An overview of the supervised machine learning methods

DOI 10.20544/HORIZONS.B.04.1.17.P05
UDC 004.85.021:519.718

Vladimir Nasteski
Faculty of Information and Communication Technologies, Partizanska bb,
7000 Bitola, Macedonia
vladimir.nasteski@gmail.com

Abstract

In the last decade a large number of supervised learning methods have been introduced in the field of the machine learning. Supervised learning became an area for a lot of research activity in machine learning. Many of the supervised learning techniques have found application in their processing and analyzing variety of data. One of the main characteristics is that the supervised learning has the ability of annotated training data. The so called labels are class labels in the classification process. There is a variety of algorithms that are used in the supervised learning methods. This paper summarizes the fundamental aspects of couple of supervised methods. The main goal and contribution of this review paper is to present the overview of machine learning and provide machine learning techniques.

Key Words: classification, supervised, machine learning

Introduction

Machine learning represents a large field presented in information technology, statistics, probability, artificial intelligence, psychology, neurobiology and many other disciplines. With machine learning the problems can be solved simply by building a model that is a good representation of a selected dataset. Machine learning has become an advanced area from teaching the computers to mimic the human brain, and has brought the field of statistic to a broad discipline that produces fundamental statistical computational theories of the learning processes.

Machine learning is all about creating algorithms that allow the computer to learn. Learning is a process of finding statistical regularities or other patterns of data. The

machine learning algorithms [1] are created to be able to represent the human approach of learning some task. These algorithms can also represent an insight into relative difficulty of learning in different environments.

These days, the development of new computing technologies in the area of Big Data, machine learning is not like machine learning was in the past. Today, many of the machine learning algorithms have been developed [2], updated and improved and the recent development in machine learning becomes the ability to automatically apply a variety of complex mathematical calculation to a big data, which calculates the results much faster.

The adaptive programming is very popular. It is used in machine learning where the applications are capable to recognize patterns, learning from experience, abstract new information from data or optimize the accuracy and efficiency of its processing and output. Also, the machine learning techniques [7] are used to work with multidimensional data which are present in diverse amount of application areas.

So, based on the desired outcome of the algorithm, the machine learning algorithms are organized in the following groups:

- **Supervised learning** - the various algorithms generate a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn (to approximate the behavior of) a function which maps a vector into one of several classes by looking at several input-output examples of the function.
- **Unsupervised learning** - models a set of inputs: labeled examples are not available.
- **Semi-supervised learning** - combines both labeled and unlabeled examples to generate an appropriate function or classifier.
- **Reinforcement learning** - the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.
- **Transduction** - similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and new inputs.
- **Learning to learn** - where the algorithm learns its own inductive bias based on previous experience.

Besides these groups of machine learning algorithms, they are basically divided into two general groups, supervised and unsupervised learning.

In supervised algorithms, the classes are predetermined. These classes are created in a manner of finite set, defined by the human, which in practice means that a certain segment of data will be labeled with these classifications. The task of the machine learning algorithm is to find patterns and construct mathematical models. These models are then evaluated based on the predictive capacity in relation to measures of variance in the data itself.

It is also useful to make difference between two main supervised models: *classification models* (classifiers) and *regression models*. Regression models map the

input space into a real-value domain. The classifiers map the input space into pre-defined classes. There are many alternatives for representing classifiers, for instance, support vector machines, decision trees, probabilistic summaries, algebraic function, etc. Along with regression and probability estimation, classification is one of the most studied models, possibly one with the greatest practical relevance. The potential benefits of progress in classification are immense since the technique has great impact on other areas, both within Data Mining and in its applications.

On the other hand, the unsupervised learning algorithms are not provided with classifications. The main task of unsupervised learning is to automatically develop classifications labels. These algorithms are searching the similarity between pieces of data in order to determinate if they can be categorized and create a group. These groups are so called clusters, and they represent whole family of clustering machine learning techniques. In this unsupervised classification (cluster analysis) the machine doesn't know how the clusters are grouped. Using the cluster analysis, there is a bigger potential for surprising ourselves. Thus, cluster analysis is a very promising tool for the exploration of relationships between many papers.

This paper is a representation of different types of supervised machine learning algorithms and their most efficient use to make decisions more efficient and to complete the task in more optimized form. In this paper, how different algorithms give the machine different learning experience and are adopting other things from the environment will be shown, and after which the machine makes a decision and performs specialized tasks.

The paper is organized as follows: Section II paper takes us into consideration the main related work that are used for completing this paper. Section III provides the overview of the supervised machine learning process. Section IV discusses the various learning algorithms used to perform learning process.

Related work

There are many research papers and articles that give us a great overview of some of the methods and algorithms that are used in the area of machine learning.

Rich Caruana, Alexandru Niculescu-Mizil [2] present a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, logistic regression, naive Bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps.

Leonidas Akritidis and Panayiotis Bozanis [5] attempt to address interesting problem where documents remain unclassified, by introducing a machine learning algorithm which combines several parameters and meta-data of a research article.

Aurangzeb Khan et al. [6]had highlighted the important techniques and methodologies that are employed in text documents classification. The paper provides a review of the theory and methods of document classification and text mining.

Pradraig Cunningham, Matthieu Cord, and Sarah Jane Delany in their chapter "Supervised learning" provide an overview of support vector machines and nearest neighbour classifiers –probably the two most popular supervised learning techniques employed in multimedia research.

S. B. Kotsiantis [16] describes various supervised machine learning classification techniques. He also points the goal of supervised learning which is to build a concise model of the distribution of class labels in terms of predictor features.

Amanpreet Singh et al [17] are discussing about the efficacy of supervised machine learning algorithms in terms of the accuracy, speed of learning, complexity and risk of over fitting measures. The main objective of their paper is to provide a general comparison with state of art machine learning algorithms.

Background: supervised learning

The learning process in a simple machine learning model is divided into two steps: training and testing. In training process, samples in training data are taken as input in which features are learned by learning algorithm or learner and build the learning model [4]. In the testing process, learning model uses the execution engine to make the prediction for the test or production data. Tagged data is the output of learning model which gives the final prediction or classified data.

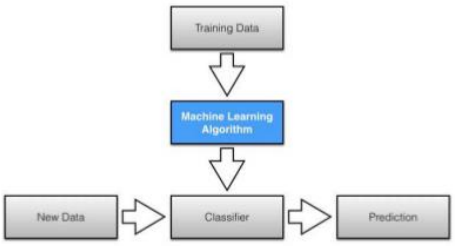

Figure 1: Supervised learning process [18]

Supervised learning (Figure 1) is the most common technique in the classification problems, since the goal is often to get the machine to learn a classification system that we've created.

Most commonly, supervised learning leaves the probability for input undefined, such as an input where the expected output is known. This process provides dataset consisting of features and labels. The main task is to construct an estimator able to predict the label of an object given by the set of features. Then, the learning algorithm receives a set of features as inputs along with the correct outputs and it learns by comparing its actual output with corrected outputs to find errors. It then modifies the model accordingly. The model that is created is not needed as long as the inputs are

available, but if some of the input values are missing, it is not possible to infer anything about the outputs.

Supervised learning is the most common technique for training for neutral networks and decision trees. Both of these are depended on the information given by the pre-determinate classification.

Also, this learning is used in applications where historical data predicts likely feature events. There are many practical examples of this learning, for instance an application that predicts the species of iris given a set of measurements of its flower.

As previously mentioned, the supervised learning tasks are divided into two categories: classification and regression. In classification, the label is discrete, while in regression, the label is continuous.

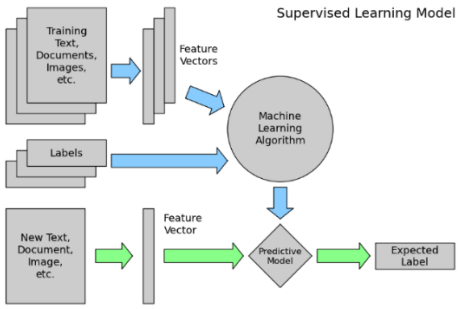

Figure 2: Supervised Learning Model [20]

As shown on Figure 2, the algorithm makes the distinction between the observed data X that is the training data, in most cases structured data given to the model during the training process. In this process, the supervised learning algorithm builds the predictive model. After its training, the fitted model would try to predict the most likely labels for a new set of samples X in the testing set. Depending on the nature of the target y, supervised learning can be classified:

- If y has values in a fixed set of categorical outcomes (integers), the task to predict y is called classification
- If y has floating point values, the task to predict y is called regression

Supervised learning algorithms

Decision trees

Decision tree [8] represents a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form so called root tree, which means that it is a distributed tree with a basic node called root with no incoming edges.

All of the other nodes have exactly one incoming edge. The node that has outgoing edges is called internal node or a test node. The rest of the nodes are called leaves. In a decision tree, each test node splits the instance space into two or more sub-spaces according to a certain discrete function of the input values. In the simplest case, each test considers a single attribute, such that the instance space is portioned according to the attribute's value. In case of numeric attributes, the condition refers to a range.

Each leaf is assigned to one class that represents the most appropriate target value. The leaf may hold a probability vector that indicates the probability of the target attribute having a certain value. The instances are classified by navigating them from the root of the tree down the leaf, according to the outcome of the tests along the path. On Figure 3 describes a simple use of the decision tree. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values.

Given this classifier, the analyst can predict the response of some potential customer and understanding the behavioral characteristics of the entire potential customers' population [9].

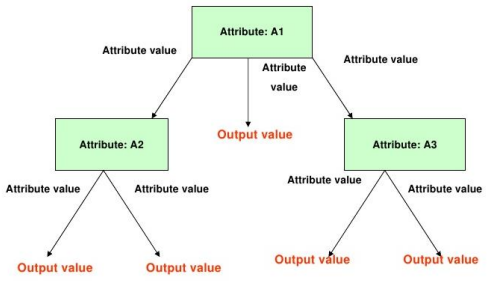

Figure 3: Decision tree example [21]

In case of numeric attributes, decision trees can be geometrically interpreted as a collections of hyper planes, each orthogonal to one of the axis. Decision-makers prefer less complex decision trees, since they may be considered more comprehensive.

Linear regression

The goal of the linear regression[1], as a part of the family of regression algorithms, is to find relationships and dependencies between variables. It represents a modeling relationship between a continuous scalar dependent variable *y* (also label or target in

[1]http://www.ess.uci.edu/~yu/class/ess210b/lecture.3.regression.all.pdf

machine learning terminology) and one or more (a D-dimensional vector) explanatory variables (also independent variables, input variables, features, observed data, observations, attributes, dimensions, data point, etc.) denoted X using a linear function. In regression analysis the goal is to predict a continuous target variable, whereas another area called classification is predicting a label from a finite set. The model for a multiple regression which involves linear combination of input variables takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + e$$

Linear regression [11] also belongs to the category of supervised learning algorithms. It means we train the model on a set of labeled data (training data) and then use the model to predict labels on unlabeled data (testing data).

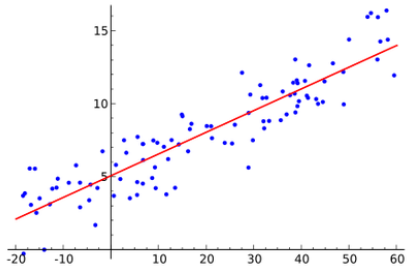

Figure 4: Visual representation of the linear regression [22]

As shown on Figure 4, the model (red line) is calculated using training data (blue points) where each point has a known label (y axis) to fit the points as accurately as possible by minimizing the value of a chosen loss function. We can then use the model to predict unknown labels (we only know x value and want to predict y value).

Naive Bayes

The Bayesian classification [15] is another method of the supervised learning methods as well as the statistical method for classification. Assumes an underlying probabilistic model and it allows capturing uncertainly about the model in a principled way by determining probabilities of the outcomes. The basic purpose of the Bayesian classification is that it can solve predictive problems.

This classification provides practical learning algorithms and can combine observed data. Bayesian classification provides useful perspective for understanding and evaluating learning algorithms. It calculates explicit probabilities for hypothesis and it robust the noise in input data.

Let's consider a general probability distribution of two values $P(x_1, x_2)$. Using Bayes rule, without loss of generality we get this equation:

$$P(x_1, x_2) = P(x_1|x_2)P(x_2)$$

Similar, if there is another class variable c, we get the next equation:

$$P(x_1, x_2|c) = P(x_1|x_2, c)P(x_2|c)$$

If the situation is generalized with two variables to a conditional independence assumption for a set of variables $x_1, \dots, x_N$ conditional on another variable c, we get the following:

$$P(x|c) = \prod_{i=i}^{N} P(x_i|c)$$

Logistic Regression

Like the naive Bayes, logistic regression [13] works by extracting some set of weighted features from the input, taking logs and combining them linearly, which means that each feature is multiplied by a weight and then added up.

The most important difference between naive Bayes and logistic regression is that the logistic regression is a discriminative classifier while the naive Bayes is a generative classifier.

Logistic regression [14] is a type of regression that predicts the probability of occurrence of an event by fitting data to a logistic function. Just as many form of regression analysis, logistic regression makes use of several predictor variables that may be numerical or categorical.

The logistic regression hypothesis is defined as:

$$h_\theta(x) = g(\theta^T x)$$

Where the function g is sigmoid function defined as:

$$g(z) = \frac{1}{1 + e^{-z}}$$

The sigmoid function has special properties that result the values in range [0,1], as visualized on Figure 5.

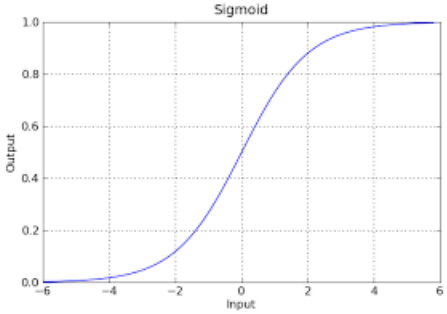

Figure 5: Visual representation of the Logistic Function [23]

The cost function for logistic regression is given as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) - (1-y^{(i)})\log(1-h_\theta\left(x^{(i)}\right))]$$

To find the minimum of this cost function, in machine learning we will use a built-in function called fmin_bfgs[2], which finds the best parameters θ for the logistic regression cost function given a fixed dataset (of x and y values). The parametars are the initial values of the parameters that need to be optimized and a function that when given the training set and a particularθ, computes the logistic regression cost and gradient with respect to θ for the dataset with x and y values. The final θ value will be used to plot the decision boundary of the training data.

Conclusion

As discussed in the paper, for the supervised learning it may be concluded that is one of the dominant methodology in machine learning. The techniques that are used are even more successful than the unsupervised techniques because the ability of labelled training data provide us clearer criteria for model optimization. The supervised learning methods contain a large set of algorithms which are improving all the time by the data scientists.

This paper provides an overview of couple of supervised learning algorithms. There is a brief explanation of the machine learning process. This paper also describes the basic structure of some various machine learning algorithms and their basic structure.

This area has the attention from many developers and has gained substantial progress in the last decade. The learning methods achieved excellent performance that would have been difficult to obtain in the previous decades. Because of the rapid progression, there is plenty of space for the developers to work or to improve the supervised learning methods and their algorithms.

References

[1] Anish Talwar, Yogesh Kumar. (2013). *Machine Learning: An artificial intelligence methodology*. In International Journal of Engineering and Computer Science.

[2]https://gist.github.com/dormantroot/4223554

[2] Rich Caruana, Alexandru Niculescu-Mizil. (2006) *An Empirical Comparison of Supervised Learning Algorithms*. In Proceeding ICML '06 Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA.
[3] Oded Maimon, Lior Rokach. (2010) *Data mining and knowledge discovery handbook - Introduction to supervised methods.* In Springer.
[4] Sandhya N. dhage, Charanjeet Kaur Raina. (2016) *A review on Machine Learning Techniques*. In International Journal on Recent and Innovation Trends in Computing and Communication, Volume 4 Issue 3.
[5] Leonidas Akritidis, Panayiotis Bozanis. (2013) *A supervised machine learning classification algorithm for research articles*. In Proceedings of the 28th Annual ACM Symposium on Applied Computing, Coimbra, Portugal.
[6] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan. (2010) *A Review of Machine Learning Algorithms for Text-Documents Classification*. In Journal of advances in information technology
[7] Taiwo Oladipupo Ayodele. (2010) *New Advances in Machine Learning - Types of Machine Learning Algorithms*. InTech.
[8] Lior Rokach, Oded Z. Maimon. (2008) *Data Mining with Decision Trees: Theory and Applications.* In World Scientific.
[9] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. (2013) *Introduction to Data Mining*. In Addison-Wesley.
[10] Pádraig Cunningham; Sarah Jane Delany (2007) *k-Nearest Neighbour Classifiers*.
[11] C.-J. Lin, R. C. Weng, S. S. Keerthi. (2008) *Trust region Newton method for large-scale logistic regression*. In Journal of Machine Learning Research, vol. 9.
[12] Theressa Hoang Diem NGO. (2012) *The steps to follow in multiple Regression Analysis.* In SAS Global forum
[13] Daniel Jurafsky & James H. Martin, (2016) *Speech and Language Processing.*
[14] Andrew Ng. (2012) *CS229 Lecture notes Machine Learning - Supervised learning.*
[15] Tom Mitchell, McGraw Hill (2015) *Machine Learning*
[16] S. B. Kotsiantis. (2007) *Supervised Machine Learning: A Review of Classification Techniques*. In Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, The Netherlands.
[17] Amanpreet Singh, Narina Thakur, Aakanksha Sharma (2016). *A review of supervised machine learning algorithms.* In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on.
[18] http://www.slideshare.net/GirishKhanzode/supervised-learning-52218215
[19] https://www.researchgate.net/publication/221907660_Types_of_Machine_Learning_Algorithms
[20] http://radimrehurek.com/data_science_python/
[21] http://www.slideshare.net/cnu/machine-learning-lecture-3

[22] http://gerardnico.com/wiki/data_mining/simple_regression
[23] http://aimotion.blogspot.mk/2011/11/machine-learning-with-python-logistic.html