Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization

Week 3: Hyperparameter tuning, Batch Normalization, Programming Frameworks.

1 Version 1: Which of the following are true about hyperparameter search?

Answer: Choosing random values for the hyperparameters is convenient since we might not know in advance which hyperparameters are more important for the problem at hand.

Comment: Different problems might be more sensitive to different hyperparameters.

1 Version 2: If searching among a large number of hyperparameters, you should try values in a grid rather than random values, so that you can carry out the search more systematically and not rely on chance. True/False?

Answer: False.

Comment: See lecture notes.

2 Version 1: If it is only possible to tune two parameters from the following due to limited computational resources. Which two would you choose?

Answer: α; The β parameter of the momentum in gradient descent.

Comment: This might be the hyperparameter that most impacts the results of a model; This hyperparameter can increase the speed of convergence of the training, thus is worth tuning.

2 Version 2: Every hyperparameter, if set poorly, can have a huge negative impact on training, and so all hyperparameters are about equally important to tune well. True/False?

Answer: False.

Comment: We've seen in lecture that some hyperparameters, such as the learning rate, are more critical than others.

3 Version 1: Using the panda strategy, it is possible to create several models. True/False?

Answer: True.

Comment: Following the panda analogy, it is possible to babysit a model until a certain point and then start again to produce a different one.

3 Version 2: During hyperparameter search, whether you try to babysit one model (panda strategy) or train a lot of models in parallel (caviar) is largely determined by:

Answer: The amount of computational power you can access.

4 If you think β (hyperparameter for momentum) is between 0.9 and 0.99, which of the following is the recommended way to sample a value for beta?

Answer:

```
r = np.random.rand()
beta = 1-10**(-r-1)
```

5 Finding good hyperparameter values is very time-consuming. So typically you should do it once at the start of the project, and try to find very good hyperparameters so that you don't ever have to tune them again. True or false?

Answer: False.

6 Version 1: In batch normalization as presented in the videos, if you apply it on the lth layer of your neural network, what are you normalizing?

Answer: $z^{[l]}$.

6 Version 2: When using batch normalization it is OK to drop the parameter $W^{[l]}$ from the forward propagation since it will be subtracted out when we compute $\tilde{z}^{[l]} = \gamma Z^{[l]}_{normalize} + \beta^{[l]}$. True/False.

Answer: False.

Comment: The parameter $W^{[l]}$ doesn't get subtracted during the batch normalization process, although it gets rescaled.

7 Version 1: When using normalization. In case σ is too small, the normalizaiton of $z^{[i]}$ may fail since division by 0 may be produced due to rounding errors. True/False?

Answer: False.

Comment: We have ϵ.

7 Version 2: In the normalizaiton formula, why do we use epsilon?

Answer: To avoid division by zero.

8 Version 1: Which of the following statements about γ and β in Batch Norm are true?

Answer: They set the mean and variance of the linear variable $z^{[l]}$ of a given layer; They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent.

8 Version 2: Which of the following is true about batch normalization?

Answer: The parameters $\gamma^{[l]}$ and $\beta^{[l]}$ set the variance and mean of $\tilde{z}^{[l]}$.

9 Version 1: A neural network is trained with Batch Norm. At test time, to evaluate the neural network on a new example you should perform the normalization using μ and σ^2 estimated using an exponentially weighted average across mini batches seen during training. True/False?

Answer: True.

Comment: This is a good practice to estimate the μ and σ^2 to use since at test time we might not be predicting over a batch of the same size, or it might even be a single example, thus using the μ and σ^2 of a single sample does not make sense.

9 Version 2: After training a neural network with Batch Norm, at test time, to evaluate the neural network on a new example you should:

Answer: Perform the needed normalizations, use μ and σ^2 estimated using an exponentially weighted average across mini batches seen during training.

10 Which of these statements about deep learning programming frameworks are true? (Check all that apply).

Answer: A programming framework allows you to code up deep learning algorithms with typically fewer lines of code than a lower level language such as python; Even if a project is currently open source, good governance of the project helps ensure that it remains open even in the long term, rather than become closed or modified to benefit only one company.