

Which of the following statements about diagnostics are true? Check all that apply.

It's hard to tell what will work to improve a learning algorithm, so the best approach is to go with gut feeling and just see what works.

Diagnostics can give guidance as to what might be more fruitful things to try to improve a learning algorithm.

Diagnostics can be time-consuming to implement and try, but they can still be a very good use of your time.

A diagnostic can sometimes rule out certain courses of action (changes to your learning algorithm) as being unlikely to improve its performance significantly.

234

---

Suppose an implementation of linear regression (without regularization) is badly overfitting the training set. In this case, we would expect:

The training error  $J(\theta)$  to be **low** and the test error  $J_{\text{test}}(\theta)$  to be **high**

The training error  $J(\theta)$  to be **low** and the test error  $J_{\text{test}}(\theta)$  to be **low**

The training error  $J(\theta)$  to be **high** and the test error  $J_{\text{test}}(\theta)$  to be **low**

The training error  $J(\theta)$  to be **high** and the test error  $J_{\text{test}}(\theta)$  to be **high**

1

---

Consider the model selection procedure where we choose the degree of polynomial using a cross validation set. For the final model (with parameters  $\theta$ ), we might generally expect  $J_{\text{cv}}(\theta)$  to be lower than  $J_{\text{test}}(\theta)$  because:

An extra parameter ( $d$ , the degree of the polynomial) has been fit to the cross validation set.

An extra parameter ( $d$ , the degree of the polynomial) has been fit to the test set.

The cross validation set is usually smaller than the test set.

The cross validation set is usually larger than the test set.

1

---

Suppose you have a classification problem. The (misclassification) error is defined as

$\frac{1}{m} \sum_{i=1}^m \text{err}(h_\theta(\mathbf{x}^{(i)}), y^{(i)})$ , and the cross validation (misclassification) error is similarly defined, using the cross validation examples  $(\mathbf{x}_{\text{cv}}^{(1)}, y_{\text{cv}}^{(1)}), \dots, (\mathbf{x}_{\text{cv}}^{(m_{\text{cv}})}, y_{\text{cv}}^{(m_{\text{cv}})})$ . Suppose your training error is 0.10, and your cross validation error is 0.30. What problem is the algorithm most likely to be suffering from?

High bias (overfitting)

High bias (underfitting)

High variance (overfitting)

Correct

High variance (underfitting)

---

In which of the following circumstances is getting more training data likely to significantly help a learning algorithm's performance?

Algorithm is suffering from high bias.

Algorithm is suffering from high variance.

Correct

$J_{\text{CV}}(\theta)$  (cross validation error) is much larger than  $J_{\text{train}}(\theta)$  (training error).

Correct

$J_{\text{CV}}(\theta)$  (cross validation error) is about the same as  $J_{\text{train}}(\theta)$  (training error).

---

Suppose you fit a neural network with one hidden layer to a training set. You find that the cross validation error  $J_{\text{CV}}(\theta)$  is much larger than the training error  $J_{\text{train}}(\theta)$ . Is increasing the number of hidden units likely to help?

Yes, because this increases the number of parameters and lets the network represent more complex functions.

Yes, because it is currently suffering from high bias.

No, because it is currently suffering from high bias, so adding hidden units is unlikely to help.

No, because it is currently suffering from high variance, so adding hidden units is unlikely to help.

4.

---

Why is the recommended approach to perform error analysis using the cross validation data used to compute  $J_{\text{CV}}(\theta)$  rather than the test data used to compute  $J_{\text{test}}(\theta)$ ?

The cross validation data set is usually large.

This process will give a lower error on the test set.

If we develop new features by examining the test set, then we may end up choosing features that work well specifically for the test set, so  $J_{\text{test}}(\theta)$  is no longer a good estimate of how well we generalize to new examples.

Correct

Doing so is less likely to lead to choosing an excessive number of features.

---