**Initialization**[**¶**](https://wkjfynxt.labs.coursera.org/notebooks/week5/Initialization/Initialization.ipynb#Initialization)

Welcome to the first assignment of "Improving Deep Neural Networks".

Training your neural network requires specifying an initial value of the weights. A well chosen initialization method will help learning.

If you completed the previous course of this specialization, you probably followed our instructions for weight initialization, and it has worked out so far. But how do you choose the initialization for a new neural network? In this notebook, you will see how different initializations lead to different results.

A well chosen initialization can:

* Speed up the convergence of gradient descent
* Increase the odds of gradient descent converging to a lower training (and generalization) error

To get started, run the following cell to load the packages and the planar dataset you will try to classify.

In [1]:

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** sklearn

**import** sklearn.datasets

**from** init\_utils **import** sigmoid, relu, compute\_loss, forward\_propagation, backward\_propagation

**from** init\_utils **import** update\_parameters, predict, load\_dataset, plot\_decision\_boundary, predict\_dec

​

**%**matplotlib inline

plt.rcParams['figure.figsize'] = (7.0, 4.0) *# set default size of plots*

plt.rcParams['image.interpolation'] = 'nearest'

plt.rcParams['image.cmap'] = 'gray'

​

*# load image dataset: blue/red dots in circles*

train\_X, train\_Y, test\_X, test\_Y = load\_dataset()

Chart, scatter chart

Description automatically generated

You would like a classifier to separate the blue dots from the red dots.

**1 - Neural Network model**

You will use a 3-layer neural network (already implemented for you). Here are the initialization methods you will experiment with:

* *Zeros initialization* -- setting initialization = "zeros" in the input argument.
* *Random initialization* -- setting initialization = "random" in the input argument. This initializes the weights to large random values.
* *He initialization* -- setting initialization = "he" in the input argument. This initializes the weights to random values scaled according to a paper by He et al., 2015.

**Instructions**: Please quickly read over the code below, and run it. In the next part you will implement the three initialization methods that this model() calls.

In [2]:

**def** model(X, Y, learning\_rate = 0.01, num\_iterations = 15000, print\_cost = **True**, initialization = "he"):

"""

Implements a three-layer neural network: LINEAR->RELU->LINEAR->RELU->LINEAR->SIGMOID.

Arguments:

X -- input data, of shape (2, number of examples)

Y -- true "label" vector (containing 0 for red dots; 1 for blue dots), of shape (1, number of examples)

learning\_rate -- learning rate for gradient descent

num\_iterations -- number of iterations to run gradient descent

print\_cost -- if True, print the cost every 1000 iterations

initialization -- flag to choose which initialization to use ("zeros","random" or "he")

Returns:

parameters -- parameters learnt by the model

"""

grads = {}

costs = [] *# to keep track of the loss*

m = X.shape[1] *# number of examples*

layers\_dims = [X.shape[0], 10, 5, 1]

*# Initialize parameters dictionary.*

**if** initialization == "zeros":

parameters = initialize\_parameters\_zeros(layers\_dims)

**elif** initialization == "random":

parameters = initialize\_parameters\_random(layers\_dims)

**elif** initialization == "he":

parameters = initialize\_parameters\_he(layers\_dims)

​

*# Loop (gradient descent)*

​

**for** i **in** range(0, num\_iterations):

​

*# Forward propagation: LINEAR -> RELU -> LINEAR -> RELU -> LINEAR -> SIGMOID.*

a3, cache = forward\_propagation(X, parameters)

*# Loss*

cost = compute\_loss(a3, Y)

​

*# Backward propagation.*

grads = backward\_propagation(X, Y, cache)

*# Update parameters.*

parameters = update\_parameters(parameters, grads, learning\_rate)

*# Print the loss every 1000 iterations*

**if** print\_cost **and** i **%** 1000 == 0:

print("Cost after iteration {}: {}".format(i, cost))

costs.append(cost)

*# plot the loss*

plt.plot(costs)

plt.ylabel('cost')

plt.xlabel('iterations (per hundreds)')

plt.title("Learning rate =" **+** str(learning\_rate))

plt.show()

**return** parameters

**2 - Zero initialization**

There are two types of parameters to initialize in a neural network:

* the weight matrices (W[1],W[2],W[3],...,W[L−1],W[L])(W[1],W[2],W[3],...,W[L−1],W[L])
* the bias vectors (b[1],b[2],b[3],...,b[L−1],b[L])(b[1],b[2],b[3],...,b[L−1],b[L])

**Exercise**: Implement the following function to initialize all parameters to zeros. You'll see later that this does not work well since it fails to "break symmetry", but lets try it anyway and see what happens. Use np.zeros((..,..)) with the correct shapes.

In [7]:

*# GRADED FUNCTION: initialize\_parameters\_zeros*

​

**def** initialize\_parameters\_zeros(layers\_dims):

"""

Arguments:

layer\_dims -- python array (list) containing the size of each layer.

Returns:

parameters -- python dictionary containing your parameters "W1", "b1", ..., "WL", "bL":

W1 -- weight matrix of shape (layers\_dims[1], layers\_dims[0])

b1 -- bias vector of shape (layers\_dims[1], 1)

...

WL -- weight matrix of shape (layers\_dims[L], layers\_dims[L-1])

bL -- bias vector of shape (layers\_dims[L], 1)

"""

parameters = {}

L = len(layers\_dims) *# number of layers in the network*

**for** l **in** range(1, L):

*### START CODE HERE ### (≈ 2 lines of code)*

parameters['W' **+** str(l)] = np.zeros((layers\_dims[l], layers\_dims[l**-**1]))

parameters['b' **+** str(l)] = np.zeros((layers\_dims[l], 1))

*### END CODE HERE ###*

**return** parameters

In [8]:

parameters = initialize\_parameters\_zeros([3,2,1])

print("W1 = " **+** str(parameters["W1"]))

print("b1 = " **+** str(parameters["b1"]))

print("W2 = " **+** str(parameters["W2"]))

print("b2 = " **+** str(parameters["b2"]))

W1 = [[ 0. 0. 0.]

[ 0. 0. 0.]]

b1 = [[ 0.]

[ 0.]]

W2 = [[ 0. 0.]]

b2 = [[ 0.]]

**Expected Output**:

|  |  |
| --- | --- |
| **W1** | [[ 0. 0. 0.] [ 0. 0. 0.]] |
| **b1** | [[ 0.] [ 0.]] |
| **W2** | [[ 0. 0.]] |
| **b2** | [[ 0.]] |

Run the following code to train your model on 15,000 iterations using zeros initialization.

In [9]:

parameters = model(train\_X, train\_Y, initialization = "zeros")

print ("On the train set:")

predictions\_train = predict(train\_X, train\_Y, parameters)

print ("On the test set:")

predictions\_test = predict(test\_X, test\_Y, parameters)

Cost after iteration 0: 0.6931471805599453

Cost after iteration 1000: 0.6931471805599453

Cost after iteration 2000: 0.6931471805599453

Cost after iteration 3000: 0.6931471805599453

Cost after iteration 4000: 0.6931471805599453

Cost after iteration 5000: 0.6931471805599453

Cost after iteration 6000: 0.6931471805599453

Cost after iteration 7000: 0.6931471805599453

Cost after iteration 8000: 0.6931471805599453

Cost after iteration 9000: 0.6931471805599453

Cost after iteration 10000: 0.6931471805599455

Cost after iteration 11000: 0.6931471805599453

Cost after iteration 12000: 0.6931471805599453

Cost after iteration 13000: 0.6931471805599453

Cost after iteration 14000: 0.6931471805599453

Shape, rectangle

Description automatically generated

On the train set:

Accuracy: 0.5

On the test set:

Accuracy: 0.5

The performance is really bad, and the cost does not really decrease, and the algorithm performs no better than random guessing. Why? Lets look at the details of the predictions and the decision boundary:

In [10]:

print ("predictions\_train = " **+** str(predictions\_train))

print ("predictions\_test = " **+** str(predictions\_test))

predictions\_train = [[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0]]

predictions\_test = [[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]

In [11]:

plt.title("Model with Zeros initialization")

axes = plt.gca()

axes.set\_xlim([**-**1.5,1.5])

axes.set\_ylim([**-**1.5,1.5])

plot\_decision\_boundary(**lambda** x: predict\_dec(parameters, x.T), train\_X, train\_Y)

Shape

Description automatically generated

The model is predicting 0 for every example.

In general, initializing all the weights to zero results in the network failing to break symmetry. This means that every neuron in each layer will learn the same thing, and you might as well be training a neural network with n[l]=1n[l]=1 for every layer, and the network is no more powerful than a linear classifier such as logistic regression.

**What you should remember**:

* The weights W[l]W[l] should be initialized randomly to break symmetry.
* It is however okay to initialize the biases b[l]b[l] to zeros. Symmetry is still broken so long as W[l]W[l] is initialized randomly.

**3 - Random initialization**

To break symmetry, lets intialize the weights randomly. Following random initialization, each neuron can then proceed to learn a different function of its inputs. In this exercise, you will see what happens if the weights are intialized randomly, but to very large values.

**Exercise**: Implement the following function to initialize your weights to large random values (scaled by \*10) and your biases to zeros. Use np.random.randn(..,..) \* 10 for weights and np.zeros((.., ..)) for biases. We are using a fixed np.random.seed(..) to make sure your "random" weights match ours, so don't worry if running several times your code gives you always the same initial values for the parameters.

In [12]:

*# GRADED FUNCTION: initialize\_parameters\_random*

​

**def** initialize\_parameters\_random(layers\_dims):

"""

Arguments:

layer\_dims -- python array (list) containing the size of each layer.

Returns:

parameters -- python dictionary containing your parameters "W1", "b1", ..., "WL", "bL":

W1 -- weight matrix of shape (layers\_dims[1], layers\_dims[0])

b1 -- bias vector of shape (layers\_dims[1], 1)

...

WL -- weight matrix of shape (layers\_dims[L], layers\_dims[L-1])

bL -- bias vector of shape (layers\_dims[L], 1)

"""

np.random.seed(3) *# This seed makes sure your "random" numbers will be the as ours*

parameters = {}

L = len(layers\_dims) *# integer representing the number of layers*

**for** l **in** range(1, L):

*### START CODE HERE ### (≈ 2 lines of code)*

parameters['W' **+** str(l)] = np.random.randn(layers\_dims[l], layers\_dims[l**-**1]) **\*** 10

parameters['b' **+** str(l)] = np.zeros((layers\_dims[l], 1))

*### END CODE HERE ###*

​

**return** parameters

In [13]:

parameters = initialize\_parameters\_random([3, 2, 1])

print("W1 = " **+** str(parameters["W1"]))

print("b1 = " **+** str(parameters["b1"]))

print("W2 = " **+** str(parameters["W2"]))

print("b2 = " **+** str(parameters["b2"]))

W1 = [[ 17.88628473 4.36509851 0.96497468]

[-18.63492703 -2.77388203 -3.54758979]]

b1 = [[ 0.]

[ 0.]]

W2 = [[-0.82741481 -6.27000677]]

b2 = [[ 0.]]

**Expected Output**:

|  |  |
| --- | --- |
| **W1** | [[ 17.88628473 4.36509851 0.96497468] [-18.63492703 -2.77388203 -3.54758979]] |
| **b1** | [[ 0.] [ 0.]] |
| **W2** | [[-0.82741481 -6.27000677]] |
| **b2** | [[ 0.]] |

Run the following code to train your model on 15,000 iterations using random initialization.

In [14]:

parameters = model(train\_X, train\_Y, initialization = "random")

print ("On the train set:")

predictions\_train = predict(train\_X, train\_Y, parameters)

print ("On the test set:")

predictions\_test = predict(test\_X, test\_Y, parameters)

/home/jovyan/work/week5/Initialization/init\_utils.py:145: RuntimeWarning: divide by zero encountered in log

logprobs = np.multiply(-np.log(a3),Y) + np.multiply(-np.log(1 - a3), 1 - Y)

/home/jovyan/work/week5/Initialization/init\_utils.py:145: RuntimeWarning: invalid value encountered in multiply

logprobs = np.multiply(-np.log(a3),Y) + np.multiply(-np.log(1 - a3), 1 - Y)

Cost after iteration 0: inf

Cost after iteration 1000: 0.6242434241539614

Cost after iteration 2000: 0.5978811277755388

Cost after iteration 3000: 0.5636242569764779

Cost after iteration 4000: 0.5500958254523324

Cost after iteration 5000: 0.544339206192789

Cost after iteration 6000: 0.5373584514307651

Cost after iteration 7000: 0.469574666760224

Cost after iteration 8000: 0.39766324943219844

Cost after iteration 9000: 0.3934423376823982

Cost after iteration 10000: 0.3920158992175907

Cost after iteration 11000: 0.38913979237487845

Cost after iteration 12000: 0.3861261344766218

Cost after iteration 13000: 0.3849694511273874

Cost after iteration 14000: 0.3827489017191917

Chart, line chart

Description automatically generated

On the train set:

Accuracy: 0.83

On the test set:

Accuracy: 0.86

If you see "inf" as the cost after the iteration 0, this is because of numerical roundoff; a more numerically sophisticated implementation would fix this. But this isn't worth worrying about for our purposes.

Anyway, it looks like you have broken symmetry, and this gives better results. than before. The model is no longer outputting all 0s.

In [15]:

print (predictions\_train)

print (predictions\_test)

[[1 0 1 1 0 0 1 1 1 1 1 0 1 0 0 1 0 1 1 0 0 0 1 0 1 1 1 1 1 1 0 1 1 0 0 1 1

1 1 1 1 1 1 0 1 1 1 1 0 1 0 1 1 1 1 0 0 1 1 1 1 0 1 1 0 1 0 1 1 1 1 0 0 0

0 0 1 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0 0 1 1 1 0 1 1 0 1 0 1 1 0 1 1 0 1 0 1

1 0 0 1 0 0 1 1 0 1 1 1 0 1 0 0 1 0 1 1 1 1 1 1 1 0 1 1 0 0 1 1 0 0 0 1 0

1 0 1 0 1 1 1 0 0 1 1 1 1 0 1 1 0 1 0 1 1 0 1 0 1 1 1 1 0 1 1 1 1 0 1 0 1

0 1 1 1 1 0 1 1 0 1 1 0 1 1 0 1 0 1 1 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1 0 1 1

0 1 1 0 1 1 1 0 1 1 1 1 0 1 0 0 1 1 0 1 1 1 0 0 0 1 1 0 1 1 1 1 0 1 1 0 1

1 1 0 0 1 0 0 0 1 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 0 1

1 1 1 0]]

[[1 1 1 1 0 1 0 1 1 0 1 1 1 0 0 0 0 1 0 1 0 0 1 0 1 0 1 1 1 1 1 0 0 0 0 1 0

1 1 0 0 1 1 1 1 1 0 1 1 1 0 1 0 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1

1 1 1 0 1 0 0 1 0 0 0 1 1 0 1 1 0 0 0 1 1 0 1 1 0 0]]

In [16]:

plt.title("Model with large random initialization")

axes = plt.gca()

axes.set\_xlim([**-**1.5,1.5])

axes.set\_ylim([**-**1.5,1.5])

plot\_decision\_boundary(**lambda** x: predict\_dec(parameters, x.T), train\_X, train\_Y)

Chart

Description automatically generated

**Observations**:

* The cost starts very high. This is because with large random-valued weights, the last activation (sigmoid) outputs results that are very close to 0 or 1 for some examples, and when it gets that example wrong it incurs a very high loss for that example. Indeed, when log(a[3])=log(0)log⁡(a[3])=log⁡(0), the loss goes to infinity.
* Poor initialization can lead to vanishing/exploding gradients, which also slows down the optimization algorithm.
* If you train this network longer you will see better results, but initializing with overly large random numbers slows down the optimization.

**In summary**:

* Initializing weights to very large random values does not work well.
* Hopefully intializing with small random values does better. The important question is: how small should be these random values be? Lets find out in the next part!

**4 - He initialization**

Finally, try "He Initialization"; this is named for the first author of He et al., 2015. (If you have heard of "Xavier initialization", this is similar except Xavier initialization uses a scaling factor for the weights W[l]W[l] of sqrt(1./layers\_dims[l-1]) where He initialization would use sqrt(2./layers\_dims[l-1]).)

**Exercise**: Implement the following function to initialize your parameters with He initialization.

**Hint**: This function is similar to the previous initialize\_parameters\_random(...). The only difference is that instead of multiplying np.random.randn(..,..) by 10, you will multiply it by 2dimension of the previous layer⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯√2dimension of the previous layer, which is what He initialization recommends for layers with a ReLU activation.

In [17]:

*# GRADED FUNCTION: initialize\_parameters\_he*

​

**def** initialize\_parameters\_he(layers\_dims):

"""

Arguments:

layer\_dims -- python array (list) containing the size of each layer.

Returns:

parameters -- python dictionary containing your parameters "W1", "b1", ..., "WL", "bL":

W1 -- weight matrix of shape (layers\_dims[1], layers\_dims[0])

b1 -- bias vector of shape (layers\_dims[1], 1)

...

WL -- weight matrix of shape (layers\_dims[L], layers\_dims[L-1])

bL -- bias vector of shape (layers\_dims[L], 1)

"""

np.random.seed(3)

parameters = {}

L = len(layers\_dims) **-** 1 *# integer representing the number of layers*

**for** l **in** range(1, L **+** 1):

*### START CODE HERE ### (≈ 2 lines of code)*

parameters['W' **+** str(l)] = np.random.randn(layers\_dims[l], layers\_dims[l**-**1]) **\*** np.sqrt(2 **/** layers\_dims[l**-**1])

parameters['b' **+** str(l)] = np.zeros((layers\_dims[l], 1))

*### END CODE HERE ###*

**return** parameters

In [18]:

parameters = initialize\_parameters\_he([2, 4, 1])

print("W1 = " **+** str(parameters["W1"]))

print("b1 = " **+** str(parameters["b1"]))

print("W2 = " **+** str(parameters["W2"]))

print("b2 = " **+** str(parameters["b2"]))

W1 = [[ 1.78862847 0.43650985]

[ 0.09649747 -1.8634927 ]

[-0.2773882 -0.35475898]

[-0.08274148 -0.62700068]]

b1 = [[ 0.]

[ 0.]

[ 0.]

[ 0.]]

W2 = [[-0.03098412 -0.33744411 -0.92904268 0.62552248]]

b2 = [[ 0.]]

**Expected Output**:

|  |  |
| --- | --- |
| **W1** | [[ 1.78862847 0.43650985] [ 0.09649747 -1.8634927 ] [-0.2773882 -0.35475898] [-0.08274148 -0.62700068]] |
| **b1** | [[ 0.] [ 0.] [ 0.] [ 0.]] |
| **W2** | [[-0.03098412 -0.33744411 -0.92904268 0.62552248]] |
| **b2** | [[ 0.]] |

Run the following code to train your model on 15,000 iterations using He initialization.

In [19]:

parameters = model(train\_X, train\_Y, initialization = "he")

print ("On the train set:")

predictions\_train = predict(train\_X, train\_Y, parameters)

print ("On the test set:")

predictions\_test = predict(test\_X, test\_Y, parameters)

Cost after iteration 0: 0.8830537463419761

Cost after iteration 1000: 0.6879825919728063

Cost after iteration 2000: 0.6751286264523371

Cost after iteration 3000: 0.6526117768893807

Cost after iteration 4000: 0.6082958970572938

Cost after iteration 5000: 0.5304944491717495

Cost after iteration 6000: 0.4138645817071794

Cost after iteration 7000: 0.3117803464844441

Cost after iteration 8000: 0.23696215330322562

Cost after iteration 9000: 0.18597287209206834

Cost after iteration 10000: 0.15015556280371806

Cost after iteration 11000: 0.12325079292273546

Cost after iteration 12000: 0.09917746546525934

Cost after iteration 13000: 0.08457055954024278

Cost after iteration 14000: 0.07357895962677369

Chart, line chart

Description automatically generated

On the train set:

Accuracy: 0.993333333333

On the test set:

Accuracy: 0.96

In [20]:

plt.title("Model with He initialization")

axes = plt.gca()

axes.set\_xlim([**-**1.5,1.5])

axes.set\_ylim([**-**1.5,1.5])

plot\_decision\_boundary(**lambda** x: predict\_dec(parameters, x.T), train\_X, train\_Y)

A picture containing chart

Description automatically generated

**Observations**:

* The model with He initialization separates the blue and the red dots very well in a small number of iterations.

**5 - Conclusions**

You have seen three different types of initializations. For the same number of iterations and same hyperparameters the comparison is:

|  |  |  |
| --- | --- | --- |
| **Model** | **Train accuracy** | **Problem/Comment** |
| 3-layer NN with zeros initialization | 50% | fails to break symmetry |
| 3-layer NN with large random initialization | 83% | too large weights |
| 3-layer NN with He initialization | 99% | recommended method |

**What you should remember from this notebook**:

* Different initializations lead to different results
* Random initialization is used to break symmetry and make sure different hidden units can learn different things
* Don't intialize to values that are too large
* He initialization works well for networks with ReLU activations.