A DEEP DIVE INTO MACHINE LEARNING

Mrs. Leena Sanu, Assistant Professor
Department of Management, Christ College
Pune, India
leenasanu.ls@gmail.com

Dr. Santosh Parakh, Associate Professor
Department of MCA, Vidya Pratishtan's Institute of Information and Technology (VIIT),
Baramati, India
santoshparakh@gmail.com

ABSTRACT

Data mining and machine learning approaches look at data from beginning to end to find hidden patterns in the dataset. A variety of study areas support the establishment of the multidisciplinary discipline of machine learning. The digital era has access to a multitude of data which includes business information, data from social media sites, Internet of Things (IoT) data, cyber security data, cellular data, health data, etc. In deep learning, multiple layers of interconnected neurons are used to process and learn complex patterns in data. These layers allow the network to automatically extract high-level features from raw input data, such as images, speech, or text. Machine learning permits the user to use a computer algorithm for a large amount of data and analyse the same with the help of a computer and make recommendations based on available data and decisions on the base of input data. Machine learning is of four kinds viz. Supervised learning, Unsupervised learning, Semi-supervised learning and Reinforced learning. We have done a deep analysis of techniques of machine learning through this research.

Keywords: Machine Learning, Algorithms, Decision tree, Neural networks, Metadata.

Introduction

Machine learning was created by Arthur Samuel, an American scientist in 1959, when he was working at IBM. According to him, machine learning is a study which gives computers an ability to learn without any systematic programming. Machine learning has a connection with artificial intelligence, which is involved in developing algorithms and statistical models, which enable computers to improve the performance in activities through experience. Machine learning is of four kinds viz. Supervised learning, Unsupervised learning, Semi-supervised learning and Reinforced learning. A recent study showed that machine learning (ML) engineers outpaced all others in salary package, demand as well as growth. In short, ML is a smart career choice for job aspirants. ML is also an important component in areas such as Big Data, Predictive Analytics, Data Mining and Computational Statistics.

The main responsibilities of an ML engineer include the following:

- Study of computer architectures, data structures and algorithms
- Designing machine learning systems
- Creating infrastructure, data and model
- Analyse large, complicated datasets
- Develop algorithms
- Build and maintain machine learning solutions in production etc.

A few examples of machine learning are as under:

- Recognition of speech and image, which help to convert speech into text.
- Google translation
- Prediction
- Extraction
- Extraction

Real-World Examples of Machine Learning (ML) are as under:

- Facial recognition
- Production recommendations
- Financial accuracy social media optimization
- Healthcare advancement
- Predictive analytics

Review of Literature

Murugan, Sathiyamoorthi (2022) aims to provide an in-depth overview to several commonly applied machine learning techniques and their applications, which are divided into supervised, unsupervised, and reinforcement learning which also serves as a concise manual for prospective researchers of data and machine learning.

Choudhary, Gianey (2017) in their study have provided a general comparison on machine learning algorithms. Unlike static programming methods, which require explicit human instruction, machine learning algorithms are designed to learn from the data and generate predictions.

Dheepak, Vaishali (2021) have provided a brief summary and viewpoint on many machine learning applications by this survey. Duarte, Ståhl (2018) in their study highlights the subdivision of the machine learning field the purpose of which is to make it possible for computers to learn.

Paluszek, Thomas (2016), in the subject of machine learning, past data are utilised to anticipate or respond to future data. It has a close connection to artificial intelligence, computational statistics, and pattern recognition. In fields like facial recognition, spam filtering, and others where it is not practical or even practicable to develop algorithms to execute a task, machine learning is vital.

Rath (2022) in his study has stated that business learning made various techniques. Every financial platform has become safer and more user-friendly for financial transactions as a result of advancements in the internet and digital marketing.

Thakur, Tembhurney & Rane (2021) aims to give readers a mature perspective on the fundamentals of ML, one of the experts' most popular recent study subjects. It outlines the problems and difficulties with machine learning (ML). The selection of features is relevant to feature engineering because feature engineering creates novel features to improve learning objectives while feature selection aims to select the features that are most appropriate.

Objectives

1. To define the study's scope by considering the traits patterns of data, which help in making predictions.
2. To detect and analyse trends which helps to solve problems.
3. To explore the ways in which the solutions based on machine learning might be used in various real-world scenarios.

Secondary Data Analysis

Real-World Data Types

Data is required for the development of machine learning models. There are different types of data, such as structured, unstructured and semi-structured data. The difference is as under:

Structured data: It is organised data, which is easy to search in any database which is rational and mostly categorised as quantitative data, which most researchers and scientists used to work with. It has fixed fields and columns in relational databases and spreadsheets. Examples are names, dates, addresses, credit card numbers and more.

Unstructured data: It has no proper format or pre-defined format and hence it is difficult for collection, formalisation and analysis. It is often qualitative data which cannot be processed and analysed using normal methods and tools. For unstructured data, in order to understand customer buying habits and timings, patterns in purchases and sentiments towards a particular product etc. can be obtained by using data mining techniques.

Semi-structured data: It is a type of structured data which stands at the mid-way between structured and unstructured data. It does not have a rational or tabular data model. Common examples of this are JSON and XML.

Metadata is "data about data," not the usual form of data. Metadata refers to data that describes other data. It provides context and information about the data, such as its format, structure, content, and relationships to other data. In other words, metadata is "data about data."

Metadata: It is used to help organize and manage data, making it easier to search, retrieve, and use. It can be stored in various forms, such as tags, labels, or fields, and can be associated with different types of data, including documents, images, videos, and databases.

Machine Learning Algorithms

There are different types of machine learning algorithms as under:

(1) **Supervised:** It is a learning function where a pair is taken, and this pair consists of input object and output value. It analyses the training data, which can be utilised for mapping new examples. Both classification and regression problems are supervised learning problems.

(2) **Semi-supervised learning:**, the algorithm is trained to use both small amounts of labelled data and large amounts of unlabelled data. The idea is that the labelled data provides information about the target labels, while the unlabeled data helps the algorithm to learn the underlying structure of the data and generalize better to unseen examples. One common technique for semi-supervised learning is to use the labelled data to train a supervised model, and then use that model to make predictions on the unlabeled data. The predictions can then be used to label some of the unlabeled data, which can be added to the labelled data for further training. This algorithm is mainly used for recognising image and speech as well detection of fraud. Generally it is used where it is very expensive to use labelled data.

(3) **Unsupervised Learning:** It is a type of machine learning algorithm, which is used to draw inferences from sets of data consisting of input data without labelled responses. Unsupervised learning algorithms are not included in observations. Since this is unsupervised, the accuracy cannot be evaluated. This method is used for cluster analysis, which is used for exploratory data analysis.

(4) **Reinforcement input data without labelled responses:** It is the problem of getting an agent to act in the world in order to maximise the rewards.

In machine learning, a learner is not informed as to what action he has to take. He has to discover which technique to use in order to derive the desired results. For example, we can teach a new trick to a dog; and we can use the technique of reward or punishment if it does right or wrong. Same method can be used in the case of computers to do many tasks such as playing chess, job scheduling and controlling robot limbs.

There are several different machine learning techniques, each with its strengths and weaknesses. Here are some comparisons of different techniques:

1. **Supervised learning vs. Unsupervised learning:**

Supervised learning is a learning function where a pair is taken, and this pair consists of input object and output value. It analyses the training data, which can be utilised for mapping new examples. Both classification and regression problems are supervised learning problems. Whereas Unsupervised Learning is a type of machine learning algorithm, which is used to draw inferences from sets of data consist of input data without labelled responses. Unsupervised learning algorithms is not included in observations. Since this is unsupervised, the accuracy cannot be evaluated. This method is used for cluster analysis, which is used for exploratory data analysis.

2. **Decision trees vs. Neural networks:**

Decision trees are a simple and interpretable machine learning technique that uses a tree-like structure to make decisions. Neural networks or artificial neural networks are more complex and can model non-linear relationships between variables. Decision trees are useful when the problem is simple and the dataset is small, while neural networks are useful for more complex problems and larger datasets.

3. **Logistic regression vs. Random forests:**

Logistic model is a statistical model, which takes into consideration the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. Logistic regression is useful when the problem involves a binary outcome and the relationship between the predictor variables and the outcome is linear, while random forests are useful for more complex problems and datasets with many features.

4. **Support vector machines vs. K-nearest neighbours:**

Support Vector Machine or SVM is one of the most famous Supervised Learning algorithms, that is used for classification and regression problems. SVMs are useful when the problem involves a large number of features and a clear separation between classes, while KNN is useful when the dataset is small and the relationship between the features and the outcome is non-linear.

5. **Clustering vs. Dimensionality reduction:**

Clustering is a technique of unsupervised learning that gathers data pointed to on the basis of its similarity. Dimensionality reduction, on the other hand, is a technique that reduces various features in a set of data, when you are retaining so much of the original information as possible. Clustering is useful when exploring the structure of a dataset or when identifying outliers, while dimensionality reduction is useful for reducing noise and simplifying complex datasets.

Table:1 gives an understanding of the different techniques which is suitable for different scenarios as per the data sets available.

Algorithm	Type	Pros	Cons
Linear Regression	Supervised Regression	Simple to implement and interpret, performs well on linearly separable data.	May underfit or overfit data if not properly tuned.
Logistic Regression	Supervised Classification	Simple to implement and interpret, performs well on linearly separable data.	May underfit or overfit data if not properly tuned.
Decision Trees	Supervised Classification/Regression	Easy to understand and visualize, can handle both categorical and numerical data.	May overfit data if not properly tuned, sensitive to small changes in the data.
Random Forest	Supervised Classification/Regression	Reduces overfitting by combining multiple decision trees, handles both categorical and numerical data.	Can be slow to train on large datasets.
Support Vector Machines (SVMs)	Supervised Classification/Regression	Performs well on high-dimensional datasets, effective in handling non-linearly separable data.	Can be slow to train on large datasets, requires careful selection of kernel function.
Naive Bayes	Supervised Classification	Simple and fast to train and classify, works well with high-dimensional data.	Assumes independence of features, may not work well with highly correlated data.
k-Nearest Neighbors (k-NN)	Supervised Classification/Regression	Simple to implement and interpret, works well with small datasets.	Can be slow to classify new data, sensitive to the choice of k.
Neural Networks	Supervised/Unsupervised Classification/Regression	Can handle complex non-linear relationships, can learn from unstructured data.	Can be difficult to interpret and may require a lot of training data.
Clustering Algorithms (e.g., k-Means, Hierarchical)	Unsupervised	Can identify underlying structure in data, can be used for anomaly detection.	Requires careful selection of the number of clusters, may not work well with high-dimensional data.

Table 1: Pros and Cons of Machine Learning Algorithms subject to the kind of data used.

Machine Learning Algorithms
Classification Analysis

• Binary classification:

• Multiclass classification:

• Multi-label classification:
Regression Analysis

Cluster Analysis

Dimensionality Reduction and Feature Learning

Association Rule Learning

Reinforcement Learning

In addition to classification analysis, there are other similar algorithms that fall under the broader umbrella of machine learning, including:

1. **Regression Analysis** - This involves predicting a continuous numerical output rather than a categorical one.
2. **Clustering Analysis** - This involves grouping similar data points together based on their features, without predefined labels or categories.
3. **Dimensionality Reduction** - This involves reducing the number of input features while retaining as much information as possible, to make it easier to analyze and visualize the data.
4. **Neural Networks** - These are a class of algorithms that can be used for both classification and regression tasks, and are modelled after the structure of the human brain.

Conclusion

The effectiveness and quality of the dataset of the algorithms of learning are needed for machine learning models in order to be successful. In this study, we have looked closely at machine learning techniques for uses that call for sophisticated data processing. Both supervised and unsupervised machine learning are possible. Unsupervised Learning typically provides superior performance and results for huge data sets, while Supervised Learning is the preferable option if you have a smaller amount of data and clearly labelled data for training. On the other hand, use deep learning approaches when you are in possession of a sizable data collection that is readily available. All of these algorithms have different merits and demerits and the algorithm merits are in line with the specific available problem as well as on the basis of data to be analysed.

References

Arul Murugan R., Sathiyamoorthi V. (2022). Introduction to machine learning and its implementation techniques. *Research Anthology on Machine Learning Techniques, Methods, and Applications*, 1-25. doi:10.4018/978-1-6684-6291-1.ch001

Choudhary, R., Gianey, H. K. (2017). Comprehensive review on supervised machine learning algorithms. *2017 International Conference on Machine Learning and Data Science (MLDS)*. doi:10.1109/mlds.2017.11

Dheepak, G., Vaishali, D. (2021). A comprehensive overview of machine learning algorithms and their applications. *International Journal of Advanced Research in Science, Communication and Technology*, 12-23. doi:10.48175/ijarsct-2301

Duarte, D., Ståhl, N. (2018). Machine learning: A concise overview. *Studies in Big Data*, 27-58. doi:10.1007/978-3-319-97556-6_3

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452-459. doi:10.1038/nature14541

Jordan, M. I., Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. doi:10.1126/science.aaa8415

Kumar, Y., Kaur, K., & Singh, G. (2020). Machine learning aspects and its applications towards different research areas. *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. doi:10.1109/iccakm46823.2020.9051502

Paluszek, M., Thomas, S. (2016). An overview of machine learning. *MATLAB Machine Learning*, 3-15. doi:10.1007/978-1-4842-2250-8_1

Rath, M. (2022). Machine learning and its use in e-Commerce and E-business. *Research Anthology on Machine Learning Techniques, Methods, and Applications*, 1193-1209. doi:10.4018/978-1-6684-6291-1.ch062

Thakur, R., Tembhurney, M., & Rane, D. (2021). Research aspects of machine learning: Issues, challenges, and future scope. *Design of Intelligent Applications Using Machine Learning and Deep Learning Techniques*, 37-60. doi:10.1201/9781003133681-3

Wang, H., Ma, C., & Zhou, L. (2009). A brief review of machine learning and its application. *2009 International Conference on Information Engineering and Computer Science*. doi:10.1109/iciecs.2009.5362936