Week5 Neural Networks_Learning

Cost Function

Suppose we want to try to minimize $J(\Theta)$ as a function of Θ, using one of the advanced optimization methods (fminunc, conjugate gradient, BFGS, L-BFGS, etc.). What do we need to supply code to compute (as a function of Θ)?

- Θ
- $J(\Theta)$
- The (partial) derivative terms $\frac{\partial}{\partial \Theta_{ij}^{(l)}}$ for every i, j, l
- $J(\Theta)$ and the (partial) derivative terms $\frac{\partial}{\partial \Theta_{ij}^{(l)}}$ for every i, j, l

正确

Backpropagation Algorithm

Suppose you have two training examples $(x^{(1)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$. Which of the following is a correct sequence of operations for computing the gradient? (Below, FP = forward propagation, BP = back propagation).

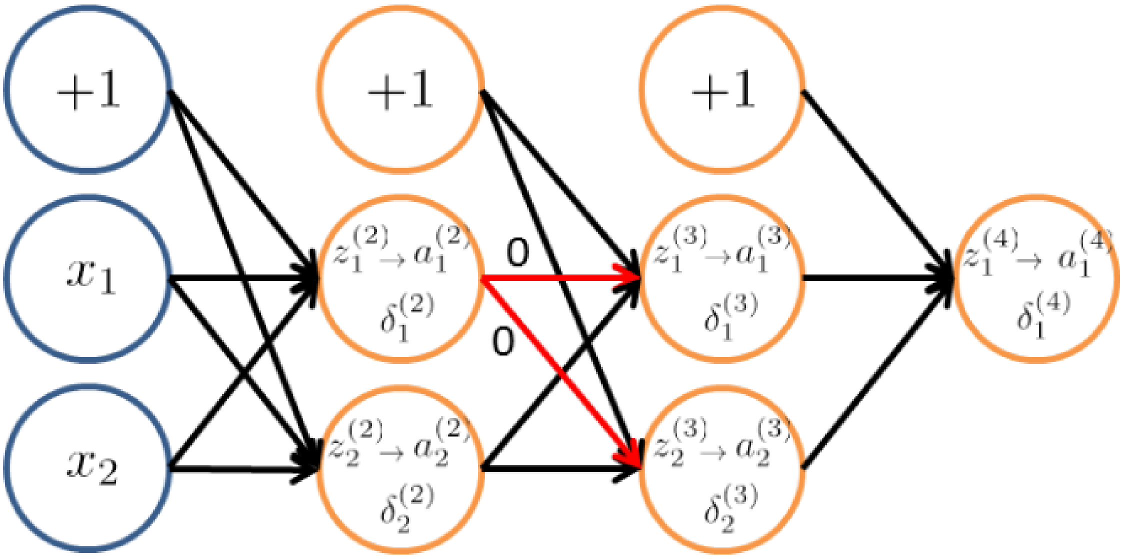
- FP using $x^{(1)}$ followed by FP using $x^{(2)}$. Then BP using $y^{(1)}$ followed by BP using $y^{(2)}$.
- FP using $x^{(1)}$ followed by BP using $y^{(2)}$. Then FP using $x^{(2)}$ followed by BP using $y^{(1)}$.
- BP using $y^{(1)}$ followed by FP using $x^{(1)}$. Then BP using $y^{(2)}$ followed by FP using $x^{(2)}$.
- FP using $x^{(1)}$ followed by BP using $y^{(1)}$. Then FP using $x^{(2)}$ followed by BP using $y^{(2)}$.

正确

前向传播和反向传播要处理同一个样本，且先进行前向传播后进行反向传播

Backpropagation Intuition

Consider the following neural network:



Suppose both of the weights shown in red ($\Theta_{11}^{(2)}$ and $\Theta_{21}^{(2)}$) are equal to 0. After running backpropagation, what can we say about the value of $\delta_1^{(3)}$?

- $\delta_1^{(3)} > 0$
- $\delta_1^{(3)} = 0$ only if $\delta_1^{(2)} = \delta_2^{(2)} = 0$, but not necessarily otherwise
- $\delta_1^{(3)} \leq 0$ regardless of the values of $\delta_1^{(2)}$ and $\delta_2^{(2)}$
- There is insufficient information to tell

正确

反向传播，第三层的误差要用第四层来计算，而非第二层。

Implementation Note: Unrolling Parameters

Suppose D1 is a 10x6 matrix and D2 is a 1x11 matrix. You set:

DVec = [D1(:); D2(:)];

Which of the following would get D2 back from DVec?

- ○ reshape(DVec(60:71), 1, 11)
- ○ reshape(DVec(61:72), 1, 11)
- ◉ reshape(DVec(61:71), 1, 11)

正确

- ○ reshape(DVec(60:70), 11, 1)

Gradient Checking

Let $J(\theta) = \theta^3$. Furthermore, let $\theta = 1$ and $\epsilon = 0.01$. You use the formula:

$$\frac{J(\theta+\epsilon)-J(\theta-\epsilon)}{2\epsilon}$$

to approximate the derivative. What value do you get using this approximation? (When $\theta = 1$, the true, exact derivative is $\frac{d}{d\theta}J(\theta) = 3$).

- ○ 3.0000
- ◉ 3.0001

正确

- ○ 3.0301
- ○ 6.0002

What is the main reason that we use the backpropagation algorithm rather than the numerical gradient computation method during learning?

- ○ The numerical gradient computation method is much harder to implement.
- ◉ The numerical gradient algorithm is very slow.

正确

- ○ Backpropagation does not require setting the parameter EPSILON.
- ○ None of the above.

Random Initialization

Consider this procedure for initializing the parameters of a neural network:

1. Pick a random number r = rand(1,1) * (2 * INIT_EPSILON) - INIT_EPSILON;
2. Set $\Theta_{ij}^{(l)} = r$ for all i, j, l.

Does this work?

- ○ Yes, because the parameters are chosen randomly.
- ○ Yes, unless we are unlucky and get r=0 (up to numerical precision).
- ○ Maybe, depending on the training set inputs x(i).
- ◉ No, because this fails to break symmetry.

正确

随机初始化意在让参数各不相同，若参数均为同一个值，无论这个值是否是随机得来都会使第二层激活单元有相同的值，不可行

Putting It Together

Suppose you are using gradient descent together with backpropagation to try to minimize $J(\Theta)$ as a function of Θ. Which of the following would be a useful step for verifying that the learning algorithm is running correctly?

- ○ Plot $J(\Theta)$ as a function of Θ, to make sure gradient descent is going downhill.
- ○ Plot $J(\Theta)$ as a function of the number of iterations and make sure it is increasing (or at least non-decreasing) with every iteration.
- ◉ Plot $J(\Theta)$ as a function of the number of iterations and make sure it is decreasing (or at least non-increasing) with every iteration.

正确

- ○ Plot $J(\Theta)$ as a function of the number of iterations to make sure the parameter values are improving in classification accuracy.

Neural Networks: Learning

1/1
分数

1. You are training a three layer neural network and would like to use backpropagation to compute the gradient of the cost function. In the backpropagation algorithm, one of the steps is to update

$$\Delta_{ij}^{(2)} := \Delta_{ij}^{(2)} + \delta_i^{(3)} * (a^{(2)})_j$$

for every i, j. Which of the following is a correct vectorization of this step?

- $\Delta^{(2)} := \Delta^{(2)} + \delta^{(3)} * (a^{(3)})^T$
- $\Delta^{(2)} := \Delta^{(2)} + \delta^{(3)} * (a^{(2)})^T$

正确
This version is correct, as it takes the "outer product" of the two vectors $\delta^{(3)}$ and $a^{(2)}$ which is a matrix such that the (i, j)-th entry is $\delta_i^{(3)} * (a^{(2)})_j$ as desired.

- $\Delta^{(2)} := \Delta^{(2)} + \delta^{(2)} * (a^{(2)})^T$
- $\Delta^{(2)} := \Delta^{(2)} + (a^{(3)})^T * \delta^{(3)}$

1/1
分数

2. Suppose $\texttt{Theta1}$ is a 5x3 matrix, and $\texttt{Theta2}$ is a 4x6 matrix. You set $\texttt{thetaVec} = [\texttt{Theta1}(:); \texttt{Theta2}(:)]$. Which of the following correctly recovers $\texttt{Theta2}$?

- $\texttt{reshape}(\texttt{thetaVec}(16:39), 4, 6)$

正确
This choice is correct, since $\texttt{Theta1}$ has 15 elements, so $\texttt{Theta2}$ begins at index 16 and ends at index 16 + 24 - 1 = 39.

- $\texttt{reshape}(\texttt{thetaVec}(15:38), 4, 6)$
- $\texttt{reshape}(\texttt{thetaVec}(16:24), 4, 6)$
- $\texttt{reshape}(\texttt{thetaVec}(15:39), 4, 6)$
- $\texttt{reshape}(\texttt{thetaVec}(16:39), 6, 4)$

3. Let $J(\theta) = 3\theta^3 + 2$. Let $\theta = 1$, and $\epsilon = 0.01$. Use the formula $\frac{J(\theta+\epsilon) - J(\theta-\epsilon)}{2\epsilon}$ to numerically compute an approximation to the derivative at $\theta = 1$. What value do you get? (When $\theta = 1$, the true/exact derivative is $\frac{dJ(\theta)}{d\theta} = 9$.)

1 / 1
分数

- (x) 9.0003

正确
We compute $\frac{(3(1.01)^3+2) - (3(0.99)^3+2)}{2(0.01)} = 9.0003$.

- ( ) 9
- ( ) 11
- ( ) 8.9997

4. Which of the following statements are true? Check all that apply.

1 / 1
分数

- [ ] Using a large value of λ cannot hurt the performance of your neural network; the only reason we do not set λ to be too large is to avoid numerical problems.

未选择的是正确的

- [ ] Gradient checking is useful if we are using gradient descent as our optimization algorithm. However, it serves little purpose if we are using one of the advanced optimization methods (such as in fminunc).

未选择的是正确的

- [x] If our neural network overfits the training set, one reasonable step to take is to increase the regularization parameter λ.

正确
Just as with logistic regression, a large value of λ will penalize large parameter values, thereby reducing the changes of overfitting the training set.

- [x] Using gradient checking can help verify if one's implementation of backpropagation is bug-free.

正确
If the gradient computed by backpropagation is the same as one computed numerically with gradient checking, this is very strong evidence that you have a correct implementation of backpropagation.

A：正则化参数过大会造成高偏差，影响神经网络的性能

B：Gradient checking是检查反向传播算法计算的偏导数的正确性，无论采用梯度下降或其他高级优化方法，都要先通过这一步确认偏导数正确

5. Which of the following statements are true? Check all that apply.

1 / 1
分数

- [ ] Suppose we are using gradient descent with learning rate α. For logistic regression and linear regression, $J(\theta)$ was a convex optimization problem and thus we did not want to choose a learning rate α that is too large. For a neural network however, $J(\Theta)$ may not be convex, and thus choosing a very large value of α can only speed up convergence.

未选择的是正确的

- [x] Suppose we have a correct implementation of backpropagation, and are training a neural network using gradient descent. Suppose we plot $J(\Theta)$ as a function of the number of iterations, and find that it is **increasing** rather than decreasing. One possible cause of this is that the learning rate α is too large.

正确
If the learning rate is too large, the cost function can diverge during gradient descent. Thus, you should select a smaller value of α.

- [ ] Suppose that the parameter $\Theta^{(1)}$ is a square matrix (meaning the number of rows equals the number of columns). If we replace $\Theta^{(1)}$ with its transpose $(\Theta^{(1)})^T$, then we have not changed the function that the network is computing.

未选择的是正确的

- [x] If we are training a neural network using gradient descent, one reasonable "debugging" step to make sure it is working is to plot $J(\Theta)$ as a function of the number of iterations, and make sure it is decreasing (or at least non-increasing) after each iteration.

正确
Since gradient descent uses the gradient to take a step toward parameters with lower cost (ie, lower $J(\Theta)$), the value of $J(\Theta)$ should be equal or less at each iteration if the gradient computation is correct and the learning rate is set properly.