Machine Learning Algorithms

Subjects: Engineering, Electrical & Electronic

Contributor: Mahtab Kokabi , Muhammad Nabeel Tahir , Darshan Singh , Mehdi Javanmard

Machine learning (ML) has grown rapidly over the past few decades and has widely used applications not only limited to healthcare problems, such as predicting drug discoveries and diagnosing diseases, but also in other fields, such as mechanics, robotics, and image recognition. In simple words, ML is a rapidly developing field of computational algorithms that aims to replicate human intelligence by adapting to their surroundings and learning from them.

biosensors impedance cytometry lab-on-a-chip cancer detection machine learning

1. Introduction

Machine learning (ML) has grown rapidly over the past few decades and has widely used applications not only limited to healthcare problems, such as predicting drug discoveries and diagnosing diseases, but also in other fields, such as mechanics, robotics, and image recognition [1][2][3][4][5]. In simple words, ML is a rapidly developing field of computational algorithms that aims to replicate human intelligence by adapting to their surroundings and learning from them [6]. There are two main types of machine learning algorithms: supervised and unsupervised learning [7]. The difference between these two main classes is the existence of labels in the training data subset, which will be discussed in the following sections.

2. Supervised Machine Learning

Supervised algorithms are a subset of machine learning models which generate a function that maps inputs to desired outputs [8]. Supervised learning is characterized by the usage of labeled datasets to train algorithms for accurate classification or outcome prediction. The model adjusts its weights as input data is fed into it, achieving proper fitting during the cross-validation process [9]. During the model training process, the predicted output is compared to the actual output, and modifications are made to decrease the overall error between the two. Supervised machine learning algorithms have a broad range of applications in biosensors and healthcare, including tasks such as distinguishing cancer from non-cancer cells, detecting circulating tumor cells (CTCs), and predicting DNA quantities [2][9][10]. In the following sections, the most well-known and commonly supervised algorithms will be discussed.

2.1. Support Vector Machines (SVMs)

The support vector machine algorithm is a popular supervised algorithm used both in classification and regression models [11]. In classification, the SVM aims to identify a hyperplane in an N-dimensional feature space, which effectively separates the data points into distinct classes, while, in regression models, the SVM aims to find a line that best fits the data [12]. The kernel-based SVM algorithm uses kernel functions to transform the input data into a higher dimensional feature space when the data cannot be separated linearly. The performance of the SVM model depends on two hyperparameters: kernel parameters and kernel types. The selection of the kernel type is determined based on the characteristics of the input data [13].

2.2. K-Nearest Neighbor (KNN)

The k-nearest neighbor (KNN) algorithm is a type of supervised machine learning algorithm that classifies objects based on the classes of their nearest neighbors [14]. It is typically used for classification but can also be applied to regression problems. The algorithm predicts the class or value of a new data point based on the k-closest data points in the training dataset. To identify the nearest neighbors, the algorithm calculates the distance between the new data point and all other data points in the dataset. For classification, the algorithm assigns the new data point to the most common class among its k-nearest neighbors, while for regression analysis, it calculates the average value of the k-nearest neighbors and assigns it to the new data point [14]. The value of k is usually determined through cross-validation or other optimization techniques, and it impacts the bias-variance trade-off of the model. Despite its simplicity, KNN is a highly effective algorithm and is widely used in many fields, including image recognition, natural language processing, and healthcare problems [15][16][17].

2.3. Decision Tree (DT)

The decision tree algorithm is a popular supervised machine learning algorithm used for classification and regression tasks. It works by constructing a tree-like model of decisions and their possible consequences based on the data [18]. The decision tree algorithm works by dividing the feature space of the training set recursively. Its goal is to identify a collection of decision rules that can partition the feature space in a way that produces a reliable and informative hierarchical classification model. In this algorithm, each node represents an attribute or feature, and each branch represents an outcome. The root node represents the entire dataset, and at each internal node, the algorithm divides the data based on a specific attribute's value. This process is repeated recursively until a stopping condition is met, such as achieving a specified level of purity or reaching a predetermined depth [18].

2.4. Gaussian Naïve Bayes (GNB)

The Gaussian naïve Bayes (GNB) algorithm is a classification technique used in machine learning that leverages a probabilistic approach and the Gaussian distribution to make predictions of input data. GNB treats each attribute variable as independent, enabling it to be trained efficiently in supervised learning and used in complex real-world scenarios. GNB is particularly effective when dealing with high-dimensional data since it assumes independence between features, making it less susceptible to the curse of dimensionality [19].

2.5. Logistic Regression (LR)

Logistic regression is a supervised machine learning algorithm designed to solve classification problems where the target variable is categorical. The primary objective of logistic regression is to establish a mapping function from the dataset's features to the target. This allows the algorithm to predict the probability of a new data point belonging to a particular class [20]. Logistic regression is a widely used algorithm in many fields, such as marketing, healthcare, and finance, as it can help identify patterns and relationships between variables that can assist in making accurate predictions and decisions [21].

2.6. Random Forest (RF)

Random forest is a supervised machine learning algorithm that builds on the concept of tree classifiers. It generates a large number of classification trees and uses them to classify new feature vectors. Each tree in the forest classifies the input vector, and the tree's classification is counted as a "vote" for that class. The forest then chooses the classification with the highest number of votes across all the trees in the forest as the final prediction. RF is a highly effective algorithm for handling complex, high-dimensional datasets. It uses ensemble learning to reduce overfitting and improve the model's accuracy by combining the outputs of multiple decision trees [22].

2.7. Artificial Neural Network (ANN)

Artificial neural networks (ANNs) are computer programs designed to mimic the way the human brain processes information. They derive their inspiration from biological neural networks and adopt a similar structure of interconnected neurons to perform complex tasks. ANNs acquire knowledge through experience by identifying patterns and relationships in data instead of relying on explicit programming to accomplish the task.

An ANN typically consists of many processing elements (PE), also known as artificial neurons, which are connected by weights. These weights constitute the neural structure of the network and are organized into layers. The structure of an ANN. Through a process of training, the network learns to adjust the weights between the neurons to produce the desired output given a specific input. ANNs can be used for a variety of tasks, such as image and speech recognition, natural language processing, predictive analytics, and healthcare [23].

3. Unsupervised Machine Learning

Unsupervised learning is a subfield of machine learning where the data provided to the machine learning algorithm is unlabeled, and it is up to the algorithm to make sense of the data on its own. In unsupervised learning, the algorithm looks for patterns and structures in the data and tries to group similar data points together based on their similarities or differences. One of the key advantages of unsupervised learning is that it can reveal insights and relationships that may not be immediately apparent to human observers. By discovering patterns and similarities in the data, unsupervised learning can help uncover the hidden relationships that can be useful for making decisions or solving problems. For example, unsupervised machine learning can be used to identify customer segments in a marketing dataset or to find anomalies or outliers in a dataset that may indicate fraudulent activity [12].

4. Machine Learning Figures of Merits

To evaluate the performance of the representative model, the following metrics are used: accuracy (ACC), true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), and false positive rate (FPR). These measures are computed using the following forms:

$$\text{Accuracy (ACC)} = \frac{TP + TN}{TN + TP + FN + FP} \quad (1)$$

$$\text{Sensitivity (TRP)} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity (TNR)} = \frac{TN}{TN + FP} \quad (3)$$

$$\text{Fallout (FPR)} = \frac{FP}{TN + FP} \quad (4)$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{TP + FN} \quad (5)$$

where the TP's and FPs refer to the number of correct and incorrect predictions of outcomes to be in the considered output class, whereas the TN's and FNs refer to the number of correct and incorrect predictions of outcomes to be in any other output classes respectively [1].

The ROC (receiver operating characteristic) curve is a graphical representation of the performance of a binary classification model. It is a graph that shows the trade-off between the true positive rate and the false positive rate. A diagonal line in the ROC curve indicates that the test has no discriminatory ability, while an ROC curve above the diagonal line indicates a test with better-than-chance discrimination ability. The area under the ROC curve (AUC) is a measure of the overall ability of the test to discriminate between the presence or absence of a condition. An AUC of 1.0 indicates perfect discrimination, and an AUC of 0.5 indicates no discriminatory ability [24].

References

1. Kokabi, M.; Donnelly, M.; Xu, G. Benchmarking Small-Dataset Structure-Activity-Relationship Models for Prediction of Wnt Signaling Inhibition. IEEE Access 2020, 8, 228831–228840.

2. Kokabi, M.; Sui, J.; Gandotra, N.; Khamseh, A.P.; Scharfe, C.; Javanmard, M. Nucleic Acid Quantification by Multi-Frequency Impedance Cytometry and Machine Learning. Biosensors 2023, 13, 316.
3. Kokabi, H.; Najafi, M.; Jazayeri, S.A.; Jahanian, O. Performance optimization of RCCI engines running on landfill gas, propane and hydrogen through the deep neural network and genetic algorithm. Sustain. Energy Technol. Assess. 2023, 56, 103045.
4. Fujiyoshi, H.; Hirakawa, T.; Yamashita, T. Deep learning-based image recognition for autonomous driving. IATSS Res. 2019, 43, 244–252.
5. Varkonyi, A.; Mosavi, A. Learning in Robotics. Int. J. Comput. Appl. 2017, 157, 8–11.
6. El Naqa, I.; Murphy, M.J. What Is Machine Learning? Springer: Berlin/Heidelberg, Germany, 2015.
7. Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. A systematic review on supervised and unsupervised machine learning algorithms for data science. In Supervised and Unsupervised Learning for Data Science; Springer: Berlin/Heidelberg, Germany, 2020; pp. 3–21.
8. Nasteski, V. An overview of the supervised machine learning methods. Horizons B 2017, 4, 51–62.
9. Berry, M.W.; Mohamed, A.; Yap, B.W. Supervised and Unsupervised Learning for Data Science; Springer: Berlin/Heidelberg, Germany, 2019.
10. Poellmann, M.J.; Bu, J.; Liu, S.; Wang, A.Z.; Seyedin, S.N.; Chandrasekharan, C.; Hong, H.; Kim, Y.; Caster, J.M.; Hong, S. Nanotechnology and machine learning enable circulating tumor cells as a reliable biomarker for radiotherapy responses of gastrointestinal cancer patients. Biosens. Bioelectron. 2023, 226, 115117.
11. Noble, W.S. What is a support vector machine? Nat. Biotechnol. 2006, 24, 1565–1567.
12. Raji, H.; Tayyab, M.; Sui, J.; Mahmoodi, S.R.; Javanmard, M. Biosensors and machine learning for enhanced detection, stratification, and classification of cells: A review. Biomed. Microdevices 2022, 24, 26.
13. Cui, F.; Yue, Y.; Zhang, Y.; Zhang, Z.; Zhou, H.S. Advancing biosensors with machine learning. ACS Sens. 2020, 5, 3346–3364.
14. Sun, S.; Huang, R. An adaptive k-nearest neighbor algorithm. In Proceedings of the 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, 10–12 August 2010; IEEE: Piscataway, NJ, USA, 2010; Volume 1, pp. 91–94.
15. Laaksonen, J.; Oja, E. Classification with learning k-nearest neighbors. In Proceedings of the International Conference on Neural Networks (ICNN'96), Washington, DC, USA, 3–6 June 1996; IEEE: Piscataway, NJ, USA, 1996; Volume 3, pp. 1480–1483.

16. Ozaki, K.; Shimbo, M.; Komachi, M.; Matsumoto, Y. Using the mutual k-nearest neighbor graphs for semi-supervised classification on natural language data. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Portland, OR, USA, 23–24 June 2011; pp. 154–162.
17. Khateeb, N.; Usman, M. Efficient heart disease prediction system using K-nearest neighbor classification technique. In Proceedings of the International Conference on Big Data and Internet of Thing, London, UK, 20–22 December 2017; pp. 21–26.
18. Myles, A.J.; Feudale, R.N.; Liu, Y.; Woody, N.A.; Brown, S.D. An introduction to decision tree modeling. J. Chemom. 2004, 18, 275–285.
19. Jahromi, A.H.; Taheri, M. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. In Proceedings of the 2017 Artificial Intelligence and Signal Processing Conference (AISP), Shiraz, Iran, 25–27 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 209–212.
20. Bisong, E.; Bisong, E. Logistic regression. In Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners; Springer: Berlin/Heidelberg, Germany, 2019; pp. 243–250.
21. Goswami, M.; Sebastian, N.J. Performance Analysis of Logistic Regression, KNN, SVM, Naïve Bayes Classifier for Healthcare Application during COVID-19. In Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2021; Springer: Singapore, 2022; pp. 645–658.
22. Kulkarni, A.D.; Lowe, B. Random Forest Algorithm for Land Cover Classification. 2016. Available online: https://scholarworks.uttyler.edu/compsci_fac/1/ (accessed on 6 May 2023).
23. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. 2000, 22, 717–727.
24. Hoo, Z.H.; Candlish, J.; Teare, D. What is an ROC curve? Emerg. Med. J. 2017, 34, 357–359.