NEURAL NETWORKS FOR APPROXIMATION

# Download the data and libraries

library(dplyr)

## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(nnet)

library(ggplot2)

library(knitr)

library (psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

#Set Working Directory  
setwd('C:\\Users\\VerkhovodTS\\Desktop\\R')  
f <- read.csv2('countries.csv', header = TRUE, encoding = 'UNICOD')  
describe(f)

## vars n mean sd median  
## Country\* 1 189 95.00 54.70 95.00  
## GDP 2 189 455572.36 1953625.16 38760.47  
## Population 3 189 40.30 146.87 9.05  
## Unemployment 4 175 6.85 5.16 5.27  
## Area..km.2 5 189 705422.93 1934498.05 129494.00  
## Average.life.expectancy..years 6 188 67.41 11.69 71.50  
## Constitutional.form\* 7 189 2.74 0.51 3.00  
## trimmed mad min max  
## Country\* 95.00 69.68 1.00 189.00  
## GDP 112681.18 53818.97 47.27 21374418.88  
## Population 15.63 12.42 0.01 1397.72  
## Unemployment 6.10 3.75 0.09 28.18  
## Area..km.2 287419.23 190911.44 2.00 17075400.00  
## Average.life.expectancy..years 68.74 9.27 37.70 82.80  
## Constitutional.form\* 2.84 0.00 1.00 3.00  
## range skew kurtosis se  
## Country\* 188.00 0.00 -1.22 3.98  
## GDP 21374371.61 8.52 80.32 142105.32  
## Population 1397.71 8.16 70.92 10.68  
## Unemployment 28.09 1.41 1.89 0.39  
## Area..km.2 17075398.00 5.40 33.80 140714.03  
## Average.life.expectancy..years 45.10 -0.93 -0.25 0.85  
## Constitutional.form\* 2.00 -1.79 2.36 0.04

Аналіз основних показників описової статистики за кожною змінною показав, що є пропущені значення в змінних – Unemployment, Average.life.expectancy..years. Змінні GDP, Population, Area..km.2 мають викиди.

# Missing data, fill n/a with average

f\_fill1 <- f  
f\_fill1$Unemployment <-ifelse(is.na(f$Unemployment), round(mean(f$Unemployment, na.rm = TRUE)),f$Unemployment)  
f\_fill1$Average.life.expectancy..years <-ifelse(is.na(f$Average.life.expectancy..years), round(mean(f$Average.life.expectancy..years, na.rm = TRUE)),f$Average.life.expectancy..years)  
head(f\_fill1)

## Country GDP Population Unemployment Area..km.2  
## 1 Afghanistan 19101.35 38.04 11.12 647500  
## 2 Angola 94635.42 31.83 6.89 1246700  
## 3 Albania 15278.08 2.85 12.33 28748  
## 4 Andorra 3154.06 0.08 7.00 468  
## 5 United Arab Emirates 421142.27 9.77 2.35 83600  
## 6 Argentina 449663.45 44.94 9.79 2760990  
## Average.life.expectancy..years Constitutional.form  
## 1 43.8 Republic  
## 2 37.7 Republic  
## 3 77.8 Republic  
## 4 82.8 Constitutional monarchy  
## 5 75.8 Absolute monarchy  
## 6 76.4 Republic

f<- f\_fill1  
describe(f)

## vars n mean sd median  
## Country\* 1 189 95.00 54.70 95.00  
## GDP 2 189 455572.36 1953625.16 38760.47  
## Population 3 189 40.30 146.87 9.05  
## Unemployment 4 189 6.86 4.97 5.56  
## Area..km.2 5 189 705422.93 1934498.05 129494.00  
## Average.life.expectancy..years 6 189 67.40 11.66 71.40  
## Constitutional.form\* 7 189 2.74 0.51 3.00  
## trimmed mad min max  
## Country\* 95.00 69.68 1.00 189.00  
## GDP 112681.18 53818.97 47.27 21374418.88  
## Population 15.63 12.42 0.01 1397.72  
## Unemployment 6.15 3.32 0.09 28.18  
## Area..km.2 287419.23 190911.44 2.00 17075400.00  
## Average.life.expectancy..years 68.73 9.04 37.70 82.80  
## Constitutional.form\* 2.84 0.00 1.00 3.00  
## range skew kurtosis se  
## Country\* 188.00 0.00 -1.22 3.98  
## GDP 21374371.61 8.52 80.32 142105.32  
## Population 1397.71 8.16 70.92 10.68  
## Unemployment 28.09 1.46 2.28 0.36  
## Area..km.2 17075398.00 5.40 33.80 140714.03  
## Average.life.expectancy..years 45.10 -0.94 -0.24 0.85  
## Constitutional.form\* 2.00 -1.79 2.36 0.04

Пропуски в змінних Unemployment, Average.life.expectancy..years заповнили середніми значеннями показника.

# Ejections (outside the three sigma).

f\_ej1 <- f  
f\_ej1$GDP<- ifelse(f$GDP < mean(f$GDP)+sd(f$GDP)\*3, f$GDP, mean(f$GDP)+sd(f$GDP)\*3)  
f\_ej1$Population<- ifelse(f$Population < mean(f$Population)+sd(f$Population)\*3, f$Population, mean(f$Population)+sd(f$Population)\*3)  
f\_ej1$Area..km.2<- ifelse(f$Area..km.2 < mean(f$Area..km.2)+sd(f$Area..km.2)\*3, f$Area..km.2, mean(f$Area..km.2)+sd(f$Area..km.2)\*3)  
describe(f\_ej1)

## vars n mean sd median  
## Country\* 1 189 95.00 54.70 95.00  
## GDP 2 189 333432.54 901991.14 38760.47  
## Population 3 189 30.76 66.25 9.05  
## Unemployment 4 189 6.86 4.97 5.56  
## Area..km.2 5 189 582618.76 1229710.52 129494.00  
## Average.life.expectancy..years 6 189 67.40 11.66 71.40  
## Constitutional.form\* 7 189 2.74 0.51 3.00  
## trimmed mad min max range  
## Country\* 95.00 69.68 1.00 189.00 188.00  
## GDP 112681.18 53818.97 47.27 6316447.85 6316400.58  
## Population 15.63 12.42 0.01 480.91 480.90  
## Unemployment 6.15 3.32 0.09 28.18 28.09  
## Area..km.2 287419.23 190911.44 2.00 6508917.07 6508915.07  
## Average.life.expectancy..years 68.73 9.04 37.70 82.80 45.10  
## Constitutional.form\* 2.84 0.00 1.00 3.00 2.00  
## skew kurtosis se  
## Country\* 0.00 -1.22 3.98  
## GDP 4.61 23.89 65610.20  
## Population 4.47 23.70 4.82  
## Unemployment 1.46 2.28 0.36  
## Area..km.2 3.72 14.49 89448.28  
## Average.life.expectancy..years -0.94 -0.24 0.85  
## Constitutional.form\* -1.79 2.36 0.04

f <- f\_ej1

Для корекції викидів обраний варіант заповнення граничними значеннями.

#Factors as numeric

f$Constitutional.form <- as.numeric(f$Constitutional.form)

Якісні змінні перетворено на кількісні.

# Features Scaling

f\_sc <- f  
f\_sc$GDP <- scale(f$GDP)  
f\_sc$Population <- scale(f$Population)  
f\_sc$Unemployment <- scale(f$Unemployment)  
f\_sc$Area..km.2 <- scale(f$Area..km.2)  
f\_sc$Average.life.expectancy..years <- scale(f$Average.life.expectancy..years)  
f\_sc$Constitutional.form <- scale(f$Constitutional.form)  
head (f\_sc)

## Country GDP Population Unemployment Area..km.2  
## 1 Afghanistan -0.34848589 0.1098424 0.857697252 0.05276139  
## 2 Angola -0.26474442 0.0161105 0.006008387 0.54003054  
## 3 Albania -0.35272459 -0.4213052 1.101324563 -0.45040744  
## 4 Andorra -0.36616599 -0.4631148 0.028156324 -0.47340472  
## 5 United Arab Emirates 0.09724012 -0.3168567 -0.908097392 -0.40580182  
## 6 Argentina 0.12886037 0.2139890 0.589908554 1.77145044  
## Average.life.expectancy..years Constitutional.form  
## 1 -2.0252442 0.5115125  
## 2 -2.5486128 0.5115125  
## 3 0.8918919 0.5115125  
## 4 1.3208825 -1.4614644  
## 5 0.7202957 -3.4344414  
## 6 0.7717745 0.5115125

Виконано шкалювання кількісних змінних.

# Splitting the scaled dataset into the TRAIN set and TEST set

set.seed(123)  
library(caTools)

## Warning: package 'caTools' was built under R version 3.6.3

split = sample.split(f\_sc$GDP, SplitRatio = 0.8)  
f\_train = subset(f\_sc, split == TRUE)  
f\_test = subset(f\_sc, split == FALSE)

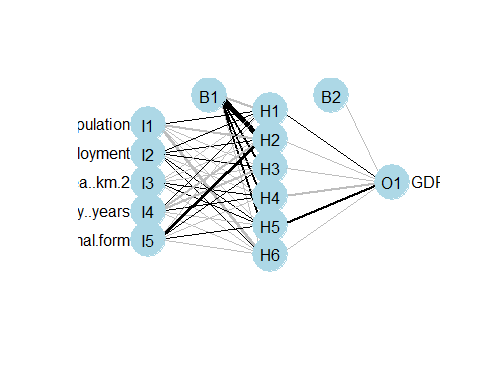
Датасет розподілений на навчальну та тестову вибірки.

# Fitting the NN

set.seed(123)  
library(nnet)  
library(scales)

library(reshape)

ff\_ap <- nnet(data = f\_train, GDP ~ Population + Unemployment + Area..km.2 + Average.life.expectancy..years + Constitutional.form, linout = TRUE ,size = 6, maxit = 10000)  
library(graphics)  
setwd('C:\\Users\\VerkhovodTS\\Desktop\\R')  
source(file = 'plot.nnet.R')  
plot.nnet(ff\_ap)



На основі усіх змінних побудовано двошарову нейронну мережу для прогнозування ВВП країни.

# Prediction

p\_ff\_ap <- predict(ff\_ap, f\_test)  
  
train\_mse\_ff\_ap <- sum((f\_train$GDP-predict(ff\_ap, f\_train))^2)/length(f\_train$GDP)  
test\_mse\_ff\_ap <- sum((f\_test$GDP-p\_ff\_ap)^2)/length(p\_ff\_ap)  
  
train\_mse\_ff\_ap

## [1] 0.006325315

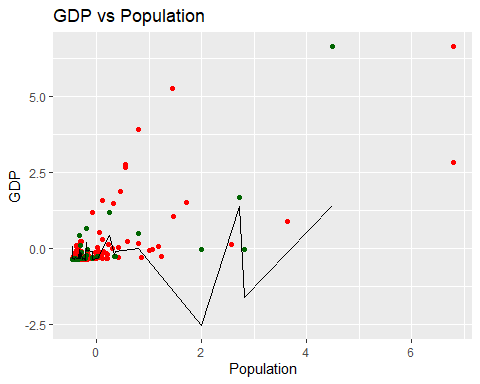
test\_mse\_ff\_ap

## [1] 0.9934499

значення середньоквадратичної помилки на навчальній вибірці – 0.006325315, на тестовій вибірці – 0.9934499.

# Visualising

library(ggplot2)  
ggplot() +  
 geom\_point(aes(f\_train$Population, f\_train$GDP),colour = 'red') +  
 geom\_point(aes(f\_test$Population, f\_test$GDP),colour = 'dark green') +  
 geom\_line(aes(f\_test$Population, p\_ff\_ap),colour = 'black') +  
 ggtitle('GDP vs Population') +  
 xlab('Population') +  
 ylab('GDP')



На графіку червоним позначені точки навчальної вибірки, зеленим – точки тестової вибірки, чорним – модельні значення.