实验报告-ex5

编程软件：MATLAB，notepad++

编程环境：Windows10

作业任务：本次实验分为上下两部分，第一部分是完成正则化的线性回归算法来根据水库中的水平线高度预测水流量。第二部分是debugging得到的算法，并且评价bias vs. variance

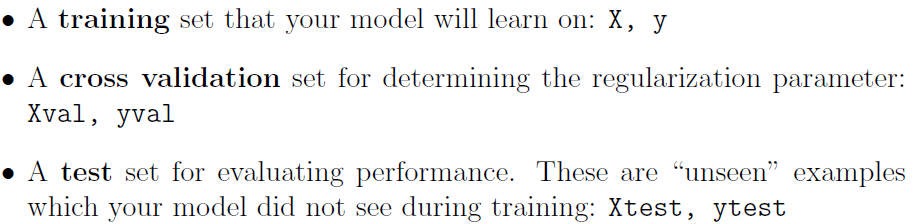
。有ex5一个函数，函数中套用了相应的函数，所以需要根据这两个目录函数的指示和实验手册，完成对应的5个部分。

1 Regularized Linear Regression

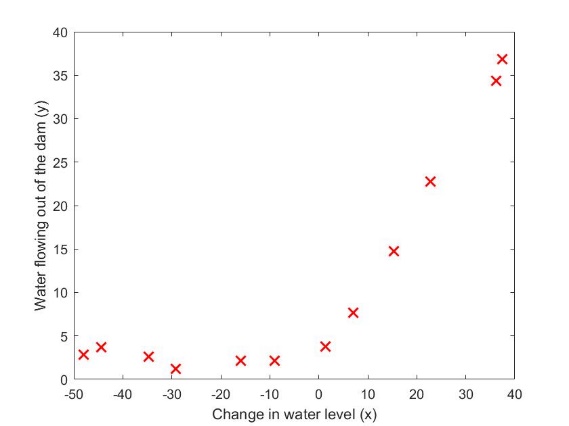
这一部分实现了带正则化的线性回归算法，为之后学习曲线的绘制奠定基础。

1.1 Visualizing the dataset

首先对dataset进行可视化操作。在本次实验中，dataset被分为三个部分：



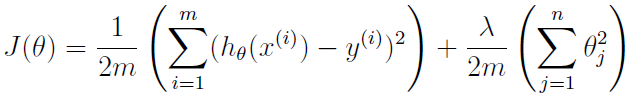
图像是针对training set进行绘制的。

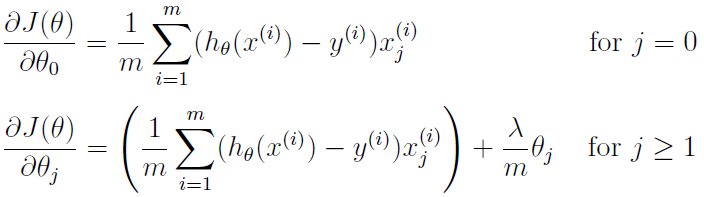


1.2 Regularized linear regression cost function

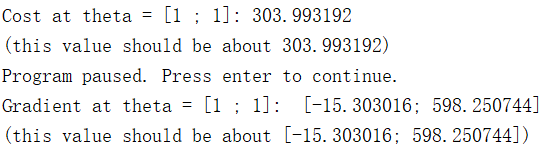
1.3 Regularized linear regression gradient

由前面已学知识可知，线性回归算法的cost function和gradient为：



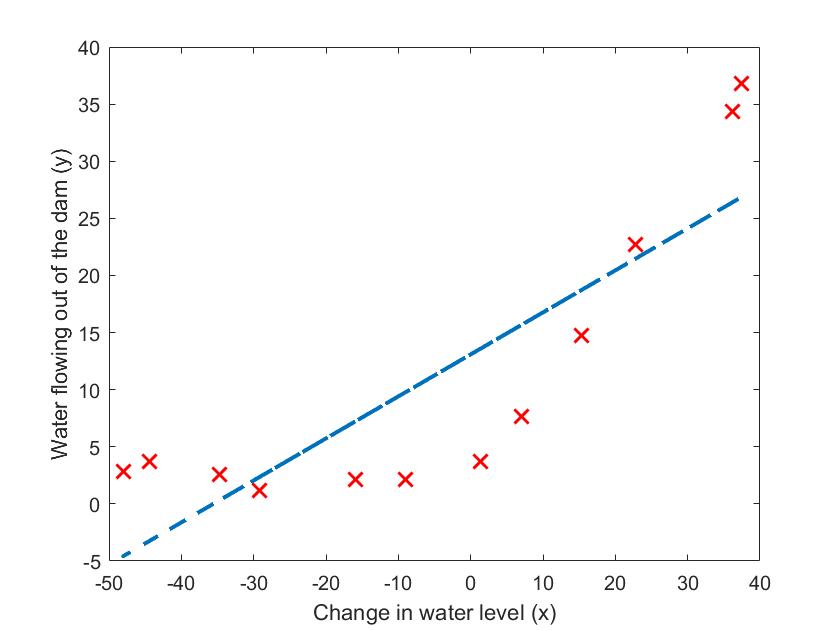


其中，λ为正则化参数，可以帮助防止overfitting问题的发生。同时要注意θ\_0不应该被正则化



1.4 Fitting linear regression

这里利用了fmincg函数对θ进行最优化处理。这里我们将λ置零，因为我们当前的线性回归算法只处理了两个维度的θ，所以λ对这么低维度的θ影响不大。得到θ后，将其可视化，如下图：



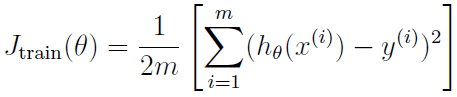
我们可以发现，这不是一个很好的模型，因为training set不是线性的，而且是high bias。但是不是所有的模型都可以用可视化来展现以判断其出现的问题，所以在后面需要画出学习曲线来帮助判断。

2 Bias-variance

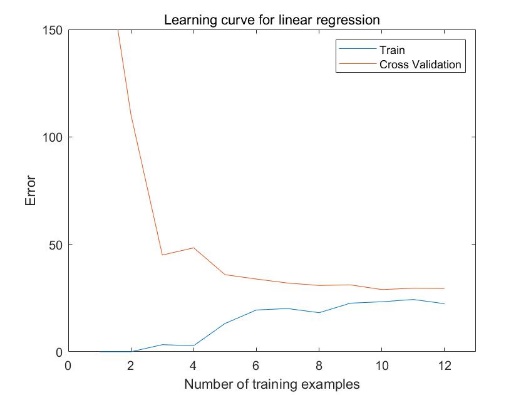
这一部分实现了机器学习中很重要的一个概念：bias-variance tradeoff。我们已知high bias意味着underfitting，high variance意味着overfitting。所以在这一部分，我们将绘制training set和test error在学习曲线上，以此评估bias-variance问题。

2.1 Learning curves

这一部分需要我们完成对学习曲线的绘制代码。由已学知识可知，学习曲线绘制了训练集和交叉验证集关于training set大小的函数曲线。我们要完成的是learningCurve函数，使其返回训练集和交叉验证集的error：

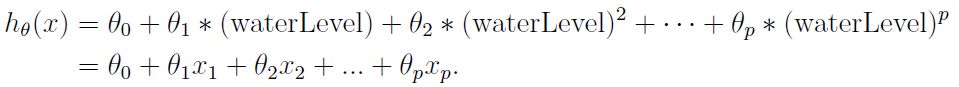


我们要做的就是对training set的大小m进行循环，每次用前i个examples计算theta，然后分别计算train和cv的error。得到的两个m大小的向量。



3 Polynomial regression

前面我们只用了二维的它和θ，所以导致太简单，underfitting。在这一部分将引入更多的变量来解决这个问题。

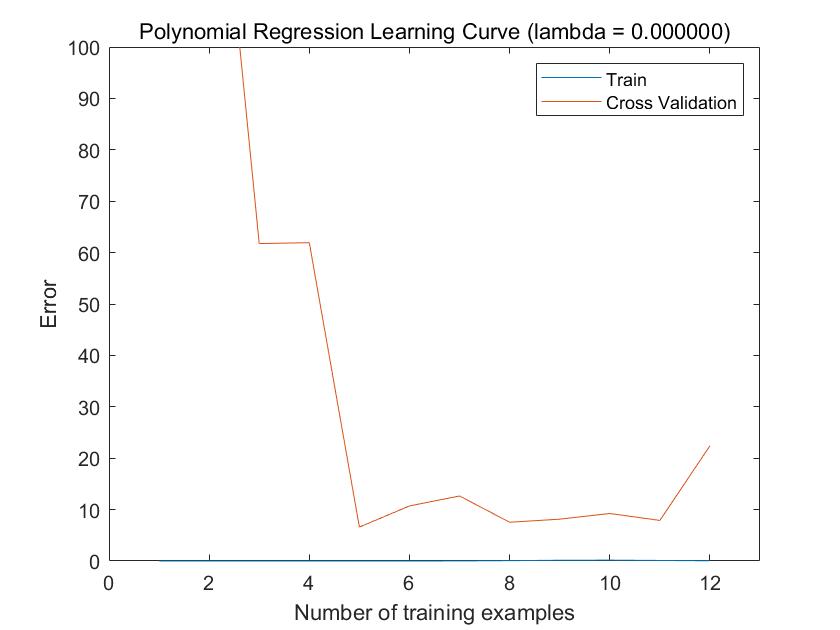
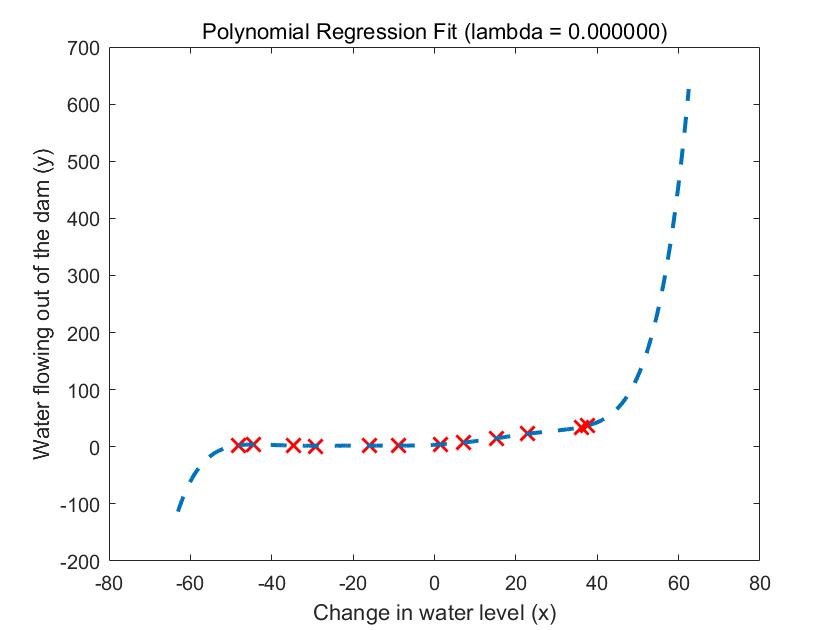


所以现在需要在polyFeatures函数中对已有的X进行扩展变量，从m×1变为m×p。

3.1 Learning Polynomial Regression

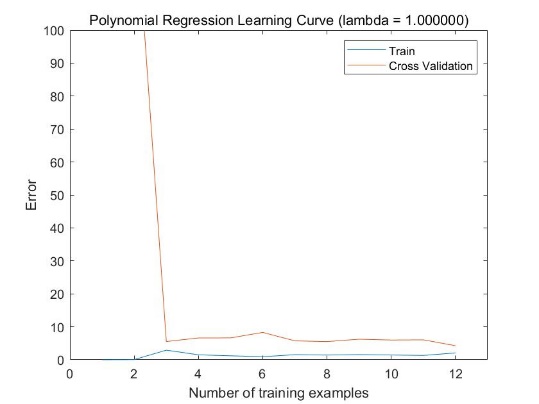
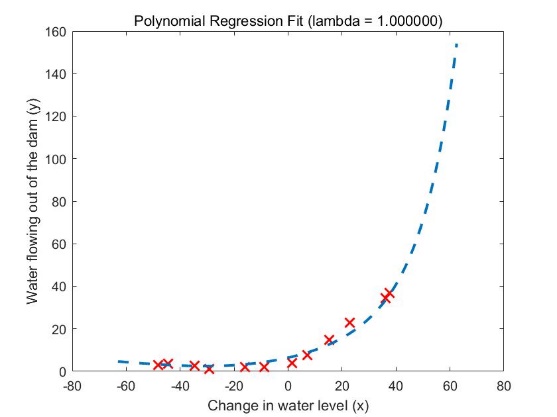
由于现在对变量进行了p次方处理，所以导致每一个变量的范围差异很大，所以就需要用到特征缩放，分别记录μ和σ。

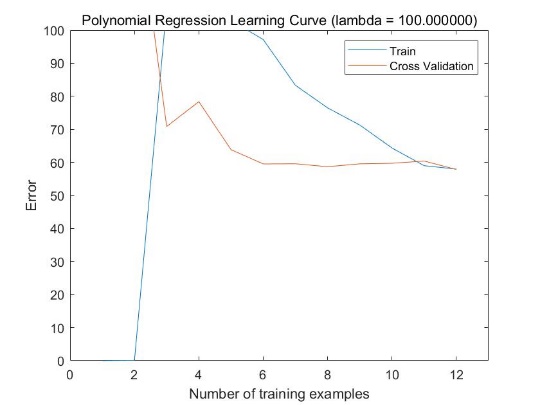
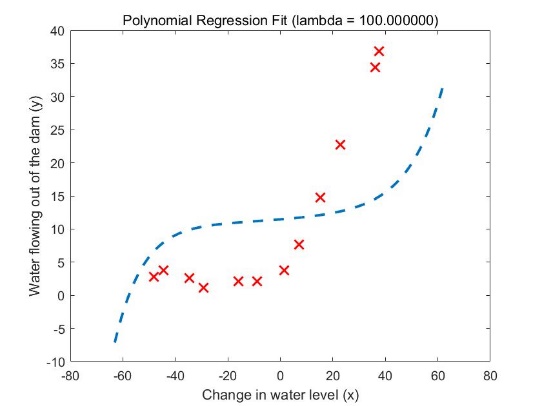
而对θ的求解和学习曲线的绘制与前面相同，不需要对代码做出修改。回归曲线和学习曲线如下图：



3.2 Optional (ungraded) exercise: Adjusting the regularization parameter

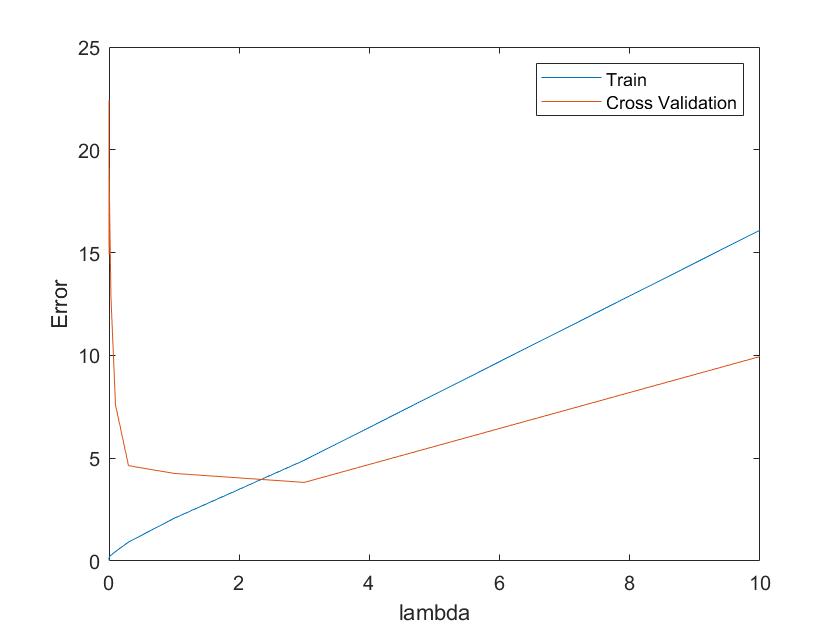
这一部分可以观察不同的λ对回归算法的影响。

λ=1：

λ=100：

3.3 Selecting λ using a cross validation set

这一部分要求对正则化参数λ进行选择。在validationCurve函数中，实现对给定可能取到数值的λ向量进行遍历，找到所有λ值对应的error。过程较简单，与绘制学习曲线的思路一致，只是此时是lamda的值随着遍历次数改变，而X和y不在改变。



根据上图可以看出，最佳的λ值应该在3附近。

○ 至此所有评分部分成功完成！

3.4 Optional (ungraded) exercise: Computing test set error

在前面的实验部分，我们已经成功地选出最佳的λ=3，但是我们仍然需要通过test error来验证该模型的正确性。

需要注意的是得到的λ应该是在学习曲线中test error和train error差别最小的。并且求解θ时使用的是X\_poly和y，而求解J时才用的X\_poly\_test和ytest



3.5 Optional (ungraded) exercise: Plotting learning curves with randomly selected examples

这一部分提出了绘制学习曲线的改进方法。不再是每次取前i个examples了，而是每次随机取i个train和cv examples。如此重复计算50次，取test error的平均值。然后画出学习曲线即可。

