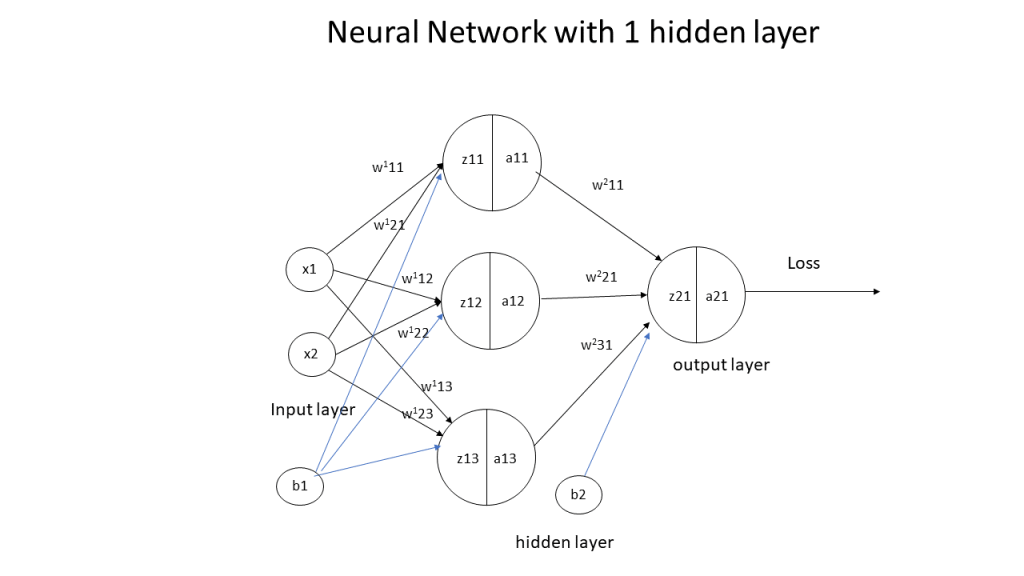
**The 3 layer Neural Network**

A simple representation of a 3 layer Neural Network (NN) with 1 hidden layer is shown below.  
  
In the above Neural Network, there are 2 input features at the input layer, 3 hidden units at the hidden layer and 1 output layer as it deals with binary classification. The activation unit at the hidden layer can be a tanh, sigmoid, relu etc. At the output layer the activation is a sigmoid to handle binary classification

# Superscript indicates layer 1  
z_{11} = w_{11}^{1}x_{1} + w_{21}^{1}x_{2} + b_{1}  
z_{12} = w_{12}^{1}x_{1} + w_{22}^{1}x_{2} + b_{1}  
z_{13} = w_{13}^{1}x_{1} + w_{23}^{1}x_{2} + b_{1}

Also a_{11} = tanh(z_{11})  
a_{12} = tanh(z_{12})  
a_{13} = tanh(z_{13})

# Superscript indicates layer 2  
z_{21} = w_{11}^{2}a_{11} + w_{21}^{2}a_{12} + w_{31}^{2}a_{13} + b_{2}  
a_{21} = sigmoid(z21)

Hence  
Z1= \begin{pmatrix}  z11\\  z12\\  z13  \end{pmatrix} =\begin{pmatrix}  w_{11}^{1} & w_{21}^{1} \\  w_{12}^{1} & w_{22}^{1} \\  w_{13}^{1} & w_{23}^{1}  \end{pmatrix} * \begin{pmatrix}  x1\\  x2  \end{pmatrix} + b_{1}  
And  
A1= \begin{pmatrix}  a11\\  a12\\  a13  \end{pmatrix} = \begin{pmatrix}  tanh(z11)\\  tanh(z12)\\  tanh(z13)  \end{pmatrix}

Similarly  
Z2= z_{21}  = \begin{pmatrix}  w_{11}^{2} & w_{21}^{2} & w_{31}^{2}  \end{pmatrix} *\begin{pmatrix}  z_{11}\\  z_{12}\\  z_{13}  \end{pmatrix} +b_{2}  
and A2 = a_{21} = sigmoid(z_{21})

These equations can be written as  
Z1 = W1 * X + b1  
A1 = tanh(Z1)  
Z2 = W2 * A1 + b2  
A2 = sigmoid(Z2)

**I) Some important results** (a memory refresher!)  
d/dx(e^{x}) = e^{x}and d/dx(e^{-x}) = -e^{-x}-(a) and  
sinhx = (e^{x} - e^{-x})/2and coshx = (e^{x} + e^{-x})/2  
Using (a) we can shown that d/dx(sinhx) = coshxand d/dx(coshx) = sinhx(b)  
Now d/dx(f(x)/g(x)) = (g(x)*d/dx(f(x)) - f(x)*d/dx(g(x)))/g(x)^{2}-(c)

Since tanhx =z= sinhx/coshxand using (b) we get  
tanhx = (coshx*d/dx(sinhx) - sinhx*d/dx(coshx))/(cosh^{2})  
Using the values of the derivatives of sinhx and coshx from (b) above we get  
d/dx(tanhx) = (coshx^{2} - sinhx{2})/coshx{2} = 1 - tanhx^{2}  
Since tanhx =z  
d/dx(tanhx) = 1 - tanhx^{2}= 1 - z^{2}-(d)

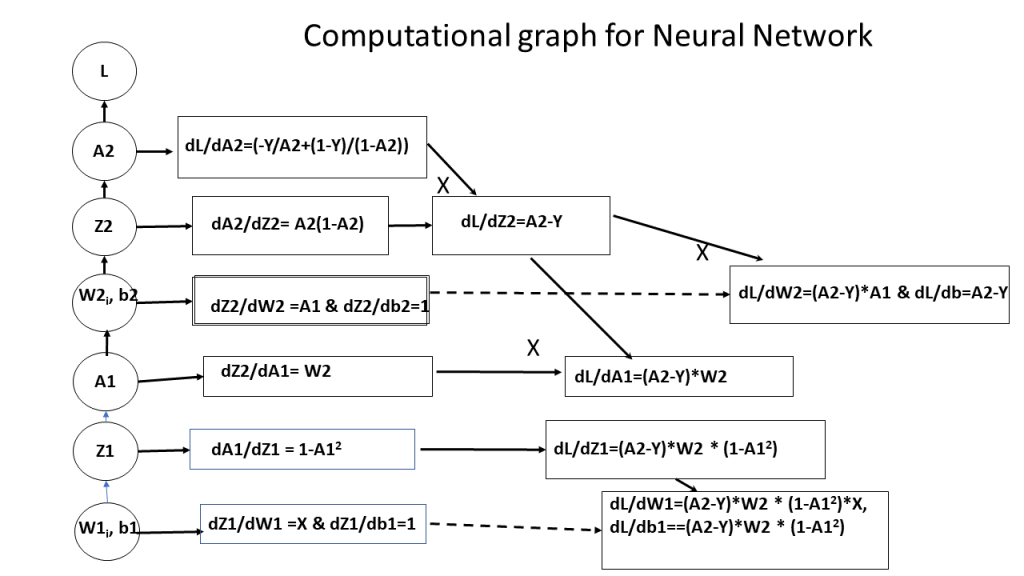
**II) Derivatives**  
L=-(Ylog(A2) + (1-Y)log(1-A2))  
dL/dA2 = -(Y/A2 + (1-Y)/(1-A2))  
Since A2 = sigmoid(Z2)therefore dA2/dZ2 = A2(1-A2)see [Part1](https://gigadom.wordpress.com/2018/01/04/deep-learning-from-basic-principles-in-python-r-and-octave-part-1/)  
Z2 = W2A1 +b2  
dZ2/dW2 = A1  
dZ2/db2 = 1  
A1 = tanh(Z1)and dA1/dZ1 = 1 - A1^{2}  
Z1 = W1X + b1  
dZ1/dW1 = X  
dZ1/db1 = 1

**III) Back propagation**  
Using the derivatives from II) we can derive the following results using Chain Rule  
\partial L/\partial Z2 = \partial L/\partial A2 * \partial A2/\partial Z2   
= -(Y/A2 + (1-Y)/(1-A2)) * A2(1-A2) = A2 - Y  
\partial L/\partial W2 = \partial L/\partial A2 * \partial A2/\partial Z2 * \partial Z2/\partial W2  
= (A2-Y) *A1-(A)  
\partial L/\partial b2 = \partial L/\partial A2 * \partial A2/\partial Z2 * \partial Z2/\partial b2 = (A2-Y)-(B)

\partial L/\partial Z1 = \partial L/\partial A2 * \partial A2/\partial Z2 * \partial Z2/\partial A1 *\partial A1/\partial Z1 = (A2-Y) * W2 * (1-A1^{2})  
\partial L/\partial W1 = \partial L/\partial A2 * \partial A2/\partial Z2 * \partial Z2/\partial A1 *\partial A1/\partial Z1 *\partial Z1/\partial W1   
=(A2-Y) * W2 * (1-A1^{2}) * X-(C)  
\partial L/\partial b1 = \partial L/\partial A2 * \partial A2/\partial Z2 * \partial Z2/\partial A1 *dA1/dZ1 *dZ1/db1  
= (A2-Y) * W2 * (1-A1^{2})-(D)

**IV) Gradient Descent**  
The key computations in the backward cycle are  
W1 = W1-learningRate * \partial L/\partial W1– From (C)  
b1 = b1-learningRate * \partial L/\partial b1– From (D)  
W2 = W2-learningRate * \partial L/\partial W2– From (A)  
b2 = b2-learningRate * \partial L/\partial b2– From (B)

The weights and biases (W1,b1,W2,b2) are updated for each iteration thus minimizing the loss/cost.

These derivations can be represented pictorially using the computation graph (from the book Deep Learning by Ian Goodfellow, Joshua Bengio and Aaron Courville)  


**3. Manually create a data set that is not lineary separable**

Initially I create a dataset with 2 classes which has around 9 clusters that cannot be separated by linear boundaries. **Note**: *This data set is saved as data.csv and is used for the R and Octave Neural networks to see how they perform on the same dataset.*

import numpy as np

import matplotlib.pyplot as plt

import matplotlib.colors

import sklearn.linear\_model

from sklearn.model\_selection import train\_test\_split

from sklearn.datasets import make\_classification, make\_blobs

from matplotlib.colors import ListedColormap

import sklearn

import sklearn.datasets

colors=['black','gold']

cmap = matplotlib.colors.ListedColormap(colors)

X, y = make\_blobs(n\_samples = 400, n\_features = 2, centers = 7,

cluster\_std = 1.3, random\_state = 4)

#Create 2 classes

y=y.reshape(400,1)

y = y % 2

#Plot the figure

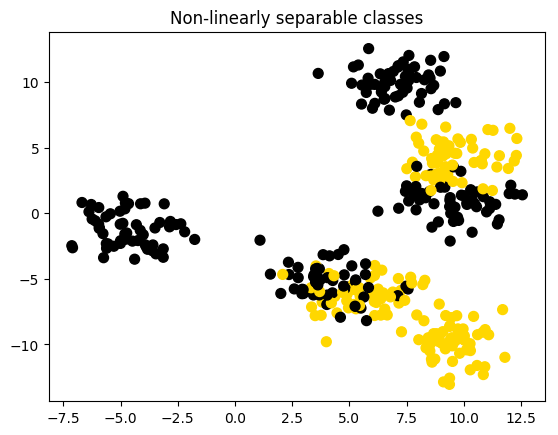
plt.figure()

plt.title('Non-linearly separable classes')

plt.scatter(X[:,0], X[:,1], c=y,

marker= 'o', s=50,cmap=cmap)

plt.savefig('fig1.png', bbox\_inches='tight')

****

**4. Logistic Regression**

On the above created dataset, classification with logistic regression is performed, and the decision boundary is plotted. It can be seen that logistic regression performs quite poorly

import numpy as np

import matplotlib.pyplot as plt

import matplotlib.colors

import sklearn.linear\_model

from sklearn.model\_selection import train\_test\_split

from sklearn.datasets import make\_classification, make\_blobs

from matplotlib.colors import ListedColormap

import sklearn

import sklearn.datasets

#from DLfunctions import plot\_decision\_boundary

execfile("./DLfunctions.py") # Since import does not work in Rmd!!!

colors=['black','gold']

cmap = matplotlib.colors.ListedColormap(colors)

X, y = make\_blobs(n\_samples = 400, n\_features = 2, centers = 7,

cluster\_std = 1.3, random\_state = 4)

#Create 2 classes

y=y.reshape(400,1)

y = y % 2

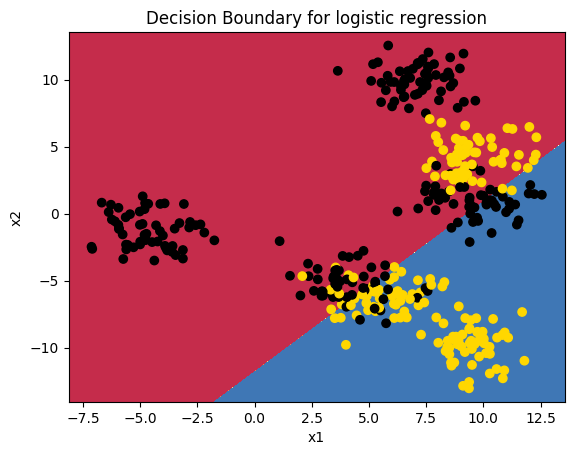
# Train the logistic regression classifier

clf = sklearn.linear\_model.LogisticRegressionCV();

clf.fit(X, y);

# Plot the decision boundary for logistic regression

plot\_decision\_boundary\_n(lambda x: clf.predict(x), X.T, y.T,"fig2.png")

****

**5. The 3 layer Neural Network in Python (vectorized)**

The vectorized implementation is included below. Note that in the case of Python a learning rate of 0.5 and 3 hidden units performs very well.

## Random data set with 9 clusters

import numpy as np

import matplotlib

import matplotlib.pyplot as plt

import sklearn.linear\_model

import pandas as pd

from sklearn.datasets import make\_classification, make\_blobs

execfile("./DLfunctions.py") # Since import does not work in Rmd!!!

X1, Y1 = make\_blobs(n\_samples = 400, n\_features = 2, centers = 9,

cluster\_std = 1.3, random\_state = 4)

#Create 2 classes

Y1=Y1.reshape(400,1)

Y1 = Y1 % 2

X2=X1.T

Y2=Y1.T

#Perform gradient descent

parameters,costs = computeNN(X2, Y2, numHidden = 4, learningRate=0.5, numIterations = 10000)

plot\_decision\_boundary(lambda x: predict(parameters, x.T), X2, Y2,str(4),str(0.5),"fig3.png")

## Cost after iteration 0: 0.692669

## Cost after iteration 1000: 0.246650

## Cost after iteration 2000: 0.227801

## Cost after iteration 3000: 0.226809

## Cost after iteration 4000: 0.226518

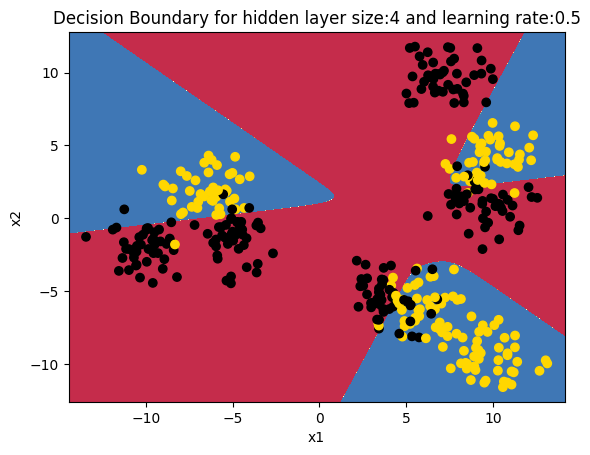
## Cost after iteration 5000: 0.226331

## Cost after iteration 6000: 0.226194

## Cost after iteration 7000: 0.226085

## Cost after iteration 8000: 0.225994

## Cost after iteration 9000: 0.225915

****

**6. The 3 layer Neural Network in R (vectorized)**

For this the dataset created by Python is saved  to see how R performs on the same dataset. The vectorized implementation of a Neural Network was just a little more interesting as R does not have a similar package like ‘numpy’. While numpy handles broadcasting implicitly, in R I had to use the ‘sweep’ command to broadcast. The implementaion is included below. Note that since the initialization with random weights is slightly different, R performs best with a learning rate of 0.1 and with 6 hidden units

source("DLfunctions2\_1.R")

z <- as.matrix(read.csv("data.csv",header=FALSE)) #

x <- z[,1:2]

y <- z[,3]

x1 <- t(x)

y1 <- t(y)

#Perform gradient descent

nn <-computeNN(x1, y1, 6, learningRate=0.1,numIterations=10000) # Good

## [1] 0.7075341

## [1] 0.2606695

## [1] 0.2198039

## [1] 0.2091238

## [1] 0.211146

## [1] 0.2108461

## [1] 0.2105351

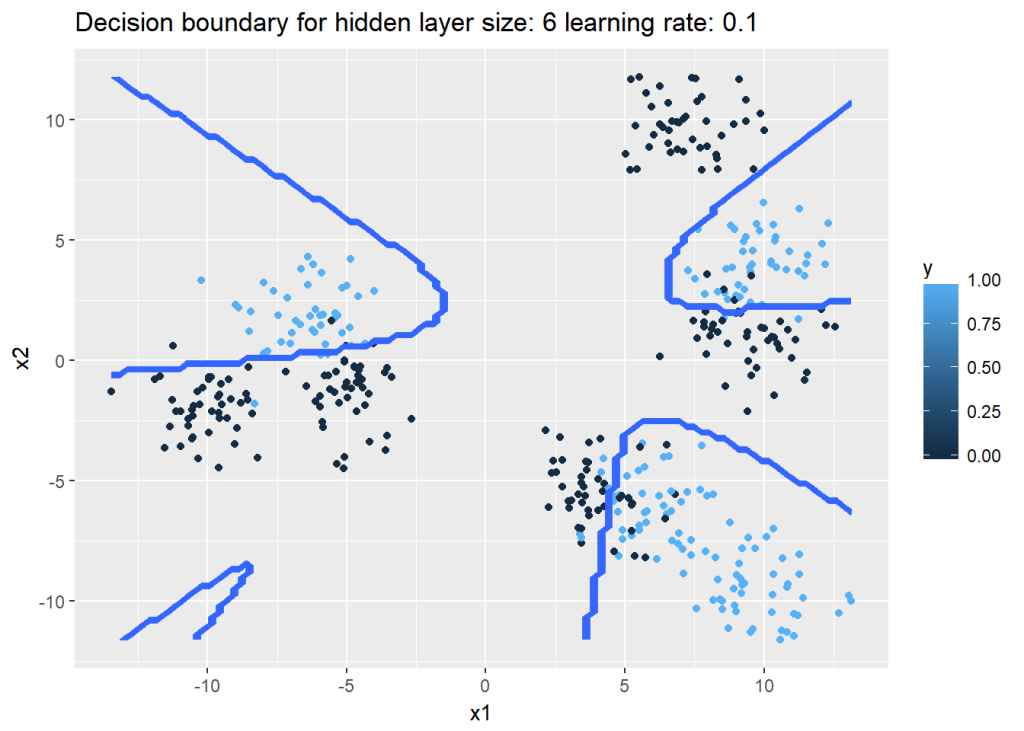
## [1] 0.210211

## [1] 0.2099104

## [1] 0.2096437

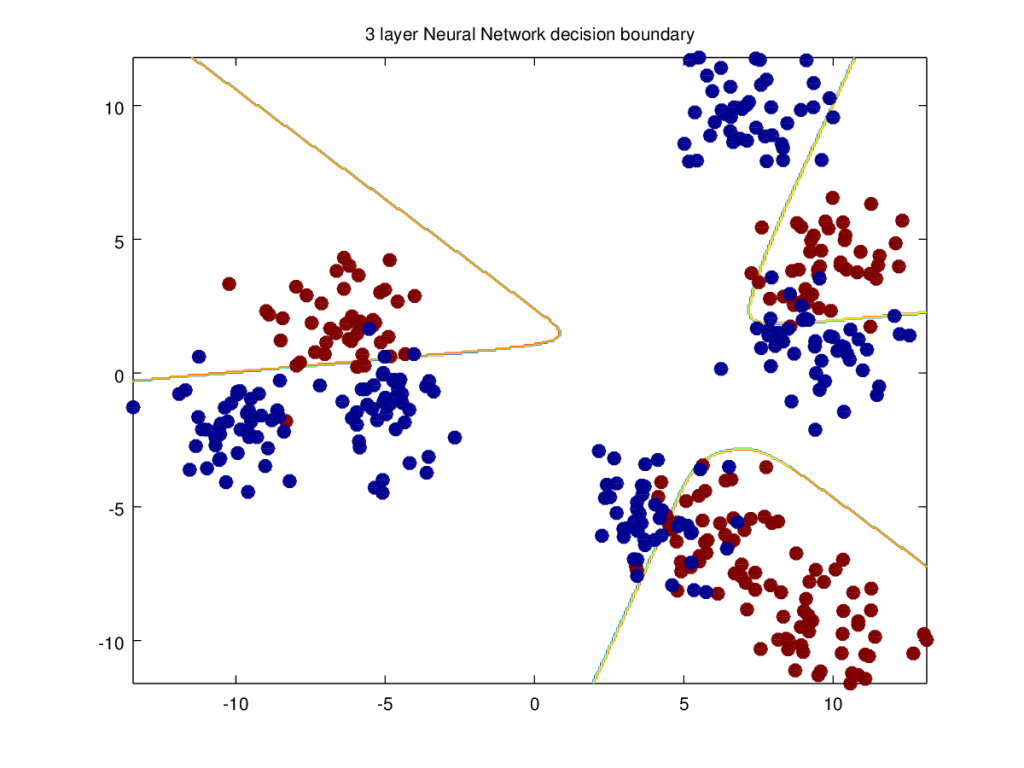
## [1] 0.209409

plotDecisionBoundary(z,nn,6,0.1)



**7.  The 3 layer Neural Network in Octave (vectorized)**

This uses the same dataset that was generated using Python code.  
source("DL-function2.m")  
data=csvread("data.csv");  
X=data(:,1:2);  
Y=data(:,3);  
# Make sure that the model parameters are correct. Take the transpose of X & Y  
#Perform gradient descent  
[W1,b1,W2,b2,costs]= computeNN(X', Y',4, learningRate=0.5, numIterations = 10000);

****

**8a. Performance  for different learning rates (Python)**

import numpy as np

import matplotlib

import matplotlib.pyplot as plt

import sklearn.linear\_model

import pandas as pd

from sklearn.datasets import make\_classification, make\_blobs

execfile("./DLfunctions.py") # Since import does not work in Rmd!!!

# Create data

X1, Y1 = make\_blobs(n\_samples = 400, n\_features = 2, centers = 9,

cluster\_std = 1.3, random\_state = 4)

#Create 2 classes

Y1=Y1.reshape(400,1)

Y1 = Y1 % 2

X2=X1.T

Y2=Y1.T

# Create a list of learning rates

learningRate=[0.5,1.2,3.0]

df=pd.DataFrame()

#Compute costs for each learning rate

for lr in learningRate:

parameters,costs = computeNN(X2, Y2, numHidden = 4, learningRate=lr, numIterations = 10000)

print(costs)

df1=pd.DataFrame(costs)

df=pd.concat([df,df1],axis=1)

#Set the iterations

iterations=[0,1000,2000,3000,4000,5000,6000,7000,8000,9000]

#Create data frame

#Set index

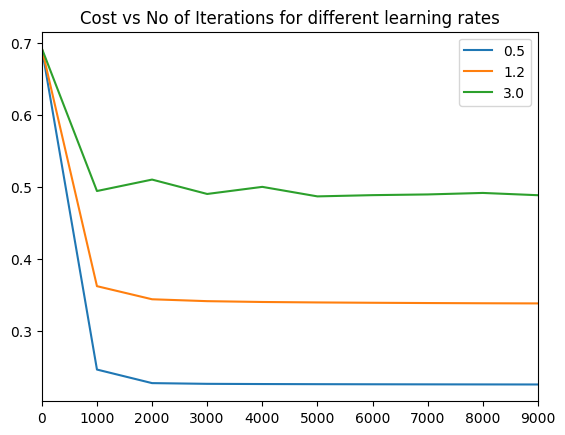
df1=df.set\_index([iterations])

df1.columns=[0.5,1.2,3.0]

fig=df1.plot()

fig=plt.title("Cost vs No of Iterations for different learning rates")

plt.savefig('fig4.png', bbox\_inches='tight')

****

**8b. Performance  for different hidden units (Python)**

import numpy as np

import matplotlib

import matplotlib.pyplot as plt

import sklearn.linear\_model

import pandas as pd

from sklearn.datasets import make\_classification, make\_blobs

execfile("./DLfunctions.py") # Since import does not work in Rmd!!!

#Create data set

X1, Y1 = make\_blobs(n\_samples = 400, n\_features = 2, centers = 9,

cluster\_std = 1.3, random\_state = 4)

#Create 2 classes

Y1=Y1.reshape(400,1)

Y1 = Y1 % 2

X2=X1.T

Y2=Y1.T

# Make a list of hidden unis

numHidden=[3,5,7]

df=pd.DataFrame()

#Compute costs for different hidden units

for numHid in numHidden:

parameters,costs = computeNN(X2, Y2, numHidden = numHid, learningRate=1.2, numIterations = 10000)

print(costs)

df1=pd.DataFrame(costs)

df=pd.concat([df,df1],axis=1)

#Set the iterations

iterations=[0,1000,2000,3000,4000,5000,6000,7000,8000,9000]

#Set index

df1=df.set\_index([iterations])

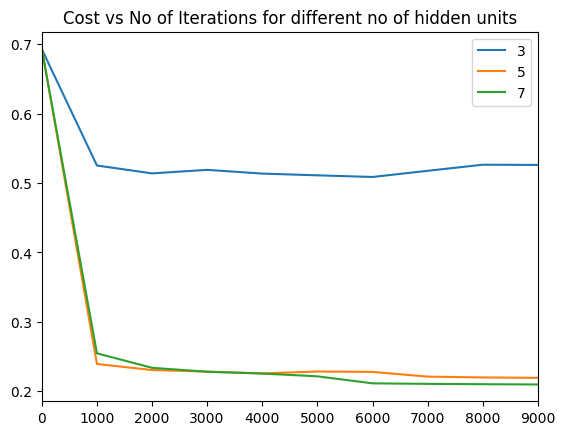
df1.columns=[3,5,7]

#Plot

fig=df1.plot()

fig=plt.title("Cost vs No of Iterations for different no of hidden units")

plt.savefig('fig5.png', bbox\_inches='tight')

****

**9a. Performance  for different learning rates (R)**

source("DLfunctions2\_1.R")

# Read data

z <- as.matrix(read.csv("data.csv",header=FALSE)) #

x <- z[,1:2]

y <- z[,3]

x1 <- t(x)

y1 <- t(y)

#Loop through learning rates and compute costs

learningRate <-c(0.1,1.2,3.0)

df <- NULL

for(i in seq\_along(learningRate)){

nn <- computeNN(x1, y1, 6, learningRate=learningRate[i],numIterations=10000)

cost <- nn$costs

df <- cbind(df,cost)

}

#Create dataframe

df <- data.frame(df)

iterations=seq(0,10000,by=1000)

df <- cbind(iterations,df)

names(df) <- c("iterations","0.5","1.2","3.0")

library(reshape2)

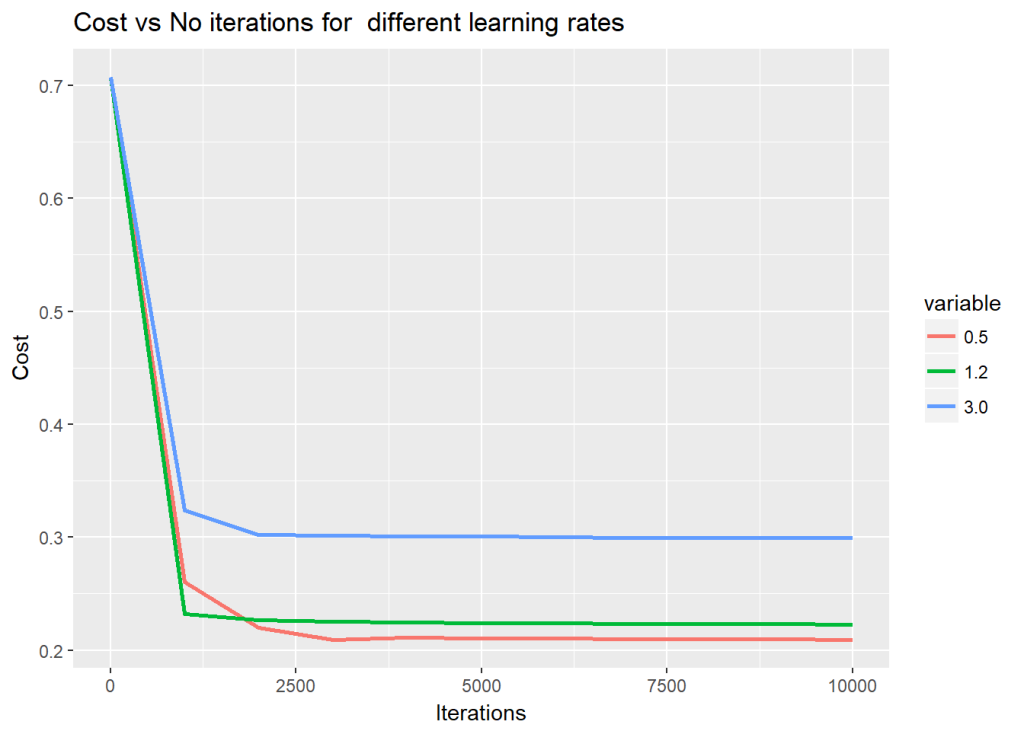
df1 <- melt(df,id="iterations") # Melt the data

#Plot

ggplot(df1) + geom\_line(aes(x=iterations,y=value,colour=variable),size=1) +

xlab("Iterations") +

ylab('Cost') + ggtitle("Cost vs No iterations for different learning rates")



**9b. Performance  for different hidden units (R)**

source("DLfunctions2\_1.R")

# Loop through Num hidden units

numHidden <-c(4,6,9)

df <- NULL

for(i in seq\_along(numHidden)){

nn <- computeNN(x1, y1, numHidden[i], learningRate=0.1,numIterations=10000)

cost <- nn$costs

df <- cbind(df,cost)

}

df <- data.frame(df)

iterations=seq(0,10000,by=1000)

df <- cbind(iterations,df)

names(df) <- c("iterations","4","6","9")

library(reshape2)

# Melt

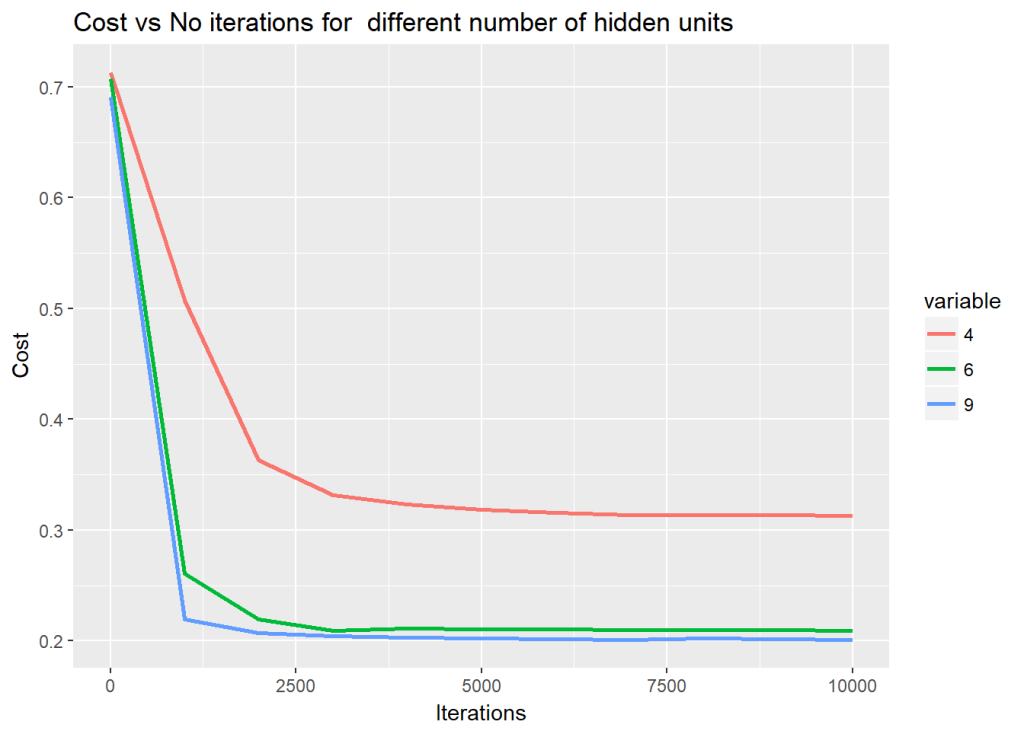
df1 <- melt(df,id="iterations")

# Plot

ggplot(df1) + geom\_line(aes(x=iterations,y=value,colour=variable),size=1) +

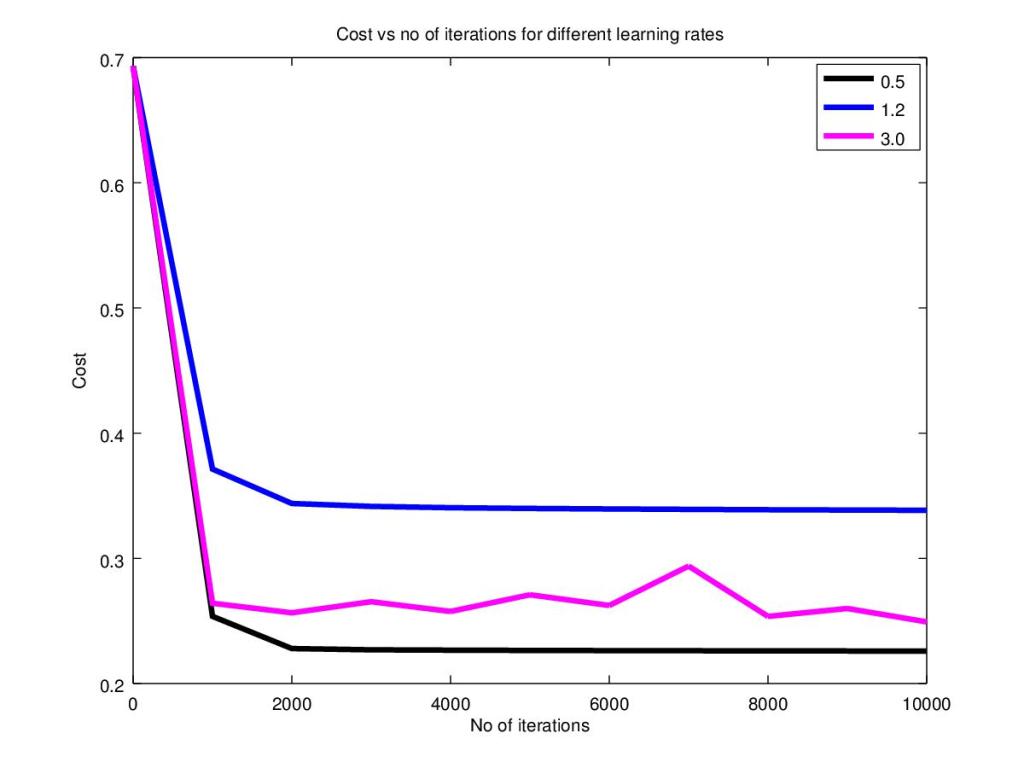
xlab("Iterations") +

ylab('Cost') + ggtitle("Cost vs No iterations for different number of hidden units")

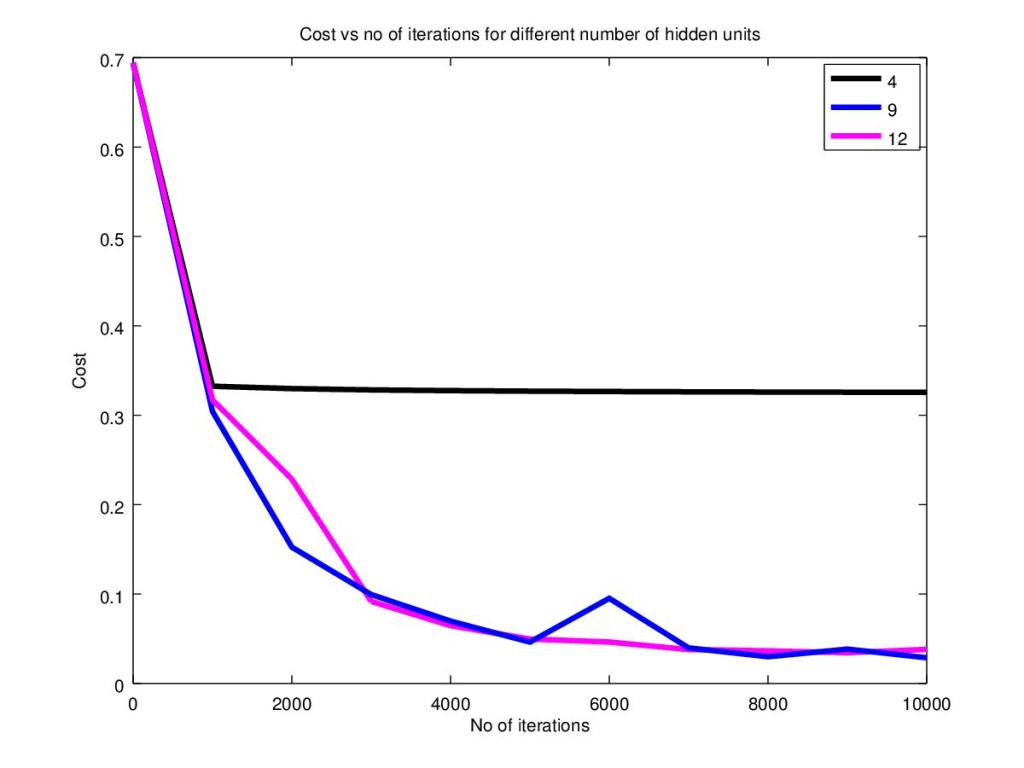


**10a. Performance of the Neural Network for different learning rates (Octave)**

source("DL-function2.m")  
plotLRCostVsIterations()  
print -djph figa.jpg



**10b. Performance of the Neural Network for different number of hidden units (Octave)**

source("DL-function2.m")  
plotHiddenCostVsIterations()  
print -djph figa.jpg  


**11. Turning the heat on the Neural Network**

In this 2nd part I create a a central region of positives and and the outside region as negatives. The points are generated using the equation of a circle (x – a)^{2} + (y -b) ^{2} = R^{2} . How does the 3 layer Neural Network perform on this?  Here’s a look! **Note**: *The same dataset is also used for R and Octave Neural Network constructions*

**12. Manually creating a circular central region**

import numpy as np

import matplotlib.pyplot as plt

import matplotlib.colors

import sklearn.linear\_model

from sklearn.model\_selection import train\_test\_split

from sklearn.datasets import make\_classification, make\_blobs

from matplotlib.colors import ListedColormap

import sklearn

import sklearn.datasets

colors=['black','gold']

cmap = matplotlib.colors.ListedColormap(colors)

x1=np.random.uniform(0,10,800).reshape(800,1)

x2=np.random.uniform(0,10,800).reshape(800,1)

X=np.append(x1,x2,axis=1)

X.shape

# Create (x-a)^2 + (y-b)^2 = R^2

# Create a subset of values where squared is <0,4. Perform ravel() to flatten this vector

a=(np.power(X[:,0]-5,2) + np.power(X[:,1]-5,2) <= 6).ravel()

Y=a.reshape(800,1)

cmap = matplotlib.colors.ListedColormap(colors)

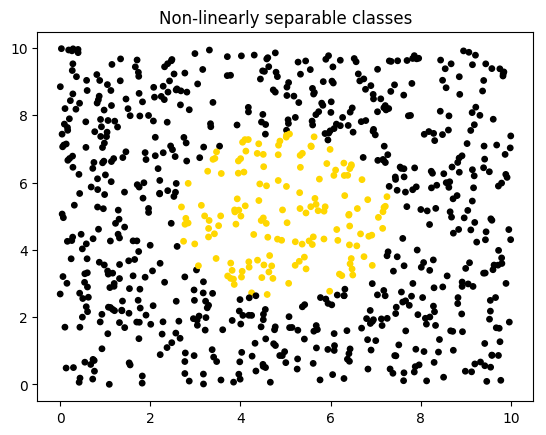
plt.figure()

plt.title('Non-linearly separable classes')

plt.scatter(X[:,0], X[:,1], c=Y,

marker= 'o', s=15,cmap=cmap)

plt.savefig('fig6.png', bbox\_inches='tight')

****

**13a. Decision boundary with hidden units=4 and learning rate = 2.2 (Python)**

With the above hyper parameters the decision boundary is triangular

import numpy as np

import matplotlib.pyplot as plt

import matplotlib.colors

import sklearn.linear\_model

execfile("./DLfunctions.py")

x1=np.random.uniform(0,10,800).reshape(800,1)

x2=np.random.uniform(0,10,800).reshape(800,1)

X=np.append(x1,x2,axis=1)

X.shape

# Create a subset of values where squared is <0,4. Perform ravel() to flatten this vector

a=(np.power(X[:,0]-5,2) + np.power(X[:,1]-5,2) <= 6).ravel()

Y=a.reshape(800,1)

X2=X.T

Y2=Y.T

parameters,costs = computeNN(X2, Y2, numHidden = 4, learningRate=2.2, numIterations = 10000)

plot\_decision\_boundary(lambda x: predict(parameters, x.T), X2, Y2,str(4),str(2.2),"fig7.png")

## Cost after iteration 0: 0.692836

## Cost after iteration 1000: 0.331052

## Cost after iteration 2000: 0.326428

## Cost after iteration 3000: 0.474887

## Cost after iteration 4000: 0.247989

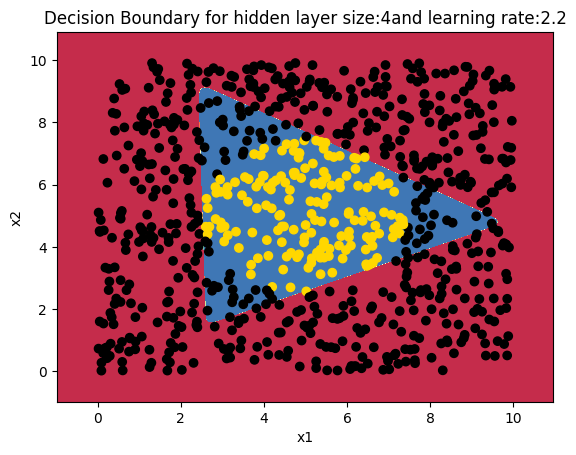
## Cost after iteration 5000: 0.218009

## Cost after iteration 6000: 0.201034

## Cost after iteration 7000: 0.197030

## Cost after iteration 8000: 0.193507

## Cost after iteration 9000: 0.191949

****

**13b. Decision boundary with hidden units=12 and learning rate = 2.2 (Python)**

With the above hyper parameters the decision boundary is triangular

import numpy as np

import matplotlib.pyplot as plt

import matplotlib.colors

import sklearn.linear\_model

execfile("./DLfunctions.py")

x1=np.random.uniform(0,10,800).reshape(800,1)

x2=np.random.uniform(0,10,800).reshape(800,1)

X=np.append(x1,x2,axis=1)

X.shape

# Create a subset of values where squared is <0,4. Perform ravel() to flatten this vector

a=(np.power(X[:,0]-5,2) + np.power(X[:,1]-5,2) <= 6).ravel()

Y=a.reshape(800,1)

X2=X.T

Y2=Y.T

parameters,costs = computeNN(X2, Y2, numHidden = 12, learningRate=2.2, numIterations = 10000)

plot\_decision\_boundary(lambda x: predict(parameters, x.T), X2, Y2,str(12),str(2.2),"fig8.png")

## Cost after iteration 0: 0.693291

## Cost after iteration 1000: 0.383318

## Cost after iteration 2000: 0.298807

## Cost after iteration 3000: 0.251735

## Cost after iteration 4000: 0.177843

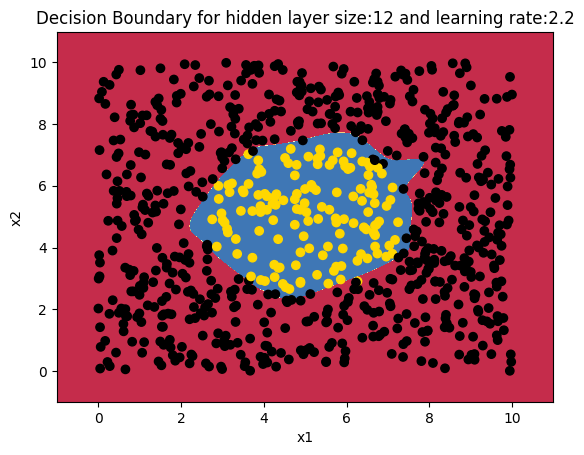
## Cost after iteration 5000: 0.130414

## Cost after iteration 6000: 0.152400

## Cost after iteration 7000: 0.065359

## Cost after iteration 8000: 0.050921

## Cost after iteration 9000: 0.039719

****

**14a. Decision boundary with hidden units=9 and learning rate = 0.5 (R)**

When the number of hidden units is 6 and the learning rate is 0,1, is also a triangular shape in R

source("DLfunctions2\_1.R")

z <- as.matrix(read.csv("data1.csv",header=FALSE)) # N

x <- z[,1:2]

y <- z[,3]

x1 <- t(x)

y1 <- t(y)

nn <-computeNN(x1, y1, 9, learningRate=0.5,numIterations=10000) # Triangular

## [1] 0.8398838

## [1] 0.3303621

## [1] 0.3127731

## [1] 0.3012791

## [1] 0.3305543

## [1] 0.3303964

## [1] 0.2334615

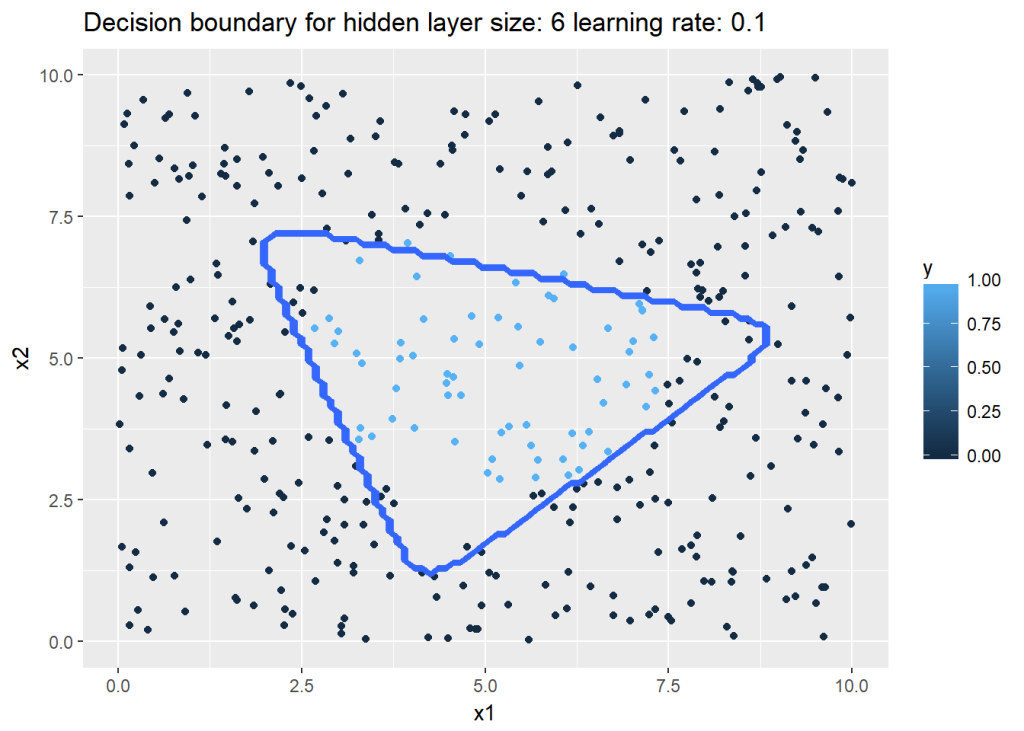
## [1] 0.1920771

## [1] 0.2341225

## [1] 0.2188118

## [1] 0.2082687

plotDecisionBoundary(z,nn,6,0.1)



**14b. Decision boundary with hidden units=8 and learning rate = 0.1 (R)**

source("DLfunctions2\_1.R")

z <- as.matrix(read.csv("data1.csv",header=FALSE)) # N

x <- z[,1:2]

y <- z[,3]

x1 <- t(x)

y1 <- t(y)

nn <-computeNN(x1, y1, 8, learningRate=0.1,numIterations=10000) # Hemisphere

## [1] 0.7273279

## [1] 0.3169335

## [1] 0.2378464

## [1] 0.1688635

## [1] 0.1368466

## [1] 0.120664

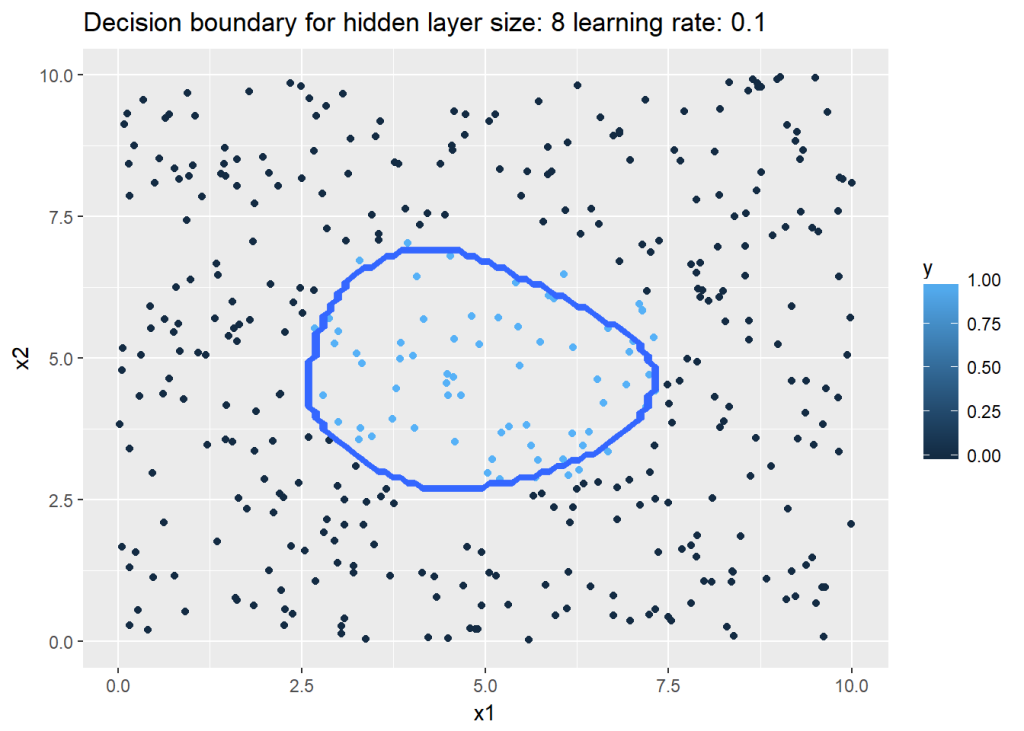
## [1] 0.111211

## [1] 0.1043362

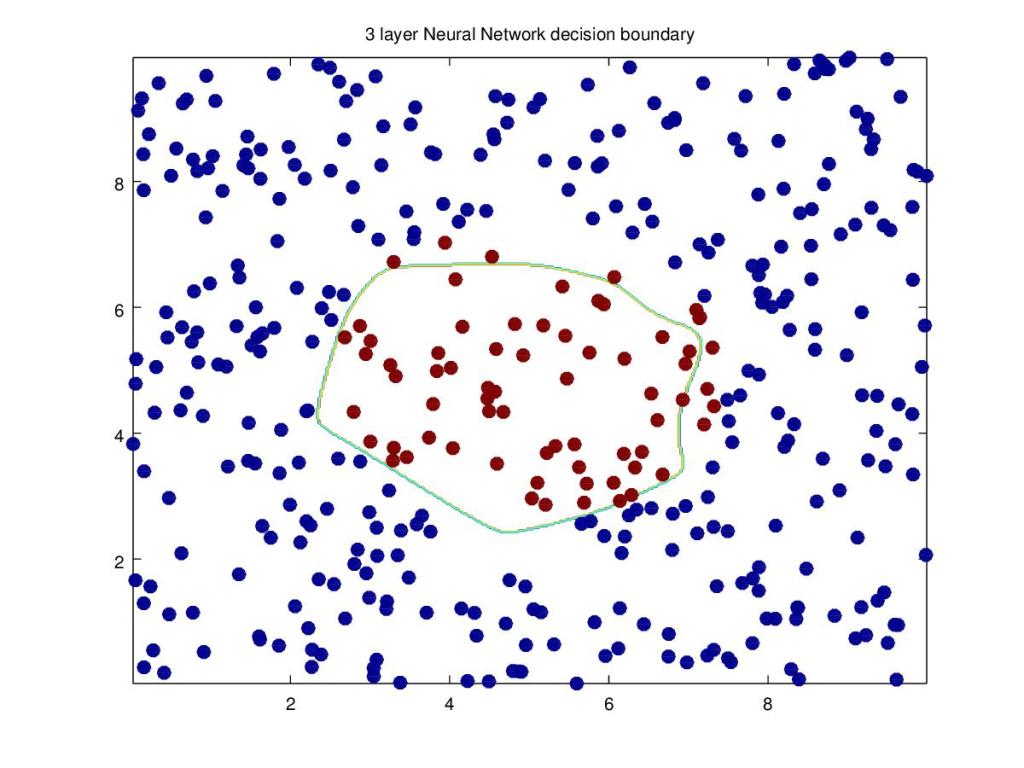
## [1] 0.09800573

## [1] 0.09126161

## [1] 0.0840379

plotDecisionBoundary(z,nn,8,0.1)

**15a. Decision boundary with hidden units=12 and learning rate = 1.5 (Octave)**

source("DL-function2.m")  
data=csvread("data1.csv");  
X=data(:,1:2);  
Y=data(:,3);  
# Make sure that the model parameters are correct. Take the transpose of X & Y  
[W1,b1,W2,b2,costs]= computeNN(X', Y',12, learningRate=1.5, numIterations = 10000);  
plotDecisionBoundary(data, W1,b1,W2,b2)  
print -djpg fige.jpg  


**Conclusion**: This post implemented a 3 layer Neural Network to create non-linear boundaries while performing classification. Clearly the Neural Network performs very well when the number of hidden units and learning rate are varied.