Classification Accuracy of SVMs, NNs, and LogReg

Brian Liu

December 2024

Abstract

In "An Empirical Comparison of Supervised Learning Algorithm" (Caruana and Niculescu-Mizil 2006), the authors performed a comprehensive empirical evaluation of ten supervised learning models with the goal of demonstrating the varying capabilities of each method, both new and old. Considering an extensive range of hyperparameters, datasets, and other relevant factors, the authors trained a significant variety of each model and reported on the average performance of each model on various metrics and datasets. Nearly 20 years later, supervised learning has taken great strides forward. In this study, I am seeking to reevaluate three of those supervised learning models to observe their modern-day performance. Namely, I will leverage modernized versions of Support Vector Machines, Neural Nets, and Logistic Regression. For consistency of results, I will similarly report average performance across various datasets and training/test partitions. However, for the purposes of this study, I will only be focusing on the classification accuracy metric.

1 Introduction

In the 18 years since Caruana and Niculescu-Mizil 2006, supervised learning algorithms have come a long way. Significant research has been poured into seemingly simple models, resulting in further optimized and fine-tuned versions of these models. This includes our selected models Support Vector Machines (SVMs), Neural Nets (NNs), and Logistic Regression (LogReg), all of which are regularly employed today for both personal and professional use through a plethora of open-source distributions.

My goal with this study is to explore the performance of these models on a variety of datasets and hyperparameters, specifically as it relates to classification accuracy, a vital metric in today's extensive machine learning adoption. Furthermore, I seek to expose any improvements or deteriorations in the performance of these models in the time that has elapsed since 2006, both in terms of literal performance (classification accuracy) and in terms of performance as compared to one another.

2 Methodology

2.1 Models

Like Caruana and Niculescu-Mizil, I will seek to thoroughly explore the parameter space and common variations of each model. Specifically, for SVMs we will consider various values of C (0.1, 10), various kernels (linear, rbf), and various values of gamma (scale, auto). For NNs, we will consider various hidden layer sizes ((50,), (100,), (50, 50), (100, 50)). Lastly, for LogReg we will consider various values of C (0.1, 10) as well as penalty types (l1, l2).

2.2 Performance Metric

As described above, I will only be focusing on classification accuracy, which is calculated as the proportion of model predictions that is correct. For example, if the model makes 100 predictions, of which 60 align with the true value, the classification accuracy of the model on those predictions is 0.6.

2.3 Training and Testing

To ensure the reproducibility of my results and minimize the effects of outliers, I will report the average classification accuracy across 3 trials per model. Furthermore, I will perform each triplet of trials on three different training/test partitions (20/80, 50/50, 80/20) for each of 3 datasets.

Throughout the training of each model, I will perform 3-fold cross validation to select optimal hyperparameters. Classification accuracy will then be calculated and reported by comparing the model's predictions on a test set to the true values.

2.4 Datasets

Each model will be evaluated across 3 diverse datasets, namely ABALONE, ADULT, and BANKMARKETING from the UCI ML repository (in order: Sellers, Talbot, Cawthorn, and Ford 1994, Becker and Kohavi 1996, and Moro, Rita, and Cortez 2014). For ABALONE, which has 4177 entries and 7 features, I will be utilizing features like sex, length, and more to predict whether an abalone's age is greater or less than or equal to the median abalone age. ABALONE is a balanced dataset (similar number of +1 (greater than median age) and -1 (less than or equal than median age) classifications). ADULT is a large dataset, with 48842 entries and 13 features. Features such as age and education will be employed to predict whether someone's yearly income is either greater than or less than or equal to $50k$. ADULT is also a balanced dataset. Lastly, BANKMARKETING, another large dataset, has 41188 entries and 20 features. I will use features like age and job to predict the success of a bank's marketing campaign, measured by whether they subscribed (+1) or not (-1). BANKMARKETING is an unbalanced dataset, with the majority of entires naturally corresponding to not subscribing (-1). All of these datasets contain categorical features, which will be one-hot encoded to facilitate training by my models.

3 Performance

The results of these trials is reported in the tables below. Let ACA stand for Average Classification Accuracy. Let AB stand for ABALONE, AD

for ADULT, and BM for BANKMARKETING. All values are rounded to 3 decimal places. Table 1 reports the average classification accuracy across all 3 test models, across all trials, partitions, and datasets. Table 2 reports the average classification accuracy across all models, per dataset. Table 3 reports the average classification accuracy across all models, per dataset, per partition (training set size).

Table 1

	SVM	ANN	LogReg
ACA	0.788	0.758	0.777

Table 2

	SVM	ANN	LogReg
AB	0.795	0.788	0.784
AD	0.670	0.606	0.655
BM	0.899	0.880	0.893

Table 3

	SVM	ANN	LogReg
AB (0.2)	0.787	0.780	0.785
AB (0.5)	0.797	0.790	0.786
AB (0.8)	0.800	0.792	0.780
AD (0.2)	0.669	0.597	0.654
AD (0.5)	0.672	0.614	0.656
AD (0.8)	0.670	0.606	0.653
BM (0.2)	0.895	0.879	0.892
BM (0.5)	0.899	0.879	0.893
BM (0.8)	0.903	0.883	0.893

From Table 1, the results do not necessarily align with the findings from Caruana and Niculescu-Mizil 2006, where SVMs and ANNs performed similar strongly, with LogReg lagging far behind. In our experiment, all 3 models perform similarly well.

From Table 2, some divergence according to dataset type is visible. SVMs outperform ANNs and LogReg across balanced/unbalanced and small/large datasets. ANNs perform noticably poorly on ADULT.

Table 3 reveals a partitioning trend, where a larger training set generally results in improved classification accuracy.

4 Conclusion

Supervised learning models have come a long way since 2006. Modern techniques and fine-tuning have supplemented the performance of these models, and yet many trends from 2006 still prevail. Previously, ANNs and SVMs performed strongly in classification accuracy, while LogReg struggled. Nowadays, open source versions of these models perform similarly well in classification accuracy. Interestingly, there is variation in performance according to dataset, with ANNs performing poorly on a large, balanced dataset while all 3 models performed extremely well on a large, unbalanced dataset.

5 Bonus Points

I believe I should receive bonus points due to the extensivity of experimenting performed. Beyond comprehensive hyperparameter tuning, my employment of a diverse range of datasets allowed for realistic and meaningful results. I was able to dive into analysis beyond that which is touched on by Caruna and Niculescu-Mizil 2006, ex-

ploring the performance of models on small/large and unbalanced/balanced datasets. Additionally, due to the heavy computational load of running cross-validation and hyperparameter tuning on 3 partitions across 3 datasets, I spent 8+ hours training these models. Specifically, I spent 6+ hours training SVMs and 2+ hours training NNs.

6 References

Becker, Barry and Ronny Kohavi. "Adult." UCI Machine Learning Repository, 1996, https://doi.org/10.24432/C5XW20.

Caruana, Rich and Niculescu-Mizil, Alexandru. "An Empirical Comparison of Supervised Learning Algorithms." Cornell University, 2006, https://www.cs.cornell.edu/ caruana/ctp/ct.papers/caruana.icml06.pdf.

Moro, S., P. Rita, and P. Cortez. "Bank Marketing." UCI Machine Learning Repository, 2014, https://doi.org/10.24432/C5K306.

Nash, Warwick, et al. "Abalone." UCI Machine Learning Repository, 1994, https://doi.org/10.24432/C55C7W.