Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

A Study on Machine Learning Techniques

K.Murali
Assistant Professor,
CMR Technical Campus,
Hyderabad.

Dr.B.Rama Subba Reddy
Professor & HOD,
Dept of CSE,
SVCE, Tirupati.

Abstract:
Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data. Machine learning is preferred approach to speech recognition, natural language processing, computer vision, medical outcomes analysis, robot control and computational biology. The success of machine learning system also depends on the algorithms. The algorithms control the search to find and build the knowledge structures. The learning algorithms should extract useful information from training examples. In this paper, we present the Machine Learning techniques. Several major kinds of machine learning method including supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning, the goal of this study is to provide a comprehensive review of different machine learning techniques.

Keywords:
Machine Learning, Supervised, Unsupervised Semi-supervised, Reinforcement, Regression, KNN, SVM.

I.INTRODUCTION:

Machine learning is a core sub-area of artificial intelligence as it enables computers to get into a mode of self-learning without being explicitly programmed. When exposed to new data, computer programs, are enabled to learn, grow, change, and develop by themselves. SAS, the North Carolina-based, American developer of analytics software comes with a definition on it: 'Machine learning is a method of data analysis that automates analytical model building'.

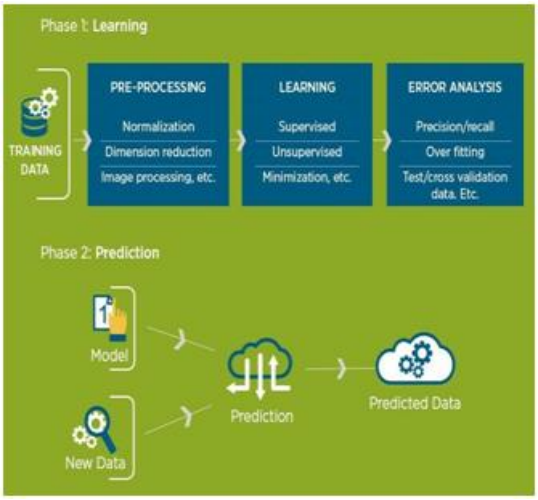
In other words, it allows computers to find insightful information without being programmed into where to look for a particular piece of information. This, it does by using algorithms that iteratively learn from data. While the concept of machine learning has been around for a long time, (one might be reminded of the notable example here – Alan Turing's famous Enigma Machine) the ability to automatically apply complex mathematical calculations to big data iteratively and quickly is gaining momentum only in recent times. This emphasizes the iterative aspect of machine learning the ability to independently adapt to new data. This is made possible as they learn from previous computations and make "pattern recognitions" to produce reliable results.

To understand better about the uses of machine learning, we might want to consider some of the instances where machine learning is applied: the self-driving Google car, cyber fraud detection, online recommendation engines like friend recommendations on Facebook, movie recommendations on Netflix and offers recommendations from Amazon are all examples of applied machine learning. All of this echoes the vitality of the role machine learning can play in today's data-rich world. A recent report from Mckinsey Global has asserted this fact by claiming that machine learning will be the driving factor behind the big wave of innovation in the coming times. Obviously, if machines can aid in filtering useful pieces of information that help in major advancements, and if machines can learn through programmed algorithms, all by themselves, then the technology is bound to find implementation in a wide variety of industries.

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

The process flow depicted here in a broader sense, is representative of how machine learning takes effect.

A.Machine Learning Process



B.Why Machine Learning?

With the constant evolution of the field, there has been a subsequent raise in the uses, demands, and importance of machine learning. The answer to the question as to why one has to adopt machine learning would be: 'High-value predictions that can guide better decisions and smart actions in real time without human intervention'(Source: SAS). Thus, if big data is gaining all the importance for the contributions it does, machine learning as a technology that helps analyze these large chunks of big data, easing the task of data scientists, in an automated process is equally gaining prominence and recognition. Machine learning has also changed the way data extraction, and interpretation is done by involving automatic sets of generic methods that have replaced traditional statistical techniques.

C.Uses Of Machine Learning

Some instances of machine learning applicability were mentioned previously. To understand the concept of machine learning better, let's consider some more examples: web search results, real-time ads on web pages and mobile devices, email spam filtering, network intrusion detection, and pattern and image recognition.

All these are by products of applying machine learning in the analysis of huge volumes of data. So, how drastically is machine learning revolutionizing the data analysis venue? Traditionally, data analysis has always been characterized by trial-and-error, an approach that becomes impossible when data sets are large and heterogeneous. It is for the very same reason, that big data was criticized as being an overhyped technology. Availability of more data is directly proportional to the difficulty of coming up with predictive models that work accurately. Also, traditional statistical solutions are focused on static analysis that is limited to the analysis of samples that are frozen in time. This could obviously result in inaccurate and unreliable conclusions. Machine learning comes as the solution to all this chaos. It proposes clever alternatives to analyzing huge volumes of data. It is a step forward from all of statistics, computer science and all other emerging applications in the industry. By developing fast and efficient algorithms and data-driven models for real-time processing of data, machine learning can produce accurate results and analysis.

II. Machine Learning Methods

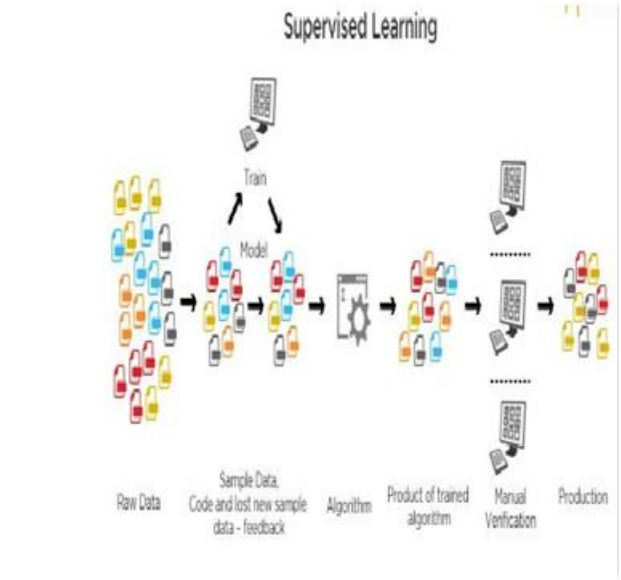
We have continually iterated the specificity about machine learning's ability to produce accurate analysis through efficient algorithms. So, how exactly do machines learn? Two popularly adopted methods of machine learning are supervised learning and unsupervised learning. It is estimated that while about 70 percent is supervised learning, unsupervised learning accounts to 10 to 20 percent. Other minor methods that are employed are semi-supervised and reinforcement learning.

1. Supervised Learning:

This kind of a learning is possible at instances when the inputs and the outputs are clearly identified, and algorithms are trained using labeled examples. To understand this better, let's consider the following example: an equipment could have data points labeled 'F' (failed) or 'R' (runs).

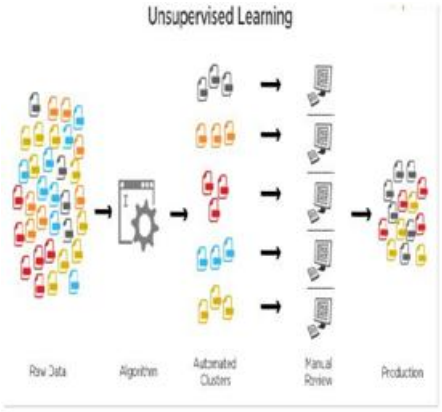
Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

The learning algorithm under supervised learning would then receive a set of inputs along with the corresponding correct output to find errors. Based on this, it would further modify the model accordingly. This is a form of pattern recognition, as supervised learning happens through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data. Supervised learning is hence more appropriate and commonly used in applications where historical data predicts future events. Examples will be: prediction of occurrences of fraudulent credit card transactions.



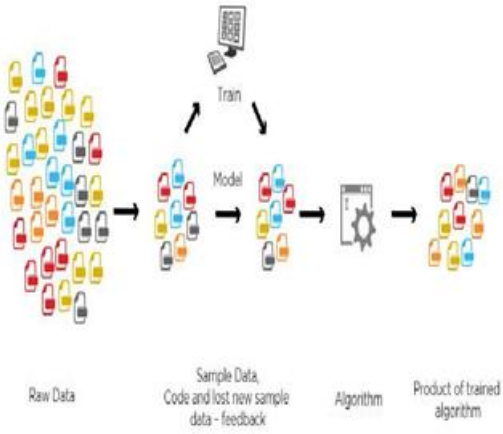
2. UNSUPERVISED LEARNING:

Unlike supervised learning, unsupervised learning is used against data that has no historical data. The algorithm must explore the surpassed data and must find the structure. This kind of learning works best in transactional data for instance, it helps in identifying customer segments and clusters with certain attributes who can be treated similarly to marketing campaigns. Popular techniques where unsupervised learning is employed, includes: self-organizing maps, nearest neighbor mapping, singular value decomposition, and k-means clustering. Basically, online recommendations, identification of data outliers, segment text topics, are all example of unsupervised learning.



3. SEMI-SUPERVISED LEARNING:

As the name suggests, semi-supervised learning is a bit of both supervised and unsupervised learning and uses both labeled and unlabeled data for training. In a typical scenario it would use small amount of labeled data with large amount of unlabeled data, the reason being that, unlabeled data is less expensive and takes less effort to acquire. This type of learning can again be used with methods such as classification, regression, and prediction. Examples of semi-supervised learning would be face and voice recognition techniques.

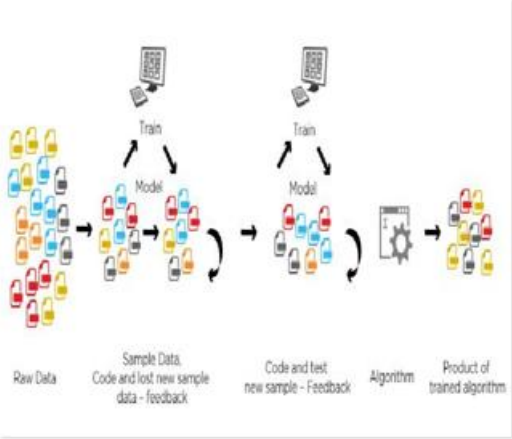


4. REINFORCEMENT LEARNING:

This is a bit like the traditional type of data analysis as the algorithm discovers through trial and error and decides which action results in greater rewards.

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Three major components can be identified in its functioning – the agent, the environment, and the actions. The agent is the learner or decision-maker, the environment includes everything that the agent interacts with, and the actions are what the agent can do.



Reinforcement learning occurs when the agent chooses actions that maximizes the expected reward over a given time. This is best achieved when the agent has a good policy in hand. Learning the best policy, hence remains to be the goal in reinforcement learning. All 3 techniques are used in this list of 10 common Machine Learning Algorithms:

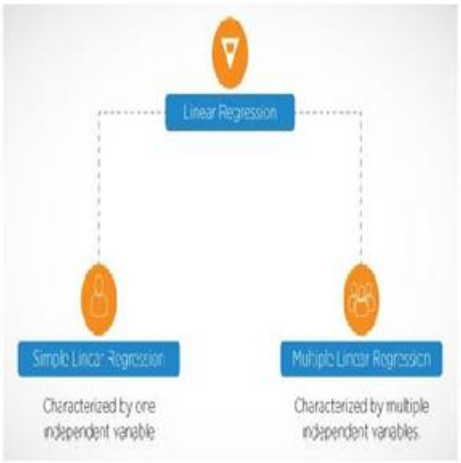
1. Linear Regression

To understand the working functionality of this algorithm, imagine how we would arrange random logs of wood in increasing order of their weight. There is a catch, however we cannot weigh each log. we must guess its weight just by looking at the height and girth of the log (visual analysis), and arrange them using a combination of these visible parameters. This is what linear regression is like. In this process, a relationship is established between independent and dependent variables by fitting them to a line. This line is known as regression line and represented by a linear equation Y= a *X + b.

In this equation:

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

The coefficients a & b are derived by minimizing the sum of the squared difference of distance between data points and the regression line.



2. Logistic Regression

Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression.

These methods listed below are often used to help improve logistic regression models:

- Include interaction terms
- Eliminate features
- Regularize techniques
- Use a non-linear model

3. Decision Tree

One of the most popular machine learning algorithms in use today, this is a supervised learning algorithm that is used for classifying problems. It works well classifying for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

4. SVM (Support Vector Machine)

SVM is a method of classification in which you plot raw data as points in an n-dimensional space (where n is the number of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data. Lines called classifiers can be used to split the data and plot them on a graph.

5. Naive Bayes

A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features are related to each other, a Naive Bayes classifier would consider all of these properties independently when calculating the probability of a particular outcome. A Naive Bayesian model is easy to build and useful for massive datasets. It's simple, and is known to outperform even highly sophisticated classification methods.

6. KNN (K- Nearest Neighbors)

This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It’s a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement. KNN can be easily understood by comparing it to real life. For example, if you want information about a person, it makes sense talk to his or her friends and colleagues! Things to consider before selecting KNN:

- KNN is computationally expensive
- Variables should be normalized, or else higher range variables can bias the algorithm
- Data still needs to be pre-processed.

7. K-Means

This is an unsupervised algorithm which solves clustering problems.

Data sets are classified into a particular number of clusters (let's call that number K) in such a way that all the data points within a cluster are homogenous, and heterogeneous from the data in other clusters.

How K-means forms clusters:

- The K-means algorithm picks k number of points, called centroids, for each cluster
- Each data point forms a cluster with the closest centroids i.e. k clusters.
- It now creates new centroids, based on the existing cluster members.
- With these new centroids, the closest distance for each data point is determined. This process is repeated until the centroids do not change.

8. Random Forest

A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is planted & grown as follows:

- If the number of cases in the training set is N, then a sample of N cases is taken at random. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<="" li="" style="box-sizing: border-box;">
- Each tree is grown to the largest extent possible. There is no pruning.

9. Dimensionality Reduction Algorithms

In today's world, vast amounts of data are being stored and analyzed by corporates, government agencies and research organizations. As a data scientist, you know that this raw data contains a lot of information - the challenge is in identifying significant patterns and variables. Dimensionality reduction algorithms like Decision Tree, Factor Analysis, Missing Value Ratio, and Random Forest can help you find relevant details.

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

10. Gradient Boosting & AdaBoost

These are boosting algorithms used when massive loads of data have to be handled in order to make predictions with high accuracy. Boosting is an ensemble learning algorithm that combines the predictive power of several base estimators to improve robustness. In short, it combines multiple weak or average predictors to a build strong predictor. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, CrowdAnalytix. These are the most preferred machine learning algorithms today. Use them along with Python and R Codes to achieve accurate outcomes.

III. Data Mining, Machine Learning and Deep Learning – The Differences Explained

Given the fact that machine learning helps in data analysis, it becomes quite unclear for learners new to the field, about the differences between data mining, machine learning and deep learning. Here's a brief explanation. To state in simple terms: machine learning and data mining use the same algorithms and techniques as data mining, except that the kind of predictions vary. While data mining discovers previously unknown patterns and knowledge, machine learning reproduces known patterns and knowledge, and further automatically apply that to data, and to decision making and actions. Deep learning on the other hand, uses advanced computing power and special types of neural networks and applies them in large amounts of data to learn, understand and identify complicated patterns. Automatic language translation, medical diagnoses are all instances of deep learning. Some Machine Learning Algorithms and Processes Although this might not make sense, now, if you have not been introduced to machine learning previously, some common machine learning algorithms and processes to familiarize oneself with, are: neural networks, decision trees, random forests, associations and sequence discovery, gradient boosting and bagging, support vector machines, self-organizing maps, k-means clustering, Bayesian networks, Gaussian mixture models and more. Other tools and processes that pair up with the best algorithms to aid in deriving the most value from big data are: comprehensive data quality and management, GUIs for building models and process flows, interactive data exploration and visualization of model results, comparisons of different machine learning models to quickly identify the best one, automated ensemble model evaluation to identify the best performers, easy model deployment so you can get repeatable, reliable results quickly, an integrated end-to-end platform for the automation of the data-to-decision process.

IV. CONCLUSION

McKinsey's report states, 'As ever more of the analog world gets digitized, our ability to learn from data by developing and testing algorithms will only become more important for what are now seen as traditional businesses.' It also quotes Google's chief economist Hal Varian who calls this "computer kaizen" and further adds by saying, "just as mass production changed the way products were assembled and continuous improvement changed how manufacturing was done," and "so continuous [and often automatic] experimentation will improve the way we optimize business processes in our organizations." It could probably be concluded that machine learning is the new avatar of big data analysis. And while Big Data has already fell off the Gartner's hype cycle, machine learning is somewhere towards the peak, in 2015. All this only emphasizes the importance of machine learning's usability in the current data-driven world. Further, it could be predicted that machine learning will evolve over the years but extinction is not a thing that will be associated with it.

V.REFERENCES

[1] T. Mitchell (1997). Machine Learning, McGraw-Hill Publishers.

[2] https://www.edx.org/course/machine-learning

[3] https://en.wikipedia.org/wiki/Machine_learning

Volume No:2, Issue No:12 (May-2017)
ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

[4] https://www.coursera.org/learn/machine-learning

[5]https://www.sas.com/en_us/insights/analytics/machine-learning.html

[6] H. Bhaskar, D. Hoyle, and S. Singh (2006). Machine Learning: a Brief Survey and Recommendations for Practitioners.