What to Do Next?

How to improve the machine learning algorithm? How can we improve it? We might think of following based on the situation:

- Get more training dataset: It helps only if we have high variance problem
- Try smaller set of features: only helps in high variance problem (overfitting)
- Try additional features: fixes high bias problem
- Building new features: fixes high bias problem
- Adding polynomial features: fixes high bias problem
- Try decreasing λ: fixes high bias problem
- Try increasing λ: fixes high variance problem

In case we are using the neural network, if we use few hidden layers, we are prone to Underfitting problem and if we use so many hidden layers, it gets more computationally expensive and prone to overfitting which we can take care of with regularization. The e=best approach for neural network is to start with1, 2,3, … layers and see how they perform on cross validation.

Hypothesis Evaluation

When we try to fit parameters, always we try to minimize the errors. Low error in training dataset, might be indication of overfitting. We can define whether algorithm is overfit or not by plotting the hypothesis, but it is not efficient for larger feature sets. Standard wsy of evaluating hypothesis is to calculate the test error. We may split dataset with 70:30 ratio for training and test dataset. Test error for classification problem is calculated by:

$$test\ error = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} error(h_\Theta(x^i_{test}),\ y^i_{test})$$

Error inside of the summation is 1 if $h_\Theta \geq 0.5$ and y=0 or $h_\Theta < 0.5$ and y=1, otherwise it's 0.

Model Selection

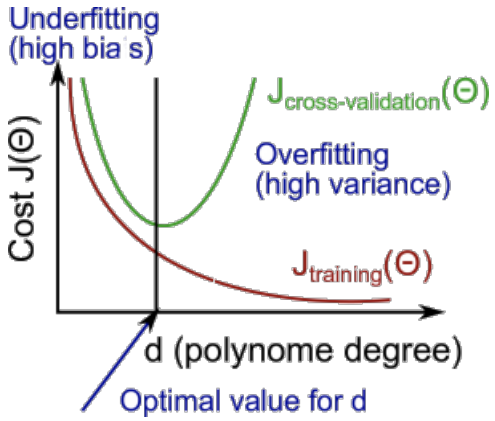
The model selection is to find the best degree of polynomial model. The best way of splitting data is to divide it into 60:20:20 for training, cross validation and test set. In this case, we can calculate the errors for 3 different set. The best way to find the best degree of polynomial for model is to have several degrees of hypothesis and evaluate them as follow:

- Test hypothesis on the cross validation set
- Pick the hypothesis with the lowest cross validation error
- Finally, estimate generalization error of model using the test set

What to Do Next?

Bias vs Variance

One of main problems in model selection is variance and bias. High variance is referred to Underfitting and high bias is a overfitting. The plot below shows higher degree of polynomial causes low training error and high cross validation error. On the other hand, low degree of polynomial results in high error is validation and training error.

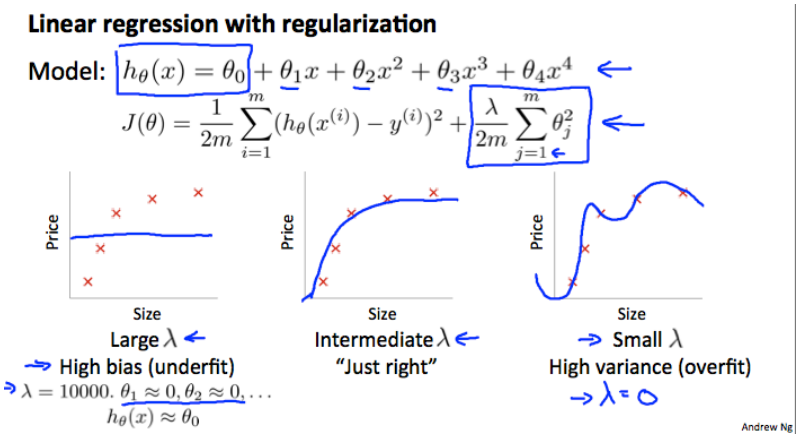


In summary:

- High bias model: high validation error and training error (cross validation error and test error trend will be similar), which occurs when we fit low degree of polynomial.
- High variance model: high validation error and low training error which happens on high polynomial model. This model shows it doesn't generalize well.

Regularization and Bias/Variance

You may ask how regularization effects the bias and variation problem?



The figure above, when we use a large λ in our model, all of features get penalized and the hypothesis ends up to be a line with 0 slop, which results in Underfitting. On the other hand, the small λ results in overfitting since we are considering every single point. The best way to choose the right λ is as follow:

What to Do Next?

- Calculate the error of training dataset of hypothesis without regularization, $\lambda = 0$ (mean squared error)
- Have a range of λ to try (like 0.01, 0.02, 0.04…) to make several hypotheses
- Iterate through the λs and for each λ go through all the models to learn some Θ.
- Select the hypothesis with lowest cross validation error
- Finally, test the hypothesis on a testing set to see how well the model generalizes to a new set.

Learning Curve

Learning curves are used to check the performance of our model and decide what needs to improve. It's a plot of average training set size vs cross validation error.

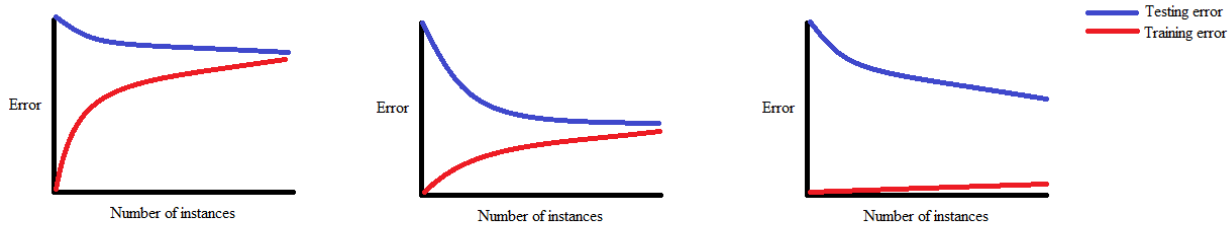


Image #1 (high bias) Image #2 (ideal) Image #3 (high variance)

It makes sense that generally by increasing the number of training set, the test error decreases and training error increases. High bias model causes a high error and adding more training set would not help improving the model. On the other hand, the high variance problem, adding more number of training set, will improve the model significantly. *when we have high bias system (too simple model) we would like to make it a high variance problem and decrease the error by adding more training dataset,* in ideal learning curve, the error and gap between testing and training error is small.