Analyzing Logistic Regression Complexity Bound Under PAC Framework

Aigerim Kushenova
Temirlan Kaiyrbekov

Abstract

In this work, logistic regression was analysed under PAC learning framework. The main goal of the project was to verify the Fundamental Theorem of Statistical Learning by examining errors of the model with respect to changing VC-dimension. By considering two cases with different number of datapoints, we managed to show how the distribution of error expands with the increasing d, which is VC-dimension. Corresponding code of with all the computations is available through the link: https://github.com/Memirlan/log-reg-complexity-bound

1. Introduction

Logistic regression is a discriminative model in supervised learning,and it works by drawing boundaries between data points in order to classify them. This characteristic allows to state the Fundamental Theorem of Statistical Learning for Binary Logistic Regression:

Let H be a hypothesis class of functions from a domain X to 0,1 and let the loss function be the 01 loss. Assume that $VCdim(H) = d < \infty$. Then, there are absolute constants C_1,C_2 such that:

$$C_1 \frac{d + log(1/\delta)}{\epsilon} \leq m_H(\epsilon, \delta) \leq C_2 \frac{dlog(1/\delta) + log(1/\delta)}{\epsilon}$$

The main idea behind this theorem is that learning model should posses finite flexibility in order to be able to actually learn. A separate notion for flexibility is VC dimension (d here), representing the maximum number of examples that a hypothesis h can "shatter". For binary logistic regression it turned out to be 3, since it is a linear model. VC dimension of any linear model is 3. The given theorem is formulated with respect to sample complexity m_H, which is the number of training samples that algorithm needs to learn the target function. Knowing this can help us in formulating PAC learning:

$Pr_D[L_D(A(s)) \leq \epsilon] > 1 - \delta$ when $m \leq m_H$

where ϵ and δ are accuracy and confidence parameters. m is the number of samples we test and m_H is a sample complexity. We can interpret it by stating that the probability that our test error is less than or equal to ϵ is greater that 1-δ if our sample size will be greater that the sample complexity. Sample complexity represents the number of samples that should be enough to train the model successfully.

2. Methodology

For applying PAC to Logistic regression setting, let us state the following:

$L_D = Pr[h(x) \neq y]$

$Pr[Pr[h(x) \neq y]] < \epsilon < C_2 \, (dlog(1/\epsilon) + \log(1/\delta))\frac{}{m} \leq C_3 \, (d + log(1/\delta))\frac{}{m}$

2.1. Complexity

$Pr[h(x) \neq y]$ is test error is this context and so we can approximate the complexity of our algorithm under PAC framework: $\dfrac{(Pr[h(x_i) \neq y_i] * m)}{d} \propto O(1)$

Considering the written above, we expect out complexity to be linear. And that if we increase m, d should decrease respectively, and vice versa.

2.2. Implementation

Theoretical complexity bound will be analyzed through the use of synthetically generated data. Different datasets will be generated based on Multivariate Gaussian distributions with different number of features d. Thus, the VC-dimension for each such dataset is going to be d+1. For each dataset, the average log-loss will be computed, since with different d, datasets will have different numbers of data entries. Finally, log-loss versus VC-dimension will be plotted to visualize findings, to notice whether there is a linear dependence and to find other insights from data.

Experimental Results

Test losses were computed for each dataset n by d dimensions where d [4, 200]— d mod 2 = 0 This procedure was held for datasets with n = 5000, and n = 15000 separately.

There is a clear linear dependence of test log-loss error from VC-dimension, where VC-dimension = d+1. Therefore, the claim that the ratio of loss over vc-dimension is of O(1) complexity was empirically proved. Interestingly, there is also a linear dependency of loss vs accuracy, since error and accuracy are inversely proportional to each other, but the latter was added more to show the performance of models.

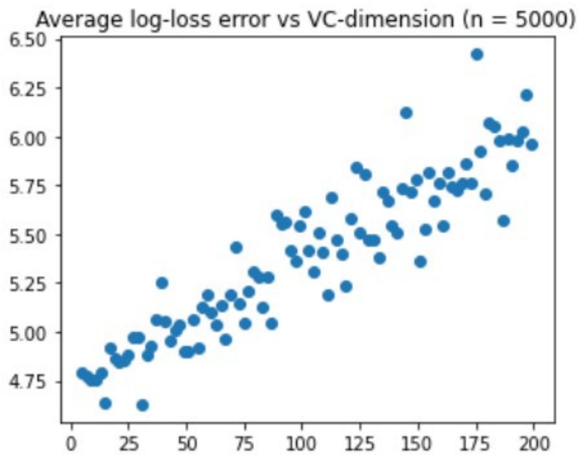

Figure 1. Figure 1. Average log-loss error vs VC-dimension (n = 5000)

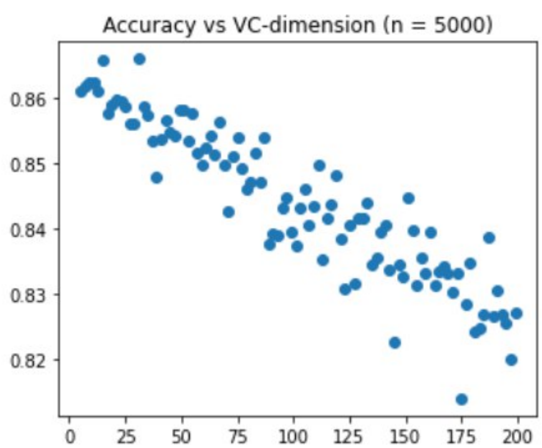

Figure 2. Figure 2. Accuracy vs VC-dimension (n = 5000)

Additionally, there is a pattern that datapoint in the Figure 1 plot: increase in variance as d increases. The line of dependency of loss on d_{VC} seems to be lower and upper-bounded by linear functions with gradients proportional to C_1 and C_2 respectively, where C_1 and C_2 are coefficients from Theorem 6.8, item 3 (Shalev-Shwartz, 2014).

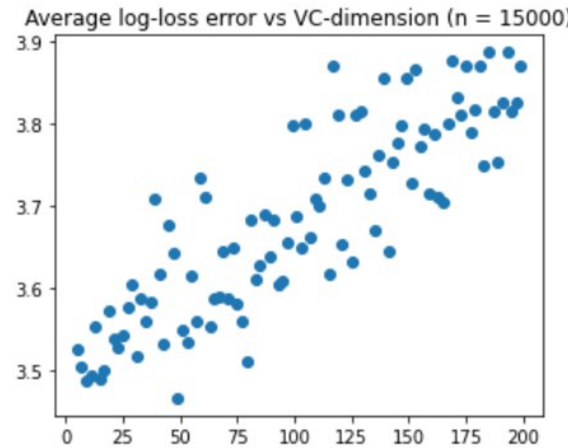

Figure 3. Figure 3. Average log-loss error vs VC-dimension (n = 15000)

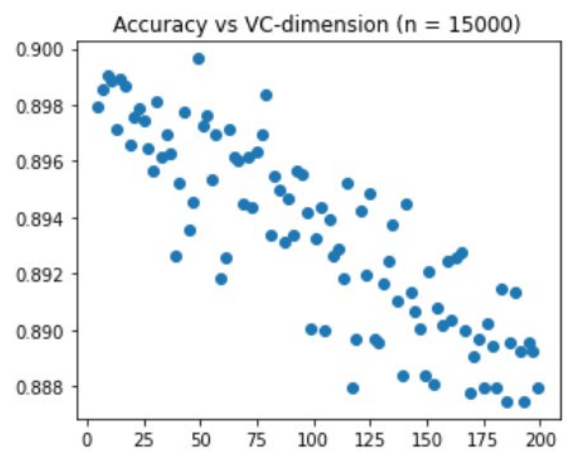

Figure 4. Figure 4. Accuracy vs VC-dimension (n = 15000)

Discussion and Conclusion

There are possible improvements that could be made such as experimenting with 1) number of data points for each dataset, 2) parameter "class_sep" that can make the classification task easier or harder, 3) "test_size" in train-test-split, since for different sizes of datasets, there can be different optimal split ratio.

Contributions

Theoretical analysis, editing by Kushenova Aigerim
Application of theory, visualization, discussion by Temirlan Kaiyrbekov
Methodology was done together

References

Shalev-Shwartz, S. B.-D. Understanding machine learning: From theory to algorithms. pp. 72–74, Cambridge, UK, 2014. Cambridge University Press.