NEURAL NETWORKS

As an introduction, we'll tackle a prediction task with a continuous variable. We'll reproduce research from the field of cement and concrete manufacturing that seeks to model the compressive strength of concrete using it's age and content. This dataset was donated to the UCI Machine Learning Repository by Professor I-Cheng Yeh from Chung-Hua University in Taiwan. This tutorial was adapted from that of Brett Lantz in Machine Learning with R (Packt Publishing 2015).

The following predictor variables, all measured in kg per cubic meter of mix (aside from age which is measured in days), are used to model the compressive strength of concrete as measured in MPa:

- Cement
- Blast Furnace Slag
- Fly Ash
- Water
- Superplasticizer
- Coarse Aggregate
- Fine Aggregate
- Age

This dataset has 1030 observations. All of the variables are numeric, taking positive values. The scales will vary quite a bit since some ingredients are more prevalent than others.

```
> load("concrete.Rdata")
> str(concrete)
```

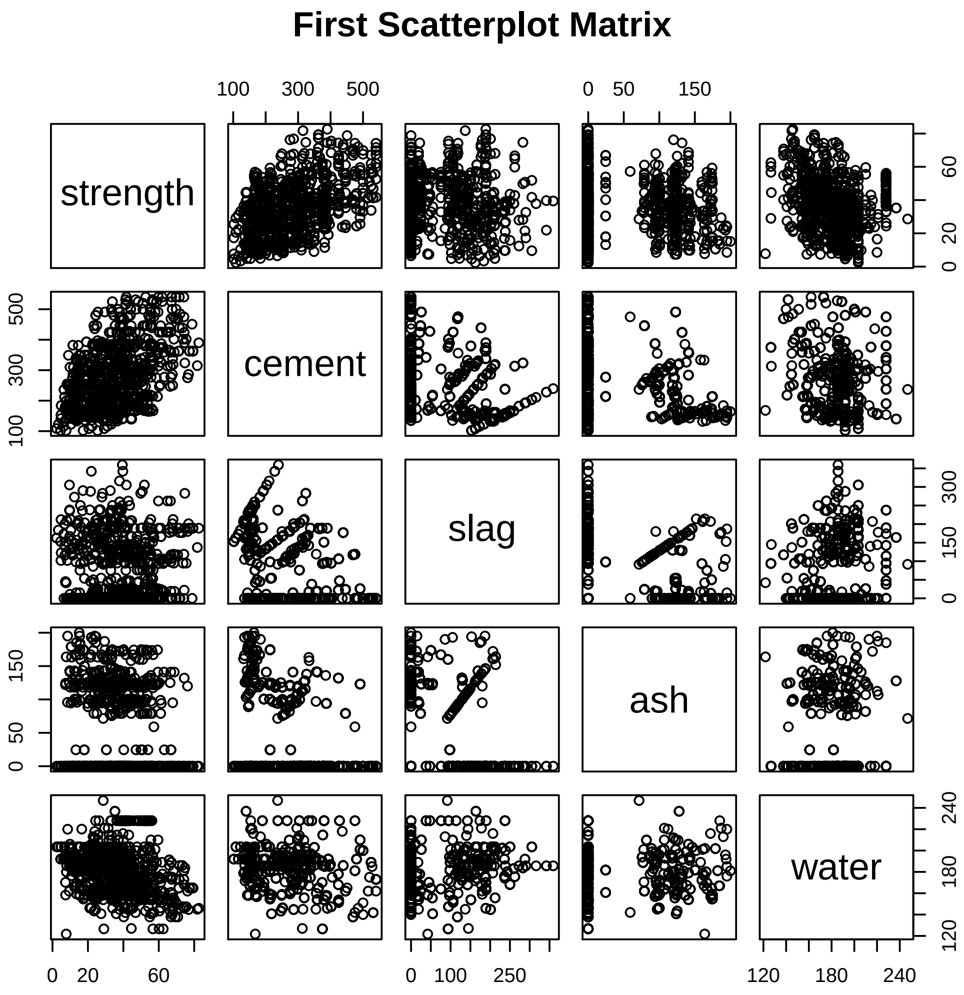
```
'data.frame':        1030 obs. of  9 variables:
 $ cement      : num  141 169 250 266 155 ...
 $ slag        : num  212 42.2 0 114 183.4 ...
 $ ash         : num  0 124.3 95.7 0 0 ...
 $ water       : num  204 158 187 228 193 ...
 $ superplastic: num  0 10.8 5.5 0 9.1 0 0 6.4 0 9 ...
 $ coarseagg   : num  972 1081 957 932 1047 ...
 $ fineagg     : num  748 796 861 670 697 ...
 $ age         : int  28 14 28 28 28 90 7 56 28 28 ...
 $ strength    : num  29.9 23.5 29.2 45.9 18.3 ...
```

Let's take an 80-20 split of the data for training and validation and then explore some preliminary relationships with the target variable.

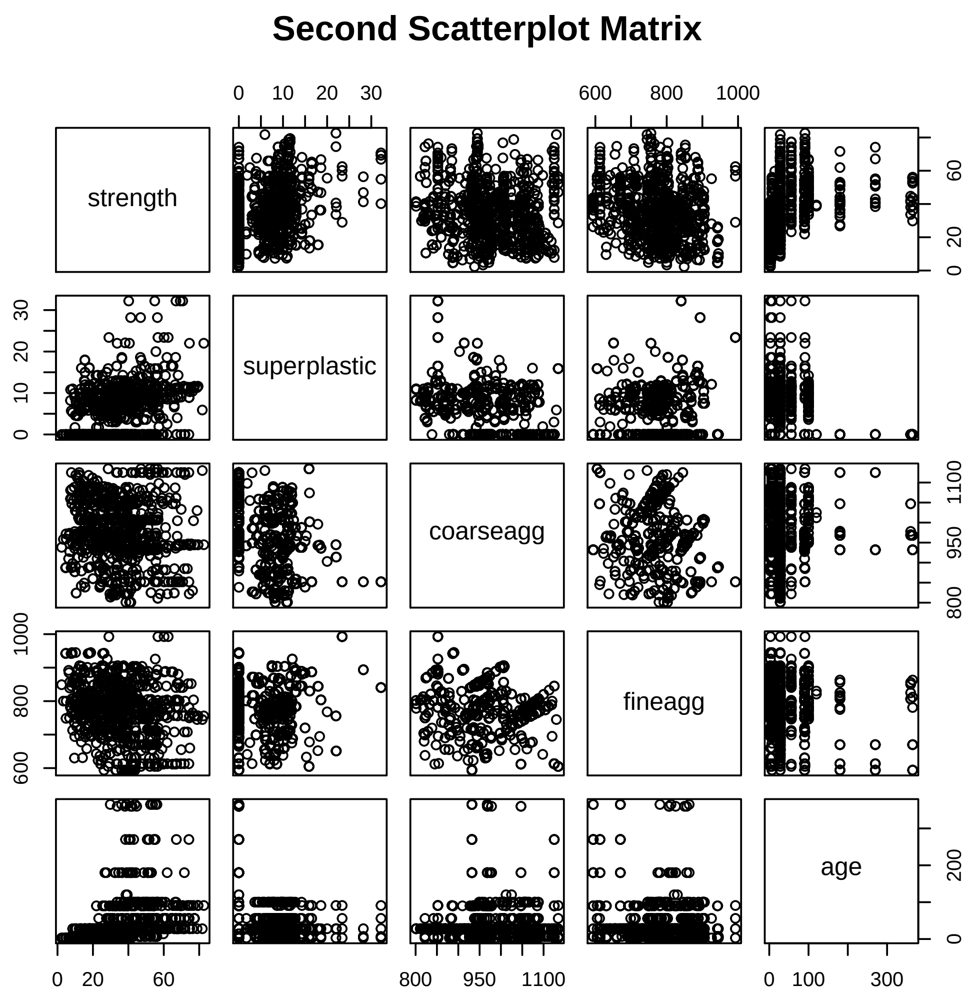
```
> TrainInd = sample(c(T,F),size=1030, replace=T,prob=c(0.8,0.2))
> train=concrete[TrainInd,]
> valid=concrete[!TrainInd,]
```

First let's check out a couple scatter plot matrices containing the target.

```
> pairs(strength~cement+slag+ash+water,data=train,
+    main="First Scatterplot Matrix")
```



```
> pairs(strength~superplastic+coarseagg+fineagg+age,data=train,
+    main="Second Scatterplot Matrix")
```



There certainly seem to be some relationships between input variables, and some variables appear to have a relationship with the target, but overall it is a little difficult to tell whether or not we will be able to make a good model for the compressive strength using these inputs. Let's try a simple linear model with all of the input variables and see how it performs.

```
> linear = lm(strength~cement+slag+ash+water+superplastic+coarseagg+fineagg+age,data=train)
> summary(linear)
```

```
Call:
lm(formula = strength ~ cement + slag + ash + water + superplastic +
    coarseagg + fineagg + age, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-28.537  -6.404   0.763   6.715  34.078

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.551600  30.777102  -0.083 0.933947
cement        0.111671   0.009617  11.612  < 2e-16 ***
slag          0.098476   0.011440   8.608  < 2e-16 ***
ash           0.077767   0.013960   5.571 3.44e-08 ***
```

```
water        -0.183113   0.048122  -3.805 0.000152 ***
superplastic  0.250140   0.109752   2.279 0.022917 *
coarseagg     0.012121   0.010702   1.133 0.257715
fineagg       0.012941   0.012271   1.055 0.291913
age           0.117658   0.006102  19.281  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.37 on 815 degrees of freedom
Multiple R-squared:  0.6136,        Adjusted R-squared:  0.6098
F-statistic: 161.8 on 8 and 815 DF,  p-value: < 2.2e-16
```

Most of the variables appear to have a significant (linear) relationship with compressive strength, but the overall performance of the model leaves much to be desired (adjusted R^2 0.6). Let's check the MAPE on our validation data.

```
> linearPred = predict(linear, valid)
> linearRsq = cor(linearPred,valid$strength)^2
> linearMAPE = mean(abs(linearPred-valid$strength)/valid$strength)
> cat("linear regression R-squared:", linearRsq)
```

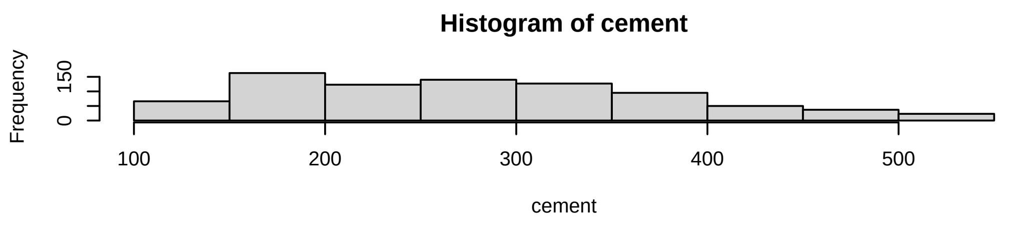
```
linear regression R-squared: 0.6190381
```

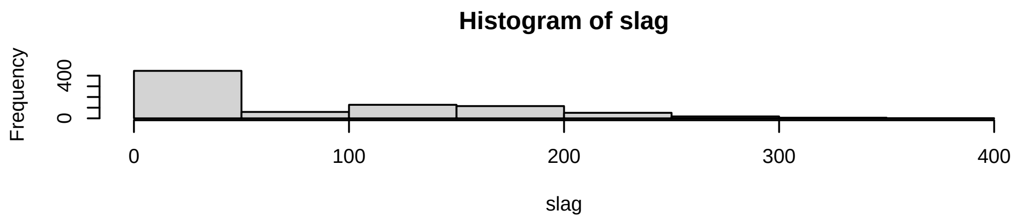
```
> cat("linear regression MAPE:",linearMAPE)
```

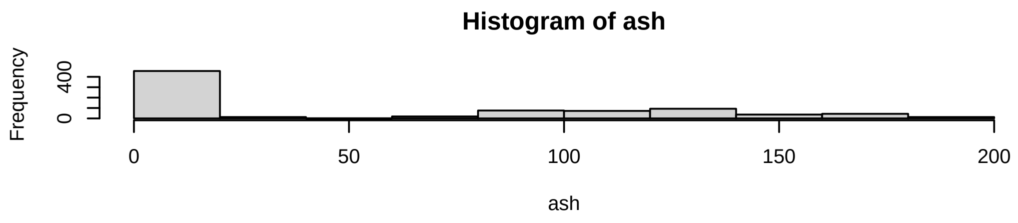
```
linear regression MAPE: 0.310473
```

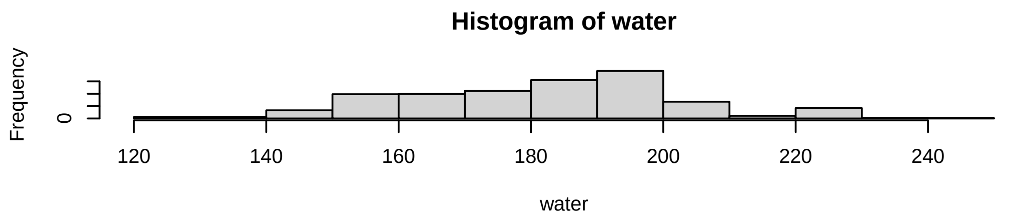
We probably desire a model with MAPE less than than 30% if we're predicting strength of concrete to build bridges! Let's see if we can improve upon this MAPE using Neural Networks. For this tutorial, we'll use the neuralnet package. If you recall, neural networks work best when the input data is standardized to a narrow range near zero. Since our data takes positive values up to 1000, we'll standardize the data. Let's first check the distribution of the variables to see if statistical standardization seems appropriate.

```
> # 3 figures arranged in 3 rows and 1 column
> attach(train)
> par(mfrow=c(4,1))
> hist(cement)
> hist(slag)
> hist(ash)
> hist(water)
```

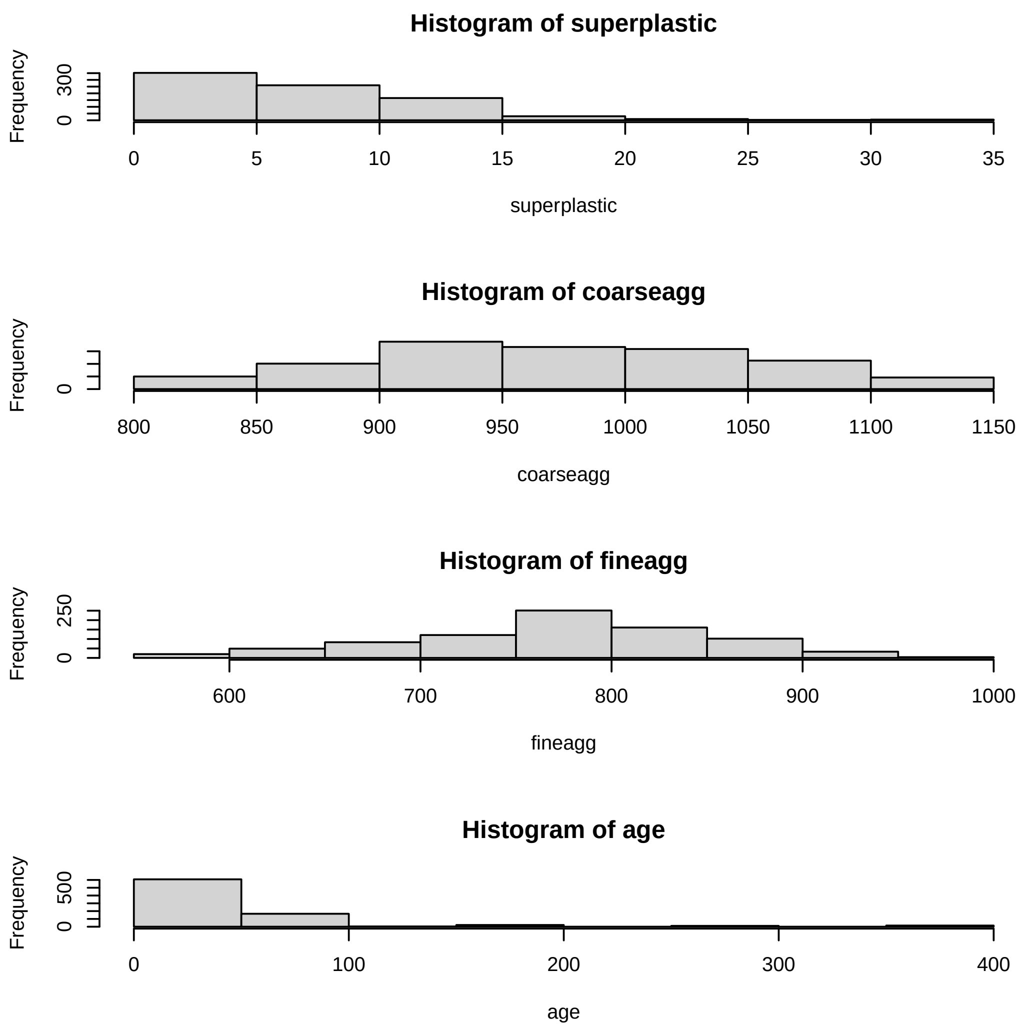








```
> par(mfrow=c(4,1))
> hist(superplastic)
> hist(coarseagg)
> hist(fineagg)
> hist(age)
```

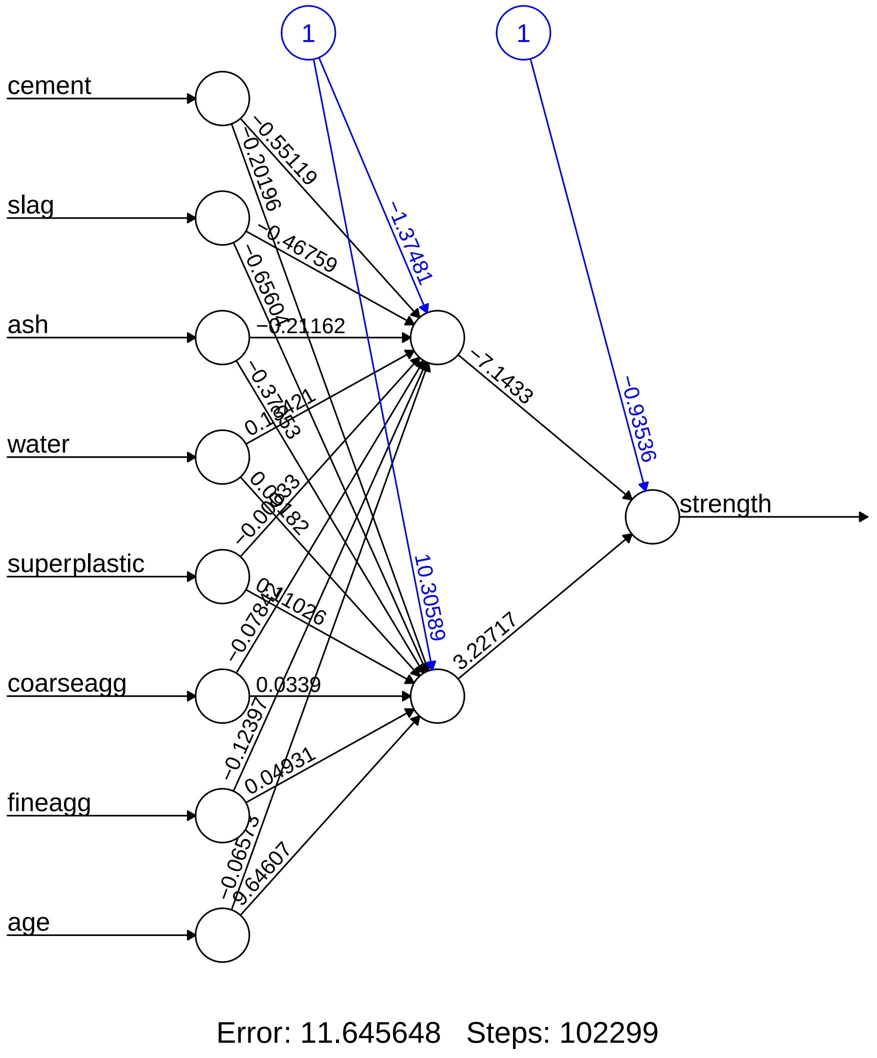


While many of the variables have bell-shaped distributions, others, like age, do not. It likely won't affect the results too much whether we choose midrange standardization or statistical standardization in this case, so you might try both. For the purposes of our tutorial, we'll go with midrange standardization. This will allow us to practice writing a function as well! This normalize function applies to variable vectors, and we'll need to apply it to every column of our data frame - including the target!

```
> scaleMidrange = function(x) {
+   midrange= (min(x)+max(x))/2
+   return(2*(x-midrange)/(max(x)-min(x)))
+ }
> concrete_norm = as.data.frame(lapply(concrete, scaleMidrange))
> train_norm=concrete_norm[TrainInd, ]
> valid_norm=concrete_norm[!TrainInd, ]
```

Let's get started with something simple, a neural network with only one hidden layer containing 2 neurons, and see how it performs compared to the linear regression. You can then plot the resulting network structure with the associated weights using the plot() command.

```
> library(neuralnet)
> set.seed(11117)
> nnet2 = neuralnet(strength~cement+slag+ash+water+superplastic+coarseagg+fineagg+age,
                        data=train_norm, rep=2, stepmax=10^6, hidden=2)
> plot(nnet2)
```



To predict the strength variable on the validation data, we'll need to use the compute() function of the neuralnet package. This function computes not only the final model output (in the $net.result component), but also the computed value at each neuron (in the $neurons component). We'll only want to retrieve the former to calculate a validation R^2 as well as a MAPE. Keep in mind that we'll have to *rescale* our data to *undo the standardization* that was done prior to prediction.

```
> # Get predictions on validation data:
> nnet2Pred = compute(nnet2, valid_norm[,1:8])$net.result
> # Rescale predictions back to normal:
> midrange = ((min(concrete$strength)+max(concrete$strength))/2)
> range = (max(concrete$strength)-min(concrete$strength))
> nnet2PredRescale = nnet2Pred*(range/2)+midrange
> nnet2Rsq = cor(nnet2PredRescale,valid_norm$strength)^2
> nnet2MAPE = mean(abs(nnet2PredRescale-valid$strength)/valid$strength)
> cat("2 Hidden Unit R-squared:", nnet2Rsq)
```

```
2 Hidden Unit R-squared: 0.8188251
```

```
> cat("2 Hidden Unit MAPE:", nnet2MAPE)
```

```
2 Hidden Unit MAPE: 0.2032028
```

The validation R^2 from our model increases dramatically to 0.9, and the MAPE reduces to approximately 20%, leaving plenty of room for improvement. This was a very simple neural network model (a good place to start) so let's see how we do if we increase the number of neurons in the hidden layer first to 3 and then to 5:

```
> set.seed(11117)
> #added additional rep= and stepmax= options because of convergence errors
> nnet3 = neuralnet(strength~cement+slag+ash+water+superplastic+coarseagg+fineagg+age,
                    data=train_norm, rep=2, stepmax=10^6, hidden=3)
> nnet4 = neuralnet(strength~cement+slag+ash+water+superplastic+coarseagg+fineagg+age,
                    data=train_norm, rep=2, stepmax=10^6,  hidden=4)
> nnet5 = neuralnet(strength~cement+slag+ash+water+superplastic+coarseagg+fineagg+age,
                    data=train_norm, rep=2, stepmax=10^6, hidden=5)
> # Score validation data on 3-neuron network
> nnet3Pred = compute(nnet3, valid_norm[,1:8])$net.result
> nnet3PredRescale = nnet3Pred*(range/2)+midrange
> nnet3Rsq = cor(nnet3PredRescale,valid$strength)^2
> nnet3MAPE = mean(abs(nnet3PredRescale-valid$strength)/valid$strength)
> cat("3 Hidden Unit R-squared:", nnet3Rsq)
```

```
3 Hidden Unit R-squared: 0.8573943
```

```
> cat("3 Hidden Unit MAPE:", nnet3MAPE)
```

```
3 Hidden Unit MAPE: 0.1722112
```

```
> # Score validation data on 4-neuron network
> nnet4Pred = compute(nnet4, valid_norm[,1:8])$net.result
> nnet4PredRescale = nnet4Pred*(range/2)+midrange
> nnet4Rsq = cor(nnet4PredRescale,valid$strength)^2
> nnet4MAPE = mean(abs(nnet4PredRescale-valid$strength)/valid$strength)
> cat("4 Hidden Unit R-squared:", nnet4Rsq)
```

```
4 Hidden Unit R-squared: 0.8551052
```

```
> cat("4 Hidden Unit MAPE:", nnet4MAPE)
```

```
4 Hidden Unit MAPE: 0.1824533
```

```
> # Score validation data on 5-neuron network
> nnet5Pred = compute(nnet5, valid_norm[,1:8])$net.result
> nnet5PredRescale = nnet5Pred*(range/2)+midrange
> nnet5Rsq = cor(nnet5PredRescale,valid$strength)^2
> nnet5MAPE = mean(abs(nnet5PredRescale-valid$strength)/valid$strength)
> cat("5 Hidden Unit R-squared:", nnet5Rsq)
```

```
5 Hidden Unit R-squared: 0.8451643
```

```
> cat("5 Hidden Unit MAPE:", nnet5MAPE)
```

```
5 Hidden Unit MAPE: 0.1723047
```

What we see is an improvement of model error in the network with 3 neurons, followed by an *increase* in validation error for the networks with 4 and 5 neurons, suggesting that more than 3 neurons might overfit the pattern in training data.

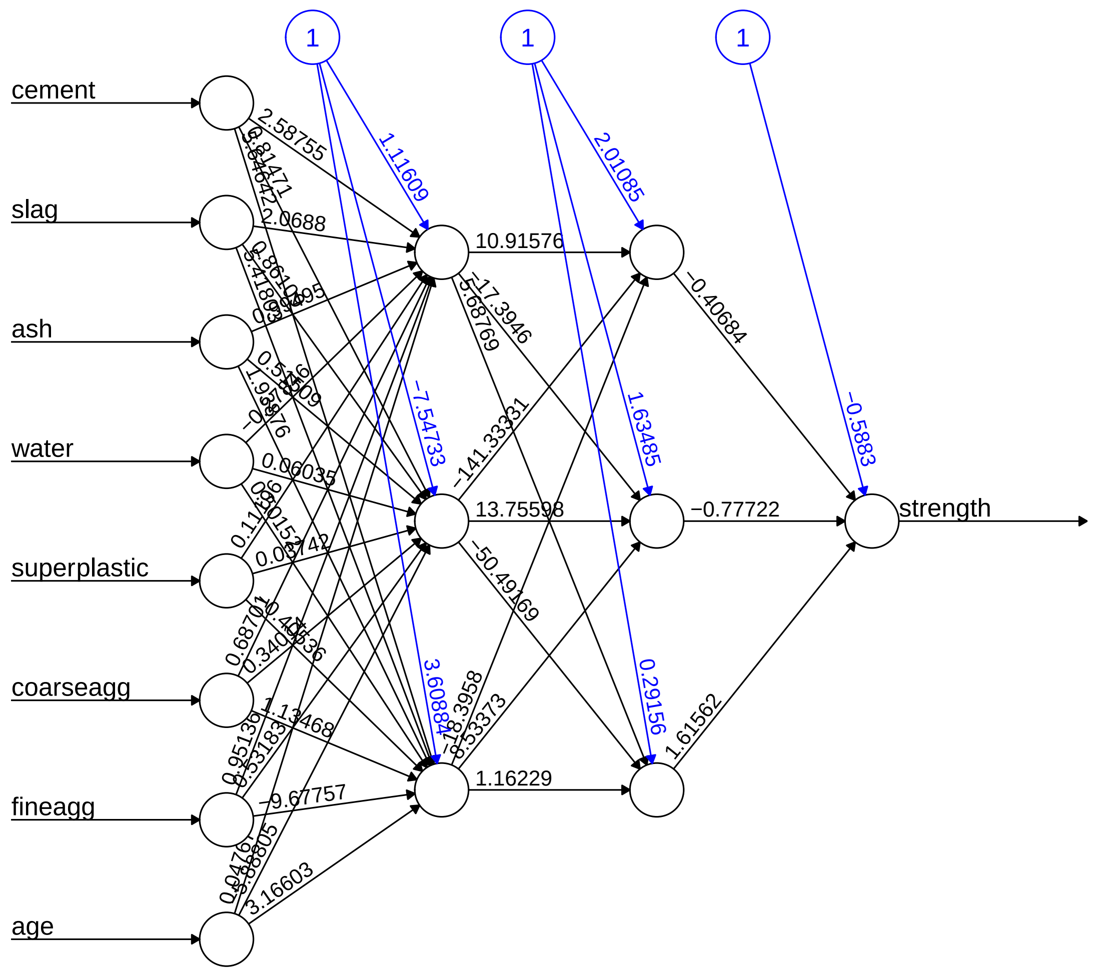
We could also attempt a multilayer network by inputting a vector for the hidden= option in the neuralnet() function as follows:

```
> set.seed(11117)
> nnet33 = neuralnet(strength~cement+slag+ash+water+superplastic+coarseagg+fineagg+age,
                      data=train_norm, rep=2, stepmax=10^6, hidden=c(3,3))
> plot(nnet33)
> # Score validation data on 3-neuron network
> nnet33Pred = compute(nnet33, valid_norm[,1:8])$net.result
> nnet33PredRescale = nnet33Pred*(range/2)+midrange
> nnet33Rsq = cor(nnet33PredRescale,valid$strength)^2
> nnet33MAPE = mean(abs(nnet33PredRescale-valid$strength)/valid$strength)
> cat("3 Hidden Unit 2 Hidden Layer R-squared:", nnet33Rsq)
```

```
3 Hidden Unit 2 Hidden Layer R-squared: 0.8590084
```

```
> cat("3 Hidden Unit 2 Hidden Layer MAPE:", nnet33MAPE)
```

```
3 Hidden Unit 2 Hidden Layer MAPE: 0.1595117
```



Error: 7.74665   Steps: 84641