第一周

1. **What does the analogy “AI is the new electricity” refer to?**

   Similar to electricity starting about 100 years ago, AI is transforming multiple industries.

2. **Which of these are reasons for Deep Learning recently taking off? (Check the two options that apply.)**

   We have access to a lot more computational power.

   We have access to a lot more data.

   //任何技术的快速发展最终都基于基础技术的发展。技术的应用很难提升技术本身的发展。

3. **Recall this diagram of iterating over different ML ideas. Which of the statements below are true? (Check all that apply.)**

   Being able to try out ideas quickly allows deep learning engineers to iterate more quickly.

   Faster computation can help speed up how long a team takes to iterate to a good idea.

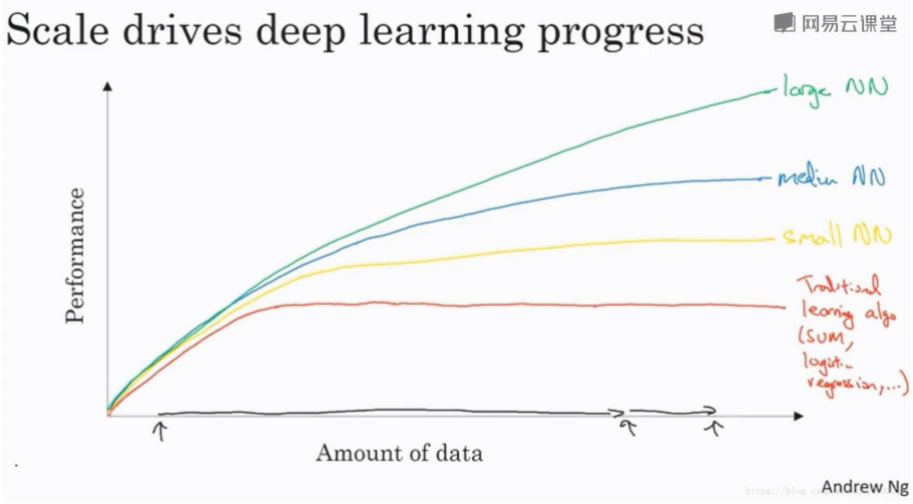
   Recent progress in deep learning algorithms has allowed us to train good models faster (even without changing the CPU/GPU hardware).

   //硬件和算法使计算速度提升，使机器学习的测试更快，发展也更快

4. **When an experienced deep learning engineer works on a new problem, they can usually use insight from previous problems to train a good model on the first try, without needing to iterate multiple times through different models. True/False?**

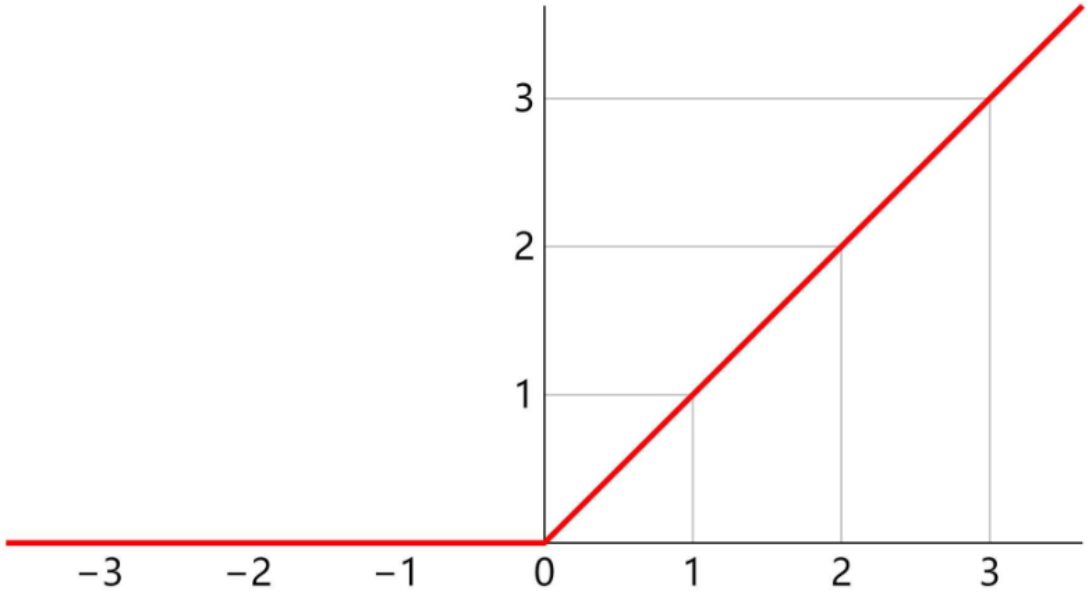
   False.

   //

   由此图：不同的模型能处理的数据集大小是不同的，能解决的问题是不同的。反过来看，解决不同的问题不一定用相同的模型，不一定能用的到以往的经验。

5. **Which one of these plots represents a ReLU activation function?**



特点是第一象限线性上升。

6. **Images for cat recognition is an example of "structured" data, because it is represented as a structured array in a computer. True/False?**

   False.因为它不是以数据库中的表或数据列的形式存在。

7. **A demographic dataset with statistics on different cities' population, GDP per capita, economic growth is an example of "unstructured" data because it contains data coming from different sources. True/False?**

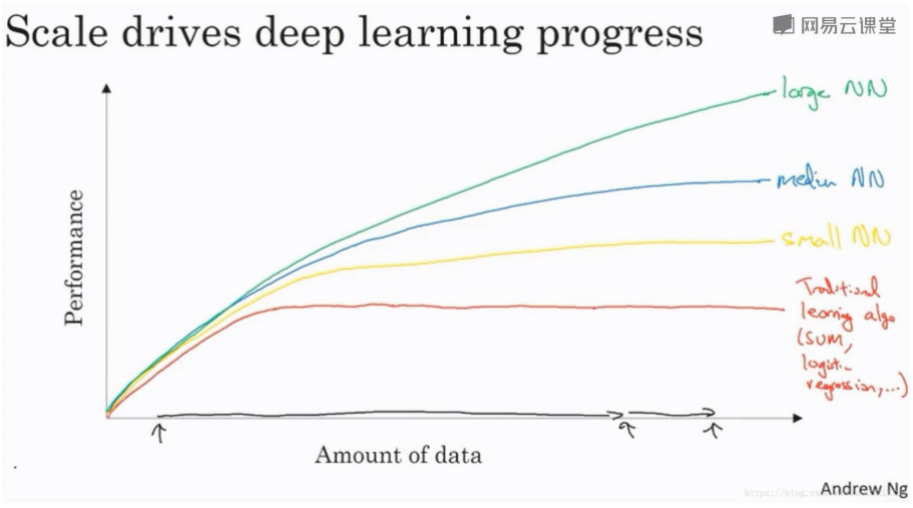
   False.理由应该是"它以数据库中的表或数据列的形式存在"。

8. **Why is an RNN (Recurrent Neural Network) used for machine translation, say translating English to French? (Check all that apply.)**

   It can be trained as a supervised learning problem.

   It is applicable when the input/output is a sequence (e.g., a sequence of words).

9. **In this diagram which we hand-drew in lecture, what do the horizontal axis (x-axis) and vertical axis (y-axis) represent?**



   x轴是数据量，y轴是算法的表现（模型的好用程度）

10. **Assuming the trends described in the previous question's figure are accurate (and hoping you got the axis labels right), which of the following are true? (Check all that apply.)**

Increasing the training set size generally does not hurt an algorithm's performance, and it may help significantly.

Increasing the size of a neural network generally does not hurt an algorithm's performance, and it may help significantly.

//数据集大小和神经网络大小决定了算法最终的表现

第二周

1. What does a neuron compute? C

   > 正确答案是B，神经元节点先计算线性函数（z = Wx + b），再计算激活函数

   - A neuron computes an activation function followed by a linear function (z = Wx + b)
   - A neuron computes a linear function (z = Wx + b) followed by an activation function
   - A neuron computes a function g that scales the input x linearly (Wx + b)
   - A neuron computes the mean of all features before applying the output to an activation function

   Note: The output of a neuron is a = g(Wx + b) where g is the activation function (sigmoid, tanh, ReLU, ...).

2. Which of these is the “Logistic Loss”?

   > 这题没找到图，但我猜测应该是给了J函数、L函数、z函数让我们区分。应该选L函数。

   - Check here.

   Note: We are using a cross-entropy loss function.

3. Suppose img is a (32,32,3) array, representing a 32x32 image with 3 color channels red, green and blue. How do you reshape this into a column vector?

   答：设img上的像素坐标为img[i][j][k],其中0≤i,j≤32,0≤k≤3，则可以构造vector[32 * 32 * 3]，其中vector[k * 32 * 32 + 32 * i + j] = img[i][j][k]

   > 正确答案为：x = img.reshape(32 * 32 * 3, 1)

4. Consider the two following random arrays “a” and “b”:

   ```
   a = np.random.randn(2, 3) # a.shape = (2, 3)
   b = np.random.randn(2, 1) # b.shape = (2, 1)
   c = a + b
   ```

   What will be the shape of “c”?

   答：c.shape = (2,3)

5. Consider the two following random arrays “a” and “b”:

   ```
   a = np.random.randn(4, 3) # a.shape = (4, 3)
   b = np.random.randn(3, 2) # b.shape = (3, 2)
   c = a * b
   ```

   What will be the shape of “c”?

答：(4,3)

> 正确答案：运算符 “*” 说明了按元素乘法来相乘，但是元素乘法需要两个矩阵之间的维数相同，所以这将报错，无法计算。

6. Suppose you have n_x input features per example. Recall that X=[x^(1), x^(2)... x^(m)]. What is the dimension of X?

   答：2

   Note: A stupid way to validate this is use the formula $Z^{(l)} = W^{(l)}A^{(l)}$ when l = 1, then we have

   - A^(1) = X
   - X.shape = (n_x, m)
   - Z^(1).shape = (n^(1), m)
   - W^(1).shape = (n^(1), n_x)

7. Recall that `np.dot(a,b)` performs a matrix multiplication on a and b, whereas `a*b` performs an element-wise multiplication.

   Consider the two following random arrays “a” and “b”:

   ```
   a = np.random.randn(12288, 150) # a.shape = (12288, 150)
   b = np.random.randn(150, 45) # b.shape = (150, 45)
   c = np.dot(a, b)
   ```

   What is the shape of c?

   答：(12288,45)

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]

```
How do you vectorize this?

答：
```

b.reshape(1,4)
c = a + b

```
> 另一种方法：c = a + b.T

9. Consider the following code:
```

a = np.random.randn(3, 3)
b = np.random.randn(3, 1)
c = a * b

```
   What will be c?

   答：(3,3)

10. Consider the following computation graph.

    //找不到图，但是看答案也能大概知道题是啥意思
```

```
J = u + v - w
  = a * b + a * c - (b + c)
  = a * (b + c) - (b + c)
  = (a - 1) * (b + c)
```

第三周

1. Which of the following are true? (Check all that apply.) **ABCD**
   - X is a matrix in which each column is one training example.
   - a^[2]_4 is the activation output by the 4th neuron of the 2nd layer（分析：A矩阵的第四行）
   - a^[2](12) denotes the activation vector of the 2nd layer for the 12th training example.（分析：A矩阵的第十二列）
   - a^[2] denotes the activation vector of the 2nd layer.

   Note: If you are not familiar with the notation used in this course, check here.

2. The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False? **True**
   - True
   - False

   Note: You can check this post and (this paper)[http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf].

   > As seen in lecture the output of the tanh is between -1 and 1, it thus centers the data which makes the learning simpler for the next layer.

3. Which of these is a correct vectorized implementation of forward propagation for layer l, where 1≤l≤L?

   //没找到备选答案

4. You are building a binary classifier for recognizing cucumbers (y=1) vs. watermelons (y=0). Which one of these activation functions would you recommend using for the output layer? **Answer:sigmoid**(because sigmoid's output is between 0 and 1)
   - ReLU

- Leaky ReLU
- sigmoid
- tanh

Note: The output value from a sigmoid function can be easily understood as a probability.

> Sigmoid outputs a value between 0 and 1 which makes it a very good choice for binary classification. You can classify as 0 if the output is less than 0.5 and classify as 1 if the output is more than 0.5. It can be done with tanh as well but it is less convenient as the output is between -1 and 1.

5. Consider the following code:

```
A = np.random.randn(4,3)
B = np.sum(A, axis = 1, keepdims = True)
```

What will be B.shape?

Answer:(4,1)

> we use (keepdims = True) to make sure that A.shape is (4,1) and not (4,). It makes our code more rigorous.

6. Suppose you have built a neural network. You decide to initialize the weights and biases（乖离率） to be zero. Which of the following statements are True? **A**
   - 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.
   - Each neuron in the first hidden layer will perform the same computation in the first iteration. But after one iteration of gradient descent they will learn to compute different things because we have “broken symmetry”.(应该是完全对称)
   - Each neuron in the first hidden layer will compute the same thing, but neurons in different layers will compute different things, thus we have accomplished “symmetry breaking” as described in lecture.（同上）
   - The first hidden layer’s neurons will perform different computations from each other even in the first iteration; their parameters will thus keep evolving in their own way.
7. Logistic regression’s weights w should be initialized randomly rather than to all zeros, because if you initialize to all zeros, then logistic regression will fail to learn a useful decision boundary because it will fail to “break symmetry”, True/False? **TRUE**

> 正确答案：FALSE
>
> 错误原因：英语不好，【it will fail to “break symmetry”】这句里的break应该翻译成动词，我翻译成了名词，当时前半句后半句自相矛盾还给我自己整蒙了

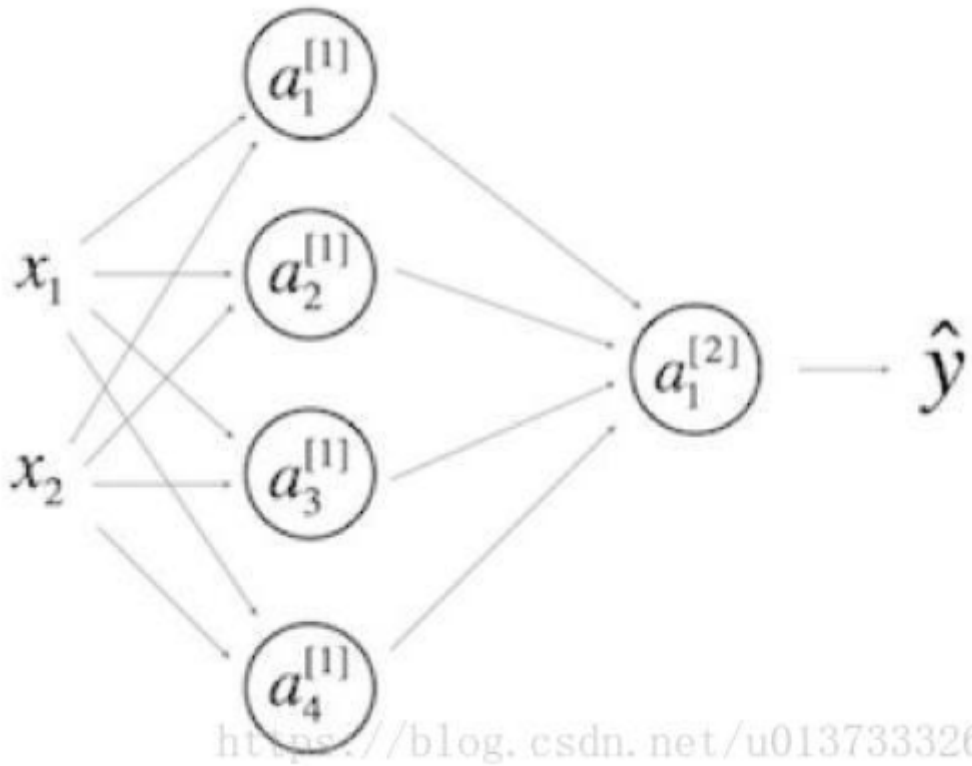
   - True
   - False

> Logistic Regression doesn’t have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there’s no hidden layer) which is not zero. So at the second iteration, the weights values follow x’s distribution and are different from each other if x is not a constant vector.

8. You have built a network using the tanh activation for all the hidden units. You initialize the weights to relative large values, using np.random.randn(..,..)*1000. What will happen? **D**
    - It doesn't matter. So long as you initialize the weights randomly gradient descent is not affected by whether the weights are large or small.
    - This will cause the inputs of the tanh to also be very large, thus causing gradients to also become large. You therefore have to set 伪 to be very small to prevent divergence; this will slow down learning.
    - This will cause the inputs of the tanh to also be very large, causing the units to be "highly activated" and thus speed up learning compared to if the weights had to start from small values.
    - This will cause the inputs of the tanh to also be very large, thus causing gradients to be close to zero. The optimization algorithm will thus become slow.

    > tanh becomes flat for large values, this leads its gradient to be close to zero. This slows down the optimization algorithm.

9. Consider the following 1 hidden layer neural network:



    判断:

    - b[1] will have shape (4, 1)对
    - W[1] will have shape (4, 2)对
    - W[2] will have shape (1, 4)对
    - b[2] will have shape (1, 1)对

    > Analyse:
    >
    > W[i] will have shape(n[i], n[i-1])
    >
    > b[i] will have shape(n[i], 1)

    Note: Check here for general formulas to do this.

10. In the same network as the previous question, what are the dimensions of Z^[1] and A^[1]?

判断：**True**

- Z[1] and A[1] are (4,m)

> Analyse:
>
> 4 means 4 neurons at this layer
>
> m means m input samples's results

Note: Check here for general formulas to do this.

第四周

1. What is the “cache” used for in our implementation of forward propagation and backward propagation? **B**
   - It is used to cache the intermediate values of the cost function during training.
   - We use it to pass variables computed during forward propagation to the corresponding backward propagation step. It contains useful values for backward propagation to compute derivatives.
   - It is used to keep track of the hyperparameters that we are searching over, to speed up computation.
   - We use it to pass variables computed during backward propagation to the corresponding forward propagation step. It contains useful values for forward propagation to compute activations.

   > the “cache” records values from the forward propagation units and sends it to the backward propagation units because it is needed to compute the chain rule derivatives.

2. Among the following, which ones are “hyperparameters”? (Check all that apply.) **ABCD**
   - size of the hidden layers n[l]
   - learning rate α
   - number of iterations
   - number of layers L in the neural network

   Note: You can check this Quora post or this blog post.

3. Which of the following statements is true? **A**
   - The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers.
     Correct
   - The earlier layers of a neural network are typically computing more complex features of the input than the deeper layers.

   Note: You can check the lecture videos. I think Andrew used a CNN example to explain this.

4. Vectorization allows you to compute forward propagation in an L-layer neural network without an explicit for-loop (or any other explicit iterative loop) over the layers l=1, 2, …,L. True/False? **False**
   - True
   - False

   Note: We cannot avoid the for-loop iteration over the computations among layers.

5. Assume we store the values for n^[l] in an array called layers, as follows: layer_dims = [n_x, 4,3,2,1]. So layer 1 has four hidden units, layer 2 has 3 hidden units and so on. Which of the following for-loops will allow you to initialize the parameters for the model?

   正解如下：

```
for(i in range(1, len(layer_dims))):
    parameter['W' + str(i)] = np.random.randn(layers[i], layers[i - 1])) *
0.01
    parameter['b' + str(i)] = np.random.randn(layers[i], 1) * 0.01123
```

6. Consider the following neural network.

   > 没找到图，但是从答案来看题应该不难

   - The number of layers L is 4. The number of hidden layers is 3.

   Note: The input layer (L^[0]) does not count.

   > As seen in lecture, the number of layers is counted as the number of hidden layers + 1. The input and output layers are not counted as hidden layers.

7. During forward propagation, in the forward function for a layer l you need to know what is the activation function in a layer (Sigmoid, tanh, ReLU, etc.). During backpropagation, the corresponding backward function also needs to know what is the activation function for layer l, since the gradient depends on it. True/False? **True**

   - True
   - False

   > During backpropagation you need to know which activation was used in the forward propagation to be able to compute the correct derivative.

8. There are certain functions with the following properties:

   (i) To compute the function using a shallow network circuit, you will need a large network (where we measure size by the number of logic gates in the network), but (ii) To compute it using a deep network circuit, you need only an exponentially smaller network. True/False? **True**

   > 这里是在考神经网络模拟中的电路观点

   - True
   - False

   Note: See lectures, exactly same idea was explained.

9. Consider the following 2 hidden layer neural network:

   Which of the following statements are True? (Check all that apply).

   没找到图，但是从答案出发图大概能猜出来

   - W^[1] will have shape (4, 4)
   - b^[1] will have shape (4, 1)
   - W^[2] will have shape (3, 4)
   - b^[2] will have shape (3, 1)
   - b^[3] will have shape (1, 1)
   - W^[3] will have shape (1, 3)

Note: See this image for general formulas.//这个打开看了一下是课件截图，已理解。

10. Whereas the previous question used a specific network, in the general case what is the dimension of W^[l], the weight matrix associated with layer l?

    - W^[l] has shape ($n^{[l]}$, n[l−1])

    > 没错这道题的答案就是这样

    Note: See this image for general formulas.