## Week 1 – Logistic Regression

1. When performing logistic regression on sentiment analysis, you represented each tweet as a vector of ones and zeros. However your model did not work well. Your training cost was reasonable, but your testing cost was just not acceptable. What could be a possible reason?
   1. The vector representations are sparse and therefore it is much harder for your model to learn anything that could generalise well to the test set.
2. Which of the following are examples of text preprocessing?
   1. Stemming.
   2. Lowercasing.
   3. Removing stopwords, punctuation, handles and URLs.
3. The sigmoid function is defined as . Which of the following is true?
   1. Large positive values of will make close to 1, and large negative values of will make close to -1.
4. The cost function for logistic regression is defined as

. Which is true about the cost function?

* 1. When , as approaches 0, the cost function approaches .
  2. When , as approaches 0, the cost function also approaches 0.

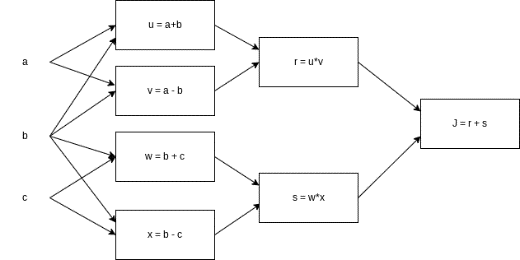
1. For what value of in the sigmoid function does .
2. When performing logistic regression for sentiment analysis, you have to:
   1. Perform data preprocessing.
   2. Create a dictionary that maps the word and the class that word is found in to the number of times that word is found in the class.
   3. For each tweet, create a positive feature with the sum of positive counts of each word in that tweet, and a negative feature with the sum of negative counts of each word in that tweet.
3. When training a logistic regression mode, what order are the operations performed in?
   1. Initialise parameters
   2. Classify/predict
   3. Get gradient
   4. Update
   5. Get loss
   6. Repeat.
4. Assuming we got the classification correct, where for some specific example . This means that . Which of the following has to hold?
   1. Our prediction, , for this specific training example is greater than
5. What is the purpose of gradient descent?
   1. It allows us to learn the parameters in logistic regression as to minimise the loss function .
   2. *grad\_theta* allows us to update the parameters by computing

.

1. What is a good metric that allows you to decide when to stop training/trying to get a good model?
   1. When your accuracy is good enough on the test set.
   2. When you plot the cost versus the number of iterations and see that your loss is converging (i.e. no longer changing a lot).

## Week 2 – Naïve Bayes

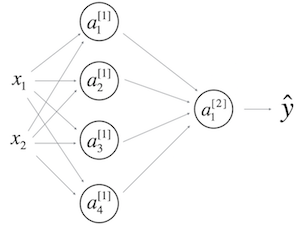
1. In logistic regression, given the input and parameters , how do we generate the output ?
   1. (In logistic regression we use a linear function followed by the sigmoid function , to get an output , referred to as , such that .
2. Suppose that and . What is the value of the “logistic loss”?
   1. 0.693. ()
3. Consider the numpy array: . What is the shape of ?
   1. (2, 2, 1) (This array has two rows and in each row it has 2 arrays of 1x1.
4. Consider the following random arrays , and :  
   #   
   #   
   What will be the shape of ?
   1. The computation cannot happen because it is not possible to broadcast more than one dimension. (It is not possible to broadcast together and . In this case there is no way to generate copies of one of the arrays to match the size of the other.)
5. Consider the following random arrays and :  
   #   
   #   
   What will be the shape of ?
   1. . (Broadcasting is invoked, so row is multiplied element-wise with each row of to create .)
6. Suppose you have input features per example. If we decide to use row vectors for the features and . What is the dimension of ?
   1. (Each has dimension , is built stacking all rows together into a array.)
7. Consider the following array: . What is the result of ?
   1. . (Recall that \* indicates element-wise multiplication and is matrix multiplication. Thus, .
8. Consider the following code snippet:  
   a.shape = (3,4)  
   b.shape = (4,1)  
     
   for i in range(3):  
    for j in range(4):  
    c[i][j] = a[i][j] \* b[j]
9. Consider the following random arrays and :  
   Which of the following arrays in stored in ?
   1. . (The array is a column vector. This is copied two times and added to the array to construct the array .)
10. Consider the following computational graph:

What is the output of J?

* 1. . (.)

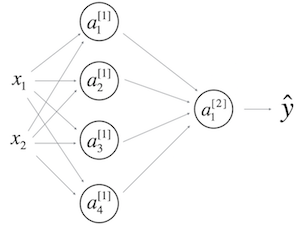
## Week 3 – Vector Space Models

1. Which of the following are true?
   1. is the activation output by the 4th neuron of the 2nd layer.
   2. is a matrix in which each column is one training example.
   3. denotes the activation vector of the 2nd layer.
   4. denotes the activation vector of the 2nd layer for the 12th training example.
2. In which of the following cases is the linear (identity) activation function most likely used?
   1. When working with regression problems. (In problems such as predicting the price of a house, it makes sense to use the linear activation function as output.)
3. Which of these is a correct vectorized implementation of forward propagation for layer , where ?
4. The use of the ReLU activation function is becoming more rare because the ReLU function has no derivative for . True/False?
   1. False. (Although the ReLU function has no derivative at this rarely causes any problems in practice. Moreover it has become the default activation function in many cases.)
5. Consider the following code:  
   A = np.random.randn(4,3)  
   B = np.sum(A, axis=1, keepdims=True)  
   What will be ?
   1. (4,1). (We use to make sure that is and not . It makes our code more robust.)
6. Suppose you have built a neural network. You decide to initialize the weights and biases to be zero. Which of the following statements is true?
   1. Each neuron in the first hidden layer will perform the same computation. So even after multiple iterations of gradient descent, each neuron in the layer will be computing the same thing as other neurons.
7. Using linear activation functions in the hidden layers of a multilayer neural network is equivalent to using a single layer. True/False?
   1. True. (When the identify or linear activation function is used, the output of composition of layers is equivalent to the computations made by a single layer.)
8. You have built a network using the tanh activation for all the hidden units. You initialize the weights to relatively large values, using . What will happen?
   1. This will cause the inputs of to also be very large, thus causing gradients to be close to zero. The optimisation algorithm will thus become slow. ( becomes flat for large values; this leads its gradient to be close to zero, which slows down this optimisation algorithm.)
9. Consider the following 1 hidden layer neural network:

Which of the following statements are true?

* 1. will have shape (1,1)
  2. will have shape (1,4)
  3. will have shape (4,1)
  4. will have shape (4,2)

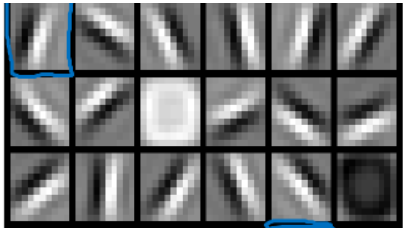
1. What are the dimensions of and ?



* 1. and are

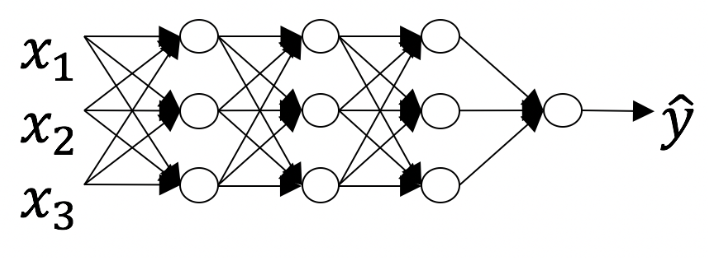
## Week 4 – Hashing and Machine Translation

1. What is stored in the 'cache' during forward propagation for later use in backward propagation?
   1. . (This value is useful in the calculation of in the backward propagation.)
2. Which of the following are “parameters” of a neural network?
   1. the bias vector
   2. the weight matrices
3. Which of the following is more likely related to the early layers of a deep neural network?



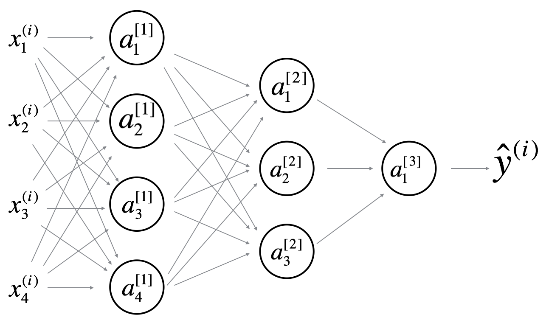
* 1. The early layer of a neural network usually computes simple features such as edges and lines.

1. We cannot use vectorization to calculate in backpropagation, we must use a for loop over all the examples. True/False?
   1. False. (We can use vectorisation in backpropagation to calculate for each layer. This computation is done over all the training examples.
2. Suppose is the array with the weights of the ith layer, is the vector of biases of the ith layer, and is the activation function used in all layers. Which of the following calculates the forward propagation for the neural network with layers?
   1. for i in range(1, L+1):  
       Z[i] = W[i] \* A[i-1] + b[i]  
       A[i] = g(Z[i])  
      (Remember that the range omits the last number; thus the range from 1 to gives the necessary values.)
3. Consider the following neural network:

  
How many layers does it have?

* 1. The number of layers is 4. The number of hidden layers is 3. (The number of layers is counted as the number of hidden layers plus 1). The input and output layers are not counted as hidden layers.)

1. If is the number of layers of a neural network then . True/False?
   1. True. The gradient of the output layer depends on the difference between the value computed during the forward propagation process and the target values.
2. For any mathematical function you can compute with an L-layered deep neural network with N hidden units there is a shallow neural network that requires only units, but it is very difficult to train.
   1. False. (On the contrary, some mathematical functions can be computed using an L-layered neural network and a given number of hidden units; but using a shallow neural network the number of necessary hidden units grows exponentially.)
3. Consider the following 2 hidden layer neural network:

  
Which of the following statements are true?

* 1. will have shape (4,4)
  2. will have shape (4,1)
  3. will have shape (1,3)
  4. will have shape (1,1)
  5. will have shape (3,4)
  6. will have shape (3,1)

(More generally, the shape of is and the shape of is .)

1. Whereas the previous question used a specific network, in the general case what is the dimension of , the bias vector associated with layer ?
   1. has shape . ( is a column vector wit the same number of rows as units in the respective layer.)