LATE DAYS USED: 2

CIS419 Machine Learning

Assignment II

Part I

James Wang
Oct. 07, 2015

PennKey: jamwang
PennID: 46576241

Part I: Problem Set

1 Gradient Descent

If α_k is a constant value for each iteration k of gradient descent, that means each "step" taken at every iteration along the cost function towards convergence at the global minimum is of the same size.

If α_k is a function of k for each iteration k of gradient descent then the implications are that as k increases, α_k will either be increasing or decreasing with each iteration, which means that θ_j will most likely not converge at the global minimum because either each successive step is too large (increasing α_k) or too small (decreasing α_k) so θ_j never reaches the global minimum.

2 Fitting an SVM by Hand

a) $< \frac{\sqrt{2}}{2}, 1 >$

b) $\sqrt{6}$

c) $\mathbf{w^T} = \left[\frac{\mathbf{1}}{\mathbf{3}}, \frac{\mathbf{1}}{\mathbf{2}}, \frac{\mathbf{-1}}{\mathbf{2}}\right]$

d) $w_0 = \frac{\mathbf{2}}{\mathbf{3}}$

e) $h(x) = w_0 + \mathbf{w^T}\emptyset(x) = \mathbf{1} + \frac{\sqrt{2}-x^2}{2}$

Part II: Programming Exercises

1 Polynomial Regression

1.1 Implementing Polynomial Regression

The graph has no regularization when $\lambda = 0$.

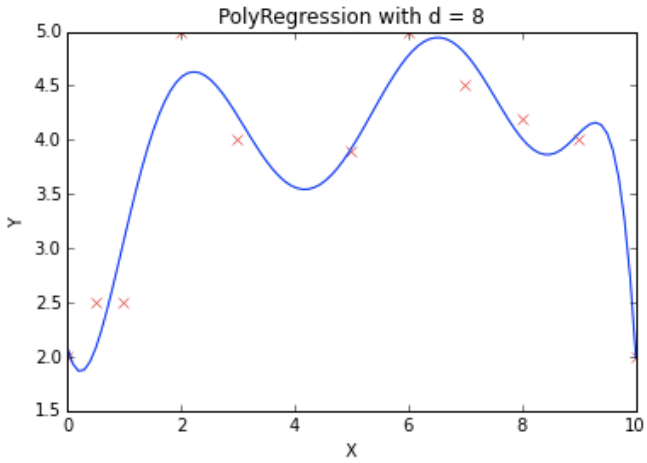


Figure 1. Lambda = 0

When we increase the amount of regularization (i.e. by increasing λ) by even the slightest, the results are immediate.

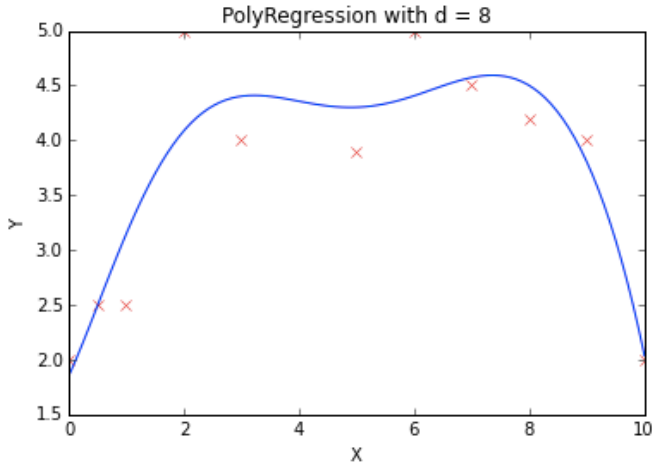


Figure 2. Lambda = 0.000001

In general, the resulting effects on the function are that it appears less volatile (in terms of the swings and dips of the function) and begins to generalize the data better.

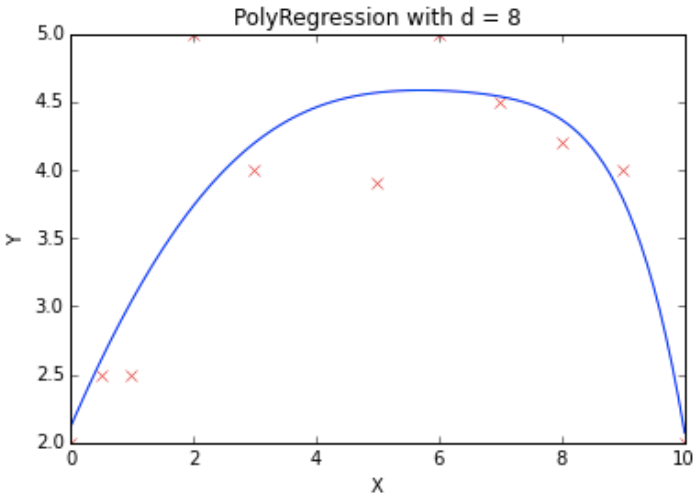
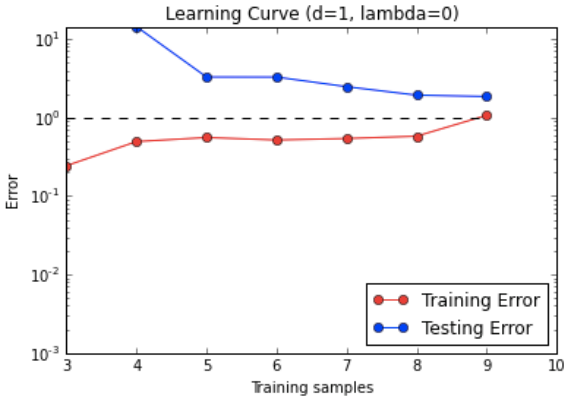


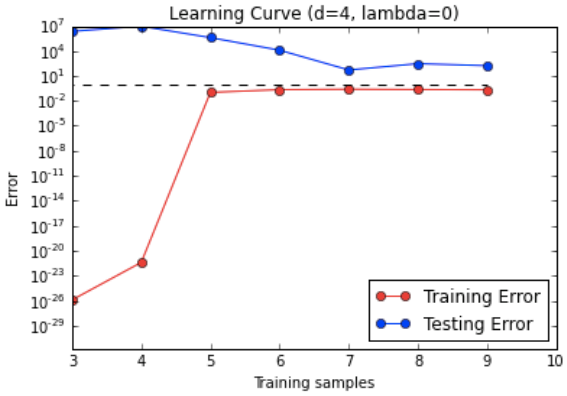
Figure 3. Lambda = 0.01

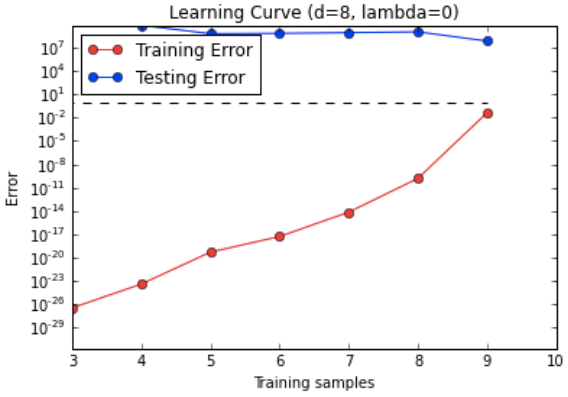
While it is ideal to have some regularization, too much is detrimental to the model too. As λ increases, bias increases and variance decreases as the trade-off. Typically, a model selection technique like cross-validation is used to select the appropriate value for λ.

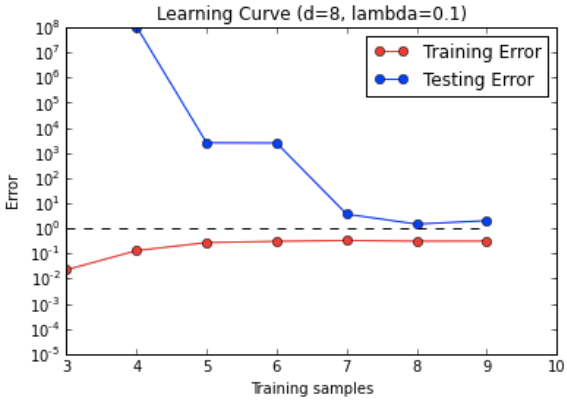
1.2 Examine the Bias-Variance Tradeoff Through Learning Curves

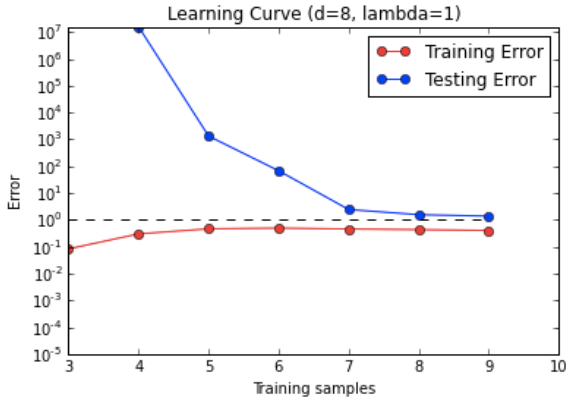
A general observation is that as the number of training samples increase, both training error and testing error decrease. Without regularization, testing error becomes very high as the degree of polynomials increases. With respect to model parsimony, this is most likely because the model started to over-fit the training data.

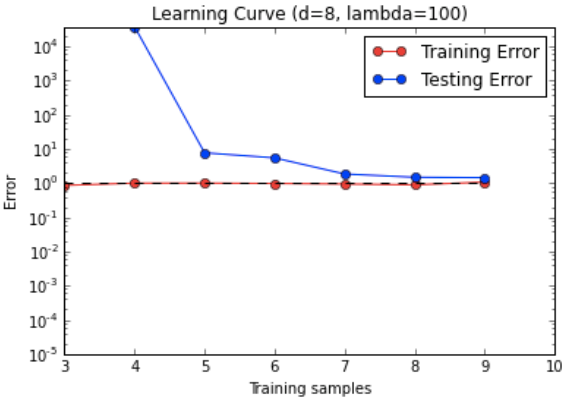




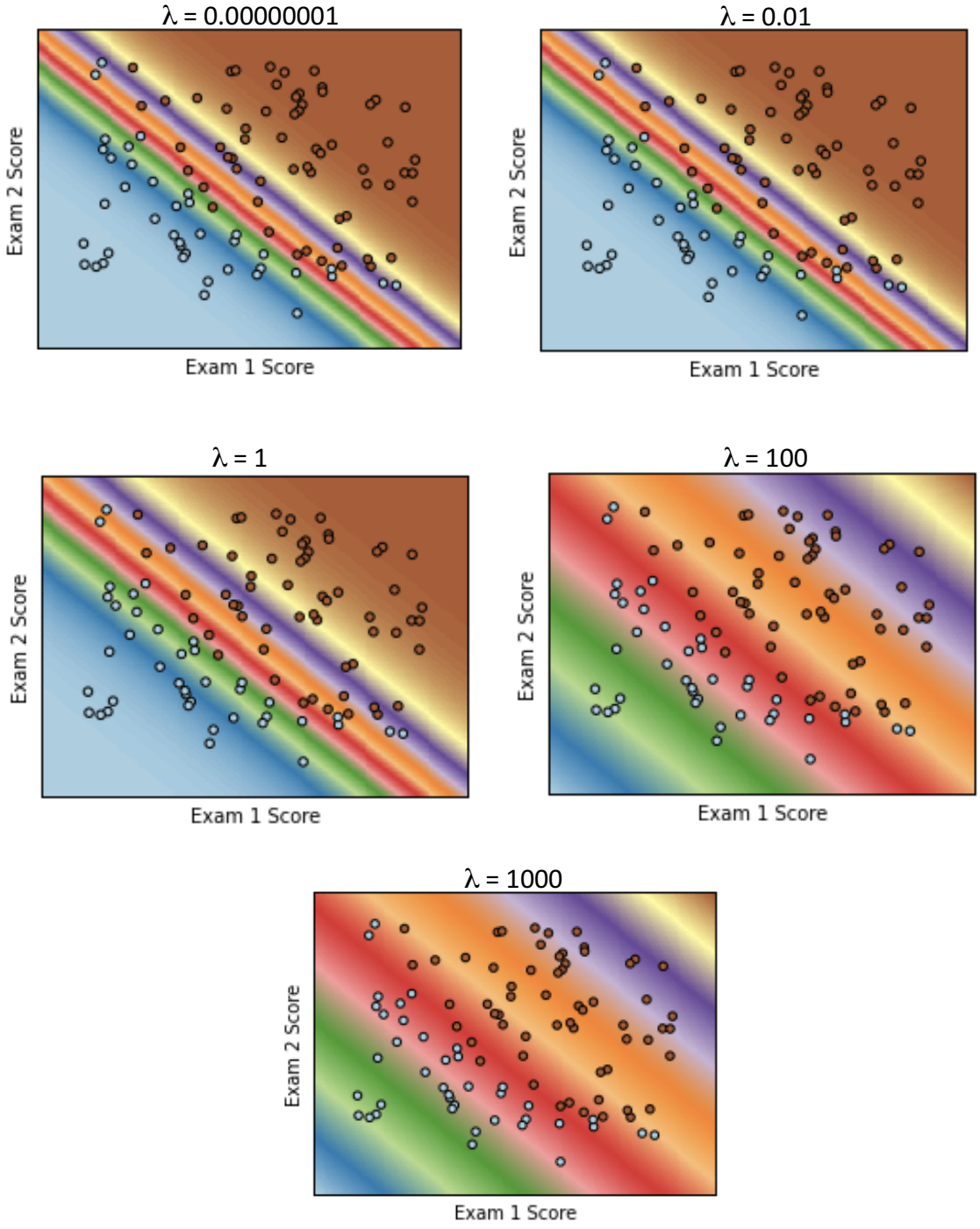








2 Logistic Regression

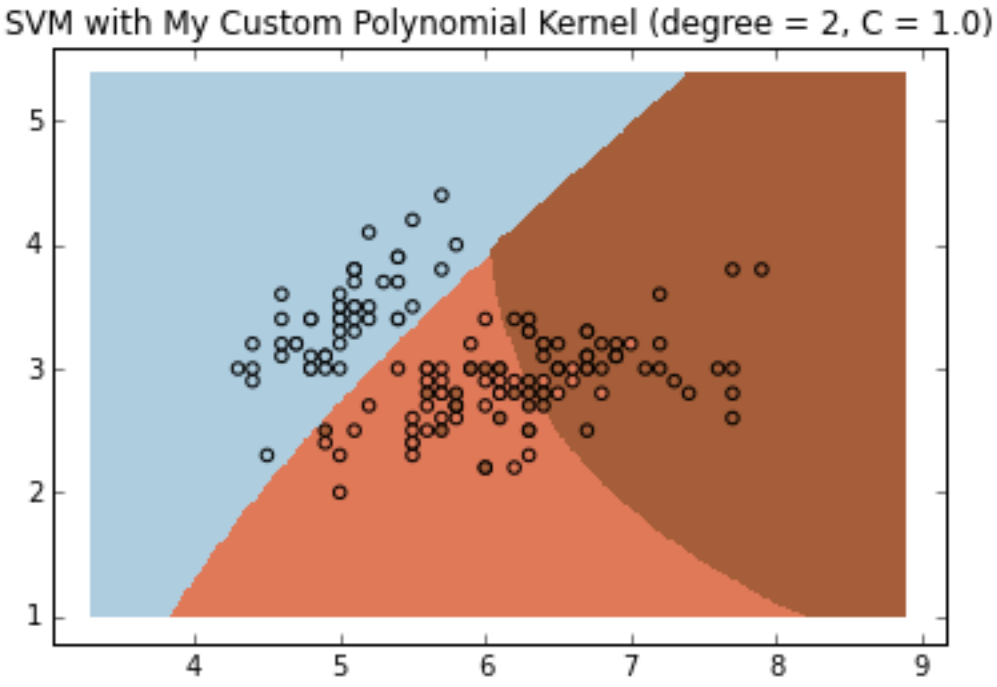


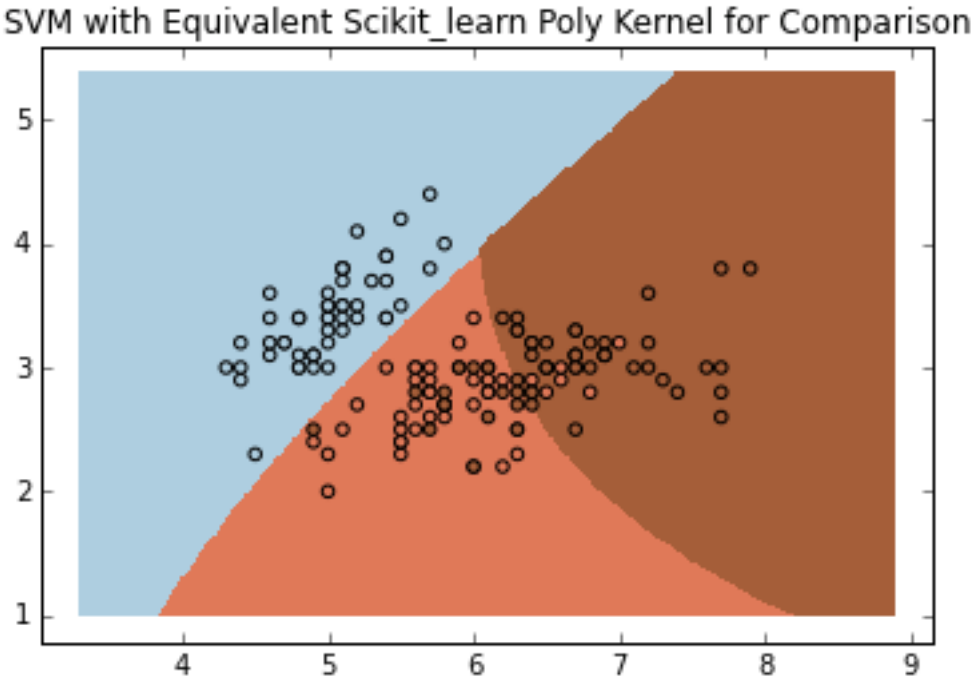
As the value of λ increases, the decision boundary (in red) becomes wider. Points near or on the decision are subject to scrutiny since the classifer may not have labeled them correctly. The regularization term λ serves to make sure the model does not over-fit,

but at the same time, as the $\lambda = 1000$ plot indicates, too high of λ can lead to over-generalization, which makes the model ineffective.

3 Support Vector Machines

3.3 Implementing the Polynomial Kernel





3.4 Implementing the Gaussian Radial Basis Function

For the polynomial kernel, an increasing C (being the inverse of lambda) has the effect of a decreasing lambda, so it may lead to overfitting. An increasing d also suggests overfitting.