Assignment 9

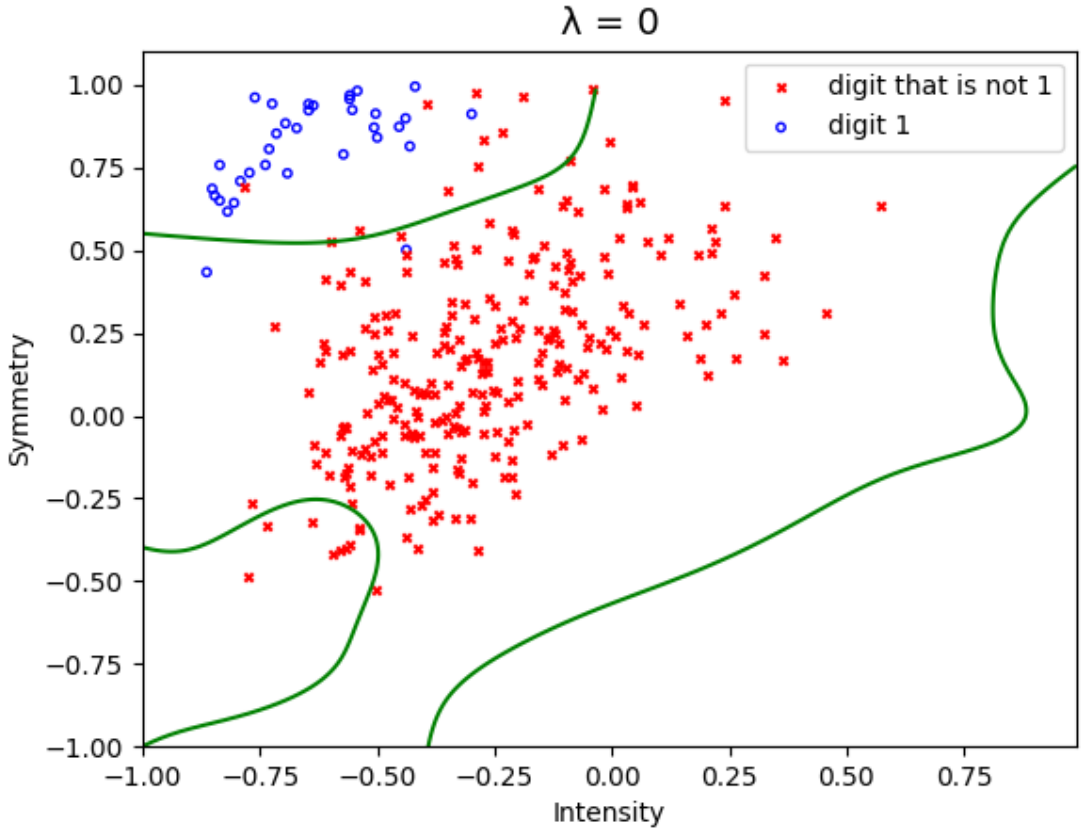
Alberto Mejia
RIN: 661514960

CSCI 4100 - Machine Learning from Data

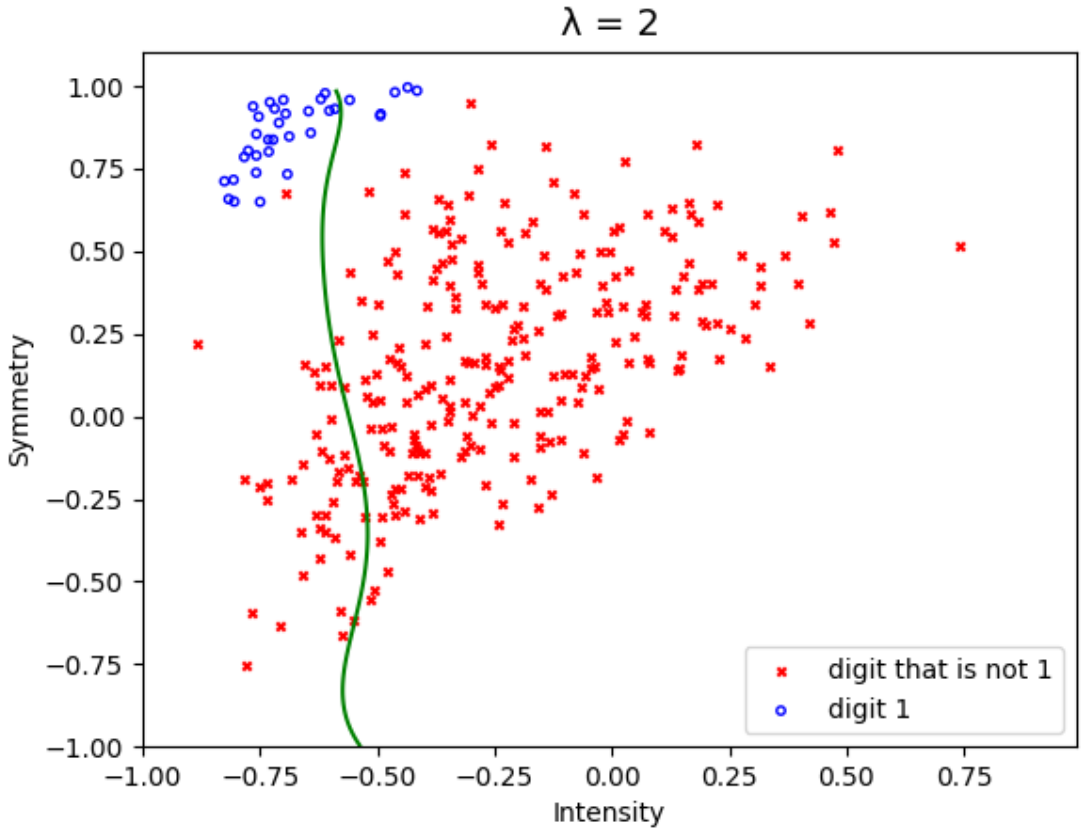
November 10, 2019

1. (100) 8^{th} order Feature Transform
The transformation for a k polynomial feature transformation is denoted by 1+2+...+(k+1). Thus, for an 8^{th} order Feature Transform, we have, 1+2+3+4+5+6+7+8+9 = 45 features. With 300 samples, the dimension of Z is $R^{300 \cdot 45}$

2. (100) Overfitting
When setting $\lambda = 0$, the algorithm overfits the data, many of the points are classified correctly but it is way too complex.

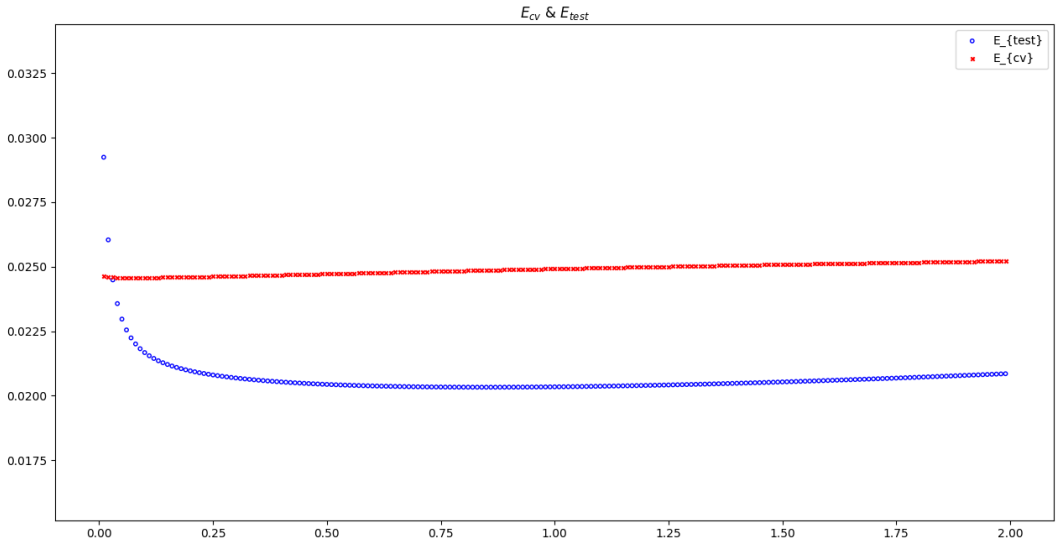


3. (100) Regularization
When setting $\lambda = 2$, the algorithm underfits the data, not many of the points are classified correctly and it is too simple.

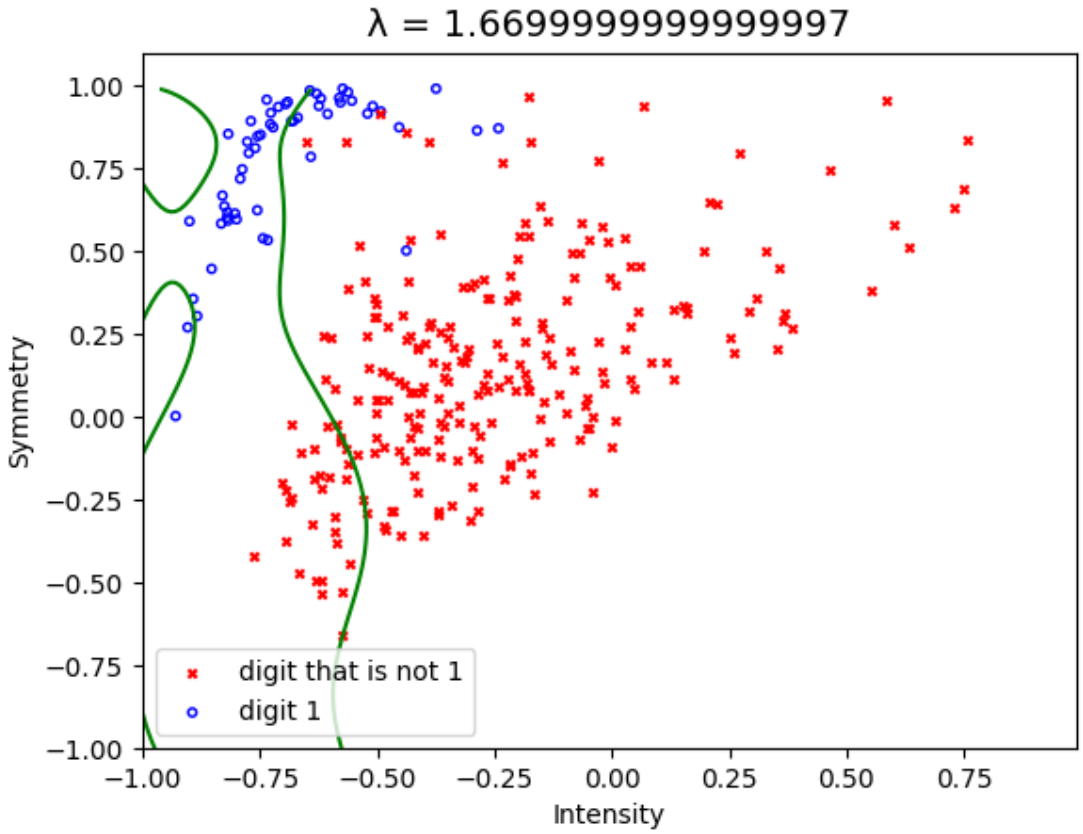


4. (100) Cross Validation

E_{cv} is always smaller than E_{test} because when E_{cv} is at its minimum, E_{test} is not at its minimum but close to it. This is because E_{cv} follows E_{test} very closely, giving it the same trend as E_{test}. So, any change in E_{test} is reflected by E_{cv}.



5. (100) Pick λ



Result: $\lambda = 1.669$

6. (100) Estimate E_{out}

Generalization Error Formula:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} ln(\frac{2M}{\delta})}$$

N = 8998 (Number of data points)
M = 1
$\delta = 0.05$

$$E_{out}(g) \leq E_{test}(g) + \sqrt{\frac{1}{2(8998)} ln(\frac{2(1)}{0.05})}$$

$= 0.0142 + 0.01431$
$= 0.0285$

7. (100) Is E_{cv} biased?

As we covered in question 4, E_{cv} is an estimate for E_{out} based on N-1 points and E_{cv} is obtained from these N-1 data point in training. $E_{cv}(\lambda^*)$ is biased as λ^* is chosen based on E_{cv}. In simpler terms, we selected λ^* based off of our results from our test set. In conclusion, we have snooped on our data, making E_{cv} a biased estimate of E_{test}.

8. (100) Data snooping

$E_{test}(w_{reg}(\lambda^*))$ is a biased estimate of $E_{out}(w_{reg}(\lambda^*))$ as it was selected randomly chosen from all data; it was chosen from a dataset that was split into a train and test set using all datapoints. The issue is that the data is normalized before we split into the training set and testing set. This would affect every datapoint in our set. And given that we chose λ^* from one of these two groups, we have data snooped. To fix this we can either use a different dataset for validation, or do the data normalization after we split into training and testing.