Evolution of Machine Learning Review Paper

Ranjan Baro

October 30, 2023

Abstract

Machine learning, a pivotal sub-field of artificial intelligence, traces its roots back to the 1950s. It wasn't until the 1990s that renewed interest and substantial advancements propelled this science into the forefront of technological innovation. This technology will improve more in future. The reason behind this development is the difficulty of analysing and processing the rapidly increasing data. Machine learning is based on the principle of finding the best model for the new data among the previous data thanks to this increasing data. As data volumes continue to surge, the trajectory of machine learning research is anticipated to surge, shaping a dynamic and promising future.In practical applications, machine learning algorithms have been indispensable in data mining, showcasing their superior performance. This paper offers a comprehensive exposition on machine learning algorithms, encompassing feature selection methodologies, dimensionality reduction techniques, and strategies for culling extraneous data.

1 What is Machine Learning ?

Learning, as defined by Simon, is a dynamic process of behavioral enhancement through the assimilation of new information over time. When this process is undertaken by machines, it is referred to as machine learning. Instead of relying on explicit programming, machine learning algorithms are designed to analyze and learn from data, identify patterns, and make informed decisions or predictions. With the advent of information technology advancements, vast amount of data accumulated that surpass the capabilities of traditional database analysis techniques. These data reservoirs are sourced from a multitude of platforms including internet applications, ATMs, and credit card terminals etc. The information collected by this way is waiting to be analysed. The aim of analysing the data collected in different fields change in accordance with their needs. Machine learning applications are used in some fields like natural language processing, image processing and computer vision, speech and handwriting recognition, automotive, aviation, production, generation of energy, calculated finance and biology.

The main aim of machine learning is to create models which can train themselves to improve, perceive the complex patterns, and find solutions to the new problems by using the previous data.

2 Types of Machine Learning Methods

There are different types of machine learning algorithms, including:

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

2.1 Supervised Learning

Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset. In a labeled dataset, each input example is associated with a corresponding target or outcome. The goal of supervised learning is to learn a mapping or relationship between the input data and the

corresponding outputs, so that the algorithm can make accurate predictions or decisions when given new, unseen data.

Supervised learning can be further divided into two main categories based on the nature of the target variable:

- **Regression :** Predicting continuous target value based on its independent features of the data. For example, predicting house prices based on features like square footage, number of bedrooms, etc., is a regression problem.
- **Classification :** Distributing the data into the categories defined on the data set according to their specific features. Examples include classifying emails as spam or not spam, or identifying handwritten digits as 0 through 9.

2.2 Unsupervised Learning

The main difference between supervised vs unsupervised learning is the need for labelled training data. Supervised machine learning relies on labelled input and output training data, whereas unsupervised learning processes unlabelled or raw data. The learning process of unsupervised learning occurs by using the relations and connections between the data. Two types of unsupervised learning can be found, they are:

- **Clustering :** Clustering is a technique in unsupervised learning that involves grouping together similar data points based on certain characteristics or features. The goal of clustering is to discover natural groupings or clusters within a dataset, where data points within the same cluster are more similar to each other than they are to data points in other clusters.
- **Association :** Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It is based on different rules to discover the interesting relations between variables in the database.

2.3 Semi-Supervised Learning

Semi-supervised learning falls in between supervised and unsupervised learning. This learning method uses a small amount of labeled data and a large amount of unlabeled data to train a model. The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning. However, unlike supervised learning, the algorithm is trained on a dataset that contains both labeled and unlabeled data.

2.4 Reinforcement Learning

Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In this learning method, the agents learn via reward system. The main objective of the agent in reinforcement learning is to learn an optimal policy, which maximizes the expected cumulative rewards over time. This can be achieved through various algorithms, including Q-Learning, Deep Q-Networks (DQN), Policy Gradients, and more. Reinforcement learning has found applications in areas such as autonomous robotics, game playing (e.g., AlphaGo), recommendation systems, and control systems for tasks like self-driving cars.

3 K-Nearest Neighbor

The K Nearest Neighbors algorithm (known as k-NN) is a simple yet powerful supervised machine learning algorithm used for both classification and regression tasks. In k-NN, the "k" stands for the number of nearest neighbors that the algorithm considers when making a prediction for a new data point. k-NN does not involve a training phase in the traditional sense. Instead, it stores the entire dataset and uses it directly during prediction.

3.1 How does K-NN work ?

The K-NN working can be explained on the basis of the below algorithm:

Step 1 : Select the number K of the neighbors

Step 1 : Calculate the Euclidean distance of K number of neighbors

Step 1 : Take the K nearest neighbors as per the calculated Euclidean distance.

Step 1 : Among these k neighbors, count the number of the data points in each category.

Step 1 : Assign the new data points to that category for which the number of the neighbor is maximum.

Step 1 : Our model is ready.

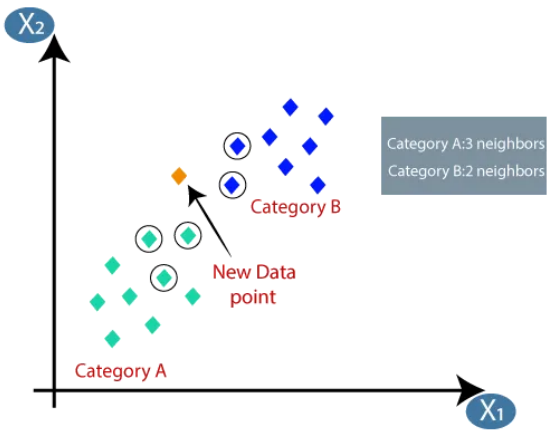

Figure 1: classification using k-NN when k = 5

3.2 Calculating Distance

The first step of k-NN algorithm is to calculate the distance between the new point and each training point. There are various methods for calculating this distance, of which the most commonly known methods are — Euclidean, Manhattan, Hamming distance, Minkowski distance etc.

- **Euclidean Distance :** Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y).

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{1}$$

- **Manhattan Distance :** This is the distance between real vectors using the sum of their absolute difference.

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \tag{2}$$

- **Hamming Distance :** It is used for categorical variables. If the value (x) and the value (y) are the same, the distance D will be equal to 0 . Otherwise D = 1.

$$D = \begin{cases} 1 & \text{if } x_i \neq x_j \\ 0 & \text{if } x_i = x_j \end{cases} \tag{3}$$

4 Regression

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. In regression, the focus is on estimating the parameters of a mathematical model that best describes the relationship between the independent variables (input features) and the dependent variable (target variable). The goal is to find a function that can accurately predict the target variable for new, unseen data.

There are two main types of regression:

- **Simple Linear Regression :** This type of regression involves one independent variable and one dependent variable. The relationship between the variables is assumed to be linear, meaning it can be approximated by a straight line. The model equation is typically represented as :

$$y = b + Wx$$

  where b and W are the regression coefficients for the intercept and slope,

- **Multiple Linear Regression :** This type of regression involves multiple independent variables and one dependent variable. The relationship between the variables is assumed to be a linear combination. The model equation is represented as :

$$y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

4.1 How to calculate Weight ?

The following table demonstrates a set of pairs x including the number of years of experience of college graduates and Y is their salary.

x years experience	y salary (in \$1000s)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Table 1: Training Dataset.

Regression coefficients can be estimated using the following equation.

$$W = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \tag{4}$$

Where $w_0 = \bar{y} - w_1\bar{x}$
Here $\bar{x}$ average $x_1, x_2, x_3, \ldots, x_{|D|}$ and $\bar{y}$ average $y_1, y_2, y_3, \ldots, y_{|D|}$ and coefficients w_0, w_1. Linear regression uses least squares methods.

The graph belongs to the table is given in below:

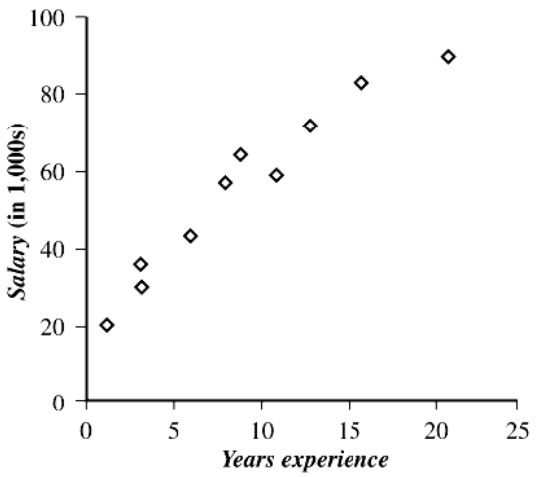

Figure 2: Years Experience vs Salary

The above graph shows relationship between X and Y.The model shows above relationship with $y = w_0 + w_1 x$ equation. According to above information we calculated $x = 9.1$ and $y = 55.4$. Substitution of these values in equations (4), gives us the following values.

$$w_1 = \frac{(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\cdots+(16-9.1)(83-55.4)}{(3-9.1)^2+(8-9.1)^2+\cdots+(16-9.1)^2} = 3.5 w_0 = 55.4-(3.5)(9.1) = 23.6 \tag{5}$$

Therefore, the linear equation of least squares with high values was achieved.

$$Y = 23.6 + 3.5X$$

Using this equation, we can predict the relationship between salary and experience. For example, one with 10 years of experience should make salary of $58,600.

5 Conclusions

The process of machine learning is divided into two stages: constructing the base model and optimizing parameter settings. Some algorithms integrate these stages for enhanced efficiency. The statement emphasizes the importance of data division for training and validation, and notes that a larger training dataset generally leads to improved model performance. Striking a balance between data quantity and quality is identified as a key challenge. Techniques like feature selection, dimensionality reduction, and data curation are discussed, with a systematic evaluation of their purposes, advantages, and disadvantages. This approach aids in selecting the most appropriate method for specific scientific domains based on algorithmic expressions and methodological considerations.

References

(1) Özer Çelik *A Research on Machine Learning Methods and Its Applications* Eskisehir Osmangazi University

(2) Ali Heydarzadegan & Yaser Nemati & Mohsen Moradi *Evaluation of Machine Learning Algorithms in Artificial Intelligence* Department of Computer Engineering, Firoozabad Branch, Islamic Azad University, Firoozabad, Iran

(3) V. Khodadadi et al. *Application Of Ants Colony System For Bankruptcy Prediction Of Companies Listed In Tehran Stock Exchange* , Business Intelligence Journal, 2010.

(4) A.Aziz and A.Humayon A predicting corporate Bankruptcy:weither do we stand? Department of Economics,Loughborough University, UK,2002.

(5) Q.Yu. Machine Learning for Corporate Bankruptcy Prediction. Information and Computer Science Department, Aalto University, 2013.

(6) Y.Chiang , et al . A Hybrid Approach Of Dea, Rough Set And Support Vector Machines For Business Failure Prediction, Expert Systems With Applications, 2010.

(7) J. Bellovary et al. A review of bankruptcy prediction studies: 1930 to present. Journal of Financial Education, 33:87–114 , 2007.

(8) J. de Andrés, M. Landajo, and P. Lorca. Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios. Knowledge-Based Systems, 30:67–77, 2012.

(9) Ming-Yuan Leon Li and Peter Miu. A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach. Empirical Finance, 17:818–833, 2010.

(10) Özkan, H. (2013). K-Means Kümeleme ve K-NN Sınıflandırma Algoritmalarının Öğrenci Notları ve Hastalık Verilerine Uygulanması Bitirme Tezi, İstanbul Teknik Üniversitesi, İstanbul.