Evaluation Metrics for Machine Learning

Miguel Ángel Banda Del Valle [1*], Jose Ángel Pertuz Montes [2*]

[1, 2] Department of Engineering, Universidad Tecnológica de Bolívar, Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 Vía Turbaco, Cartagena, Colombia
[*]mbanda@utb.edu.co
[*]pertuzj@utb.edu.co

1. Introduction

Currently, we live in an era in which information is found in an innumerable amount of data, from which an immense benefit can be obtained. The human being has limitations in terms of processing this available information and therefore makes use of tools that allow both processing, storage and evaluation, in order to obtain the benefits, they provide.

In the branch of Artificial Intelligence, Machine Learning, which gives rise to machine learning by means of patterns within a set of data, it can be stated that the correct operation of these algorithms is based on an adequate evaluation of the results, this is achieved using statistical metrics that can identify errors, uncertainty, approximation to real models, among other parameters that define a correct operation of an algorithm with respect to the predictions that it makes, in the present the evaluation metrics of greater use within the implementation of prediction algorithms are described.

2. Confusion Matrix

A confusion matrix is an N x N table that is based on the number of actual and predicted values obtained from a classification model to estimate the performance of a classification model, in other words, the matrix consists of several correct and incorrect predictions, comparing them with each other through the machine learning model.

This matrix is represented as a square matrix where the column represents the actual values, and the row represents the predicted value of the model.

The following basic terminologies allow to understand the metrics to be obtained:

- True Positive (TP): occurs when the actual value is positive, and the prediction is also positive.

- True negative (TN): occurs when the actual value is negative, and the prediction is also negative.
- False Positive (FP): also called as type 1 error, it occurs when the true value is negative, but the prediction is Positive.
- False Negative (FN): also called as type 2 error, occurs when the actual result is positive, but the prediction is negative. [1]

3. F1-Score

F1-score is an evaluation metric that adjusts model accuracy and recall, defined as the harmonic mean. It is used to evaluate machine learning model types specifically in natural language processing. If you need to compare two or more machine learning algorithms containing the same data, the F1-score would be helpful.

Mathematically it is expressed as follows,

$$\textit{\textbf{F - measure}} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

[2]

4. Gain and Lift Charts

Gain and lift charts is a visual way by which the performance of different models can be estimated, this type of evaluation metric shows the actual lift, which is known as the ratio between the results obtained by the prediction model compared to those obtained without a model. In addition, it can also be observed how the response rate of some x group differs compared to that of a selected group.[3]

5. Kolmogorov Smirnov Chart

Also known as the K-S test, the Kolmogorov-Smirnov goodness-of-fit test performs a comparison of data distribution with a known distribution, and allows to identify the similarity between these distributions, this test in principle does not assume distributions, and is used in most cases to check if there is a normal distribution between data.

Defining the distribution generated by the data by an empirical distribution function (EDF), this statistic measures the largest distances between the EDF and the theoretical function that alludes to the known hypothetical probability distribution. Which makes it easier to identify whether the data set comes from a population with a specific distribution. [4]

Many statistical procedures assume the assumption of normality that entails generalized and non-parametric analysis techniques, hence, this type of analysis on a data set is important since applying analysis techniques that are based on specific distribution assumptions generate better results than if more robust and non-parametric techniques were applied. [5]

6. ROC Curve & AUC

The Receiver Operating Characteristic Curve (ROC Curve) allows to identify through true positive rates (TPR) and false positive rates (FPR) the performance of a classification model.

TPR being that which measures the degree to which the model was able to correctly predict the positive cases, and FPR being the inverse of the positive case, since it measures the degree to which the model was able to correctly identify the true negative cases.

This curve illustrates TPR as a function of FPR with different classification thresholds and facilitates the determination of the appropriate threshold for which to proceed with model predictions in algorithms.

For the calculation of points on this curve, logistic regression models can be implemented by varying their classification thresholds, however, by calculating the area under this curve (AUC) such information can be obtained, this being the probability that the model will best classify the positive data from the samples.[6]

7. Logarithmic Loss

This ranking metric is based on eventual probabilities according to certain evaluation parameters, Megha Setia mathematically interprets this metric as:

"Log Loss is the negative average of the log of corrected predicted probabilities for each instance." [7]

It is the fundamental basis of logistic regression algorithms, log loss is based on statistical entropy, which is a measure of the uncertainty associated with a certain distribution, but this can only be computed if the distribution model is known, otherwise the cross-entropy between two distributions or more is used. [8]

8. Gini Coefficient

The Gini Coefficient is a type of metric used to compare the performance of different models and their ability to predict. This model is usually seen in banking, insurance, and targeted marketing activities. Everything related to the financial sector, because thanks to its efficiency, it can distinguish between a good and a bad borrower.[9]

9. Root Mean Square Error

Root Mean Square Error (RMSE) is a standard way of measuring the error of a model when predicting quantitative data.

In order to be able to say that a regression model fits a data set, it is necessary to calculate the root mean square error (RMSE) which, thanks to it, we can know the average distance between the actual values of the set and the predicted values of the model. The smaller this RMSE is, the better an objective model can be fitted to a data set. [10]

10. References

Works Cited

[1] G. Giannakopoulos, P. Mavridi , G. Paliouras, G. Papadakis and K. Tserpes, “Representation Models for Text Classification:a comparative analysis over three Web document types,” Athens, 2012.

[2] S. Alias, S. K. Mohammad, G. K. Hoon and T. T. Ping, “A text representation model using Sequential Pattern-Growth method,” 2017.

[3] M. W. Berry, Ed., Survey of Text Mining, Clustering, Classification, and Retrieval, 2nd ed., New York: Springer-Verlag New York, 2004.

[4] R. Guerrero Álvarez, “Aplicación del modelo word2vec para el análisis de sentimientos en tweets en idioma inglés,” UCLV, Santa Clara, 2019.

[5] A. Rodríguez Blanco, A. Simón Cuevas, E. Guevara Martínez and W. Hojas Mazo, “Modelo de representación de textos basado en grafo para la minería de texto,” *Red de Revistas Científicas de América Latina, el Caribe, España y Portugal,* vol. 46, pp. 63 - 71, 2015.

[6] Merours, “Stack Overflow,” 6 June 2014. [Online]. Available: https://stackoverflow.com/questions/24073030/what-are-co-occurence-matrixes-and-how-are-they-used-in-nlp. [Accessed September 2021].

[7] A. Sehgal, “Medium,” 11 July 2020. [Online]. Available: https://medium.com/@imamitsehgal/nlp-series-distributional-semantics-co-occurrence-matrix-31283629951e. [Accessed September 2021].

[8] T. Bach, “Codegram,” 9 July 2020. [Online]. Available: https://www.codegram.com/blog/finding-similar-documents-with-transformers/. [Accessed September 2021].

[9] P. Joshi, “Analytics Vidhya,” 11 March 2019. [Online]. Available:

https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/. [Accessed September 2021].