Exploring Machine Learning Model Performance across Diverse Datasets: A Comparative Analysis

Yan Mazheika   Polina Petrova

May 8, 2024

Introduction

This report provides an analysis of the application of various machine learning projects on several datasets. The goal is to identify the best-use scenarios of model variants given the nature of input data. In this report, the authors analyze the

- Decision Tree, *provided by Polina Petrova*

- k Nearest Neighbors, *provided by Polina Petrova*

- Neural Network, *provided by Yan Mazheika*

- Random Forrest, *provided by Yan Mazheika*

against the handwritten digits, titanic survival, loan eligibility, and Parkinson's classification datasets.

It is our aim to explain the nature of the data, our models, and the performance of these classifiers. Throughout our report, we justify our algorithm choice for the given dataset and our choice of hyper-parameters for the tuning of the algorithm. These insights are for the reader's benefit; we also aim to provide insight into how to solve novel machine problems, which algorithms may work best, and how to adjust their hyper-parameters for optimal performance.

Approach

The authors have chosen to coordinate on every dataset. For the first two algorithms of every dataset, one of them comes from *Yan Mazheika* while the other comes from *Polina Petrova*. These algorithms may have been modified to handle a new type of data but their fundamental logic stays consistent from previous use cases.

We use cross-validation of k=10 folds when testing the performance of our models (except for kNN). We also transform the data into the [0,1] range when using the neural network or kNN classifiers.

Digits Dataset

The digits dataset come's from the sci-kit learn library, available in Python. We have chosen to analyze the performance of a neural network classifier and the k-Nearest Neighbors classifier. Again, we mention that the neural network tuning was performed by *Yan Mazheika* while the kNN tuning was done by *Polina Petrova*.

We chose a neural network as one of our algorithms for this dataset because of the algorithm's ability to handle complex patterns and identify non-linear patterns in these data. Anecdotally, neural networks have shown impressive results in image classification tasks, making them a top contender for handwritten digit recognition, the focus of this dataset.

On the other hand, the k-NN algorithm is a simple yet effective classifier that makes predictions based on the closeness of past training instances in the feature space. We figured that similar pixel activations in the handwritten digits would translate into less distance in the 64-dimensional feature space. A concern we had when choosing this algorithm is precision loss; kNN performs poorly with a large number of features, which is 64 in our case. The distance measurement between two instances converges to zero as we add more features and normalize them. This (has/hasn't) proved to be a problem.

Performance Metrics

	Neural Network
Learning Rate α	0.0500
Regularization λ	0.0100
Architecture	[64, 32, 32, 10]
Mean Accuracy	0.9962
Mean F1-score	0.9809
Mean Test Cost	59.3100

	kNN
Neighbors k	k
Mean Accuracy	?
Mean F1-score	?

Tuning Hyper-parameters

For the neural network, we've chosen an α of 0.05 and a regularization constant of $\lambda = 0.01$. At first, with a $\lambda = 0$ and $\alpha = 0.50$, the model had good performance but a quick convergence in the training curve. We found increased performance and a slower convergence by decreasing α and increasing λ to a final accuracy of 99.62%. This indicates a misclassification of 6 total instances out of 1,797 total. The architecture was chosen because it had the lowest accuracy variance between each fold out of the models tested.

...

Analysis

...

Titanic Dataset

This dataset required some extra pre-processing. Originally, the data had a 'Name' feature which contained

Analysis

Loan Eligibility Dataset

Analysis

Oxford Parkinson's Disease Dataset

Analysis