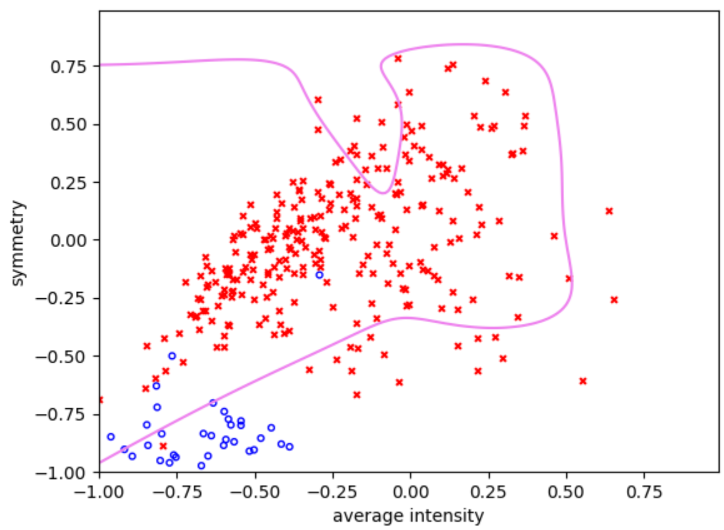
Machine Learning from Data CSCI 4100
Assignment 8
Jae Park (RIN: 661994900)

1. 8th order Feature Transform.

$dim(Z) = 300 \times (1 + 2 + \cdots + 8) = 300 \times 45.$

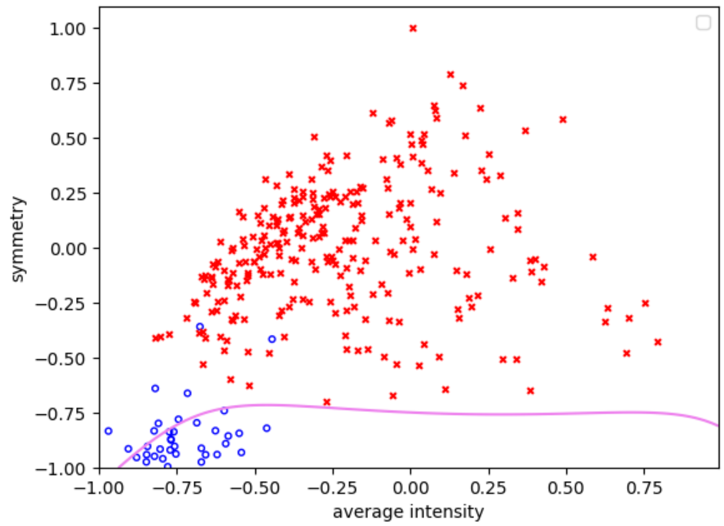
2. Overfitting.

Linear regression without any regularization ($\lambda = 0$).



Based on the figure, there seems to be an overfitting to the data set.

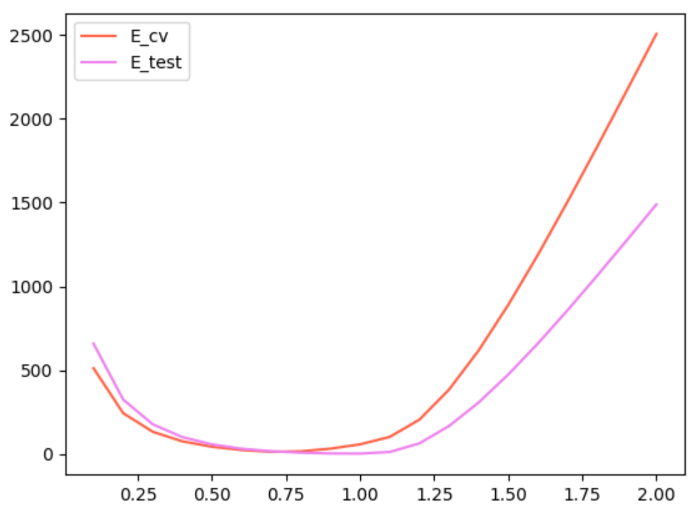
3. Regularization.
Linear regression with regularization ($\lambda = 2$).



Based on the figure, it did a moderate job but still underfitted to the data set.

4. Cross Validation.

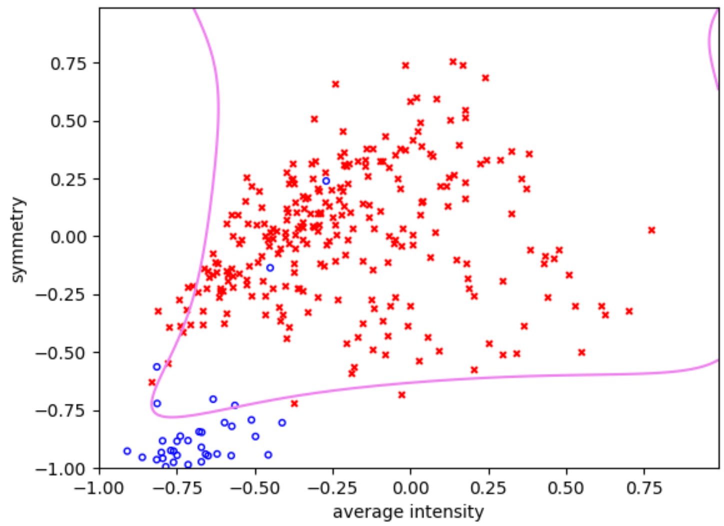
With $\lambda \in \{0, 0.01, 0.02, \dots, 2\}$, the cross validation error and the test set regression error are plotted against each other.



E_{cv} seems to estimate E_{test}, and they intersect at the point where E_{cv} has a minimum.

5. Pick λ.

The best value of λ, λ^*, is found to be 0.7. Plot of the decision boundary for the weights $\mathbf{w}_{\text{reg}}(\lambda^*)$ is the below



6. Estimation of the Classification Error.

There are 300 points in the train data and 1707 points in the test data. 54 of the test data are misclassified points, so $E_{test} = 0.031634$.

$$
\begin{aligned}
E_{out} &\leq E_{test} + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \\
&\leq \frac{54}{1707} + \sqrt{\frac{1}{2 \times 1707} \ln \frac{2 \times 1}{0.01}} \\
&\leq 0.071029\ldots
\end{aligned}
$$

$\therefore$ estimate for E_{out} is 0.071029.

7. Is E_{cv} biased?

No, E_{cv} is not an unbiased estimate of E_{test}. Although E_{test} is estimated by applying a hypothesis on just the test set, the λ^* itself is given from choosing the best of λ based on E_{cv}. This gives us a biased estimation of E_{test}.

8. Data Snooping.

No, E_{test} is not an unbiased estimate of E_{out}. E_{test} is estimated using λ^* which is given from the training data set. This training set, however, is not independent from the testing data set because we first combined them, normalized, and randomly chose from the mix. The test set is affected by the partial shift and scale of the training set. To prevent data snooping, we must first separate the training set and test set. Also when separating the training set, we should use a really good random algorithm. After obtaining the two independent sets, first normalize the training set. With the shift and scale values from this, we normalize the test set. We could throw away data points that are off the range [-1, 1].