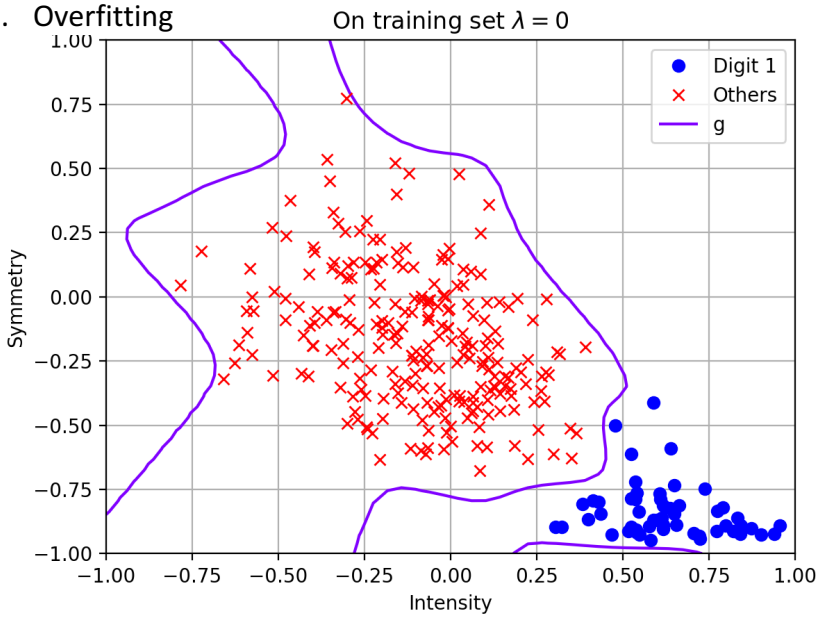
CSCI 4100 Assignment 9

Boliang Yang 661541863

1. 8^{th} order Feature Transform

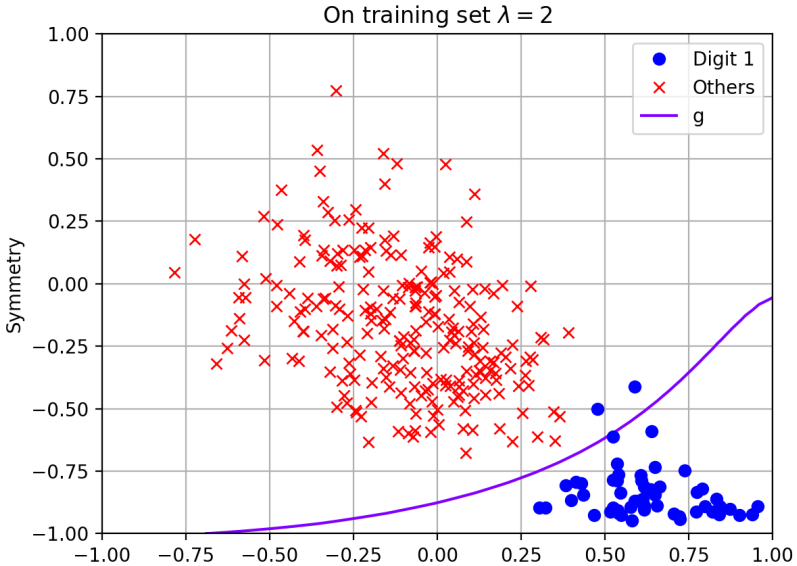
   We have 300 training points, and 45 dimensions for single point. Then, the dimension of Z is 300×45.

2. Overfitting



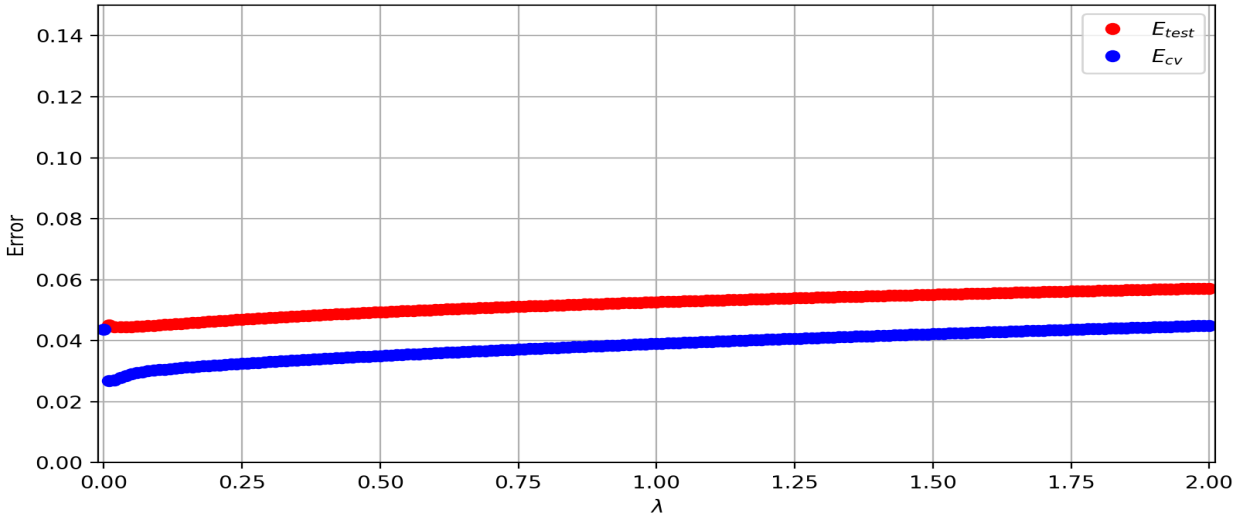
   I think there is overfitting. As we can see at the bottom of the right, there is some boundary lies on unnecessary part.

3. Regularization



   I think there is underfitting since there are some blue points outside of the boundary.

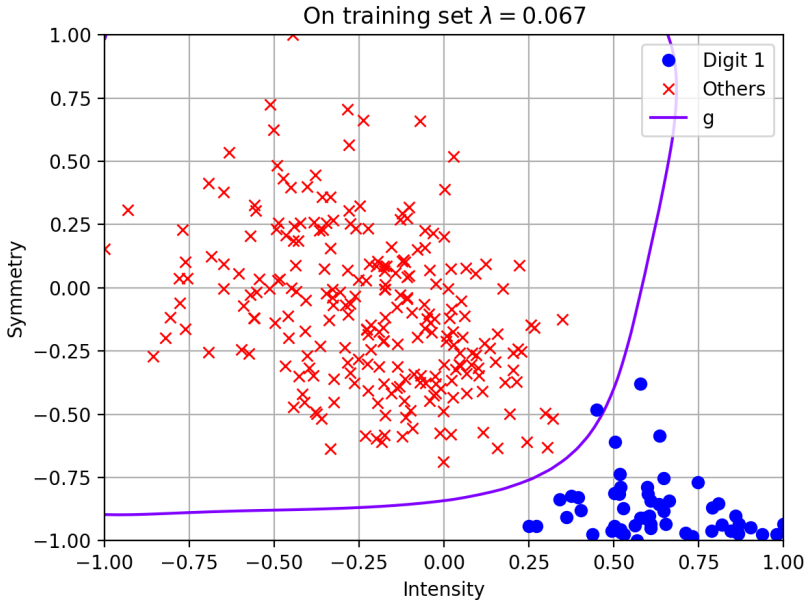
4. Cross Validation

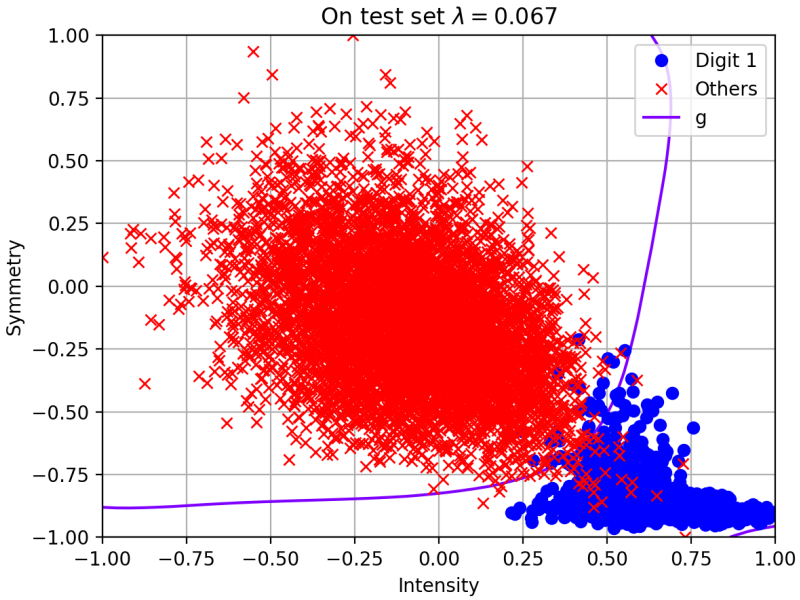


E_{cv} is always smaller than E_{test}, which is consistent with what the book says. Also, both errors first decrease and then increase slowly as λ gets larger. This is also consistent with the behavior of regularization: some amount of regularization is important to reduce overfitting, but as we increase regularization, it will underfit the data.

5. Pick λ

The minimum E_{cv} is found at $\lambda = 0.067$. The following are the plots of the decision boundary for the weights $\mathbf{w}_{reg}(\lambda^*)$.





6. Estimate E_{out}

   Since $E_{test}(\mathbf{w}_{reg}(\lambda^*)) = 0.057199,\ N = 8998,\ M = 1$, then we can estimate E_{out} by: $E_{out}(g) \leq E_{test}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} = 0.057199 + \sqrt{\frac{1}{2\times 8998}\ln\frac{2}{0.05}} = 0.071516$

7. Is E_{cv} biased?

   E_{cv} is biased. Since we choose the best λ^* based on the minimum of E_{cv}, this selection makes $E_{cv}(\lambda^*)$ a biased estimate of $E_{test}(\mathbf{w}_{reg}(\lambda^*))$.

8. Data snooping

   No, $E_{test}(\mathbf{w}_{reg}(\lambda^*))$ is not an unbiased estimate of $E_{out}(\mathbf{w}_{reg}(\lambda^*))$. The problem is in pre-processing: when normalization, we used both the training set and test set. Thus, the results would be in a way dependent on the test set.

   In order to fix the problem, we should remove the test set when doing the normalization on training set, then rescale the test set after we complete training.