Programming Assignment 3 – Bias-Variance Tradeoff and Learning Curve

Due date: Friday, November 11, 2016 by midnight

Language: Java, Perl, Python, etc.

An important concept in machine learning is the bias-variance tradeoff. Models with high bias are not complex enough for the data and tend to underfit, while models with high variance overfit to the training data. In this programming assignment, you will plot training and test errors on a learning curve to diagnose bias-variance problems.

You have a dataset (Figure 1) containing historical records on the change in the water level, x, and the amount of water flowing out of the dam, y.

This dataset is divided into 2 parts:  
A training set that your model will learn on: training\_x, training\_y  
A cross validation set for calculate cross validation error: cv\_x, cv\_y

**PART 1:**

For linear regression (degree = 1):

our hypothesis has the form:

Apply linear regression (we have done in programming assignment 1) to our training data to get θ parameter. Then compute the error on the training and cross validation sets. Note that the training error for a dataset is defined as

To plot the learning curve, we need a training and cross validation set error for different training set sizes.

As you have 12 rows of training data set, to obtain error with different training set sizes, you should repeat 11 times linear regression with different subsets of the original training set. When you are computing the training set error, make sure you compute it on the training subset.

For instance:

Round 1: use training subset (row: 1, 2) to train your linear regression model. Compute the training error on the training subset. Compute the cross-validation error over the entire cross validation set.

Round 2: use training subset (row: 1, 2, 3)

Round 3: use training subset (row: 1, 2, 3, 4)

…

At last, you will produce a plot similar to Figure 2.

In Figure 2, you can observe that both the training error and cross validation error are high when the number of training examples is increased. This reflects a high bias problem in the model - the linear regression model is too simple and is unable to fit our dataset well.

**PART 2:**

The problem with our linear model was that it was too simple for the data and resulted in underfitting (high bias). In this part of the exercise, you will address this problem by adding more features.

For use polynomial regression (degree = p, p > 1), our hypothesis has the form:

If the degree is too high, you will observe in learning curve figure that low training error is low, but the cross validation error is high. There is a gap between the training and cross validation errors, indicating a high variance problem.

Your tasks:

1. Follow the above instruction, generate the learning curve with degree 1 and degree 6.
2. Answer questions: what do you observe in each learning curve figure? is that underfitting or overfitting for each degree?

Send your code file, learning curve and answer to zz296@njit.edu

Reference:   
<https://www.youtube.com/watch?v=Iz61geqni7g&index=59&list=PLJ1-ciQ35nuiyL1PX6O4NdF5CjjaDdnVC>  
<https://github.com/everpeace/ml-class-assignments/blob/master/ex5.Regularized_Linear_Regression_and_Bias-Variance/ex5.pdf>