Review Question

(1/1 point)

One of the following statements is actually a lie. Select the statement that is inaccurate from the list below:

- (●) Overfitting means your machine learning algorithm is performing at 100% and has been over trained ✔
- ( ) In supervised learning, you provide the algorithm the correct answers while training it
- ( ) If you split off too many samples for testing, your training is going to suffer as a consequence
- ( ) There shouldn't be an overlap in your training and testing dataset, because your algorithm already has the answers to the training data

EXPLANATION

The incorrect statement above is the overfitting one.

Lab

Overfitting actually means that you've trained your model to the point where it classifies even the noise in your training dataset, as opposed to the underlying relationship being modeled. You can have 100% accuracy on your training dataset and not be overfit; but you can have 80% accuracy on your training set and be overfitted.

Overfitting usually happens when your model is overly complex because the number of features is greater than the number of observations. It can also happen if you train against your entire dataset without saving a testing set to **validate** if your data has truly become overfit or not. Such a model would perform strong on the training dataset, but would not generalize well at all, performing badly on unseen samples. Even small fluctuations in the training data causes the model to react to it in an attempt to classify it.

You have used 1 of 2 submissions