

Research on algorithms of machine learning

Binyan Yu^{1,3,4}, Yuanzheng Zheng²

¹College of Big Data and Artificial Intelligence, Xiamen Chain Construction Technology Co., LTD, Xiamen, 361001, China

²School of Artificial Intelligence and Data Science, Hebei University of Technology, Hebei, 300130, China

³2318802305@aliyun.com

⁴Corresponding author

Abstract. Machine learning has endless application possibilities, with many algorithms worth learning in depth. Different algorithms can be flexibly applied to a variety of vertical fields, such as the most common neural network algorithms for face recognition, garbage classification, picture classification, and other application scenarios image recognition and computer vision, the hottest recent natural language processing and recommendation algorithms for different applications are from it. In the field of financial analysis, the decision tree algorithm and its derivative algorithms such as random forest are the mainstream. As well as support vector machines, naive Bayes, K-nearest neighbor algorithms, and so on. From the traditional regression algorithm to the hottest neural network algorithm. This paper discusses the application principle of the algorithm and lists some corresponding applications. Linear regression, decision trees, supervised learning, etc., while some have been replaced by more powerful and flexible algorithms and methods, by studying and understanding these foundational algorithms in depth, neural network models can be better designed and optimized, and a better understanding of how they work can be obtained.

Keywords: machine learning, algorithms, neuron, decision tree.

1. Introduction

The ability of a computer program or machine to think and learn is called artificial intelligence (AI) [1]. It is a marginal subject that belongs to the intersection of natural science and social science. John McCarthy coined the name "artificial intelligence" in 1955, a field of research that aims to give computers the ability to learn and solve problems on their own, without the need for humans to code. In everyday life, we often define artificial intelligence as a program that mimics human cognition. At the very least, we can think that computers are capable of achieving similar effects to humans in certain aspects related to thinking, such as learning and problem-solving, although in different ways [2]. According to Andreas Kaplan and Michael Haenlein, AI is the ability to enable systems to correctly interpret external data and utilize that data to achieve specific goals through flexible adaptation and learning [3].

In the modern era of technological marvels, AI stands as a testament to the relentless pursuit of creating machines that can simulate human-like intelligence and cognitive abilities. From envisioning intelligent robots in science fiction to witnessing the tangible advancements in our daily lives, AI has

emerged as a multidisciplinary field that amalgamates computer science, mathematics, and cognitive psychology. The ever-expanding landscape of AI encompasses a myriad of facets, including expert systems, natural language processing, computer vision, robotics, and the topic that takes center stage in this discourse - Machine Learning. The scope of Applications of artificial intelligence is very wide, including medicine, diagnosis, financial trade, robot control, law, scientific discovery, and toys. Many kinds of AI applications run deep into the foundations of every industry.

In financial transactions, artificial intelligence can analyze data, predict market trends, and improve the accuracy and safety of trading decisions. In terms of robot control, artificial intelligence realizes autonomous navigation, target recognition, and object grasping, and is applied in manufacturing, logistics, and service fields.

In the legal world, artificial intelligence accelerates the processing of legal documents, conducts research, predicts case outcomes, and improves efficiency. In terms of scientific discovery, artificial intelligence analyses data, assists drug development, and genetic research, and promotes scientific innovation. Children's toys and smart toys interact with children, provide educational games and learning experiences, and cultivate cognition and creativity.

In the 1990s and early 2000s, artificial intelligence technology became an element of a large system; but only a small portion of people think that this is an achievement in the field of artificial intelligence. Its prospects are very bright. Next, our attention turns to a crucial aspect of AI - machine learning, delving into its fundamental concepts and historical progression. Subsequently, in the following section, we will explore the practical applications of various foundational methods within machine learning. Examples encompass well-known algorithms like decision trees, along with their diverse real-world applications. Concluding this exploration, we will then engage in a discourse regarding the promising future prospects within the realm of machine learning.

2. Development of Machine Learning

Machine learning means giving the computer the chance to learn without being truly commanded [4], It constitutes a division and specialized area within the broader field of computer science. [5]. This opinion comes from Artificial Intelligence(AI)[6], AI refers to making machines as intelligent as the human brain. Learning and predicting data can be explored by studying algorithms for machine learning [7]. An algorithm operates based on commands beforehand while also capable of making predictions or decisions using data. They use other sample inputs to build models [8], guided by well-crafted algorithms, and each iteration of calculation culminates in a trustworthy and efficient yield of results. It has a rich history spanning decades, even potentially reaching back centuries. Fundamental tools such as Bayesian techniques, Laplace's least squares derivations, and Markov chains have been pivotal in its evolution. The journey began in the 1950s, when Alan Turing proposed the concept of a learning machine, and progress persisted until the early 2000s, marked by breakthroughs like the practical application of deep learning, exemplified by Alex Net in 2012.

During the early phase, spanning from the mid-1950s to the mid-1960s, the primary emphasis lay in the pursuit of "learning, whether with or without preexisting knowledge." This era was dedicated to augmenting the execution prowess of systems, a feat achieved through the manipulation of the system's surroundings and performance parameters. Much like the intricacies of programming, fine-tuning its vacant space function engendered shifts in its very structure. The overarching aspiration was for the system to seamlessly acclimate and flourish within its designated milieu. A particularly noteworthy example hailing from this period materializes in the form of a chess program devised by Samuel. Regrettably, this methodology proved insufficient in attaining a standard akin to human-level proficiency [9]. The second phase extended from the mid-1960s to the mid-1970s. It involved embedding diverse knowledge domains into systems to simulate human learning processes. Graph and logical structures were employed to describe the systems, primarily using symbols to represent machine language. Researchers discovered that learning was a gradual process, prompting the incorporation of expert knowledge into the system to achieve meaningful results. This phase saw significant contributions from Hayes-Roth and Winson in the realm of structure learning systems. The revival period emerged

from the mid-1970s to the mid-1980s, marked by a shift from learning single concepts to multiple ones. This era explored various learning strategies and methods, linking learning systems with practical applications, yielding substantial successes. The demand for knowledge acquisition in expert systems spurred machine learning research, with example-based inductive learning becoming prominent. The 1980 International Symposium on Machine Learning at Carnegie Mellon marked a global rise in research. Machine Learning's second volume, co-authored by Simon and AI experts, along with the launch of the international magazine Machine Learning in 1984, underscored the rapid advancement. Guided learning by Mostow, Lenat's program for discovering mathematical concepts and Langley's BACON program, and its iterations were notable endeavors during this phase [9].

3. Basic algorithms for machine learning

3.1. Linear regression

As the most basic machine learning algorithms, linear regression and logistic regression algorithms contain some of the most key basic ideas in machine learning. Many powerful nonlinear models can be obtained by introducing a hierarchical structure of high dimensional mapping on the basis of linear models.

Linear regression is an extension of the most basic mathematical problems in the algorithm:

$$f(x) = w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_dx_d + b \quad (1)$$

Simply put, we need the same number of weights(w) as the variables to get a linear combination for prediction.

Mean squared error is the most commonly used performance measure to find an optimal set of weights to fit a linear model with higher accuracy:

$$(w^*, b^*) = \arg \min_{(w, b)} \sum_{i=1}^m (f(x_i) - y_i)^2 \quad (2)$$

This versatile method is called the least square method in mathematics, and its geometric meaning is to try to find a line that minimizes the sum of the Euclidean distances from all samples to the line.

3.2. Neural networks

Neural networks are the best-known and most important structure of machine learning. In fact, it is already a fairly broad subject area, including a variety of interdisciplinary subjects. In biology, neurons are connected to each other and send chemical messages to change the electric potential to a certain threshold to activate the neuron. In neural networks, activation functions are used to deal with the input from other neurons, and how to choose the optimal activation function becomes a key part of building the model. Rectified linear Unit is the most used single neuron in modern deep learning models because their nonlinearity helps us train deeper neural networks. The experimental results show how the new modular activation function is better than the benchmark activation function in the field of image classification [10].

3.3. Decision tree

A decision tree is a common classification and regression algorithm in machine learning, which makes branch decisions to reach the final decision with a tree-like structure. A decision tree (DT) is a directed graph, a graph consisting of a root node with finite connections. The dataset after feature engineering is segmented into subsets recursively so that the cases in each subset belong to the same class in each tree branch node for classification calculation. How to choose the optimal partition attribute is the key to decision tree learning. The higher the "purity" of the node, the better the model effect is obviously

CART decision trees use the "Gini index" as a standard selection to divide the dataset. About a variable C , it was defined as:

$$\text{gini}(C) = 1 - \sum_{k=1}^K p_k^2 \quad (3)$$

In this way, we can define the split criterion [11]

$$G_{ix}(\mathcal{C}, \mathcal{X}) = \text{gini}(\mathcal{C}|\mathcal{X}) - \text{gini}(\mathcal{X}) \quad (4)$$

3.4. Ensemble learning

If we look at these algorithms up front as some different single learners, ensemble learning can combine multiple individual models to improve predictive performance. In many fields, it is often possible to obtain significantly better generalization performance than a single learner. According to data collected and collated by Ammar Mohammed, ensemble deep learning outperforms traditional ensemble learning in several fields including image classification, natural language processing (NLP), etc. In the field of medical statistics and NLP, stacking or voting methods based on multiple deep learning models (CNN, LSTM, GRU, etc.) perform better than the most advanced integration methods [12].

3.5. Random forest

Bagging is the most famous representative of Ensemble learning, roughly divided into three types – Bagging, Boosting, and Stacking – which are ensemble learning algorithms. Random Forest (Breiman, 2001a), one of the most popular and commonly used algorithms, is a kind of extended variant of Bagging.

Compared with its based learner -- Decision Trees, the framework of Random Forest applies a multitude of classification and regression trees to develop utilizing randomly chosen training datasets and arbitrary selections of predictor variables for predicting results [13]. According to different relevant experimental data collected by the writer from 2011 to 2015, we can reach an obvious conclusion that VSURF, varSelRF, Boruta, and Altmann are the methods with the lowest error rate in the absence of missing values.

The criteria for optimal algorithms are rarely the same across domains. However, the random forest algorithm is simple and easy to implement, and can still show strong performance in many real businesses while having low computational overhead.

3.6. Application

Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are all forms of Machine learning divided into four different forms: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. The difference between supervised learning and unsupervised learning is whether labeled samples are used, whereas semi-supervised learning relies on only a few labeled samples, and reinforcement learning requires a feedback mechanism. For example, in the field of environmental ecology, Shixuan Cui proposed application methods of machine learning such as ecological factor identification and prediction, ecological risk assessment, remote sensing and image recognition, high throughput screening of pollutants, toxicity and potential mechanism prediction, biomarkers, etc. [14]

In contrast, Jan Grudniewicz, in a study on the application of machine learning to algorithmic investment strategies in global stock markets, investigated the profitability of quantitative investment strategies based on machine learning. By comparing the performance models of these strategies with the risk and return metrics of the benchmark buy-and-hold returns, the study in the original paper shows that a polynomial support vector machine model based on the WIG20 index (Poland), a linear support vector machine model based on the DAX(Germany), and a polynomial support vector machine model based on the S&P 500 Index (USA) is the best-performing policy Slight. Linear support vector machine strategies yield the best risk-adjusted returns on average. In every metric case, the strategy based on machine learning techniques yielded better returns than the benchmark.[15]

4. Conclusion

In addition to this, research related to machine learning translation between computer vision and natural language processing (NLP) proposes multimodal machine learning in one direction from image to natural language or from natural language to image.

Advances in multimodal machine learning can help artificial intelligence more closely resemble human intelligence, whether it is a world model or a non-general AI, sensing the world from multiple modes is an important training method. If all the AI models in all fields are stacked into a pyramid, the one that shines most brightly at the top of its tower must be the large language model. Because it is like an AI brain center, it is only a matter of time before wisdom appears on the basis of emergent phenomena, similar to the appearance of thinking in the human brain, a large number of individual neurons will change quantitatively and produce qualitative changes, and under the support of sufficient hardware in the future, the large language model will likely move from logical reasoning to independent consciousness. After the emergence of the large language model, the knowledge graph will not disappear but become an "external brain" combined with the vertical domain and the language model, which has become a reality.

References

- [1] Andreas Kaplan. (2022) Artificial Intelligence, Business and Civilization: Our Fate Made in Machines, Routledge,"
- [2] Russell, Stuart J. & Norvig, Peter. (2003) Artificial intelligence: a modern approach. 2nd ed, Upper Saddle River, New Jersey: Prentice Hall. ISBN 0-13-790395-2
- [3] Kaplan, Andreas; Haenlein, Michael. (2019) "Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence". Business Horizons. 62 (1): 15–25. doi:10.1016/j.bushor.2018.08.004. S2CID 158433736.
- [4] John McCarthy & Edward Feigenbaum. (1990) In Memoriam Arthur Samuel: pioneer in machine learning. AI Magazine. AAAI. 11 (3).[1] Archived 2018-01-22 at the Wayback Machine
- [5] "Machine Learning | Data Basecamp". 2021-11-26. Retrieved 2022-08-14.
- [6] "Machine learning | artificial intelligence | Britannica".
- [7] Ron Kohavi; Foster Provost. (1998) "Glossary of terms". Machine Learning. 30: 271–274. doi:10.1023/A:1007411609915. S2CID 36227423.
- [8] Christopher Bishop. (1995) Neural networks for pattern recognition. Oxford University Press. ISBN 0-19-853864-2
- [9] Chen Haihong, Huang Biao, Liu Feng, Chen Wenguo. (2017) Principles and applications of machine learning. Chengdu: University of Electronic Science and Technology Press: 2-19
- [10] Iván V, Emilio S, Marcelino M, et al. (2023) Empirical study of the modulus as activation function in computer vision applications[J]. Engineering Applications of Artificial Intelligence, 120.
- [11] Morcillo G L, Poyo C J F, Maldonado L G. (2014) Using Decision Trees for Comparing Different Consistency Models[J]. Procedia - Social and Behavioral Sciences, 160.
- [12] Ammar M, Rania K. (2023) A comprehensive review on ensemble deep learning: Opportunities and challenges[J]. Journal of King Saud University - Computer and Information Sciences, 35(2).
- [13] Speiser L J, Miller E M, Tooze J, et al. (2019) A comparison of random forest variable selection methods for classification prediction modeling[J]. Expert Systems With Applications, 134.
- [14] Shixuan Cui, Yuchen Gao, Yizhou Huang, Lilai Shen, et al. (2023) Advances and applications of machine learning and deep learning in environmental ecology and health [J] Environmental Pollution, 122358.
- [15] Jan G, Robert S. (2023) Application of machine learning in algorithmic investment strategies on global stock markets[J]. Research in International Business and Finance, 66.