A REVIEW OF MACHINE LEARNING AND IT’S METHOD

SC STEMINA¹, J.Beschi Raja²

¹Department of Computer Science and Engineering, Sri Krishna College of Technology Coimbatore
²Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology Coimbatore

ABSTRACT

There is a large amount of data available everywhere. Therefore, it is very important to analyse this data in order to extract some useful information and to develop an algorithm based on this analysis. This can be achieved through data mining and machine learning. Machine learning is an integral part of artificial intelligence, which is used to design algorithms based on the data trends and historical relationships between data. Machine learning is used in various fields such as bioinformatics, intrusion detection, information retrieval, game playing, marketing, malware detection, image deconvolution and so on. The goal of machine learning is to program computers to use example data or past experience to solve a given problem. Many successful applications of machine learning exist already, including systems that analyze past sales data to predict customer behavior, optimize robot behavior so that a task can be completed using minimum resources, and extract knowledge from bioinformatics data. This paper presents an overview of the machine learning techniques currently in use.

Keywords: Supervised, unsupervised, reinforcement, database, neural network

1. INTRODUCTION

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. Machine Learning is used anywhere from automating mundane tasks to offering intelligent insights, industries in every sector try to benefit from it. Machine learning is a tool for turning information into knowledge. In the past years, there has been an explosion of data. This mass of data is useless unless we analyse it and find the patterns hidden within. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover. The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making. It’s a must-have skill for all aspiring data analysts and data scientists, or anyone else who wants to wrestle all that raw data into refined trends and predictions. Data mining is about searching large stores of data to uncover patterns and trends. Data mining is an automated analytical method to extract usable information from massive sets of raw data. One of the defining characteristics of this method of analysis is its automation, which involves machine learning and database tools to expedite the analytical process and find information that is more relevant to users. Due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Thus, we use information and knowledge for applications ranging from market analysis. This is the reason why data mining, known as knowledge discovery from data.

Fig 1. Machine learning frame work

2. TYPES OF MACHINE LEARNING

2.1 SUPERVISED LEARNING

Supervised learning is a learning model built to make prediction, given an unforeseen input instance. The term supervised learning comes from the idea that an algorithm is learning from a training dataset. It is mostly associated with retrieval-based AI but they may also be capable of using a
generative learning model. Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. The goal is to produce an accurate enough mapping function that when new input is given, the algorithm can predict the output. This is an iterative process, and each time the algorithm makes a prediction, it is corrected or given feedback until it achieves an acceptable level of performance. It provides you with a powerful tool to classify and process data using machine language. A learning algorithm then trains a model to generate a prediction for the response to new data or the test dataset. Supervised learning uses classification algorithms and regression techniques to develop predictive models. There are two types of Supervised Learning techniques: Regression and Classification. Classification separates the data, Regression fits the data.

2.2 UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Unsupervised learning algorithms allows you to perform more complex processing tasks compared to supervised learning. The best time to use unsupervised machine learning is when you do not have data on desired outcomes, such as determining a target market for an entirely new product that your business has never sold before. Although, unsupervised learning can be more unpredictable compared with other natural learning methods. The patterns you uncover with unsupervised machine learning methods may also come in handy when implementing supervised machine learning methods. Transactional data serves as a good source of data set for unsupervised learning. Algorithms are left to their own devises to discover and present the interesting structure in the data.

2.3 REINFORCEMENT LEARNING

Reinforcement Learning is defined as a Machine Learning method that allows machines and software agents to automatically determine the ideal behaviour within a specific context, in order to maximize its performance. It differs from supervised learning in not needing labelled input/output pairs be presented, and in not needing sub-optimal actions to be explicitly corrected. Instead the focus is on finding a balance between exploration and exploitation. It is particularly well-suited to problems that include a long-term versus short-term reward trade-off. It has been applied successfully to various problems, including robot control, elevator scheduling, telecommunications, backgammon and checkers. Reinforcement learning is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. Time plays a crucial role in Reinforcement problems. In reinforcement learning feedback is always delayed, not instantaneous and agent's actions determine the subsequent data it receives. The primary idea of reinforcement learning is to identify the best policy or the method that helps businesses in achieving the goals faster. While humans can create a few good models in a week, machine learning is capable of developing thousands of such models in a week.

3. ADVANTAGES OF MACHINE LEARNING

The value of machine learning technology has been recognized by companies across several industries that deal with huge volumes of data. By leveraging insights obtained from this data, companies are able work in an efficient manner to control costs as well as get an edge over their competitors. Companies in the financial sector are able to identify key insights in financial data as well as prevent any occurrences of financial fraud, with the help of machine learning technology. Government agencies like utilities and public safety have a specific need as they have multiple data sources, which can be mined for identifying useful patterns and insights. Marketing and Sales use machine learning technology to analyze the purchase history of their customers and make personalized product recommendations for their next purchase. Based on the travel history and pattern of traveling across various routes, machine learning can help transportation companies predict potential problems that could arise on certain routes, and accordingly advise their customers to opt for a different route. There are search engines available while searching to provide the best results to customers. There are many machine learning algorithms created for searching the particular user query like for Google. Whatever the page is being opened by the users for the particular topic frequently that will remain at the top of the page for a long time. Most of the organizations are using applications of machine learning and investing in it a lot of money to make the process faster and smoother. It is one of the widely used and adopted language or technology in today’s world.

4. KEY FACTORS OF MACHINE LEARNING AND DEEP LEARNING

Machine learning and deep learning are two subsets of artificial intelligence. Machine Learning involved with the creation of algorithms which can modify itself without human intervention to produce desired output- by feeding itself through structured data. A subset of machine learning where algorithms are created and function similar to those in machine learning, but there are numerous layers of these algorithms- each providing a different interpretation to the data it feeds on. While basic machine learning models do become progressively better at whatever function is, but they still need some guidance. If an AI algorithm returns an inaccurate prediction, then an engineer has to step in and make adjustments. With a deep learning model, an algorithm can determine on its own if a prediction is accurate or not through its own neural network. Machine learning algorithms almost always require structured data, whereas deep
learning networks rely on layers of the ANN (artificial neural networks). Machine learning algorithms are built to “learn” to do things by understanding labelled data, then use it to produce further outputs with more sets of data. However, they need to be retrained through human intervention when the actual output isn’t the desired one. In the case of machine learning, the algorithm needs to be told how to make an accurate prediction by providing it with more information, whereas, in the case of deep learning, the algorithm is able to learn that through its own data processing. As we see a continuous growth in the popularity of machine learning and deep learning, it will become increasingly competitive for organizations to survive in the industry if they are not part of this bandwagon soon.

5. ALGORITHM OF MACHINE LEARNING

5.1 NAIVE BAYES CLASSIFIER ALGORITHM

In machine learning, Naive Bayes classifiers are a family of simple “probabilistic classifiers” based on applying Bayes’ theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models. Naive Bayes classifier is a straightforward and powerful algorithm for the classification task. Even if we are working on a data set with millions of records with some attributes, it is suggested to try Naive Bayes approach. Naive Bayes classifier gives great results when we use it for textual data analysis. Such as Natural Language Processing. Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP), Naive Bayes classifier assumes that all the features are unrelated to each other. Naive Bayes model is easy to build and particularly useful for very large data sets. Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes. It is easy and fast to predict class of test data set. It also perform well in multi class prediction. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data. It perform well in case of categorical input variables compared to numerical variable(s).

5.2 LINEAR REGRESSION

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, rather than trying to classify them into categories. Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly. The goal of regression analysis is to create a trend line based on the data you have gathered. Linear Regression can then be used to draw a trend line which can then be used to confirm or deny the relationship between attributes. Linear Regression’s power lies in its simplicity, which means that it can be used to solve problems across various fields. At first, the data collected from the observations need to be collected and plotted along a line. If the difference between the predicted value and the result is almost the same, we can use linear regression for the problem.

5.3 LOGISTIC ALGORITHM

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. The binary logistic regression model has two levels of the dependent variable: categorical outputs with more than two values are modeled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model.

5.4 DECISION TREE LEARNING ALGORITHM

Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model. A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface. Decision tree learning is the construction of a decision tree from class-labeled training tuples. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost
5.5 K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures. Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase. The k-nearest neighbors (KNN) algorithm easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. It uses test data to make an "educated guess" on what an unclassified point should be classified as. It is a method preferred by many in the industry because of its ease of use, low calculation time and does not make assumption about the data. KNN is an algorithm that is considered both non-parametric and an example of lazy learning. KNN is often used in simple recommendation systems, image recognition technology, and decision-making models. k-nearest neighbors is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-nearest neighbour algorithm is that it is sensitive to the local structure of the data.

6. CONCLUSION

Machine learning techniques are being widely used to solve real-world problems by storing, manipulating, extracting and retrieving data from large sources. Machine learning approaches applied in systematic reviews of complex research fields such as quality improvement may assist in the title and abstract inclusion screening process. Machine learning approaches are of particular interest considering steadily increasing search outputs and accessibility of the existing evidence is a particular challenge of the research field quality improvement. Increased reviewer agreement appeared to be associated with improved predictive performance.

REFERENCE


Advanced Technology (IJEAT), Volume-8, page(s) 954-959, August 2019.


35. E. Laxmi Lydia, Gorapalli Chandra Sekhar, Madhu Babu Chevuru, Dasari Ramya, K. Vijaya Kumar, “Text Mining with Apache Hadoop over different Hadoop Clusters Architectures”, International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-2, Page(s) 1252-1256, July 2019

36. Dr.E. Laxmi Lydia, Dr.K. Vijaya Kumar, P. Amaranatha Reddy, D. Ramya, “Text Mining with Hadoop: Document Clustering with TF_IDF and Measuring Distance Using Euclidean” Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 14-Special Issue, Page(s) 1784 - 1792, 2018


51. Dr.E. Laxmi Lydia, Dr.K. Somasundaram,., Dr.A. Krishan Mohan,., “Data Analytics through Hadoop Eco-Components”, Jour of Adv Research in Dynamical & Control Systems, Vol. 9, No. 5, 2017


