Machine Learning Algorithms

Hayden Bartolovich, Anthony Ponzani, Rachel Sutor, Joe Wolf

Introduction

With the rise in machine learning and its development beginning to be used in research. Understanding the tradeoffs of the types of learning algorithms is crucial for its effective use for different applications.

Abstract

Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, much like humans do. With many companies collecting corporate and customer data, machine learning is the key to processing and using the massive quantity of data to identify profitable opportunities or risks. Besides data mining, machine learning is used to compute complex tasks that humans could not feasibly perform, including facial recognition, network surveillance, automation, online search and healthcare. With the infinite amount of uses cases for machine learning, there are a multitude of algorithms that are used by programmers. Each algorithm differ in their approach and the type of problem that they are built to solve. We will explore some of the most common algorithms including, supervised learning algorithms, semi-supervised learning algorithms, unsupervised learning algorithms, reinforcement learning, and compare their strengths and weaknesses. This effort should help optimize the use of machine learning algorithms.

Methods

- To understand certain algorithms we used the python programming language and the scikit-learn library to understand a few specific algorithms

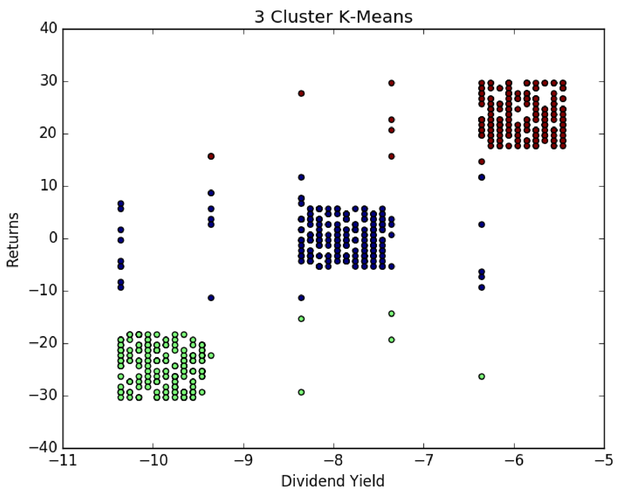

Figure 1. K-means cluster of stock returns according to their dividend yield (Unsupervised)

Types of Learning Algorithms

Machine learning has four classifications of learning algorithms:

Supervised Learning – algorithms that have input variables (x) and an output variable (Y) and the algorithm learns the mapping function from the input to the output.

Unsupervised learning – algorithms must learn relationships between elements in a data-set and classify the raw data without "help"

Semi-supervise Learning – algorithms are trained on a combination of labeled and unlabeled data. It falls between supervise and unsupervised learning

Reinforcement Learning – concerned with how software agents ought to take actions in an environment to maximize some notion of cumulative reward

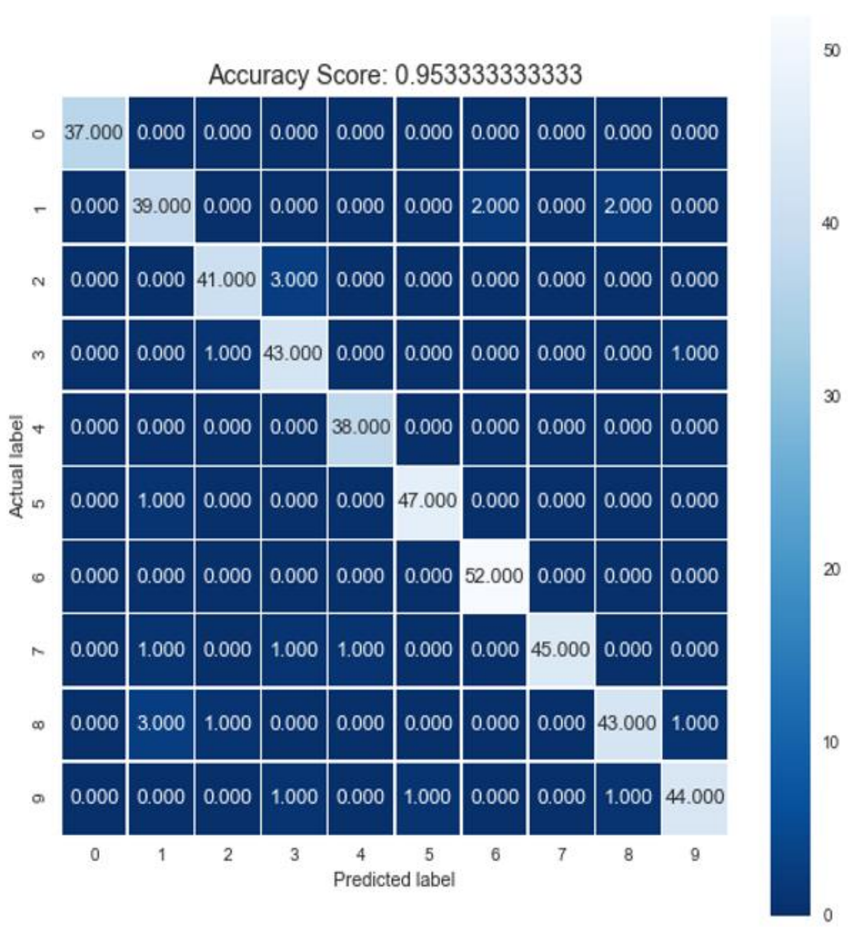

Figure 2. Confusion matrix for the classification of images of digits using logistic regression (Supervised)

Conclusion

When using machine learning algorithms, determining what learning algorithm you use is based on the dimensionality, complexity, and the method which you collect your data. Semi-supervised algorithms can be the most effective for different applications, but reinforcement algorithms cannot be substituted.

Future Work

- How do individual algorithms affect the results of a single type of application?
- For a specific application how does changing the bias vs variance affect the effectiveness of an algorithm?
- What are the effects of using a semi-supervised algorithm vs a supervised or unsupervised

RESULTS

There a few aspects of each type of learning algorithms that have to be taken into consideration when comparing the use for a specific application. The complexity of the classifier affects how much training data needs to feed into the algorithm. Training sets' dimensionality can affect the variance of the output. If an algorithm has a simple function, but the inputted training data has a very high dimensionality it can confuse the learning algorithm. This can be dampened by affecting the bias. Overfitting and underfitting are issues of the incorrect outputs based on the fitting of the data.

References

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798-1828. doi:10.1109/tpami.2013.50

Louridas, P., & Ebert, C. (2016). Machine Learning. IEEE Software, 33(5), 110-115. doi:10.1109/ms.2016.114

Resources used for python and sklearn

https://scikit-learn.org/stable/documentation.html

Ng, A. (n.d.). Machine Learning. Retrieved from

https://www.coursera.org/learn/machine-learning

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000). The Analysis of a Simple k-Means Clustering Algorithm. doi:10.21236/ada458738

Acknowledgments

A special thanks to Haodong Wang, (BS/MS/PhD) Dept. of Engineering, for his help on this project