## Congratulations! You passed!

Grade received 80% | Latest Submission Grade 80% | To pass 80% or higher

Go to next item

**1.** Which of the following are true about hyperparameter search? 1 / 1 point

- When using random values for the hyperparameters they must be always uniformly distributed.
- When sampling from a grid, the number of values for each hyperparameter is larger than when using random values.
- (selected) 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.
- Choosing values in a grid for the hyperparameters is better when the number of hyperparameters to tune is high since it provides a more ordered way to search.

Expand

Correct
Correct. Different problems might be more sensitive to different hyperparameters.

**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
- (selected) False

Expand

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

**3.** Even if enough computational power is available for hyperparameter tuning, it is always better to babysit one model ("Panda" strategy), since this will result in a more custom model. True/False? 1 / 1 point

- True
- (selected) False

Expand

Correct
Correct. Although it is possible to create good models using the "Panda" strategy, obtaining better results is more likely using a "caviar" strategy due to the number of tests and the nature of the deep learning process of ideas, code, and experiment.

**4.** Knowing that the hyperparameter $\alpha$ should be in the range of $0.001$ and $1.0$. Which of the following is the recommended way to sample a value for $\alpha$? 0 / 1 point

- r = 4*np.random.rand()
  alpha = 10**r
- r = np.random.rand()
  alpha = 0.001 + r*0.999
- r = -3*np.random.rand()
  alpha = 10**r
- r = -5*np.random.rand()
  alpha = 10**r

Expand

**Incorrect**
No. This will pick a random value from a uniform scale, which is not the recommended way to choose $\alpha$.

**5.** Finding new values for the hyperparameters, once we have found good ones for a model, should only be done if new hardware or computational power is acquired. True/False? **1 / 1 point**

- True
- False

Expand

**Correct**
Correct. As the data changes for the model, it might be beneficial to tune some of the hyperparameters again.

**6.** 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 $\tilde{z}^{[l]} = \gamma z^{[l]}_{\text{normalize}} + \beta^{[l]}$. True/False? **1 / 1 point**

- False
- True

Expand

**Correct**
Correct. The parameter $W^{[l]}$ doesn't get subtracted during the batch normalization process, although it gets re-scaled.

**7.** When using normalization: **0 / 1 point**

$z^{(i)}_{norm} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \varepsilon}}$

In case $\sigma$ is too small, the normalization of $z^{(i)}$ may fail since division by 0 may be produced due to rounding errors. True/False?

- True
- False

Expand

**Incorrect**
Incorrect. The normalization formula uses a smoothing parameter $\epsilon$ so in $z^{(i)}_{\text{norm}} = \frac{z^{(i)}-\mu}{\sqrt{\sigma^2+\epsilon}}$ use of the $\epsilon$ parameter prevents that the denominator be 0.

**8.** Which of the following statements about $\gamma$ and $\beta$ in Batch Norm are true? **1 / 1 point**

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

  **Correct**

- [ ] There is one global value of $$\gamma \in \Re$$ and one global value of $$\beta \in \Re$$ for each layer, and these apply to all the hidden units in that layer.
- [ ] $$\beta$$ and $$\gamma$$ are hyperparameters of the algorithm, which we tune via random sampling.
- [ ] The optimal values are $$\gamma = \sqrt{\sigma^2 + \varepsilon}$$, and $$\beta = \mu$$.
- [x] They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent.

  **Correct**

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/Fraktur-Regular.js momentum, or RMSprop, not just with gradient descent.

Expand

**Correct**
Great, you got all the right answers.

**9.** 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 $\mu$ and $\sigma^2$ estimated using an exponentially weighted average across mini-batches seen during training. True/false? **1 / 1 point**

- ○ False
- ◉ True

Expand

**Correct**
Correct. This is a good practice to estimate the $\mu$ and $\sigma^2$ 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 $\mu$ and $\sigma^2$ of a single sample doesn't make sense.

**10.** Which of the following are some recommended criteria to choose a deep learning framework? **1 / 1 point**

- ○ It must run exclusively on cloud services, to ensure its robustness.
- ○ It must be implemented in C to be faster.
- ◉ Running speed.
- ○ It must use Python as the primary language.

Expand

**Correct**

Correct. The running speed is a major factor, especially when working with large datasets.