Final Project: Study on Error bounds for Hard-margin Support Vector Machine

Aigerim Kalizhanova, Olzhas Shortanbaiuly

MATH 540 Statistical Learning. Fall 2022

Abstract

This project analyses the performance of Hard-margin SVM both on synthetic and real data, by finding and comparing training and test errors of the classifier output to the theoretical error bound. The validity of the theoretical error bound was proved based on this model.

1. Introduction

Support Vector Machine (SVM) is a supervised learning model, which aims to find a (d-1)-dimensional hyperplane in d-dimensional space that best classifies given data points. Data points might be linearly separable or not. In the latter case, the Kernel function is used to transform the data. Taking linearly separable data, the hyperplane "divides" the data set into two classes, each half-space representing one class of data points. The objective is to learn a set of parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ which satisfy the following constraints:

$$\forall i,\ y_i(w^T x_i - b) \geq 0$$

The hyperplane is of the form

$$H = x : w^T x - b = 0$$

There might be several hyperplanes, which perfectly classify the training data. However, some of them might misclassify the test data points lying very close to them. Hence there is a Hard-margin SVM which does not allow it, requiring the linear separability of data. Its aim is to find the unique hyperplane that maximizes the margin γ such that the violation of the boundary is impossible. Here, the margin is the minimum distance from the hyperplane to the training data point. There is at least one point on one side and one on the other with equal margins γ, which are called Support vectors. The optimization problem for Hard-margin SVM is

$$\min_{w,b}\ \frac{1}{2}\|w\|_2^2$$
$$\text{s.t. } y_i(w^T x_i - b) \geq 1\ \ \forall i$$

In this project, Hard-margin SVM on the 2-dimensional data set and the error bound of its outputs are of interest.

Algorithm 1 Hard-SVM on noisy data

Input: m_vals,c_vals
for $i = 1$ **to** 9 **do**
  for $j = 1$ **to** 9 **do**
    X_train, X_test, y_train, y_test = generate_data (m_vals[i], c_vals[j])
    clf = SVC (kernel = 'linear')
    clf = clf.fit(train_data)
    $\gamma[i][j] = \frac{1}{\|clf.coef\|}$
  end for
end for

The **margin γ - error bound** and **sample size m - error bound** relationships are to be studied. The model is to be trained on the data generated randomly using Gaussian distribution and tested on the real data sample from **Red Wine Quality Dataset**. Section 2 describes the generated data, and model training algorithm and presents the theory on the error bound. Section 3 discusses the results of error analysis and tests on real data.

2. Methodology

For this project, a linearly separable, normally distributed synthetic data of the form $X \times \{0, 1\}$ was generated using *make_blobs* function of *scikit-learn* library, and assigning various sample sizes and centers of two clusters of data. To satisfy the distribution condition for Y, 0 values were changed to -1. *cluster_std* (standard deviation of the distribution of clusters) parameter of *make_blobs* was arbitrarily picked as $\frac{3c}{4}$ in order to get reasonable margin γ values.

2.1. Model Training

In order to train and test the model, *train_test_split* and *SVC* functions from *scikit-learn* library were used. Sample data set were split into *train* and *test* samples with ratios $\frac{2}{3}$ and $\frac{1}{3}$ respectively. For varying sample size m and centers c, which also vary the margin γ, the classifier was trained and tested by algorithm 1. It was written supposing that appropriate libraries were already imported.

Here, m_vals represents an array of possible

sample sizes, which was artificially defined: $[10, 50, 100, 500, 1000, 5000, 10000, 50000]$; and c_vals is an array of various centers for data generation: $[2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.59, 9.5, 10, 10.5, 11, 11.5, 12]$. A function *generate_data(m_vals, c_vals)* was defined out of the nested loop presented in algorithm 1, and generates data based on passed arguments, splits the data into train/test samples and returns them.

2.2. Theory on the Error Bound

In order to evaluate the performance of a classifier, the following theorem on the error bounds was used (1). It was assumed that if the data is separable with margin γ, then the complexity of the data is upper-bounded by a function of $\frac{1}{\gamma^2}$. In addition, it is possible to show that the complexity might still stay low with a high-dimensional sample while the above condition is satisfied.

Theorem 1 *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (γ, ρ)-separability with margin assumption using a homogeneous halfspace. Then, with the probability of at least $1 - \delta$ over the choice of a training set of size m, the $0 - 1$ error of the output of Hard-SVM is at most*

$$\sqrt{\frac{4(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

According to the theorem, we decided that the probability of at least $1 - \delta$ and ρ should be fixed while varying sample size m and margin γ. For this project we fixed $\delta = 0.2$, $\rho = max\{\|X_i\|\}$ (for $i = 1, 2, ..., |m\_vals| \times |c\_vals|$) individually for each data set, such that $\|x\| \leq \rho$. In addition, according to the requirements stated in the theorem, we artificially changed labels y from $\{0, 1\}$ to $\{\pm 1\}$.

The values of different sample sizes, margins, ρ's, training, and test errors were stored into a matrix called "value_matrix" in the .ipynb file for further calculations of error bounds and comparison.

3. Discussion

In this section we examine the performance of the classifier in terms of its complexity, given synthetic data. Following this the classifier is evaluated on the real data on wine quality taken from **UCI Machine Learning Repository**. (2)

3.1. Error Analysis

To examine the relationships **margin γ - error bound** and **sample size m - error bound**, the values stored in the "value_matrix" were used. To illustrate, the following is a table of values for fixed $m = 1000$.

Table 1. Values of a matrix for $m = 1000$

MARGIN γ	ρ	TRAINING ERROR	TEST ERROR
0.84	7.98	0.0254	0.0303
0.79	9.12	0.0313	0.0121
1.06	10.41	0.0254	0.0333
1.18	13.51	0.0269	0.0333
1.35	15.52	0.0433	0.0212
1.12	16.77	0.0239	0.0394
1.60	18.93	0.0313	0.0152
2.35	21.78	0.0313	0.0485
1.69	23.49	0.0289	0.0515
2.68	28.11	0.0373	0.0333
2.44	25.63	0.0239	0.0303
2.73	27.99	0.0298	0.0333
2.59	32.39	0.0209	0.0242
3.10	31.79	0.0343	0.0394
2.84	34.09	0.0269	0.0394
2.98	38.04	0.0288	0.0182
2.90	40.28	0.0298	0.0212
2.99	41.64	0.0284	0.0182
2.38	46.49	0.0224	0.0424
4.31	43.91	0.0343	0.0212
3.61	47.22	0.0224	0.0242

3.1.1. ERROR - MARGIN

To evaluate the dependence of training and test errors on the margin γ and see if the error does not exceed the theoretical bound, the values of errors of the model outcome versus the margin γ were plotted. Mentioned relationship was checked for all possible values of m, by fixing each of them one by one. Here the result with the lowest error bound was presented in Figure 1 with $m = 50000$. The theoretical bound was calculated using the formula given in Theorem 1.

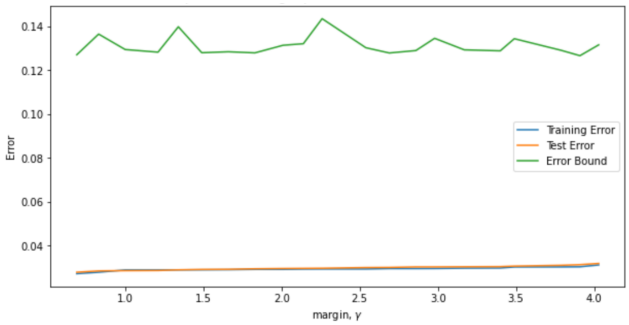

Figure 1. Relationship between margin γ and error (with theoretical error bound)

As can be observed from Figure 1, the graph of theoretical error bound (green line) versus γ stays above the lines describing training error (blue line) and test error (yellow line). Hence, the 0-1 error of the output of Hard-SVM does not

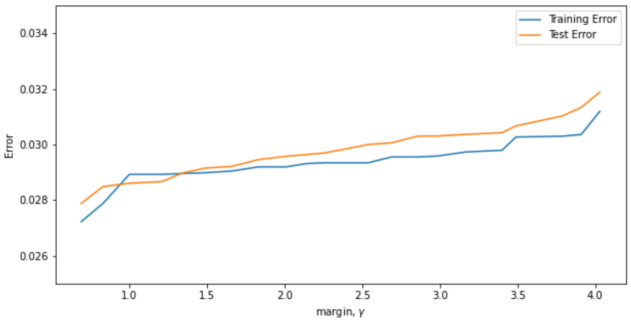

Figure 2. Relationship between margin γ and error (without theoretical error bound)

exceed the bound. The error bound fluctuates with an insignificant upward trend as the value of γ increases. Figure 2 illustrates the scaled plot of training and test error versus the margin. It can be concluded that the error increases with γ.

3.1.2. Error - sample size

Next, to observe the relationship between errors and sample size m, we plotted the values of errors after training and testing the classifier versus sample sizes m. In addition, the dependence of an error bound on m is plotted on the same coordinate plane. Since the margin γ also affects the error bound, according to Theorem 1, we fixed $\gamma = 3$. As we randomly generated values, it was an attempt to find an optimal one for each sample size m.

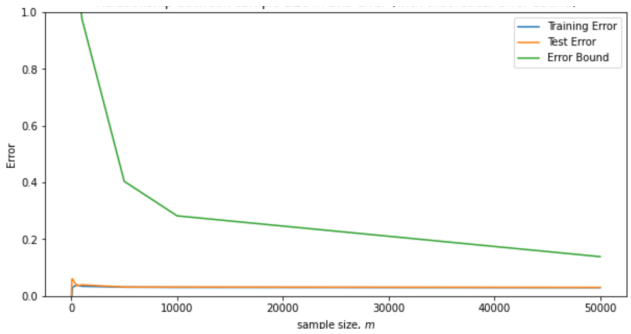

Figure 3. Relationship between sample size m and error (with theoretical error bound)

As it can be observed from Figure 3 the graph of error bound versus sample size m stays above the lines describing training error and test error. Hence, the 0-1 error of the output of Hard-SVM does not exceed the bound. Moreover, according to the theorem, the error bound decreases as m increases. To be more precise, the relationship Error bound $\propto \frac{1}{m}$ was verified.

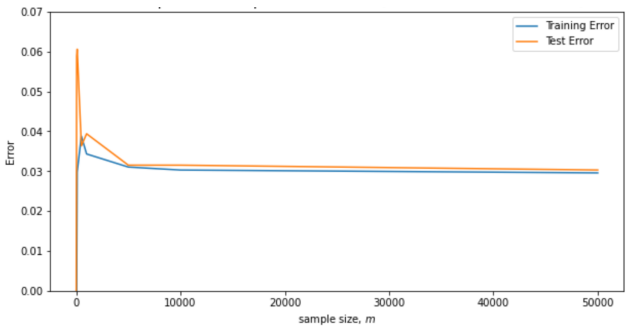

Figure 4. Relationship between sample size m and error (without theoretical error bound)

Figure 4 is a scaled portion of Figure 3, which clearly illustrates the downward trend of training and test errors depending on the increase in m. The jump at $m = 0$ is reasoned by the relationship given above: Error bound $\propto \frac{1}{m}$. Since $m \to 0$, the error goes to $-\infty$, returning to normal as it gets non-zero values of m.

3.2. Evaluation on Real Data

The model was tested on the real data set obtained from **some repository**, and the same methods as in section 3.1 were applied to evaluate the outcome of the Hard-SVM.

Checking dependencies between margins and errors, sample sizes and errors, similar relationships as from synthetic data can be observed.

3.2.1. Error - margin

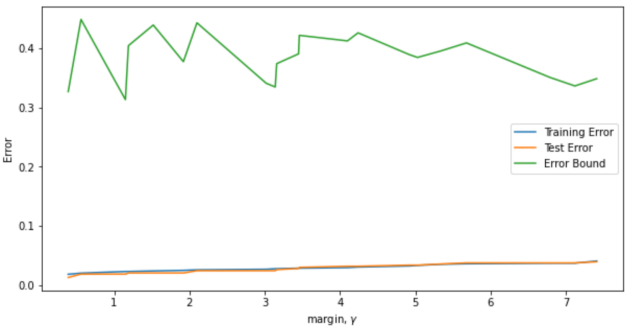

Figure 5. Relationship between margin γ and error (with theoretical error bound)

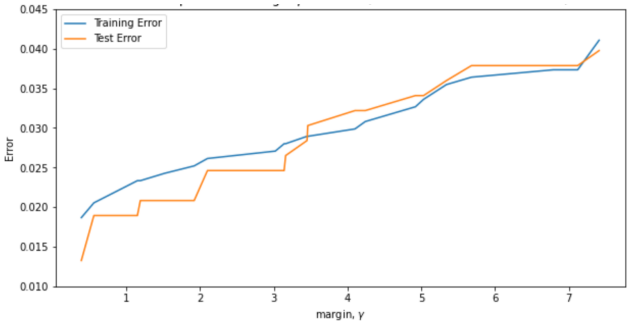

Figure 6. Relationship between margin γ and error (without theoretical error bound)

3.2.2. Error - sample size

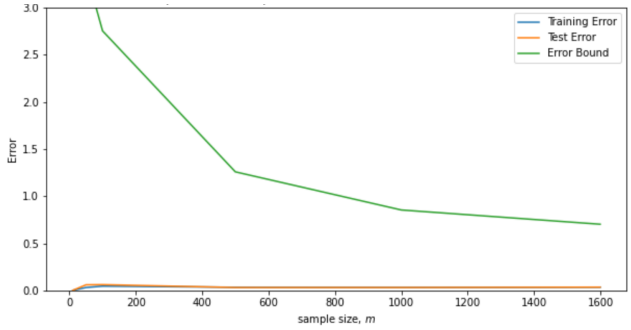

Figure 7. Relationship between sample size m and error (with theoretical error bound)

4. Conclusion

In conclusion, we trained a Hard-Margin SVM using synthetic linearly separable data generated from Gaussian distribution. After this, we evaluated the performance of the model and verified the theoretical error bound by varying the sample size and the centers, which led to changes in margins for each sample of size m. The sample data fulfilled the requirements needed to find the theoretical error bound of a 0-1 probability of the output of the classifier. As a result of the analysis, it was verified that the train and test errors are indeed bounded above by the theoretical error bound, and the errors increase with increasing margin, and decreasing factor $\frac{1}{m}$, where m is a sample size. The model behaves similarly when tested on the real data set.

The weakest point of our work was obtaining the relationship between error and margin γ, which should give us completely opposite relationship, it is the main point to improve on.

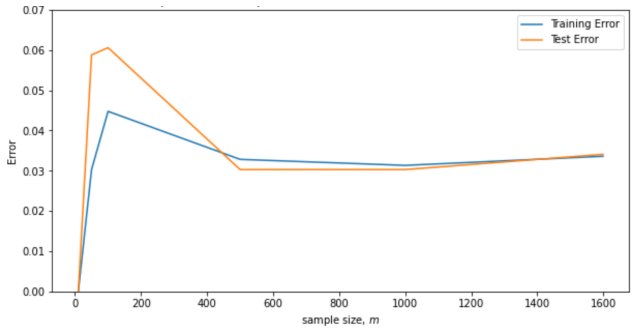

Figure 8. Relationship between sample size m and error (with theoretical error bound)

Due to the high run-time cost, we fixed the variance of the sample distribution. However, later it would be interesting to see how margins act under different variances. During the experiments, we found that the variance which is greater than c (centers) affects the linear separability of the data, by not letting us apply Hard-SVM. In addition, it is suggested to try even larger sample sizes and test the model on larger real data sets, since the real-world data are significant.

References

[1] Shalev-Shwartz, S. & Ben-David, S. Understanding machine learning: From theory to algorithms. Cambridge University Press, 2014.

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.