Deep Learning in Python

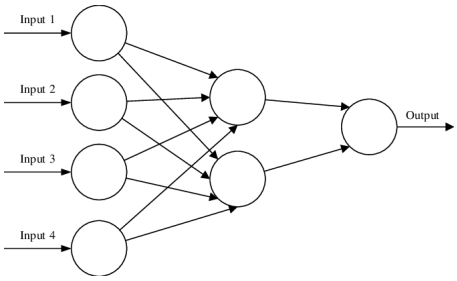
Contents

1 Neural Networks and Deep Learning

1.1 Logistic Regression as a Neural Network

1.1.1 Introduction to Logistic Regression

A single **neural network** can be built off of an input (x), an activation layer known as a *neuron*, and producing an output (y). A larger neural network is then formed by taking many of the single neurons and stacking them together. Each *feature* $(x_1, x_2, ..., x_n)$ can be used as an input to the activation layers to produce our output y. For the below example, we say that the layer in the middle is *densely connected* since every feature is in input.



To store an image, your computer stores three different matrices corresponding to the red, green, and blue channel (RGB values). So if your input image is 64x64 pixels, you will have three 64x64 matrices. To unroll these values into a **feature vector**, we will add values from all 3 vectors into a single x vector, where n_x is the number of features in the vector (in this case, 12288).

In **binary classification**, our goal is to learn a classifier that can input an image represented by feature vector x and predict whether the corresponding label y is a 1 or 0 (1 for cat, 0 for non cat).

Here is some common notation we will be using:

- Single training example: (x,y) where $x \in \mathbb{R}^{n_x}$, y $\in \{0, 1\}$
- M training examples: $\{(x^1, y^1), ..., (x^m, y^m)\}$
- Matrix X: $\begin{bmatrix} | & | & ... & | \\ x^1 & x^2 & ... & x^m \\ | & | & ... & | \end{bmatrix}$ where rows = n_x, columns = m. In Python, X.shape = (n_x, m).
- Matrix Y: $\begin{bmatrix} y^1, y^2, ..., y^m \end{bmatrix}$ where Y.shape = (1, m)

Given x, we want an estimate known as $\hat{y}$ = P(y=1|x) given the following parameters: $x \in \mathbb{R}^{n_x}$, $w \in \mathbb{R}^{n_x}$, and $b \in \mathbb{R}$. We want out output to be $0 \leq \hat{y} \leq 1$, so we will use the **sigmoid function** to find our output, which will be:

$$\hat{y} = \sigma(z) \text{ where } z = w^T x + b \text{ and } \sigma(z) = \frac{1}{1 + e^{-z}}$$

If z is large, then $\sigma(z)$ will be very close to 1. But if z is a large negative number, then $\sigma(z)$ will be very close to 0. So given $\{(x^1, y^1), ..., (x^m, y^m)\}$ we want $\hat{y}^i \approx y^i$

1.1.2 Logistic Regression Cost Function

We will associate x^i, y^i, and z^i with the i^{th} training example of our data. We will need to define a **loss function**, with respect to a single training example, to measure how good our output ($\hat{y}$) is when the true label is y. Since we are using gradient descent, we will define the following loss function:

$$\mathcal{L}(\hat{y}, y) = -(y\, log(\hat{y}) + (1 - y)\, log(1 - \hat{y}))$$

We want this loss function to be as small as possible. Lets look at the two cases:

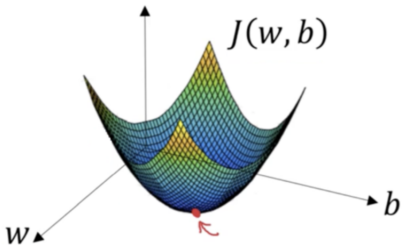
If y=1: $\mathcal{L}(\hat{y}, y) = -y\, log(\hat{y})$ and we want this to be as small as possible ($\hat{y}$ large).
If y=0: $\mathcal{L}(\hat{y}, y) = -\, log(1 - \hat{y})$ and we want this to be large ($\hat{y}$ small).

To train the parameters w and b, we need to define a **cost function**, which measures how well your doing on an entire training set (cost of the parameters). We will define this as:

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\hat{y}^i, y^i) = -\frac{1}{m}\sum_{i=1}^{m}[(y^i\, log(\hat{y}^i) + (1-y^i)\, log(1-\hat{y}^i))]$$

1.1.3 Gradient Descent

We want to find the values of w and b that will *minimize* the cost function J(w,b). Our cost function that we defined is convex (only one minimum). So we will initialize (w,b) to a random value, typically zero, and will take steps downhill in the steepest direction it can. Eventually it will converge to a minimum value and find our parameter values.



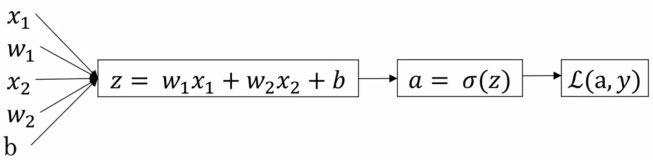
Gradient descent will repeatedly update the value of w and b with the formula:

$$w = w - \alpha\frac{\partial J(w,b)}{\partial w}\ ,\ b = b - \alpha\frac{\partial J(w,b)}{\partial b}$$

where α is our learning rate that we set multiplied by the partial derivative (since there are two variables) of the cost function with respect to the given parameter.

We want to modify out w and b parameters in order to reduce the loss when performing gradient descent on our Logistic Regression. We can set up a computation graph to find the derivatives through **backpropagation**. We will do this for a *single* training example, lets remind ourselves of our equations and the graph:

$z = w^T x + b$
$\hat{y} = a = \sigma(z)$
$\mathcal{L}(a,y) = -(y\, log(a) + (1-y)\, log(1-a))$



The first back step is to compute "da" $= \frac{\partial\mathcal{L}(a,y)}{\partial a} = \frac{-y}{a} - \frac{1-y}{1-a}$

Next we step back again and compute "dz" $= \frac{\partial\mathcal{L}(a,y)}{\partial z} = \frac{\partial\mathcal{L}}{\partial a} * \frac{\partial a}{\partial z} = a(1-a) * (\frac{-y}{a} + \frac{1-y}{1-a}) = a - y$

The final step back is to find how much to change our w_1, w_2, and b values. We can do this by:

- Calculating: $\frac{\partial\mathcal{L}}{\partial w_1}$ = "dw_1" $= x_1 * dz$, "dw_2" $= x_2 * dz$, and "db" $= dz$
- Then update our variables: $w_1 = w_1 - \alpha * dw_1$, $w_2 = w_2 - \alpha * dw_2$, and $b = b - \alpha * dz$

Now we want to **perform gradient descent on m examples**. This will use the cost function (not the loss function like we did on a single example). We will write sudo-code for Python that implements this m example gradient descent (assume that $n_x = 2$):

Initialize: J=0, dw_1=0, dw_2=0, db=0
for i=1 to m:
$z^i = w^T x^i + b$
$z^i = \sigma(z^i)$
$J+ = -[(y^i\, log(a^i) + (1-y^i)\, log(1-a^i))]$
$dz^i = a^i - y^i$
$dw_1+ = x_1^i * dz^i$
$dw_2+ = x_2^i * dz^i$
$db+ = dz^i$

After the loop, we then take the average and update our varaibles:

J /= m
dw_1 /= m
dw_2 /= m
db /= m
$w_1 = w_1 - \alpha * dw_1$
$w_2 = w_2 - \alpha * dw_2$
$b = b - \alpha * db$

1.2 Vectorization in Python

We often find ourselves training on big data sets and we need our code to run as fast as possible. This is where **vectorization** comes into play. One example is when we want to calculate $z = w^T x + b$. We can use *np.dot(w, x) + b* in Python rather than a *for loop* to decrease our run time by a significant amount. The takeaway from this section is that in deep learning, we want to avoid for loops and use vectorized code (such as NumPy methods) to save time when using large data sets.

1.2.1 Vectorizing Logistic Regression

Remember previously we had defined a matrix, X = $\begin{bmatrix} | & | & ... & | \\ x^1 & x^2 & ... & x^m \\ | & | & ... & | \end{bmatrix}$, that contains the training data.

In order to vectorize finding $z^i = w^T x^i + b$ for m training examples, we can instead create a vector of these z values, denoted by Z.

$$Z \ = \ [z^1, \ z^2, ..., \ z^m] = w^T X + [b, \ b, \ ..., \ b] = [w^t x^1 + b, \ w^t x^2 + b, \ ..., \ w^t x^m + b]$$

We will also want to vectorize our sigmoid function, which we will see in our programming assignment where we find A = $[a^1, a^2, ..., a^m] = \sigma(z)$. These are the forward propogations.

Next we will want to vectorize the remaining steps in order to speed up our code. Lets begin with $dz^i = a^i - y^i$. Instead of looping through each training example, we can use vectors we already created, A = $[a^1, a^2, ..., a^m]$ and Y = $[y^1, y^2, ..., y^m]$. We can define:

$$dZ = A - Y = [a^1 - y^1, a^2 - y^2, ..., a^m - y^m] = [dz^1, dz^2, ..., dz^m]$$

Now we want to vectorize dw for all training examples. We know that $dw^i += x^i_m * dz^i$ must be updated fo each example and then divided by the total number of examples (m), so we define this as:

$$dw = \frac{1}{m} X \, dz^T = \frac{1}{m} \begin{bmatrix} | & | & ... & | \\ x^1 & x^2 & ... & x^m \\ | & | & ... & | \end{bmatrix} \begin{bmatrix} dz^1 \\ | \\ dz^m \end{bmatrix} = \frac{1}{m} [x^1 dz^1 + ... + x^m dz^m]$$

Finally, we see db is the sum of dz^i divided by the total number of examples (m), we can write this as:

$$db \ = \ \frac{1}{m} \sum_{i=1}^{m} dz^i = np.sum(dZ)$$

Note that the below steps are only one step of gradient descent, you would still need a for loop to perform multiple steps. The new vectorized gradient descent can now be written as:

$Z = w^T X + b$
$A = \sigma(z)$
$dZ = A - Y$
$dw = \frac{1}{m} X \, dz^t$
$db = \frac{1}{m} \ np.sum(dZ)$
$w = w - \alpha dw$
$b = b - \alpha db$

Some quick notes on **broadcasting** in Python:

- When computing a mathematical operation on an (m,n) matrix with either a (1,n) or (m,1) vector, Python with automatically manipulate it to convert it into an (m,n) matrix to match.
- When computing a mathematical operation on a row vector (m,1) with a real number, Python will copy the number m times in order to match the sizes of the two vectors.
- To simplify your code, don't use "rank 1" arrays (m,). Instead use either column vectors (m,1) or row vectors (1,m) to avoid any errors. You can use assert statements to ensure they are the correct dimensions and reshape() to change any dimensions needed.

Lets take an in depth look at the math behind the **Logistic Regression cost function**:

Remember that $\hat{y} = \sigma(w^T x + b)$ where $\sigma(z) = \frac{1}{1+e^{-z}}$. We interpret $\hat{y} = P(y = 1 \mid x)$ such that:

If y=1 : P(y|x) = $\hat{y}$

If y=0 : P(y|x) = 1-$\hat{y}$

We know that P(y|x) = $\hat{y}^y * (1 - \hat{y})^{(1-y)}$ and by taking the log of this, we get:

$$log(p(y|x)) \; = \; log(\hat{y}^y * (1 - \hat{y})^{(1-y)})$$

$$log(p(y|x)) \; = \; ylog(\hat{y}) + (1 - y)log(1 - \hat{y})$$

We know that the above equation for log(p(y|x)) can be denoted as -$\mathcal{L}(\hat{y}, y)$, which is our cost function. This is only for one example, but we need to find this for m examples. We will use **maximum likelihood estimation** to find the parameters that maximize this equation:

$$log(p(m\ examples)) \; = \; log(\prod_{i=1}^{n} p(y^i|x^i))$$

$$log(p(m\ examples)) \; = \; \sum_{i=1}^{m} log(p(y^i|x^i))$$

$$log(p(m\ examples)) \; = \; -\sum_{i=1}^{m} \mathcal{L}(\hat{y}^i, y^i)$$

This justifies out cost function, and because now we want to minimize the cost we drop the negative sign and to make sure our quantities are scaled, we add the $\frac{1}{m}$. Note that minimizing the loss below corresponds to maximizing log(p(y|x)).

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\hat{y}^i, y^i) = \frac{1}{m} \sum_{i=1}^{m} [(y^i \, log(\hat{y}^i) \; + \; (1 - y^i) \, log(1 - \hat{y}^i))]$$

1.2.2 Programming Assignment

Problem Statement: You are given a dataset ("data.h5") containing:

- A training set of m_train images labeled as cat (y=1) or non-cat (y=0).
- A test set of m_test images labeled as cat or non-cat.
- Each image is of shape (num_px, num_px, 3) where 3 is for the 3 channels (RGB). Thus, each image is square (height = num_px) and (width = num_px).

You will build a simple image-recognition algorithm that can correctly classify pictures as cat or non-cat. We added "_orig" at the end of image datasets (train and test) because we are going to preprocess them. After preprocessing, we will end up with train_set_x and test_set_x.

```
# Loading the data (cat/non-cat)
train_set_x_orig, train_set_y, test_set_x_orig, test_set_y, classes = load_dataset()
```

Lets find the dimensions of our data. Remember that train_set_x_orig is a numpy-array of shape (m_train, num_px, num_px, 3).

```
m_train = train_set_x_orig.shape[0] # 209 examples
m_test = test_set_x_orig.shape[0] # 50 examples
num_px = train_set_x_orig.shape[1] # 64
```

Note that each image is of size (64, 64, 3), the training set has shape (209, 64, 64, 3) with corresponding labels (1, 209), and the test set has shape (50, 64, 64, 3) with corresponding labels (1,50).

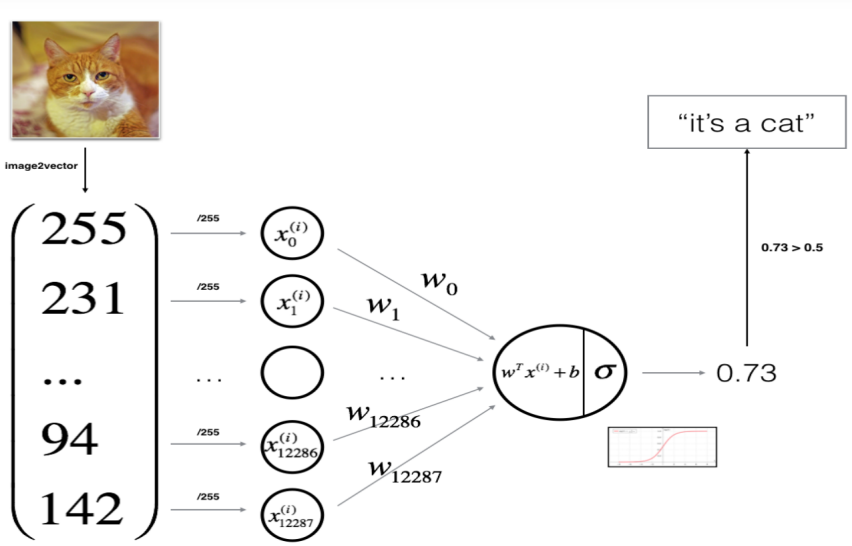
We now want to reshape the training and test data sets so that images of size (num_px, num_px, 3) are **flattened** into single vectors of shape (num_px*num_px*3, 1). To flatten a matrix X of shape (a,b,c,d) to a matrix X_flatten of shape (b*c*d, a) we use: X_flatten = X.reshape(X.shape[0], -1).T

```
train_set_x_flatten = train_set_x_orig.reshape(train_set_x_orig.shape[0], -1).T
test_set_x_flatten = test_set_x_orig.reshape(test_set_x_orig.shape[0], -1).T

train_set_x_flatten.shape # (12288, 209)
test_set_x_flatten.shape # (1288, 50)
```

Now we will **standardize** our dataset. One common practice is to substract the mean of the whole numpy array from each example, and then divide each example by the standard deviation. But for picture datasets, it is simpler and more convenient and works almost as well to just divide every row of the dataset by 255 (since the range for the vector values of an image are [0, 255]).

```
train_set_x = train_set_x_flatten/255.
test_set_x = test_set_x_flatten/255.
```



Above is a depication of how our Logistic Regression Neural Network will operate (refer to the previous section for an explanation of the mathematical expressions). We will carry out the following steps:

- Initialize the parameters of the model.
- Learn the parameters for the model by minimizing the cost.
- Use the learned parameters to make predictions (on the test set).
- Analyse the results and conclude.

Lets begin building the parts of our algorithm. There are 3 main steps for **building a Neural Network**:

1. Define the model structure (such as number of input features).
2. Initialize the model's parameters.
3. Loop:
   - Calculate current loss (forward propagation).
   - Calculate current gradient (backward propagation).
   - Update parameters (gradient descent).

```python
def sigmoid(z):
    # Compute sigmoid of z = w.T * x + b
    s = 1 / (1+np.exp(-z))
    return s

def initialize_with_zeros(dim):
    # Creates a vector of zeros of shape (dim, 1) for w and initializes b=0.
    w = np.zeros((dim,1))
    b = 0
    return w, b
```

Now that your parameters are initialized, you can do the **forward/backward propagation** steps for learning the parameters. The steps for forward propagation are:

- You get X
- You compute $A = \sigma(w^T X + b) = (a^1, a^2, ..., a^{m-1}, a^m)$
- You calculate the cost function: $J = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(a^i) + (1 - y^i)\log(1 - a^i)$

The steps for back propagation are:

- Compute $\frac{\partial J}{\partial w} = \frac{1}{m} X(A - Y)^T$
- Compute $\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}(a^i - y^i)$

```python
def propagate(w, b, X, Y):
    """
    Implement the cost function and its gradient for the propagation explained above
    Arguments:
    w -- weights, a numpy array of size (num_px * num_px * 3, 1)
    b -- bias, a scalar
    X -- data of size (num_px * num_px * 3, number of examples)
    Y -- true "label" vector (containing 0 if non-cat, 1 if cat) of size (1, m)

    Return:
    cost -- negative log-likelihood cost for logistic regression
    dw -- gradient of the loss with respect to w, thus same shape as w
    db -- gradient of the loss with respect to b, thus same shape as b
    """

    m = X.shape[1]

    # FORWARD PROPAGATION (FROM X TO COST)
    A = sigmoid(np.dot(w.T, X) + b) # compute activation
    cost = -(1/m)*np.sum(Y*np.log(A) + (1-Y)*np.log(1-A)) # compute cost

    # BACKWARD PROPAGATION (TO FIND GRAD)
    dw = (1/m) * np.dot(X, (A-Y).T)
    db = (1/m) * np.sum(A-Y)

    cost = np.squeeze(cost)
    grads = {"dw": dw,
             "db": db}

    return grads, cost
```

Now that we have intialized our parameters and can compute a cost function and its gradient, we can create an **optimize function** to update the parameters using gradient descent. The goal is to learn w and b by minimizing the cost function J . For a parameter θ , the update rule is $\theta = \theta - \alpha d\theta$, where α is the learning rate.

We can also create a **predict** function with our learned parameters to make predictions for a dataset X. We will calculate $\hat{Y}$ = A, and then convert entries to 0 or 1 based on probabilities.

```
def optimize(w, b, X, Y, num_iterations, learning_rate, print_cost = False):
  """
  Arguments:
  w -- weights, a numpy array of size (num_px * num_px * 3, 1)
  b -- bias, a scalar
  X -- data of shape (num_px * num_px * 3, number of examples)
  Y -- true "label" vector (containing 0 if non-cat, 1 if cat), of shape (1, m)
  num_iterations -- number of iterations of the optimization loop
  learning_rate -- learning rate of the gradient descent update rule
  print_cost -- True to print the loss every 100 steps

  Returns:
  params - dictionary containing the weights w and bias b
  grads - the gradients of the weights and bias with respect to the cost function
  costs - list of all the costs computed during the optimization (graphing)
  """
  costs = []

  for i in range(num_iterations):
    grads, cost = propagate(w, b, X, Y)

    dw = grads["dw"] # Retrieve derivatives from grads
    db = grads["db"] # Retrieve derivatives from grads

    w = w - learning_rate * dw # update rule
    b = b - learning_rate * db # update rule

    if i % 100 == 0: # Record the costs
      costs.append(cost)

    if print_cost and i % 100 == 0: # Print the cost every 100 training iterations
      print ("Cost after iteration %i: %f" %(i, cost))

  params = {"w": w, "b": b}
  grads = {"dw": dw, "db": db}
  return params, grads, costs

def predict(w, b, X):
'''
Predict whether the label is 0 or 1 using learned  parameters (w, b)

Arguments:
w -- weights, a numpy array of size (num_px * num_px * 3, 1)
b -- bias, a scalar
X -- data of size (num_px * num_px * 3, number of examples)

Returns a numpy array (vector) containing all predictions (0/1) for the examples in X
'''
m = X.shape[1]
Y_prediction = np.zeros((1,m))
w = w.reshape(X.shape[0], 1)

# Compute "A" predicting the probabilities of a cat being present in the picture
A = sigmoid(np.dot(w.T, X) + b)

for i in range(A.shape[1]): # Convert prob. A[0,i] to actual predictions p[0,i]
  Y_prediction[0,i] = A[0,i] > 0.5

assert(Y_prediction.shape == (1, m))

return Y_prediction
```

The final step is to **merge all functions into a model** by putting together the previous parts in the correct order.

```
def model(X_train, Y_train, X_test, Y_test, num_iterations = 2000,
          learning_rate = 0.5, print_cost = False):
    """
    Arguments:
    X_train -- training set represented by a np array (num_px * num_px * 3, m_train)
    Y_train -- training labels represented by a np array (1, m_train)
    X_test -- test set represented by a np array (num_px * num_px * 3, m_test)
    Y_test -- test labels represented by a np array (1, m_test)
    num_iterations -- hyperparameter used to optimize the parameters
    learning_rate -- hyperparameter used in the update rule of optimize()
    print_cost -- Set to true to print the cost every 100 iterations

    Returns:
    d -- dictionary containing information about the model.
    """
    w, b = initialize_with_zeros(X_train.shape[0]) # initialize parameters with zeros

    # Gradient descent
    parameters, grads, costs = optimize(w, b, X_train, Y_train, num_iterations,
                                        learning_rate, print_cost)

    # Retrieve parameters w and b from dictionary "parameters"
    w = parameters["w"]
    b = parameters["b"]

    Y_prediction_test = predict(w, b, X_test) # Predict test set examples
    Y_prediction_train = predict(w, b, X_train) # Predict test set examples

    # Print train/test Errors
    print("train accuracy: {} %".format(100 - np.mean(np.abs(Y_prediction_train -
                                                             Y_train)) * 100))
    print("test accuracy: {} %".format(100 - np.mean(np.abs(Y_prediction_test -
                                                            Y_test)) * 100))
    d = {"costs": costs,
         "Y_prediction_test": Y_prediction_test,
         "Y_prediction_train" : Y_prediction_train,
         "w" : w,
         "b" : b,
         "learning_rate" : learning_rate,
         "num_iterations": num_iterations}

    return d

d = model(train_set_x, train_set_y, test_set_x, test_set_y, num_iterations = 2000,
          learning_rate = 0.005, print_cost = True)
```

```
Cost after iteration 0: 0.693147
Cost after iteration 100: 0.584508
Cost after iteration 200: 0.466949
Cost after iteration 300: 0.376007
Cost after iteration 400: 0.331463
Cost after iteration 500: 0.303273
Cost after iteration 600: 0.279880
Cost after iteration 700: 0.260042
Cost after iteration 800: 0.242941
Cost after iteration 900: 0.228004
Cost after iteration 1000: 0.214820
Cost after iteration 1100: 0.203078
Cost after iteration 1200: 0.192544
Cost after iteration 1300: 0.183033
Cost after iteration 1400: 0.174399
Cost after iteration 1500: 0.166521
Cost after iteration 1600: 0.159305
Cost after iteration 1700: 0.152667
Cost after iteration 1800: 0.146542
Cost after iteration 1900: 0.140872
train accuracy: 99.04306220095694 %
test accuracy: 70.0 %
```

From our model, we can see that the training accuracy is 99% while the test accuracy is 70%, which is a cause of overfitting. We can also see that the cost is decreasing per onehundred iterations (meaning the parameters are being leaned). We can continue to idecrease the cost by increasing the number of iterations, but this will also cause more offerfitting to occur. Given the small dataset and the fact that Logistic Regression is a linear classifier, overall it is not a bad simple model. In the future, we will learn to avoid overfitting and increase the test accuracy.

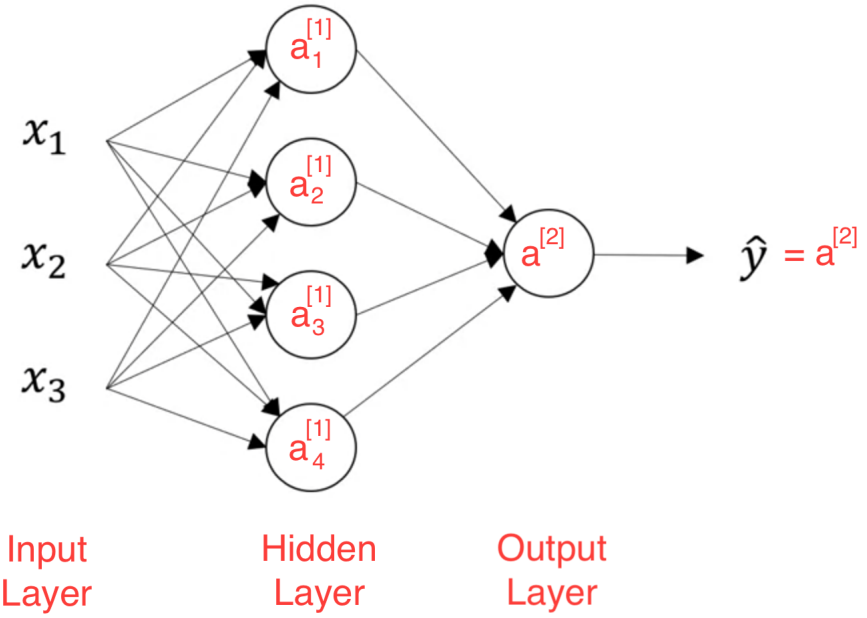
1.3 Shallow Neural Network

1.3.1 Overview and Representation

Previously, we only had a single sigmoid function in our Logistic Regression model. Now, we will stack multiple sigmoids ontop of one another, followed by then feeding these into another sigmoid to create a Nerual Network. Some new notation we are introducing:

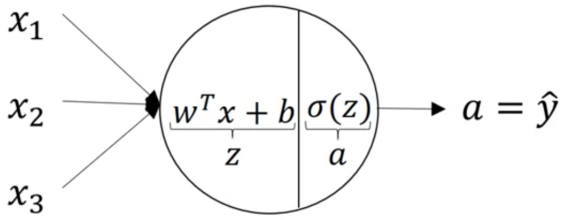
- [#] will refer to quanities associated with a given layer.
- (i) will refer to the ith training example (similar to the previous section).
- a$^{[0]}$ will be the activation layer (same as our vector x that has the input features).
- a$^{[1]}$ will be the values for our hidden layer.
- a$^{[2]}$ will be the output layer values (our $\hat{y}$).
- a$_i^{[l]}$ will be $[l]$ = layer, i = node in the layer.

We will be focusing on a **Two Layer Neural Network**, which has an input layer, one hidden layer, and an output layer. We don't count the input layer as an "official" layer. The *hidden layer* will have parameters $w^{[1]}$ and $b^{[1]}$, while the output layer has parameters $w^{[2]}$ and $b^{[2]}$ associated with it.



1.3.2 Computing a Neural Network Output

Each node in our hidden layer will take all of the input values from x, compute z, and input this value into an activation function (sigmoid). Since each node has to compute z, we will vectorize this process by using matrix multiplication in python. This gives us the following equation to find our vector $z^{[1]}$ and our activation vector $a^{[1]}$ for the hidden layer.



$$z^{[1]} = \begin{bmatrix} - w_1^{[1]T} - \\ - w_2^{[1]T} - \\ - w_3^{[1]T} - \\ - w_4^{[1]T} - \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \\ b_4^{[1]} \end{bmatrix} = \begin{bmatrix} w_1^{[1]T} x + b_1^{[1]} \\ w_2^{[1]T} x + b_2^{[1]} \\ w_3^{[1]T} x + b_3^{[1]} \\ w_4^{[1]T} x + b_4^{[1]} \end{bmatrix} = \begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \\ z_3^{[1]} \\ z_4^{[1]} \end{bmatrix}, \quad \text{also let } a^{[1]} = \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \\ a_3^{[1]} \\ a_4^{[1]} \end{bmatrix} = \sigma(z^{[1]})$$

Note: Let the matrix with the w values be denoted as $W^{[1]}$ and the vector holding b values be $b^{[1]}$.

For the **hidden layer**, this gives us the general formulas (remember x = $a^{[0]}$):

$z^{[1]} = W^{[1]}a^{[0]} + b^{[1]}$ with shapes: $z^{[1]} = (4,1),\ W^{[1]} = (4,3),\ a^{[0]} = (3,1),\ b^{[1]} = (4,1)$

$a^{[1]} = \sigma(z^{[1]})$ with shapes: $a^{[1]} = (4,1),\ z^{[1]} = (4,1)$

For the **output layer**, this gives us the following formulas (ouput $a^{[1]}$ used as input "x"):

$z^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$ with shapes: $z^{[2]} = (1,1),\ W^{[2]} = (1,4),\ a^{[1]} = (4,1),\ b^{[2]} = (1,1)$

$a^{[2]} = \sigma(z^{[2]})$ with shapes: $a^{[2]} = (1,1),\ z^{[2]} = (1,1)$

1.3.3 Vectorizing Across Multiple Examples

When we want to compute predictions for all of our training examples (not just a single example like we did above), we need to vectorize a method in order to compute all of these at once. Some new notation we will use:

- ex: $a^{[2](i)}$ refers to the layer 2 value for the i^{th} training example.

We will define the following matrices to work with n_x training examples where X = m training examples, $Z^{[1]}$ = is all of the $z^{[1]}$ values for m training example, and $A^{[1]}$ = all of our $a^{[1]}$ values for m training examples. Note that the format is the same for $Z^{[2]}$ and $A^{[2]}$ with their corresponding values.

$$X = \begin{bmatrix} | & | & \dots & | \\ x^1 & x^2 & \dots & x^m \\ | & | & \dots & | \end{bmatrix} \quad Z^{[1]} = \begin{bmatrix} | & | & \dots & | \\ z^{[1](1)} & z^{[1](2)} & \dots & z^{[1](m)} \\ | & | & \dots & | \end{bmatrix} \quad A^{[1]} = \begin{bmatrix} | & | & \dots & | \\ a^{[1](1)} & a^{[1](2)} & \dots & a^{[1](m)} \\ | & | & \dots & | \end{bmatrix}$$

This gives us the following **vectorized formulas** to find all predicted values for m examples:

$Z^{[1]} = W^{[1]}X + b^{[1]}$

$A^{[1]} = \sigma(Z^{[1]})$

$Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$

$A^{[2]} = \sigma(Z^{[2]})$

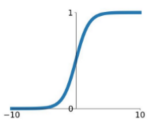
1.3.4 Activation Functions

Previously, we have used the sigmoid function as our activation function. However, there are much better functions that we can use instead, denoted by a general symbol g. For a Two Layer Neural Network, we can denote the activation function for the **hidden layer** as $g^{[1]}(z^{[1]})$ and the **output layer** as $g^{[2]}(z^{[2]})$.

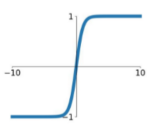
One exception is when you are doing binary classification to use a sigmoid function for the output layer. This is because a a sigmoid function will output a value between 0 and 1, which works perfectly since our true labels (y) will either be a 0 or 1.

Another option for an activation function is tanh(z). This is often much more useful than the sigmoid function, and will output a value in the range [-1,1]. A downfall to both the sigmoid and tanh function are that when z is very large or small, the gradient is near 0 and can cause gradient descent to slow. The most commonly used activation function is the ReLU function (or the Leaky ReLU function to avoid having a 0 gradient for negative numbers).

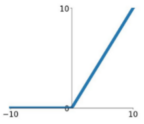
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$



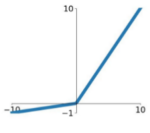
tanh
$\tanh(x)$



ReLU
$\max(0, x)$



Leaky ReLU
$\max(0.1x, x)$



1.3.5 Gradient Descent for Neural Networks

A Neural Network with one hidden layer will have the following:

- Parameters: $W^{[1]}$ with shape $(n^{[1]}, n^{[0]})$.
  $b^{[1]}$ with shape $(n^{[1]}, 1)$.
  $W^{[2]}$ with shape $(n^{[2]}, n^{[1]})$.
  $b^{[2]}$ with shape $(n^{[2]}, 1)$.
- $n^{[0]}$ input features (known as n_x), $n^{[1]}$ hidden layer units, and $n^{[2]}$ output units.
- Cost Function: J($W^{[1]}$, $b^{[1]}$, $W^{[2]}$, $b^{[2]}$) = $\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\hat{y}, y)$
- Gradient descent (repeat the following steps):
  1) Compute predicts $(\hat{y}^{(i)}, ..., \hat{y}^{(m)})$
  2) Compute $dW^{[1]} = \frac{dJ}{dw^{[1]}}$, $db^{[1]} = \frac{dJ}{db^{[1]}}$, ... and similary for $dW^{[2]}$ and $db^{[2]}$
  3) Compute $W^{[1]} = W^{[1]} - \alpha\, dw^{[1]}$
  4) Compute $b^{[1]} = b^{[1]} - \alpha\, db^{[1]}$
  5) Compute $W^{[2]} = W^{[2]} - \alpha\, dw^{[2]}$
  6) Compute $b^{[2]} = b^{[2]} - \alpha\, db^{[2]}$

Recall the formulas for **forward propagation** (where g is the generalized activation function):

$Z^{[1]} = W^{[1]}X + b^{[1]}$

$A^{[1]} = g^{[1]}(Z^{[1]})$

$Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$

$A^{[2]} = g^{[2]}(Z^{[2]})$

This means that we can perform **back propagation** (gradient descent) with the following steps:

$dz^{[2]} = A^{[2]} - Y$

$dW^{[2]} = \frac{1}{m}\, dz^{[2]}\, A^{[1]T}$

$db^{[2]} = \frac{1}{m}\, np.sum(dz^{[2]}, axis = 1, keepdims = True)$

$dz^{[1]} = W^{[2]T}\, dz^{[2]} * g^{[1]'}(Z^{[1]})$, which is an element-wise product of two $(n^{[1]}, m)$ matrices.

$dW^{[1]} = \frac{1}{m}\, dz^{[1]}\, X^T$

$db^{[1]} = \frac{1}{m}\, np.sum(dz^{[1]}, axis = 1, keepdims = True)$, which is an $(n^{[1]}, 1$ vector).

Random Initialization:

For a Neural Network, we want to initialize the weights to random values instead of zeros (initializing to zero will cause all of the calculations to be symmetric). To randomly initialize our weights, we can:

- Set $W^{[1]}$ = np.random.randn((2,2)) * 0.01 to create small Gaussian random variables.
- Initialize $b^{[1]}$ = np.zero((2,1)) because b does not have the symmetry problem that w can have.
- Similarly, we can do the same for $W^{[2]}$ and $b^{[2]}$.

1.3.6 Programming Assignment