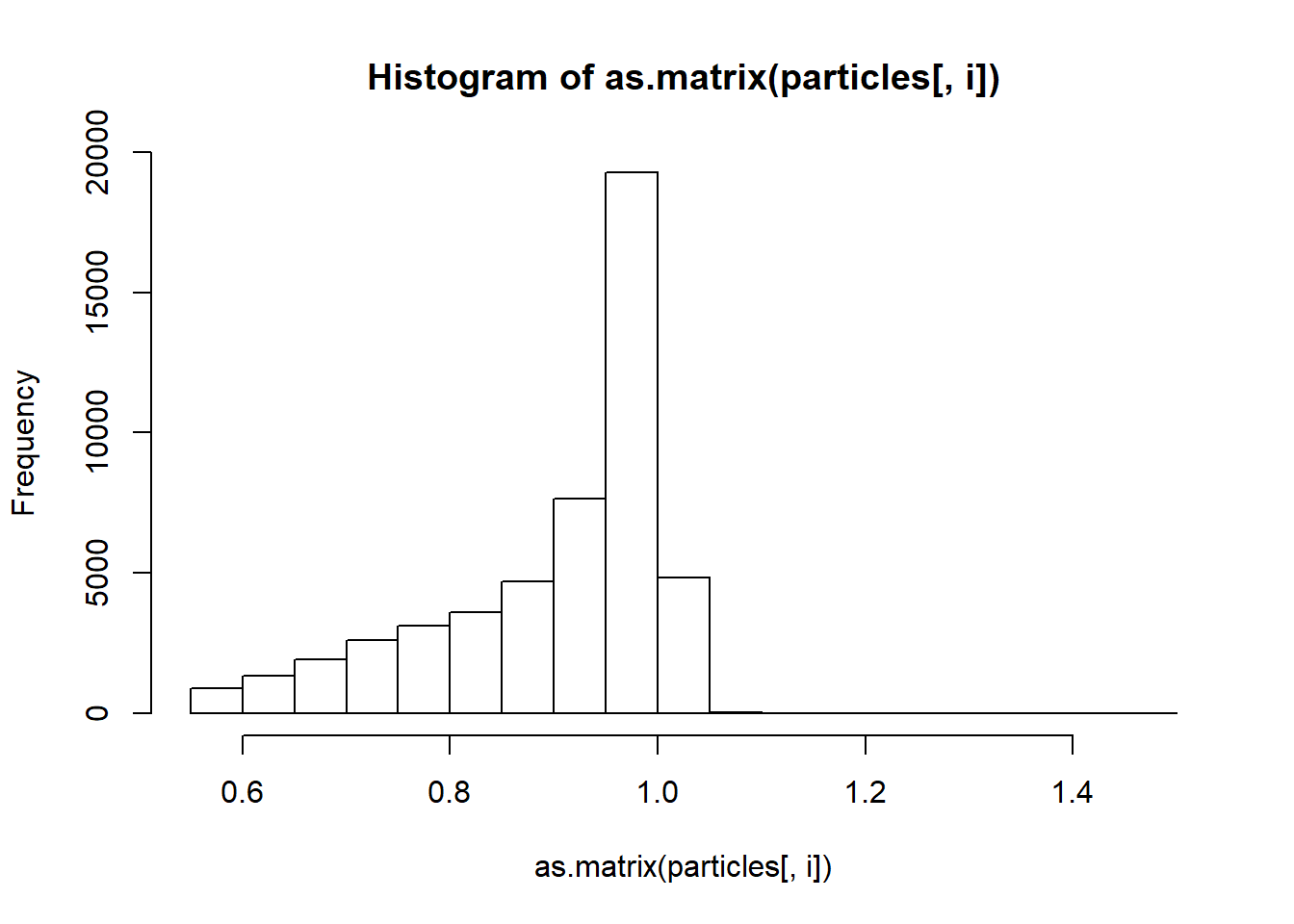
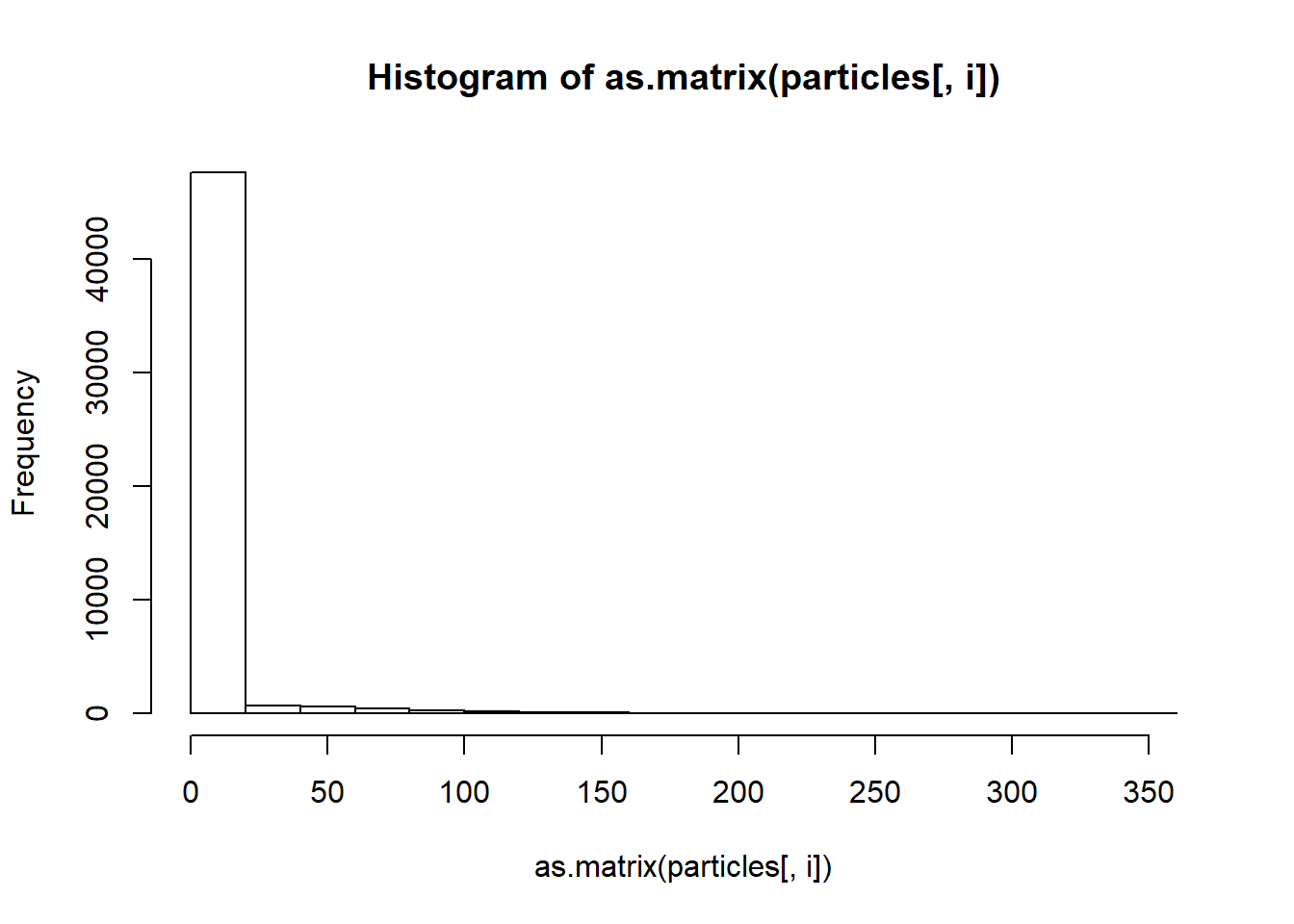
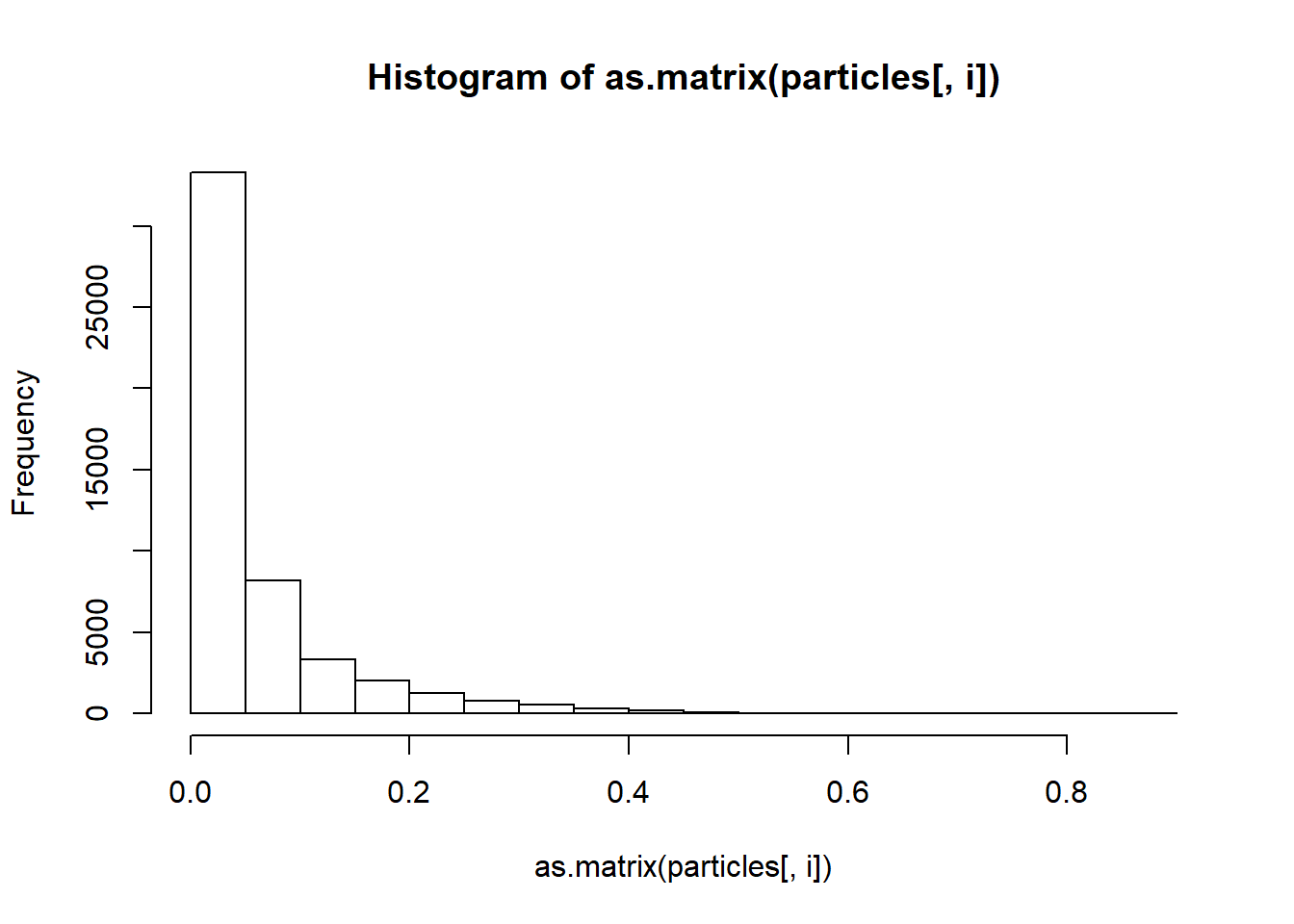
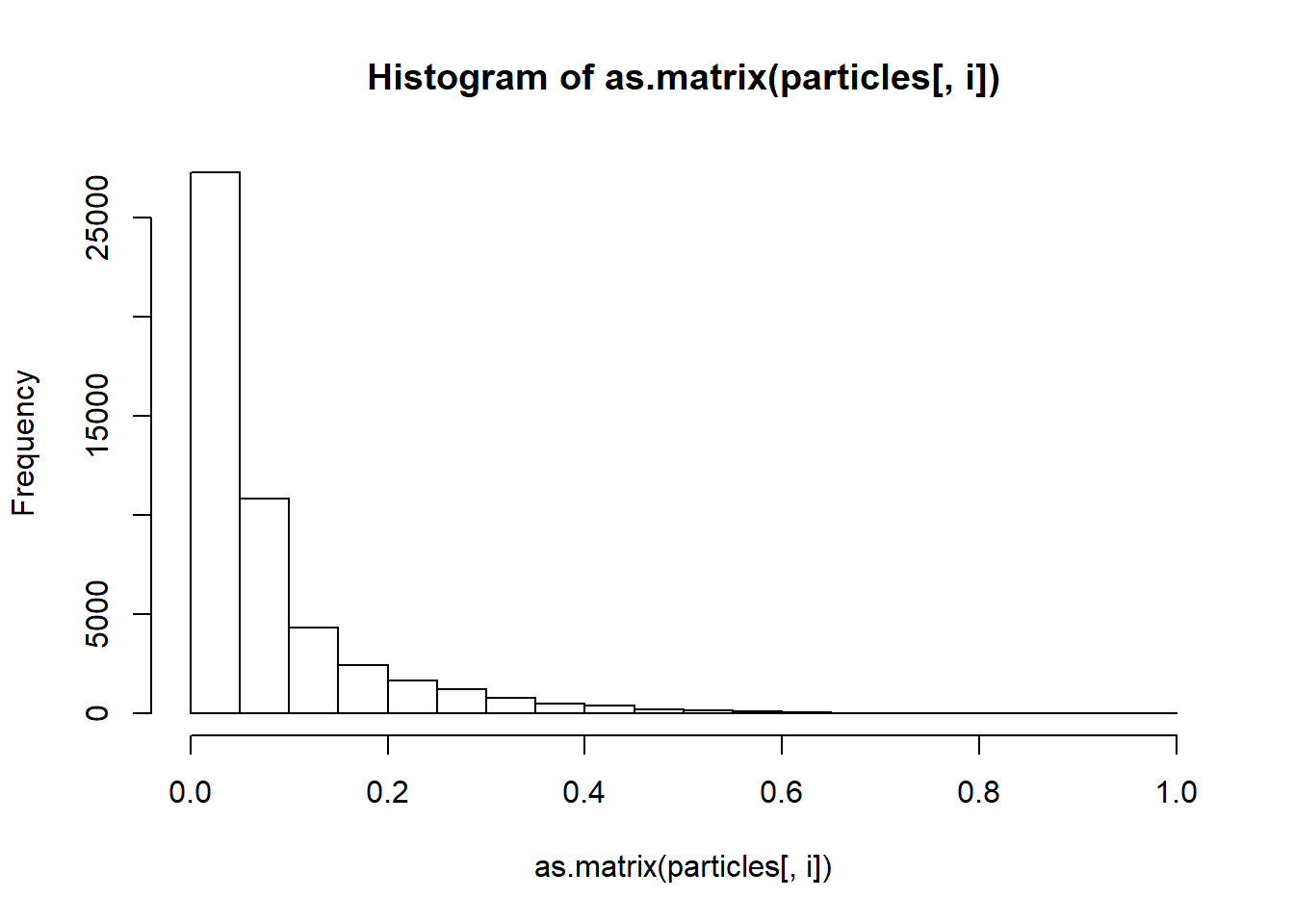
Step 9: Repeat steps 4-7 but with different activation functions this time and compare results.



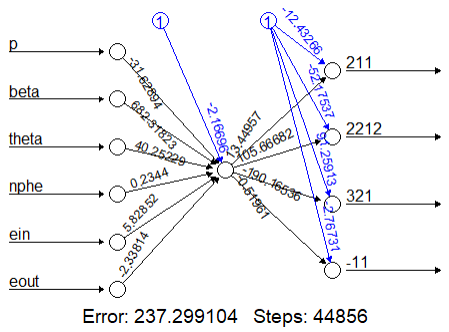
Here are the standard deviations of the variables in order of above left to right then top down.



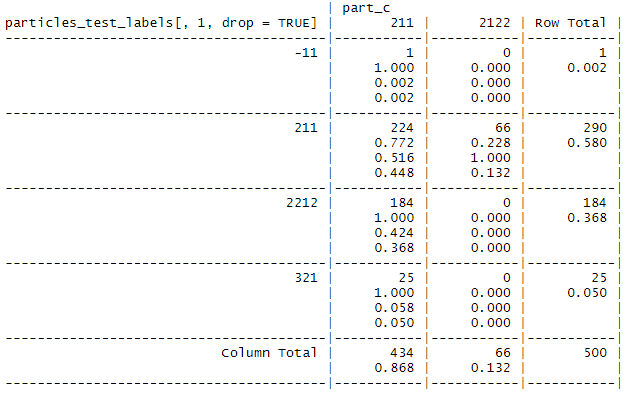
Here the corresponding histograms and the aforementioned order.

Results

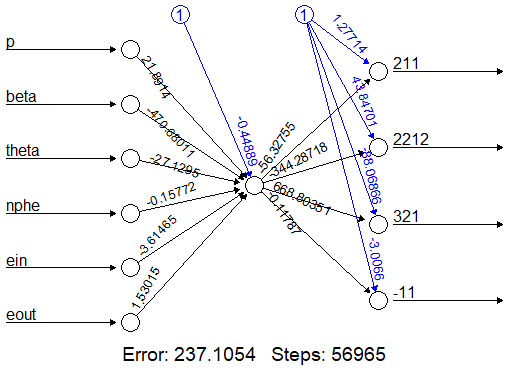


The accuracy is 0.448

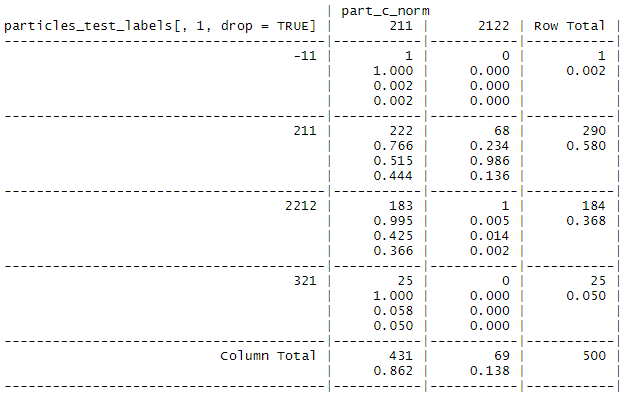


^Columns are predicted values and rows are actual values

With Normalization:



The accuracy is 0.444



^Columns are predicted values and rows are actual values

Interpretation of the Results

Note 1: If particle -11 was of particular interest, a supercomputer (and thus ability to process a test size significantly more 500) would be of use, as 1 data points is not reliable.

Note 2: The summaries and histograms are about the same as those of the prior projects bigger samples.

Note 3: After standardizing, neural net did not converge

Note 4: The results for one node with and without normalization are the roughly same.

Note 5: The accuracy is terrible for both

Note 6: Attempting with different numbers of nodes (like hidden = 5 and hidden = c(3, 2)) lead to various errors including not converging.

Note 6: I think the poor accuracy and lack of convergence may be due to how small I was forced to sample and the one in 500 count of particle -11. (I tried to find a way to run this online, but I ran into even more complications.)

Note 7: It may also be worth noting that the book intended this neuralnet function for numeric prediction and not classification. So, maybe trying this all over again with alternative packages (like nnet) would be a good idea.