Congratulations! You passed!

TO PASS 80% or higher

Keep Learning

GRADE
100%

Hyperparameter tuning, Batch Normalization, Programming Frameworks

LATEST SUBMISSION GRADE

100%

1. 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 or False? **1 / 1 point**

   - ○ True
   - ◉ False

   Correct

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 or False? **1 / 1 point**

   - ○ True
   - ◉ False

   Correct

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

3. 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: **1 / 1 point**

- Whether you use batch or mini-batch optimization
- The presence of local minima (and saddle points) in your neural network
- The amount of computational power you can access
- The number of hyperparameters you have to tune

Correct

4. If you think β (hyperparameter for momentum) is between on 0.9 and 0.99, which of the following is the recommended way to sample a value for beta? **1 / 1 point**

```
r = np.random.rand()
beta = r*0.09 + 0.9
```

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

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

```
r = np.random.rand()
beta = r*0.9 + 0.09
```

Correct

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 revisit tuning them again. True or false? **1 / 1 point**

- True
- False

False

Correct

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

- $a^{[l]}$
- $z^{[l]}$
- $W^{[l]}$
- $W^{[l]}$
- $b^{[l]}$

Correct

7. In the normalization formula $z^{(i)}_{norm} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \varepsilon}}$, why do we use epsilon? 1 / 1 point

- To avoid division by zero
- In case μ is too small
- To have a more accurate normalization
- To speed up convergence

Correct

Correct

8. Which of the following statements about γ and β in Batch Norm are true? 1 / 1 point

- [ ] The optimal values are $\gamma = \sqrt{\sigma^2 + \varepsilon}$, and $\beta = \mu$.

- [ ] There is one global value of $\gamma \in \Re$ and one global value of $\beta \in \Re$ for each layer, and applies to all the hidden units in that layer.

- [ ] β and γ are hyperparameters of the algorithm, which we tune via random sampling.

- [x] They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent.

Correct

- [x] They set the mean and variance of the linear variable $z^{[l]}$ of a given layer.

Correct

- [x] They set the mean and variance of the linear variable $z^{[l]}$ of a given layer.

Correct

7. In the normalization formula $z^{(i)}_{norm} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \varepsilon}}$, why do we use epsilon? **1 / 1 point**

- (•) To avoid division by zero
- ( ) In case μ is too small
- ( ) To have a more accurate normalization
- ( ) To speed up convergence

Correct

8. Which of the following statements about γ and β in Batch Norm are true? **1 / 1 point**

- [ ] The optimal values are $\gamma = \sqrt{\sigma^2 + \varepsilon}$, and $\beta = \mu$.

- [ ] There is one global value of $\gamma \in \Re$ and one global value of $\beta \in \Re$ for each layer, and applies to all the hidden units in that layer.