Analysis of Different Statistical Indicators for Machine Learning

Tiexuan Zhu *

Chongging Bashu Secondary School, Chongqing, China

* Corresponding Author Email: tiexuan.zhu.biec@bashuschool.cn

Abstract. As a matter of fact, machine learning has been widely adopted in various fields on account of rapid development of computing ability in recent years. In reality, with advances in computer hardware and the explosion of data, machine learning has become an ideal choice for processing this data. With this in mind, processing such large amounts of data requires a lot of computing power and algorithms, and also offers a wide range of applications for machine learning. On this basis, this article mainly considers the formulation and application of statistical measures from his three aspects: classifier, regression, and clustering. To be specific, different metrics are presented and real examples are discussed in the various situation in detail. In addition, current application scenarios and limitations of machine learning are presented. At the same time, the prospects for further study are proposed. Overall, these results shed light on guiding further exploration of machine learning.

Keywords: Machine learning; statistic; evaluation metrics.

1. Introduction

From a temporal perspective, machine learning uses the past to predict the future. From the perspective of information flow processing, machine learning is to compress and extract information according to certain rules. From the perspective of neuron parameters, the process of machine learning is to establish connections between neurons. Patterns that appear repeatedly in learning samples will establish associations with heavy weights, and patterns that appear rarely will have associations with small weights. From an academic point of view, machine learning is the generalization of spatial search and functions. From the perspective of application: machine learning can be roughly interpreted as data mining + artificial intelligence. From a philosophical point of view, machine learning is "the process of recreating human understanding of the world.

Machine learning automatically analyses models from data and uses them to make predictions about unknown data. Machine learning also illustrates the importance and necessity of interdisciplinarity. The origins of machine learning in the modern sense are linked to the name of Frank Rosenblatt, a psychologist at Cornell University, who created a group that built a machine called the Perceptron. This machine is the prototype of the modern artificial neural network (ANN) [1]. As soon as electronic computers came into use in the 1950s and 1960s, they developed algorithms that could model and analyze large amounts of data [2]. Currently, there are three main areas of machine learning: supervised learning (based on unsupervised learning) and reinforcement learning (based on reinforcement learning). The advent of the big data era has improved the standards of data processing. Additionally, the industry's needs and concerns have changed significantly as companies' focus has shifted to data and the computing sector has evolved into a true information industry.

Computing speed will shift to focusing on big data processing capabilities, and software will also shift from programming to data processing [3]. Machine learning helps metamaterials and metasurfaces in forward spectral prediction and reverse structure design, predicts the phase response of anisotropic digital coding units, and quickly screens candidate structures for metasurface holographic imaging, etc. In this way, the combination of machine learning algorithms and the propagation mechanism of metamaterials improves the accuracy of machine learning algorithms and further promotes the development of new mechanisms of metamaterials [4]. The application of machine learning in landslide disasters has promoted the development of disaster prevention and control in the direction of intelligence. By analyzing the internal and external factors that affect landslides to assess the level of danger, efficiently solve problems and prevent them in advance [5]. Different machine learning algorithms are also used in environmental microbial research to summarize patterns from microbial big data and create models to identify, classify and predict new data [6].

In order to truly understand machine learning and lay a foundation for future cross-industry cooperation, this paper aims to analyze different statistical indicators of machine learning. The aim of this study is to gain a comprehensive understanding of the different statistical metrics used to analyze machine learning. Specifically, this study aims to:

- Introduce the types of machine learning and some commonly used models.
- Introduce the formulas of Classifier, Regression and Clustering, and some results of specific application.
- Point out the current limitations in this field.
- Think about the future.

Overall, the aim of this research is to better understand the different statistical metrics of machine learning and provide insights into how to solve existing problems.

2. Analysis of Metrics

2.1. Basic Descriptions

Machine learning can be classified into regression, classifier, and clustering according to task type. Regression is the use of regression analysis technology in mathematical statistics to determine the dependence between more than two variables. There are some Types of Regression Analysis Techniques: Linear Regression, Logistic Regression, Ridge Regression, Lasso Regression, Polynomial Regression, Bayesian Linear Regression. Classifier is the most common class of tasks in machine learning, such as image classification, text classification, etc. Major classification techniques have decision tree induction, Bayesian Network (BN), K-nearest neighbor (KNN), Support Vector Machines, etc. [7]. Cluster is also called group analysis. The goal is to divide the sample into closely related subsets or clusters. Simply put, it is hoped to use the model to gather sample data into several categories, and common applications: market analysis, community analysis, etc.

2.2. Classifier

Classifier is a trick school algorithm designed to assign class labels to input. Machine learning can be evaluated in terms of accuracy, precision, recall, F1 score, ROC curve, PR curve and AUC. In this method, pre-defined binary classifiers (PBCs) are connected in a linear fashion, with each label prediction becoming a property of the other PBCs. These techniques have demonstrated their scalability and robustness. They have achieved the best results in real life situations across many datasets and evaluation measures for multiple categories. A classifier is a classification decision function or classification model obtained by machine learning. Classify generally means you get a function that you enter a value into and get a yes or no result to classify. The accuracy_score task computes the precision of the fit expectation as a division (default) or number (normalize=False). This work returns subset accuracy for multi-label classification. If the total set of names expected by the test is exactly the same as the actual set of names, the subset accuracy is 1.0, otherwise, it is 0.0.

If $\hat{y}_i$ represents the predicted value of the i-th sample, and y_i represents the actual value, then the percentage of valid predictions is as follows:

$$accuracy = \frac{\Sigma_i 1(\hat{y}_i = y_i)}{n} \tag{1}$$

Precision answers the question that when the model predicted the positive class, what percentage of the predictions were correct:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2}$$

Recall answers the question what the actual percentage of positive examples is correctly identified:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

ROC analysis investigates and employs the relationship between sensitivity and specificity of a binary classifier [8].

An artificial classifier's score was calculated for two classes based on a random sample of 25 points from a Gaussian distribution with a mean of 0 a standard deviation of 2 and a unit variance. On the graph, the raw scores are presented on the x-axis, as well as a normalized histogram derived from the standard deviation using a uniform five-bin discretization method. The dashed line on the right represents the result of the comparison of scores and the selection of a specific point. The smoothed curve presented in the histogram is the expected performance of a process divided into five components. The dotted line indicates the expected performance associated with a particular distribution. It is an expectation that a (consistently) randomly selected positive score will be higher than a randomly selected negative score [8].

The MCC is a way to measure how well a prediction does in different categories. It gives a good score only when the prediction does well in all four categories. The score changes depending on how many positive and negative things are in the data set. In addition, precision is not applicable when there are multiple names in the dataset. In any event, when the data set is heterogeneous (i.e., there are many more tests in one class than in the other classes), accuracy is not a dependable measure. F1-Score may be a more complete assessment file compared to ACC. In order to fully necessitate a show preparing handle to achieve a far better; much better; higher; stronger; model impact, F1-score is preferred. When there is no relevant perplexity framework advantage, one recommends that peruses evaluate their twofold assessment by analyzing their PR AUC (procedural error rate) and ROC.

2.3. Regression

Regression is the process by which a data set (x) is approximated to its true value (y1) by a functional relationship (y2 = g(x)). Due to the fact that the fitting cannot be perfect, there may be errors. The error value for this regression is Y2-Y1, i.e., the fitted value is less than the true value. Regression is commonly measured by R^2 score, mean absolute error, and mean squared error, and mean squared logarithmic error. R2-score indicates the magnitude of the change (y) that has been elucidated by the model’s free factors. The magnitude of the clarified fluctuation provides an indication of the degree to which the test is expected to be well-fit and in this manner an indication of the extent to which the show is likely to be well-understood. R2 score is not restricted to the situation in which the true objective is consistent. It can be expressed as either a NaN value (idealized expectations) or a -Inf value (false predictions). The formulae are as follows:

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y_i})^2}{\Sigma(y_i - \bar{y})^2} \tag{4}$$

The mean_absolute_error (MAE) is a helpful way to measure how well a model performs [12]. It is commonly used in evaluating models. The MAE function calculates the average absolute difference between the predicted value and the actual value. It is a measure of how far off the predictions are from the real values. According to our experiment, it is necessary to round the decimal predicted scores to the proper integers, as this not only reduces the MAE values, but also improves the precision and quality of recommendation. Few preliminary works have observed this fact [9,10]. The average squared error function calculates the risk metric of the expected squared error or loss value using the mean square error function. The MAE values for distinct preparing set proportions between 0.1 and 0.9 are presented in Fig. 1 for the MovieLens dataset and EachMovie dataset. As indicated in Fig. 1, the MAE values decrease significantly when the anticipated evaluations are adjusted using the Essential Circular approach for all distinct settings. This demonstrates that the adjustment of the anticipated rating can reduce the MAE value for all tested suggestion calculations. The decrease of the MAE values for EachMovie information set is particularly pronounced.

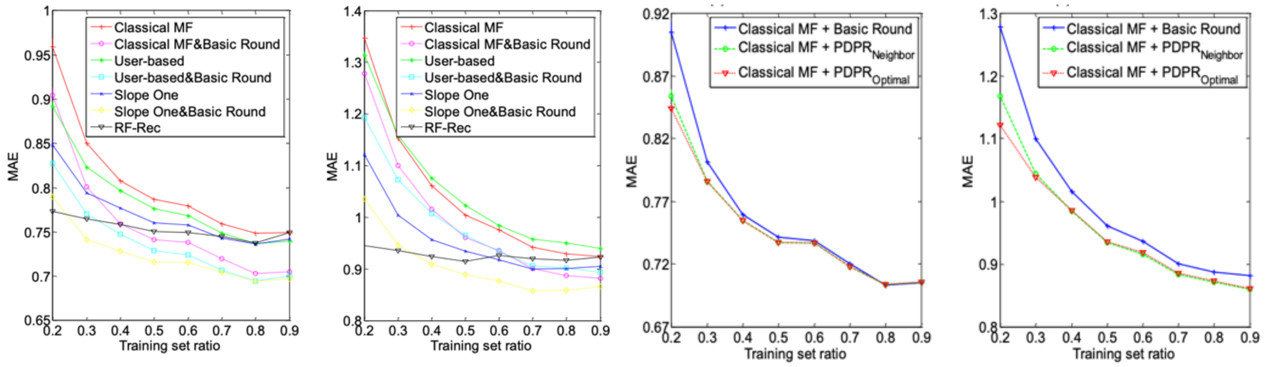

Fig. 1 A typical sketch of MAE evaluation.

This risk-measurement function predicts the average value of the squared logarithmic error or loss of the means of the logarithmic logarithm function. The means_squared_error function computes mean squared error, a risk metric that corresponds to the expected value of squared (squared) error or loss:

$$MSE = \frac{1}{n}\Sigma(y_i - \widehat{y_i})^2 \tag{5}$$

2.4. Clustering

Clustering can be understood as an unsupervised learning process in which the test check is unclear, the test is divided into several disjoint subgroups agreeing on a particular standard or the inherent law of the information; each subgroup is referred to as a cluster; each cluster contains at least one protest; each protest has a corresponding place to and a corresponding place to a cluster; the information similarity within clusters is high, but the information similarity between clusters is low. Cluster can also be employed to prepare for other forms of learning, such as categorization. V-measure is an entropy-based degree which unequivocally measures how effectively the criteria of homogeneity and completeness have been satisfied [11]:

$$v = \frac{(1+\beta)Homegenity \times completeness}{\beta \times homegenity + completeness} \tag{6}$$

In order to ensure consistency within a cluster, one only includes data that is part of the same group within that cluster. Completeness is the number of identical samples in the same group. If all identical samples are included in the same group, the integrity value is 1. In contrast, if the identical samples are divided into different groups (i.e., subgroupings), the integrity value is less than 1. Conditional empirical entropy (CE) is calculated by the following equation: H(k|C). Homogeneity, completeness and V-measure can boundeded score but can notnormalized with regards to random labeling. Examples of matching problems are presented in Fig. 2. For the purposes of this discussion, the matching problem is represented by the F value. However, these problems also relate to actions that necessitate the mapping of clusters to classes for assessment. The Fowlkes-Mallows index is a tool you can use if you know the correct class assignments for the samples. The Fawkes-Mallows score (FMI) is a number determined by calculating the average of precision and recall:

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{7}$$

One of the advantages of random label assignment is that it can have a near-zero FMI score, i.e., 0 with a maximum value of 1, since there are no pre-determined cluster structures. This allows us to compare two different algorithms for grouping. The first algorithm pre-determines that the clusters are round in shape, whereas the second algorithm can find clusters with folded paper-like shapes. However, FMI based measures require information about the actual classes, which are often not available or must be assigned manually by humans.

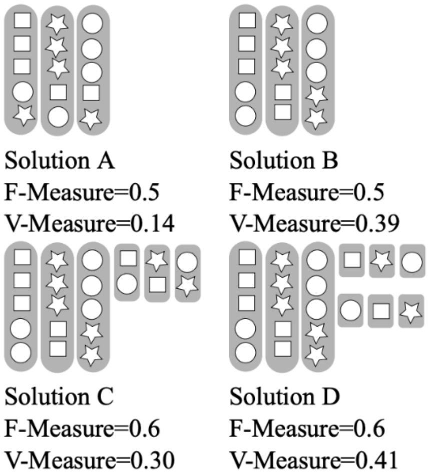

Fig. 2 Example of the Problem of Matching

3. Limitations and Prospects

Machine learning has certain limitations. Because the results of machine learning algorithms are derived from the data they are taught and trained on and do not contain human ethics, the results may be contrary to moral responsibility or unfair. Machine learning requires a large amount of data to train him. For one thing, a large amount of data is likely to contain sensitive information, and if the information is leaked, it will have a huge impact on personal security. When personal interests are affected, most people may refuse to provide relevant information, resulting in the loss of part of the data, and eventually getting the wrong model. In addition, for some areas with little data or where the cost of information collection is too expensive, existing technologies cannot support machine learning to use very little data to reach a highly accurate conclusion. The continuous development of machine learning in recent years has had a huge impact on various industries, and this technology will play a huge role in the future and bring great benefits to society. In recent years, machine learning has contributed to the rapid growth of the Internet industry and intelligent technology. But it also brings opportunities and challenges. Machine learning can be fine-tuned in different industries or fields such as natural language processing, computer vision, intelligent recommendation, and financial risk control to help humans improve efficiency and reduce errors. It is hoped that in the future, one can further explore ways to solve algorithmic bias and unfairness. Enable machine learning to develop on the basis of ensuring fairness, trustworthiness, and the security of individual privacy.

4. Conclusion

To sum up, this study has considered various statistical indicators of machine learning. This study ha briefly explained the development history and overview of machine learning. Understand the differences when using different metrics or methods in different situations. It points out the shortcomings of current machine learning and hopes for the future. Machine learning has many drawbacks and is not safe for everyone to use today, but it may become safe in the future. The purpose of this article is to summarize machine learning.

References

[1] Fradkov A L. Early history of machine learning. IFAC-PapersOnLine, 2020, 53(2): 1385-1390.

[2] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 2001, 23 (1): 89-109.

[3] He Q, Li N, Luo W, et al. Overview of machine learning algorithms based on Big Data. Pattern Recognition and Artificial Intelligence, 2014, 27(4): 327-336.

[4] Liu W, Xiong J, Fan Y, et al. Research status of machine learning in metamaterial intelligent design. Journal of Xinyang Normal University (Natural Science Edition), 2021, 34(3): 501-509.

[5] Dou J, Xiang Z, Xu Q, et al. Application and development trend of machine learning in landslide intelligent disaster prevention and reduction. Earth Science, 2023, 48(5): 1657-1674.

[6] Chen C. Application of machine learning algorithm in environmental microbiology research. Acta Microbiologica Sinica, 2022, 62(12): 4646-4662.

[7] Soofi, A A, Arshad A. Classification techniques in machine learning: applications and issues. Journal of Basic & Applied Sciences, 2017, 13(1): 459-465.

[8] Flach P A. ROC analysis. Encyclopedia of machine learning and data mining. Springer, 2016: 1-8.

[9] Hodson T O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development 2022, 15(14): 5481-5487.

[10] Wang W, Lu Y. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. IOP conference series: materials science and engineering. IOP Publishing, 2018, 324: 012049.

[11] Rosenberg A, Hirschberg J V. Measure: A conditional entropy-based external cluster evaluation. DBLP, 2007.