Analysis of Statistic Metrics in Different Types of Machine Learning

Jiayi Chen *

Shanghai Qibao Dwight High School, Shanghai, China

* Corresponding Author Email: jychen_selina@qibaodwight.org

Abstract. As a matter of fact, machine leading techniques has been attracting a lot of attention as one of the most promising areas of development in recent years on account of the rapid development of computing ability. Among other things, these autonomous learning mods need to go through a series of evaluations to determine their utility and usability and other features. In addition, this is where statistic metrics become particularly important. With this in mind, this paper will systematically analyze the use of different evaluation metrics and evaluation methods from three aspects, i.e., regression, classification and clustering. To be specific, some examples and formulae will be interspersed with practical examples. The purpose of this paper is to help better understand the different statistic metrics, so that they can apply them to different models of machine learning. Overall, these results shed light on guiding further exploration of machine learning implementation in various aspects.

Keywords: Machine learning; evaluation metrics; regression; clustering; classification.

1. Introduction

Machine learning as one of the fastest growing technology fields nowadays, it mainly covers the methods and research for developing autonomous learning machines. Not only is it at the intersection of statistics and computer science, but it is also central to the fields of data science and artificial intelligence. Many machines learning research started in the last half of the twentieth century, which include the study of support vector machines in 1974 and boosting in 1990 [1]. However, the development of machine learning is not simple, nor is it an overnight process. There are three phases: the first phase is logic learning from 1956 to 1960s, the main achievement: Automated theorem proving (e.g., the "Logic Theorist" system) [2]; second is knowledge engineering from 1970s to 1980s, which invented the Expert System [2]; and from the 1990s to the present, machine learning is in its third step, known as "machine learning" [2]. Despite the increasing interaction and research between scientists and the field of machine learning over the past nearly decade, this discipline of analyzing algorithmic features in samples and computational complexity by describing them and learning from them is only just entering into the diversity of experimental methods free from tedious procedures (e.g., The developers of many AI systems have now realized that it is significantly easier to train a system by providing it with instances of the required behavior of its inputs-outputs than it is to program it manually by predicting the required responses to every potential input) [3].

Through continuous research and deepening, there has been a substantial growth of machine learning systems. They correspond to different models and methods, and researchers and applicators need to evaluate the quality and goodness of these models through a variety of metrics. These metrics determine the final outcome of a machine learning system and the feasibility of its operational capabilities. For example, in the Azure Machine Learning Studio's designated module, the Evaluation Module, it is common to utilize MAE (Mean absolute error), RMSE (Root mean squared error), RSE (Relative squared error), CoD (Coefficient of determination) and RAE (Absolute error) to evaluate the prediction error [4]. Among the metrics related to categorization and segmentation, classification metrics can be divided in three groups: binary, multiclass, and multilabel. Confusion matrix is one of the examples in binary classification. It can be used to calculate many other metrics, such as false-positive rate, false-negative rate, sensitivity, positive predictive value, negative predictive value, F1 score and accuracy [5]. In order to evaluate the model, researchers need to focus on the amount of

difference between the predictor variables and the actual variables. There are several variants of such error calculation metrics such as MSE and MAE [6].

The aim of this paper is to contribute to the readers' more comprehensive understanding of the use of knowledge at the intersection of statistics and computer science by presenting the ways in which analytical statistical metrics are judged in different types of machine learning, their role, accuracy, and so on. In the latter part of the paper, Section 2 introduces the classifications and definitions related to machine learning. Section 3 talks about the commonly used metrics in regression and how to use them to judge the quality of machine learning. And Section 4 illustrates Evaluation metrics for classification. In Section 5, Metrics of clustering are discussed by assessing the standards and discussing some examples about clustering metrics. Limitations of the types of metrics the author have discussed and future perspectives in this field are placed in Section 6. And finally, Section 7 is the summary of the whole paper.

2. Description of Machine Learning

Machine learning is one of the subdivisions of computer science and artificial intelligence that is used to study how data and corresponding algorithms can be used to imitate human learning and gradually increase the accuracy of this behavior. Depending on the amount and type of supervision received during training, there are four forms of artificial intelligence: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, transduction, and learning to learn. The six groups can also be divided into two main classes, supervised and unsupervised. Supervised Learning refers to the generation of a function by various algorithms that map the source to the requested result. Because the purpose of guided (supervised) learning is to predict/classify specific results of interest, it is used to describe the prediction mission. A classic formulation as a supervised learning task comes in the form of a classification question: the learner is asked to study a functionality that allows researchers to make a mapping of a vector to one of multiple categories by observation of a few input-output instances of that functionality [7]. Classification models and regression models are the two main branches of such model. The regression model converts the input space into a domain with actual values. Classification models divide the input space into preset categories [7]. In supervised learning, a few of the most common algorithms are Decision Trees (a type of classifier), Linear Regression (part of regression), Naive Bayes (a statistical method for classification) and Logistic Regression. Unsupervised learning refers to modeling a group of inputs that is not labeled with examples. This type of machine learning is called unsupervised because it lacks responders that can be supervised and analyzed [8]. The main task of unsupervised learning is to automate the generation of categorization labels, so they are particularly useful in descriptive tasks. An algorithmic program will determine whether a data structure can be categorized into a cluster and create them by looking for similarities between the data structures without the need to measure the results. These are clusters which represent the family of entire clustering. In a word, supervised and unsupervised learning are two of the most researched models and the most relevant in the current state of research.

3. Evaluation Metrics for Regression

Many kinds of metrics are utilizing to evaluate regression. In this session, three of the classical ones will be mainly introduced. Mean absolute error is used widely to assess the accuracy of recommender systems. The equation of it is:

$$MAE = \frac{\Sigma|e_i|}{n} \tag{1}$$

It calculates the absolute distance between the true value and the regression prediction, taking the measure of the average of all observations, and the smaller this value, the more accurate the model being tested. One of the advantages of MAE is that it is most stable to abnormal values. Its

disadvantage is that the graphs it generates cannot be differentiable, so people need to apply optimizer. To solve the problem, MSE was invented. Mean squared error (MSE) is one of the most commonly used and very simple metrics to find the squared difference between actual and predicted values.

$$MSE = \frac{\Sigma(y-\hat{y})^2}{N} \tag{2}$$

MSE's graph is differentiable which make it easy to use it as a loss function. MSE is an ideal performance metric for models that predict continuous variables because it is related to the cross-entropy idea in information theory [10]. So, this metric is best employed in normal distributions because the minimization of the MSE is equal to the minimization of the cross-entropy (which is equal to maximizing the data's probability) [10]. In the following, Fig. 1 lists a regression that utilizes MSE.

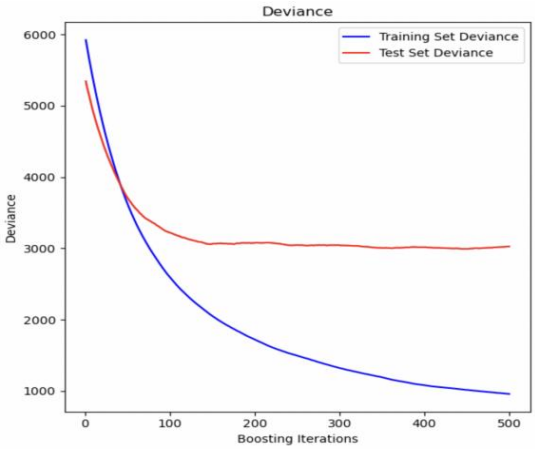

Fig. 1 Gradient Boosting Regression

R2 score can tell the model's performance, it is not a symmetric function:

$$R^2 = 1 - \frac{\Sigma(\hat{y}_i - y_i)}{\Sigma(\bar{y} - y_i)} \tag{3}$$

In the best case, the possible score scenario for this metric is 1.0. However, the meaning of the result tells researchers that it is a perfect prediction which is impossible to occur in the real world. Thus, one can conclude that as our regression line converges to perfection, the R2 score converges to 1. The model performance improves. Otherwise, when this model yields an R2 of 0.0, it implies that the regression model cannot account for any variability in the response data around its average value [9]. Since the model can vary arbitrarily, it can be a negative number which mean the poor performance of regression [9]. One of the more classic examples of using R2 for evaluation is Linear Regression. As can be seen in Fig. 2, the linear regression attempts to plot a straight line minimizing the sum of squared residuals between response as observed in the dataset and predicted by linear approximation. In addition, coefficients, coefficients of determination, and residual sums of squares were calculated.

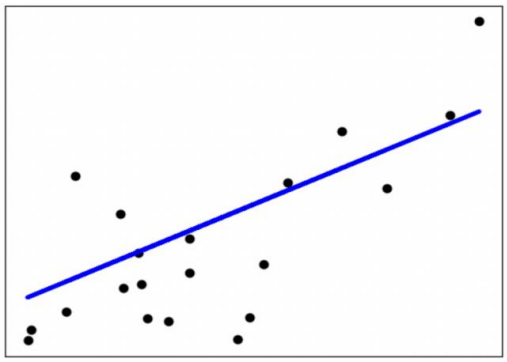

Fig. 2 Linear Regression

4. Evaluation Metrics for Classification

In general data classification, evaluation metrics are used in the studying and testing phases. In the first one, they are used to distinguish and pick the best solution and to make more accurate predictive evaluations of the classifier [11]. Whereas in the second phase, they are used to evaluate the classifier's performance when evaluated on unseen data [11]. The first one in this paper would introduce is accuracy. It is the most popular evaluation metric used in real applications (classification). Accuracy is a quality assessment (evaluation) of a solution according to the percentage of positive predictions as a percentage of the number of total cases. The advantages are that the indicator is simple, easy to calculate and understand by humans. The drawbacks, on the other hand, are limitations, one of the main ones being that it produces less obvious values. While this metric is quite interpretable, high accuracy does not always feature a good classifier [12]. Fig. 3 is an example.

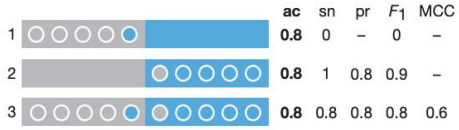

Fig. 3 FNs or FPs are more frequent.

The second one is precision. This type of metric is the capability of the classification to not label negative samples as positives. In other words, it is used to measure which of the predicted patterns from a positive category is predicted correctly [11]. The best and worst values is 1 and 0. Precision-Recall is a case that use precision (Fig. 4). Precision-recall is a helpful indicator of the predictive success when the class is very unbalanced. A high area represents high recall and precision, indicating that it returns accurate results, as well as returning most of all positive results. A high recollection but poor accuracy one will provide several outcomes whose predictive labels are mostly wrong, vice versa.

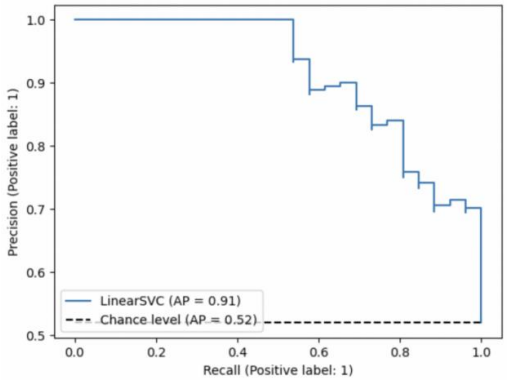

Fig. 4 Binary classification setting.

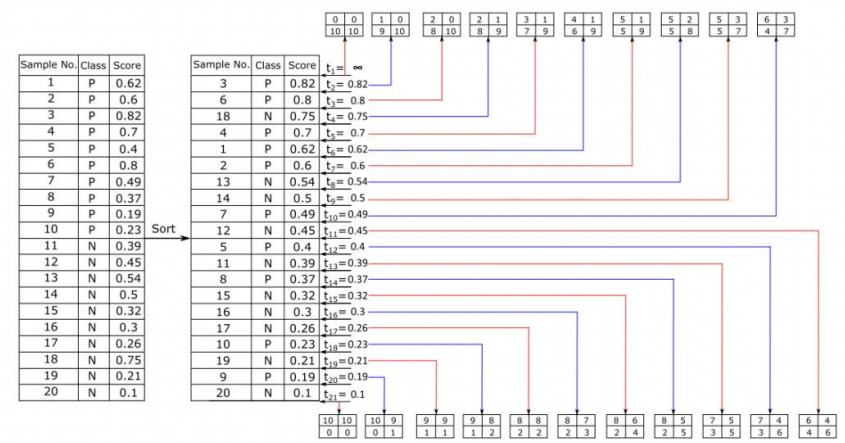

Fig. 5 Data category.

ROC is also one of the metrics that the researchers use:

$$ROC = ((P(i)-P(n)/P(n)) \times 100\% \quad (4)$$

It is a two-dimensional graph that evaluates performance better than Accuracy. Two of the advantages in ROC are that has clear logic of calculation and clear signal interpretation. However, highly sensitive of changing prices and requirement of continuous optimize are its disadvantage. It is used to balance true positives and false positives. As an example (seen from Fig. 5) if the scores of the samples from the test set are more than and equivalent to the set threshold (minimum value), they are categorized as positive; if the scores are less than and equal to the threshold, they are negative [13].

5. Evaluation Metrics for Clustering

A clustering problem is judged on the basis of whether it possesses three features which are scale invariance, consistency and richness [14]. However, none of them can satisfy all three. Thus, they can be designed to violate one of the axioms and can usually satisfy the properties of scale invariance and consistency by relaxing their richness [14]. Evaluation of the results of clustering algorithms is a very significant part of the data clustering process. Adjusted Mutual Information (AMI) is a kind of metric that usually uses between two clusterings. It can be defined as:

$$\Delta I(X,Y) = I(X,Y) - E(I(X,Y_\sigma)) \quad (5)$$

In general, AMI is used when the reference clusters are unbalanced and there are small clusters [15]. AMI can be adjusted to take into account the occurrence of chance by adjusting the MI, which is essentially higher for clusterings with more two clusters. The metric is not dependent on the labels with absolute value and is symmetric in nature. AMI can validate clustering solutions based on ground truth clustering, and it is also useful when the real situation is not known [15]. Completeness indicates that all given categories have all their members assigned to the same cluster [16]. Alternatively, individuals in different categories should be contained in different groups. One can model this notion using Fig. 6. Suppose D1 is a distribution in which two types of clusters, A1 and A2, contain merely a few items from the identical category H [17]. Suppose D2 is one equivalent distribution, with the exception of A1 and A2 have been combined into a cluster [17]. Therefore, D2 is superior.

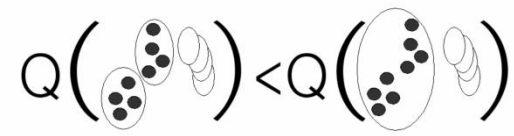

Fig. 6 Modeling of clustering.

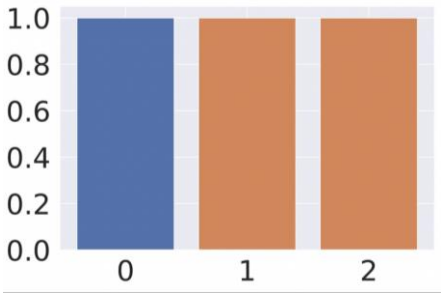

Fig. 7 AMI example.

As with AMI, this metric is uncorrelated with the label's absolute value, but in the meantime, it is not a symmetric metric. Homogeneity metric in cluster represents that the clusters contain all data points of one type. It measures how similar the samples in a cluster are. The researchers can determine how much proximity a specific clustering is to the desired ideal by checking the conditional entropy of the proposed clustering's class distributions [18]. Like the first two metrics, it is irrelevant to the

absolute value of the labels. And, like the Completeness metric, is not symmetric. However, they are symmetric to each other [18]. Here is an example for Homogeneity in this paper allows readers to fully understand the feature of the metric (Fig. 7). Each of the three clusters contains one color. The situation of the example is not quite well because the orange one is separated into different parts.

6. Limitations and Prospects

Machine learning evaluation metrics are critical for measuring the performance and quality of the models. However, they also have some limitations that may affect your results and decisions. There is no single metric that can assess all aspects of a mod and cannot provide comprehensive coverage. If a researcher wants to come and test different aspects of a mod that needs to be evaluated, then several different metrics are needed to calculate them, thus causing inconvenience and complexity. What's more, sometimes different metrics give conflicting results. In addition, they are not absolute metrics, but relative and subjective. Depending on the problem area, the baseline and the objectives, the range of indicators can deviate. For example, if 0.5 is an indicator range, it may be considered good or bad in different contexts of indicators. Finally, the degree to which evaluation metrics can be visualized and communicated effectively and transparently is also a limitation; the complexity of the metrics can result in others not being able to maximize their comprehension of the execution and quality of the model being evaluated.

In the future, it is believed that the accuracy of Evaluation metrics will continue to improve. Through constant experimentation and the creative ability of scientists, more comprehensive and complete metrics of decreasing complexity will be obtained, which will help us to select the best model for a given problem more efficiently by contrasting the performance of various models. They help identify fields where models are underperforming, so that humans can focus on improving those things without being blind. Meanwhile, the ability to make more accurate predictions on unseen data will help humans understand the actual performance of different, novel models in the future.Machine learning as a very promising discipline, evaluation metrics are indispensable to assist it and promote it. It is believed that the field of evaluation metrics will grow in the future.

7. Conclusion

To sum up, the paper has studied and analyzed different types of statistic metrics in machine learning. Based on the study, this paper has come up with the evaluation methods and roles of different evaluation metrics in three major categories of models: regression, classification and clustering and Characterization. Understood the different situations where different metrics are applicable and the differences between these metrics. It also uses formulas and examples to visually maximize the reader's understanding. Although these metrics still have a lot of limitations, being unable to provide comprehensive coverage, being relatively subjective, and being complex... It can still help human beings to go further and further on the road of machine learning in the future. This paper investigates statistic metrics to provide the reader with a more intuitive feeling for the impact of different metrics on machine learning, how they are evaluated, how they are used, and their pros and cons. It will help readers to utilize these evaluation metrics in the realm of ML (machine learning) more easily.

References

[1] Molnar C, Casalicchio G, Bischl B. Interpretable machine learning–a brief history, state-of-the-art and challenges. Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer International Publishing, 2020: 417-431.

[2] Zhou Z. Machine learning: Future and Development. Communication of Association for Computer Science of China, 2017, 13 (1): 44-51.

[3] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349(6245): 255-260.

[4] Botchkarev A. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. Available at SSRN 3177507, 2018.

[5] Erickson B J, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. Radiology: Artificial Intelligence, 2021, 3(3): e200126.

[6] Handelman G S, Kok H K, Chandra R V, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. American Journal of Roentgenology, 2019, 212(1): 38-43.

[7] Nasteski V. An overview of the supervised machine learning methods. Horizons. b, 2017, 4: 51-62.

[8] Jiang T, Gradus J L, Rosellini A J. Supervised machine learning: a brief primer. Behavior Therapy, 2020, 51(5): 675-687.

[9] Chicco D, Warrens M J, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 2021, 7: e623.

[10] Hodson T O, Over T M, Foks S S. Mean squared error, deconstructed. Journal of Advances in Modeling Earth Systems, 2021, 13(12): e2021MS002681.

[11] Hossin M, Sulaiman M N. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 2015, 5(2): 1.

[12] Lever J. Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. Nature methods, 2016, 13(8): 603-605.

[13] Tharwat A. Classification assessment methods. Applied computing and informatics, 2020, 17(1): 168-192.

[14] Palacio-Niño J O, Berzal F. Evaluation metrics for unsupervised learning algorithms. arXiv preprint arXiv:1905.05667, 2019.

[15] Romano S, Vinh N X, Bailey J, et al. Adjusting for chance clustering comparison measures. The Journal of Machine Learning Research, 2016, 17(1): 4635-4666.

[16] Majumdar J, Naraseeyappa S, Ankalaki S. Analysis of agriculture data using data mining techniques: application of big data. Journal of Big data, 2017, 4(1): 20.

[17] Amigó E, Gonzalo J, Artiles J, et al. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval, 2009, 12: 461-486.

[18] Rosenberg A, Hirschberg J. V-measure: A conditional entropy-based external cluster evaluation measure. Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). 2007: 410-420.