Deciding what to Try next

Debugging a learning algorithm:

* Get more training examples
* Try smaller sets of features to prevent overfitting
* Try getting additional features
* Try adding polynomial features
* Try decreasing or increasing lambda

Evaluating a Hypothesis

Machine learning diagnostic:

* Diagnostic:
  + A test that you can run to gain insight what is/isn’t working with a learning algorithm, and gain guidance as to how best to improve its performance.

Training/testing procedure for linear regression

* Learn parameter theta from training data (minimizing training error J(theta) (70%))
* Compute test set error:
  + Jtest(theta) = (1/2mtest) \* (summation from i = 1 to mtest)(htheta(xitest) – yitest)2

Training/testing procedure for logistic regression

* Learn parameter theta from training data
* Compute test set error
* Misclassification error (0/1 misclassification error):
* Error(htesta(x), y)
  + = 1 if htheta(x) >= 0.5, y = 0 or if htheta(x) < 0.5, y = 1
  + = 0 otherwise
* Test error = (1/mtest)(summation i = 1 to mtest)err(htheta(xitest), ytesti)

Model Selection

* Once parameters theta0, theta1, etc… were fit to some set of data (training set), the error of the parameters as measured on that data (the training error J(theta)) is likely to be lower than the actual generalization error.
* Data 60%, Cross Validation 20%, Test 20%
* Use cross validation to choose model, and then use test to evaluate model error.

Diagnosis bias v.s. variance

* Suppose your learning algorithm is performing less well than you were hoping (Jcv(theta) is Jtest(theta) is high). Is it a bias or variance problem?
  + If Jcv(theta) and Jtrain(theta) are high , then bias (underfit)
  + If Jcv(theta) is high and Jtrain(theta) is low, then bias