Assignment 1 in DD2424 Deep Learning

C. Chou

Kungliga Tekniska Högskolan

chchou@kth.se

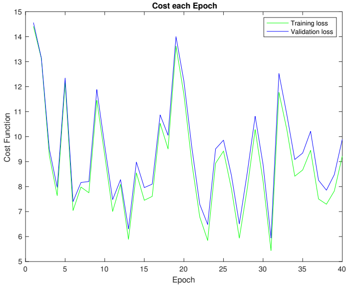
April 10, 2019

Gradient Computation

I successfully managed to compute the gradient analytically by using following code:

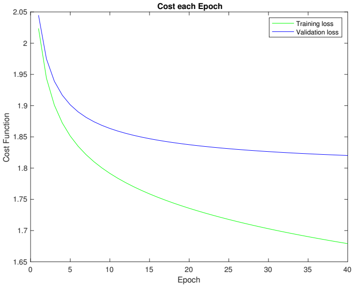
[K,n] = size(Y)
[K,d] = size(W) $\texttt{grad}_w = zeros(K, d)$
$\texttt{grad}_b = zeros(K, 1)$
g = -(Y-P)'
$\texttt{grad}_b = sum(g, 1)'./n$
$\texttt{grad}_w = g' * X'./n + 2 * lambda * W$

And the result was tested by computing the absolute difference between this result and the result from centered difference formula. The difference is less than 1e-6.



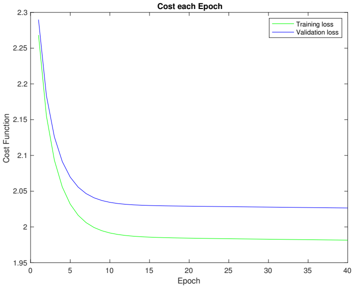
The accuracy is 26.42 %

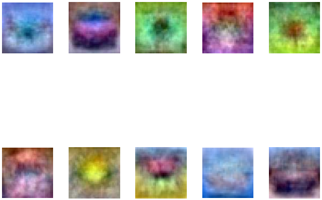
case 2: $\lambda = 0$, $n_{epochs} = 40$, $eta = 0.01$, and $n_{batch} = 100$



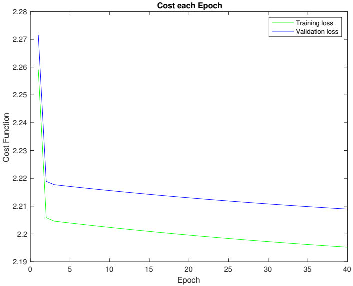
The accuracy is : 36.73 %

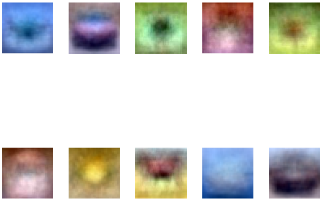
case 3: $\lambda = 0.1$, $n_{epochs} = 40$, $eta = 0.01$, and $n_{batch} = 100$





The accuracy is : 33.37 %





The accuracy is : 21.94 %

Conclusion - *lambda*

When λ is 0, means that there is no regularization, overfit might happen. From the Cost plot in case 2 we found that, when λ is 0, the cost of training data continuously decreases after 20 epoch iteration, however in the same time, the cost of validation data doesn't decrease as training data behaves. Thus in this case, our model overfits.
On the other hand, when λ is large, in case 4, for example , λ is 1, although the model is not overfitting, but the cost for both validation data and training data decrease little after 40 epoch iteration, which means that our model with $\lambda = 1$ generalizes too much that it doesn't catch all the important information,so the accuracy is low.

Conclusion - *eta*

When *eta* is not proper, our parameters w and d might converge either too fast or too slow. From the plot in case 1, we can see that after the first five epoch iteration, the cost function of both training data and validation data decrease, but after the fifth epoch iteration, the cost jumps up and down, unlike in case 2, the cost function gradually decreases. Also, comparing the accuracy of case 1 and case 2, we know that the accuracy is only 26.42% in case 1, which is lower than 36.73% in case 2.