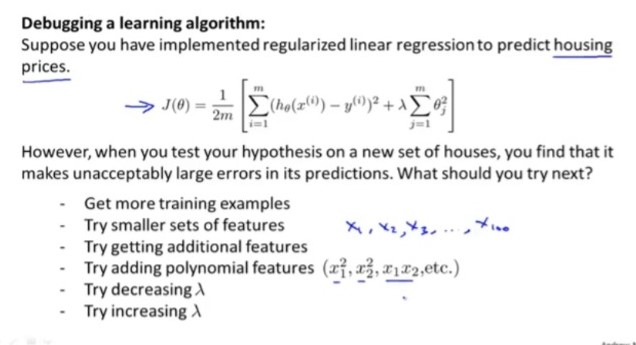
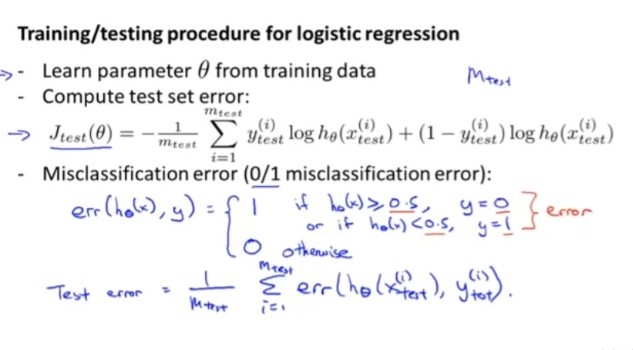
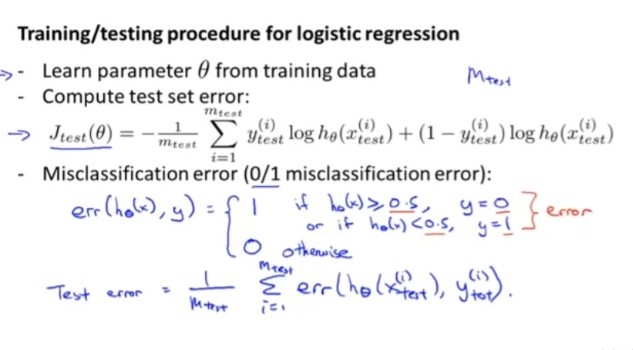
Advice for Applying Machine Learning



Evaluating Your Hypothesis



Model Selection

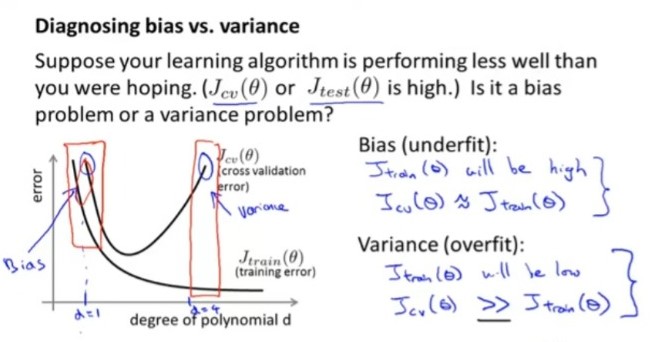
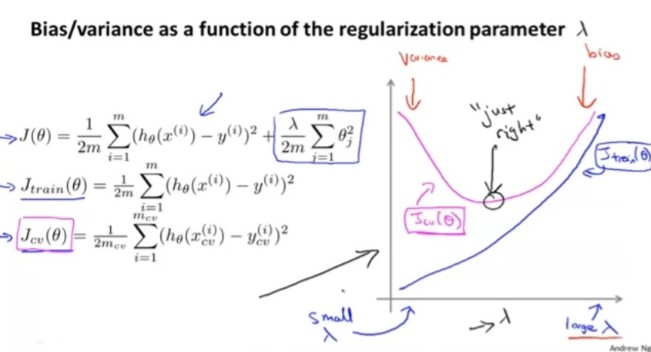
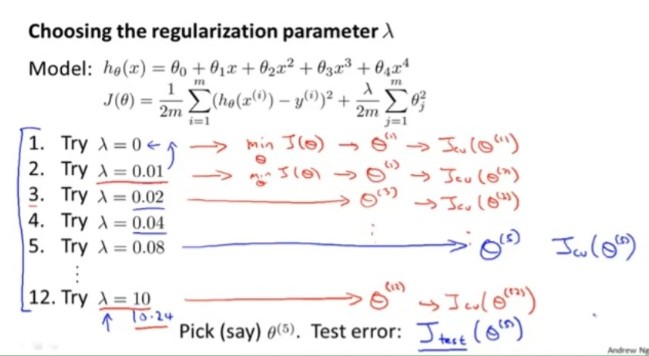
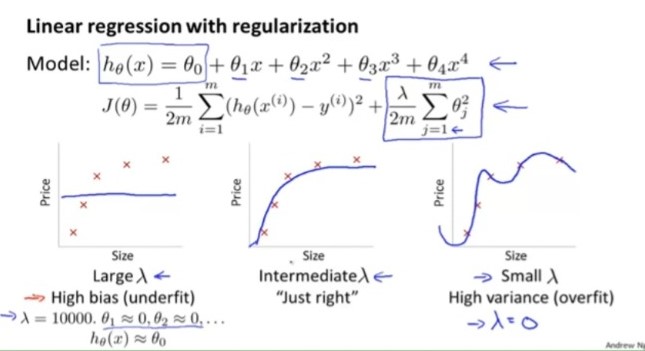


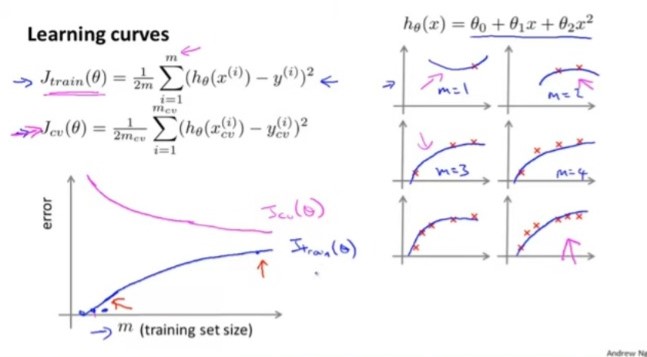
Instead of Splitting data into training and testing split it into

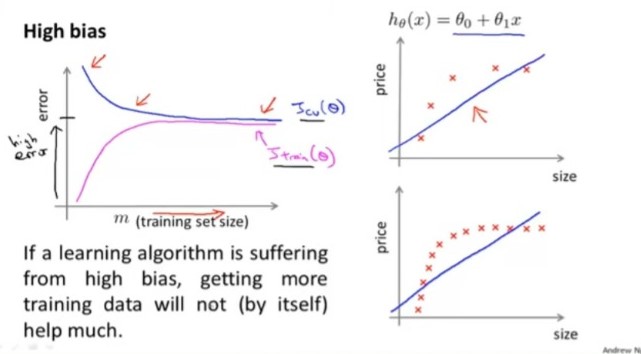
Training data Cross Validation set Testing data and

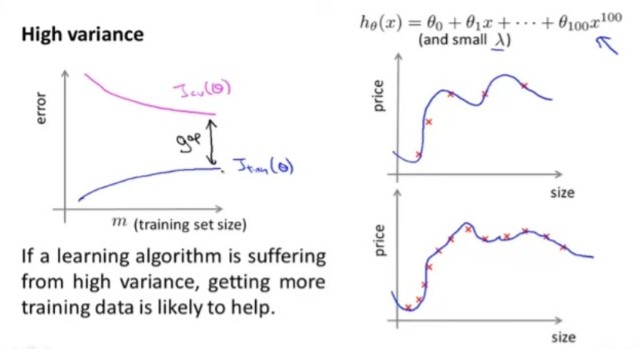
Use cross validation set for evaluating your model and choose the degree which gives generalized one

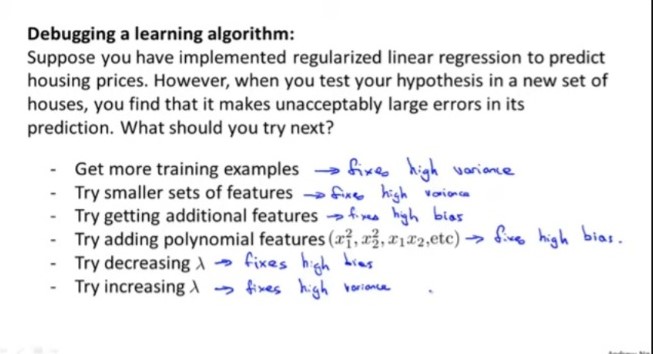
Diagnosing Bias/Variance

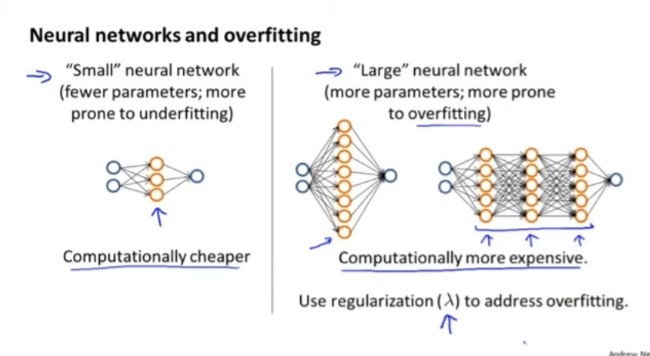
`











ERROR ANALYSIS

Skewness

It refers to a situation where distribution of classes or categories in the dataset are highly imbalanced. The term "skewed" is used because the distribution is not balanced or symmetrical.

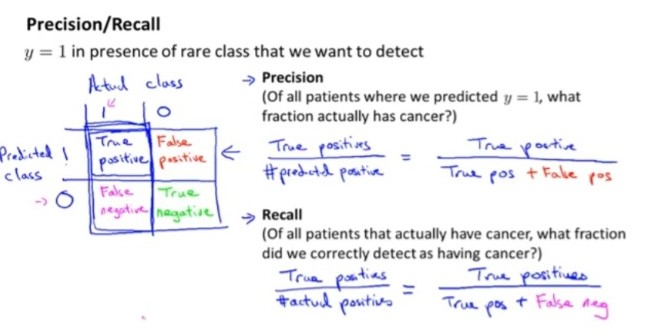
For example, consider a dataset used for predicting customer churn in a telecom company. If only 5% of the customers in the dataset have churned while the remaining 95% have not, then the class distribution is skewed. In this case, the "churned" class is the minority class, and the "not churned" class is the majority class.

Skewed classes can pose challenges in machine learning tasks because models trained on such datasets tend to be biased towards the majority class.

In Case of Skewed class the evaluation metrics we ae going to use are

Precision and Recall

**EVALUATION METRICS**

****

**Precision and Recall**

The ideal values for precision and recall in a classifier depend on the specific requirements and priorities of the problem at hand. The trade-off between precision and recall is often represented by the precision-recall curve.

Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives). It indicates the classifier's ability to avoid false positives.

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives + false negatives). It indicates the classifier's ability to identify all positive instances.

In some cases, precision is more important, such as when the cost of false positives is high, and you want to minimize the number of false alarms. For example, in medical diagnostics, you may want to ensure that positive predictions are highly reliable to avoid unnecessary treatments or surgeries.

In other scenarios, recall may be the priority, particularly when the cost of false negatives is high, and you want to minimize the number of missed positive instances. For instance, in spam email detection, you would want to capture as many spam emails as possible to prevent them from reaching users' inboxes.

In general, a good classifier should aim to achieve a balance between precision and recall. However, the optimal trade-off point depends on the specific problem and its associated costs and consequences. A high precision value indicates low false positive rates, while a high recall value indicates low false negative rates.

In addition to precision and recall, other performance metrics like accuracy, F1 score, and area under the precision-recall curve (AUPRC) can also provide insights into the classifier's overall performance and trade-offs between different evaluation measures. The choice of appropriate metrics should consider the specific goals and requirements of the classification problem.

