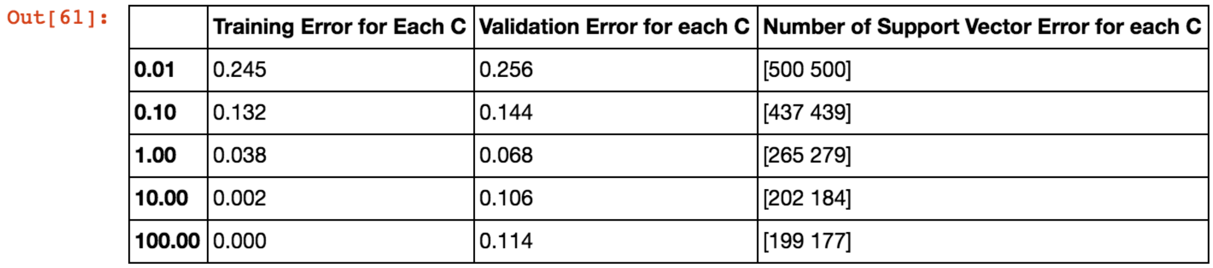
Support Vector Machine Detector of Generated Domain Names
Statistical Machine Learning
William Galindez Arias

Assignment 1

Implementation and training of the SVM classifier was done using suggested library through the import of Scikit module. The imports, training, validation and test stages of the project can be seen in the attached Python notebook in different formats (.py, ipnb and html).
The below table shows the results obtained for different values of the regularization **C** constant.



Out[61]:

	Training Error for Each C	Validation Error for each C	Number of Support Vector Error for each C
0.01	0.245	0.256	[500 500]
0.10	0.132	0.144	[437 439]
1.00	0.038	0.068	[265 279]
10.00	0.002	0.106	[202 184]
100.00	0.000	0.114	[199 177]

fig. Error values for different regularization constant and number of support vectors

Using the obtained values, is possible to find out that the lowest Validation error is for the constant regularization C **= 1**, where the error is **0.068.** The behavior of training error and validation error can be seen graphically in fig2.

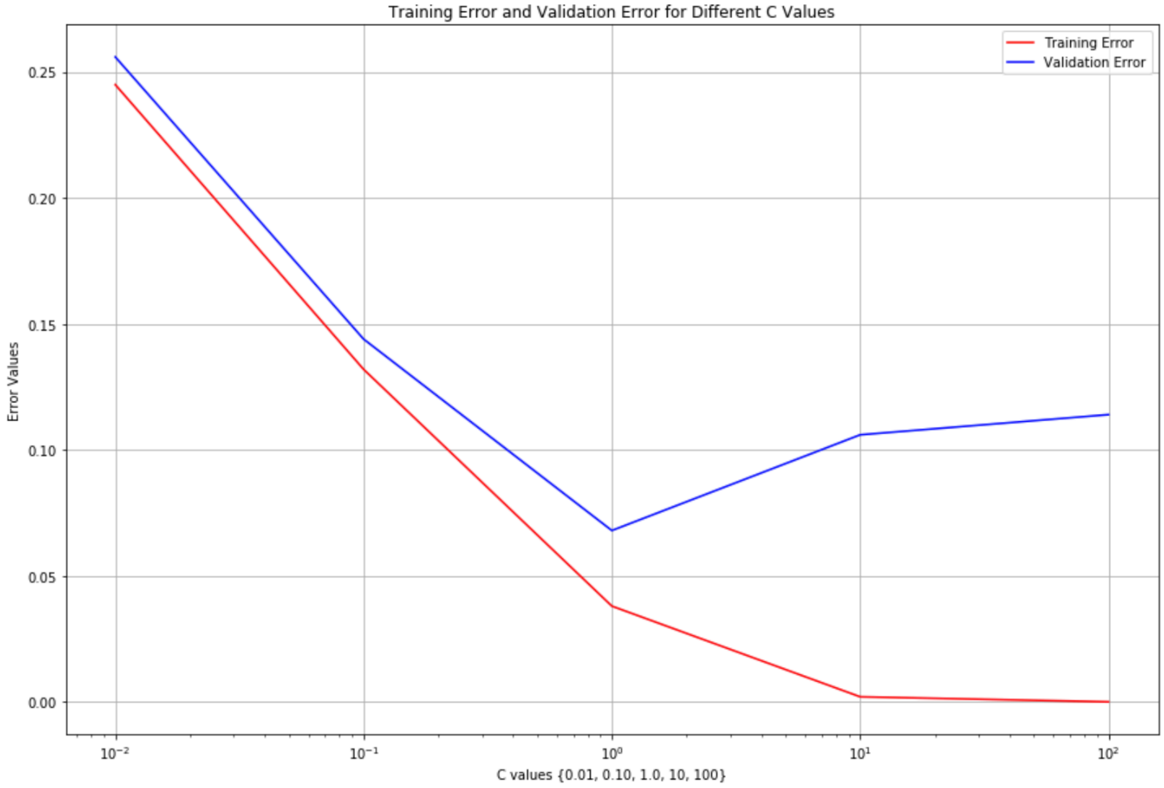


fig2. Training and validation error

Assignment 2

True error classification

By getting the lower validation error with the C = 1. And using the Test Data with this C=1 constant set. Rs(tst) error obtained is 0.0785. Using the Hoeffding inequality concepts. The ε can be found by solving:

$$epsilon = |b - a|\sqrt{\log(2) - \log(1 - conf)/2l}$$

Where absolute value of b-a = 1 then "vanishes", *l* is the number of samples in the test data set, and using the values provided of δ=0.99, the equation above gives:

ε≈ 0.036394

All the values obtained above can be seen as a report too in the Jupyter notebook provided with this report.