Scientia Africana, Vol. 21 (No. 3), December, 2022. Pp 49-56 *https://dx.doi.org/10.4314/sa.v21i3.4*
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

RELATING STATISTICAL METHODS TO MACHINE LEARNING PREDICTIVE MODELS

Agu, S. C[1] and Elugwu, F.[2]
[1]Department of Computer Science, Madonna University Nigeria, Elele Campus.
[2]Department of Computer Science, Delta State Polytechnic, Otefe, Nigeria.
Email: [1] sndyaguu@gmail.com , [2] felixelugwu@gmail.com

Received: 18-11-2022
Accepted: 31-11-2022

ABSTRACT

The paper reviewed the probabilistic feature of binomial distribution in the operation of machine learning (ML) classifications. It also examined a normal distribution and the concepts for approximating the binomial distribution to a normal distribution in estimating generalization error and its role in machine learning model selection. Again, it studied the confident interval and hypothesis testing and their estimations in the evaluation and comparison of the Performance metrics (Accuracy) of the learning algorithms. The paper highlighted their statistical significance to the ML models and classifiers as well as the differences in their utilization in statistics and machine learning.

Keywords: Machine Learning Models, Classifiers, Approximation, Estimation, Binomial distribution, Normal distribution, Confident interval, Hypothesis testing

INTRODUCTION

Probabilistic machine learning (ML) provides rich tools for modeling uncertainty, carrying out probabilistic inferences, and making predictions or decisions in an uncertain environment [1]. As ML task is being modeled, one or two statistics are inevitably encountered. After modeling and evaluating the performance of the machine learning projects, different types of estimations and statistical tests are also needed for comparing the performance of the ML models and classifiers. The type of estimations and statistical tests that are applied depends on the type of report that is required. This study explored four statistics: binomial distribution, normal distribution, confident interval, hypothesis testing, and aimed at showing how they are approximated in the evaluation and comparison of the performance of machine learning models and classifiers.

Distribution

Distribution of a statistical dataset (or a population) is a listing that shows all the possible values (or intervals) of data [2]. Data distribution also known as a distribution function is a mathematical expression that describes the probability that a system will take a specific value or set of values. Examples of such functions are binomial function, normal function, uniform function, etc [3].

Binomial Distribution

In statistics, a binomial distribution is simply the probability of "success" or "failure" outcome in an experiment or a survey [4] as indicated in equation 1.

$$b(x, n, p) = {}_nc_x * p^x * (1-p)^{n-x} \qquad (1)$$

Equation 1 shows that the parameters of the binomial function are $x, n, and\ p$, where b = binomial probability, x = total number of

"successes", p = probability of success on an individual trial and n = number of trials.

Binomial Distribution in Machine Learning Classification

Machine learning classifiers naturally follow a binomial distribution. Thus, machine learning classifications operate based on the fraction of instances that are correctly classified in a dataset. As a result, the training error of a learning algorithm can be determined by the probability of the correctly classified proportion of the instances in the dataset [5]. The classification accuracy or the classification error is a proportion of a ratio that describes the proportion of correct or incorrect predictions made by a model, where each prediction is a binary decision called Bernoulli trial and the proportion in a Bernoulli trial has a specific distribution called binomial distribution [6].

Normal Distribution

A normal distribution is a distribution of data around a central value, the mean, with no left or right bias [7]. It occurs in many natural situations such as exam scores, heights, salaries where many data in the set of observation cluster around the mean with a few of the data clustering at both extremes. Plotting the distribution on a graph will show a bell shape.

However, there are cases where the data obtained from the observation, survey, or experiment are below expectations, and making decisions based on Normal distribution will negatively affect the entire system for which the observation is deployed. To salvage such occurrences, the Normal distribution is converted to Standard Normal distribution (known as Standard score or Z-score) as shown in equation 2, upon which new decisions are made or new mathematical equations are re-recalculated.

$$z = \frac{x - \mu}{\sigma} \qquad (2)$$

In equation 2, z = z-score (standard score), x = data to be standardized (i.e., each of the data in the Normal distribution), μ = the mean, and σ = the standard deviation

Estimating Generalization Error

In machine learning, the gap between predictions and observed data is induced by model inaccuracy, sampling error, or noise which is reducible. Choosing the right algorithm and tuning parameters can improve accuracy. By using Mean Square Error (MSE) in regression, generalization error is estimated to determine the optimal model capacity where bias and variance are low [8]. According to [5] however, the problem is establishing a normal distribution for estimating the generalization error of a model. Being that the generalization error tends to be higher than the training error; a statistical correction is usually computed as an upper bound to the training error. Thus, calculating the generalization error means determining the upper limit to the observed training error. By approximating a binomial distribution with a normal distribution, the upper bound of the error rate e can be derived as in equation 3

$$e_{upper}(N, e, \alpha) = \frac{e + \frac{z_{\alpha/2}^2}{2N} + z_{\alpha/2}\sqrt{\frac{e(1-e)}{N} + \frac{z_{\alpha/2}^2}{4N^2}}}{1 + \frac{z_{\alpha/2}^2}{N}} \qquad (3)$$

Where α is the confidence level, $z_{\alpha/2}$ the standardized value from a standard normal distribution, and N, the total number of training records used to compute e.

Scientia Africana, Vol. 21 (No. 3), December, 2022. Pp 49-56 https://dx.doi.org/10.4314/sa.v21i3.4
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

Remark I

In statistics, standardization allows for comparison between different types of observations, surveys, and experiments based on where each observation falls within its distribution [9].

In ML, the complexity of a model has an impact on the model overfitting. The ideal complexity is the model that produces the lowest generalization error on a test set. Thus, selecting the right model for learning algorithms means producing a model with the lowest generalization error.

Confident Interval

Confident Interval is a statistical test that gives a range of values within which the uncertainty of the true result of an experiment or survey is confirmed. It is related to a confident level which is expressed in percentage [12]. The statistical test function for the confidence interval for sample size and a population proportion is given in equations 4 and 5.

$$CI = \bar{x} \pm z * \frac{\sigma}{\sqrt{n}} \qquad (4)$$

In equation 4, $\bar{x}$ is the population mean, $z =$ z score, σ = standard deviation, and n = number of observations.

$$CI = \hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \qquad (5)$$

In equation 5, $\hat{p}$ is the fraction of the given population to the sample size, z = z-score, and N = the sample size.

Estimating a Confident interval for a Model Accuracy

With a large dataset of 30 instances or above, binomial distribution can be approximated with a Gaussian. Using the assumption of Gaussian distribution of the proportion, that is, the classification accuracy or the classification error, the confident interval of the model can be calculated [6]. For the classification error, the radius of the interval can be calculated as shown in equation 6, while the radius of the interval for classification accuracy is shown in equation 7

$$Interval = z * sqrt\left(\frac{\left(error * (1-error)\right)}{n}\right) \qquad (6)$$

$$Interval = z * sqrt\left(\frac{\left(accuracy * (1-accuracy)\right)}{n}\right) \qquad (7)$$

According to [5] however, the issue is estimating the confident interval of given model accuracy. With a test set containing N records, letting x be the number of correctly predicted records by a model and p be the true accuracy of the model. By modeling the prediction task as a binomial experiment, x has a binomial distribution with mean N and variance $Np(1-p)$. Then, the empirical accuracy, $acc = x/N$ has a binomial distribution with mean p and variance $Np(1-p)$. At this point, the confidence interval for acc can be estimated using the binomial distribution. However, the binomial distribution is approximated with a normal distribution when N is sufficiently large. Hence, the confident interval for acc can be derived as shown in equation 8

$$p\left(-z_{\alpha/2} \leq \frac{acc-p}{\sqrt{\frac{p(1-p)}{N}}} \leq z_{1-\alpha/2}\right) = 1-\alpha \qquad (8)$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the upper and lower bound obtained from a standard normal distribution at a confident level$(1-\alpha)$. Being that the standard normal distribution is symmetric around $z = 0$, it means that $z_{\alpha/2}$ = $z_{1-\alpha/2}$. Thus, repositioning the inequality

gives the confidence interval for true accuracy p, as shown in equation 9.

$$\frac{2 * N * acc + z_{\alpha/2}^2 \pm z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4Nacc - 4Nacc}}{2(N + z_{\alpha/2}^2)} \quad (9)$$

Remarks II

In statistics, the confidence interval compares experiments not based on their results but on how much confidence is placed on each experiment [12].

In machine learning, after evaluating the performance of two different models based on accuracy or error rates, a conclusion cannot be drawn on which of the models has a better performance until the confident intervals of both models are compared to determine which of the model outperforms the other. Hence, Confident Interval shows the level of confidence in the performance of a model.

Hypothesis

In statistics, a hypothesis is a prediction about the result of research, which is testable through scientific research methods such as experiments, observations, or statistical analysis of data. And if the hypothesis is required in research, it is stated before data is collected for the experiment [10].

In machine learning, the hypothesis is a model that uses available data to approximate a target function and performs the mapping of inputs to outputs on all possible observations from the possible domain. And the choice of algorithms such as neural networks and the configuration such as network topology and hyperparameter define the space of possible hypotheses that the model may represent [13].

Hypothesis testing

Hypothesis testing is simply a confirmation of the prediction written in the Hypothesis statements. Scientifically, it determines the validity of a result after an experiment or survey by confirming that the result is not by chance. Results obtained by chance indicate that the experience is not repeatable and not acceptable as a scientific proof [11]. Research involving hypothesis testing requires a null and alternative hypothesis statement.

Hypothesis Statements in Statistics

In statistics, the null hypothesis statement denoted by H_0 is an already known and accepted fact that is nullifiable, while an alternative hypothesis statement denoted by $H_1$is the new fact, that is, the prediction to be established to nullify the null hypothesis statement as exemplified in statements 1 and 2.

H_0: Blood glucose levels for Covid-19 patients have a mean of 100.

Statement 1

H_1: Diet high in raw starch will have a positive or negative effect on blood glucose levels.

Statement 2

The hypothesis statements 1 and 2 can be stated mathematically as shown in statements 4 and 5:

$H_0: \mu = 100$ statement 4

$H_1: \mu \neq 100$ statement 5

Hypothesis Testing Process in Statistics

The hypothesis statements determine the type of test to perform: a one-tail test or a two-tail test, followed by a chosen alpha level which is used in finding the critical value from a Z-table or a t-table by utilizing the selected alpha value. Then, the test statistic for the hypothesis is applied as shown in equation 10.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad (10)$$

Scientia Africana, Vol. 21 (No. 3), December, 2022. Pp 49-56 https://dx.doi.org/10.4314/sa.v21i3.4
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

where z is the statistical test for the Hypothesis, $\bar{x}$ the random sample mean, μ_0 the population mean, σ the population standard deviation, and n the sample size.

Hypothesis Statements and Approximation for Comparing the Performance of two ML Models

According to [13], a single hypothesis statement denoted by lowercase h is a given specific hypothesis and a set hypothesis statement denoted by uppercase H is the hypothesis space that is being searched as exemplified in statements 5 and 6.

h(hypothesis single): An instance or specific candidate model that maps inputs to output and can be evaluated and used to make predictions. *Statement 5*

H(hypothesis set): A space of possible hypotheses for mapping inputs to outputs that can be searched. *Statement 6*

The hypothesis set statement is often constrained by the choice of the framing of the problem, the choice of model, and the choice of model configuration. Although [6] used statsmodels function, proportion_ confint(), an implementation of a Binomial Proportion Confident Interval for hypothesis testing, it did not show the theoretical procedures.

A pertinent question, however, is the possibility of explaining the difference in the accuracy of two different models as a result of the variations in the composition of their test sets. The question relates to the issue of testing the statistical significance of the observed deviation [5]. Considering two different models that are evaluated on two independent test sets, letting n_1 and n_2 denote the respective numbers of records in the test sets, and e_1 and e_2 denote the respective error rates, the aim is to test whether the observed difference between e_1 and e_2 is statistically significant. Assuming that n_1 and n_2 are sufficiently large, then e_1 and e_2 can be approximated using a normal distribution. If the observed difference in the error rate is $d = e_1 - e_2$, then d is also normally distributed with mean d_t, the true difference and the variance σ_d^2, of d, is computed as in equation 11.

$$\sigma_d^2 \simeq \hat{\sigma}_d^2 = \frac{e_1(1-e_1)}{n_1} + \frac{e_2(1-e_2)}{n_2} \qquad (11)$$

Where $e_1(1-e_1)/n_1$ and $e_2(1-e_2)/n_2$ are the variances of the error rates. Then, at the $(1-\alpha)\%$ confidence level, the confidence interval for the true difference d_t is shown in equation 12.

$$d_t = d \pm z_{\alpha/2}\hat{\sigma}_d \qquad (12)$$

Remarks III

In statistics, hypothesis testing allows for determining whether the null hypothesis will be rejected or not rejected. In a one-tail test, if the test statistics is greater than the critical value, the null hypothesis is rejected. In a two-tail test, if the test statistics is lesser than the negative critical value or greater than the positive critical value, the null hypothesis is rejected.

In machine learning, following hypothesis statements 4 and 5, a two-tail test can be performed to check whether $d_t = 0$ or $d_t \neq 0$. If the confidence interval for d_t spans the value zero, it can be concluded that the observed difference is not statistically significant assuming the confidence level is 95%. Another pertinent question is, at which confidence level can the hypothesis $d_t = 0$ be rejected. This issue is resolved by determining the value of $z_{\alpha/2}$ such that the confidence interval for d_t does not span value zero. As a result, the computation can be reversed to look

for the value of $z_{\alpha/2}$ such that $d > z_{\alpha/2}\hat{\sigma}_d$. Hence, the result which is the value of the confidence level or lower at which the null hypothesis can be rejected could then be suggested.

Hypothesis Statements and Approximation for Comparing the Performance of two ML Classifiers

Most often, an ML task requires comparing the performance of two Classifiers that are evaluated on the same test set using the K-fold cross-validation method. Considering a model denoted as M_{ij} induced by a classification technique L_i during the j^{th} iteration, and noting that each pair of models M_{1j} and M_{2j} are tested on the same partition j, and also letting e_{1j} and e_{2j} be their respective error rates, the difference between their error rates during j^{th} fold can be written as $d_j = e_{1j} - e_{2j}$ [5]. If k is sufficiently large, then d_j is normally distributed with mean d_t^{cv}, the true difference in error rates, with variance σ^{cv}, and the overall variance in the observed difference is estimated as in equation 13.

$$\hat{\sigma}_{d^{cv}}^2 = \frac{\sum_{j=1}^{k}(d_j - \bar{d})^2}{k(k-1)} \qquad (13)$$

In this method, the t-distribution is used to compare the confidence interval for the true difference d_t^{cv} as in equation 14.

$$d_t^{cv} = \bar{d} \pm t_{(1-\alpha),k-1}\hat{\sigma}_{d^{cv}} \qquad (14)$$

where $\bar{d}$ is the average difference, the coefficient $t_{(1-\alpha),k-1}$ is obtained from a probability table with two input parameters, the confidence level $(1-\alpha)$, and the number of degrees of freedom is $k-1$.

Remarks IV

If the confidence interval of equation 14 does not span the value zero, the observed difference between the classifier is statistically significant. According to [13], (1) Learning means searching the hypotheses space to find the best hypothesis that approximate the target function even on new inputs beyond the training set. (2) A training set is used to learn hypothesis and a test set is used to evaluate the hypothesis. (3) Unlike the statistical hypothesis, machine learning hypothesis reflects the broader scientific hypothesis based on the following characteristics (a) by being an explanation that covers available evidence: the training set (b) Is falsifiable: a test is used to estimate the performance of a model and compare it with a baseline model to check the skillfulness. (c) can be used in new situations: make predictions on new data.

CONCLUSION

The study introduced two vital statistic distributions – binomial and normal distributions, and two statistical test – confident interval and hypothetical test and reviewed their relationships in the evaluation and comparison of performance metrics (accuracy matrices) in machine learning classifiers and models.

REFERENCES

Jun, Z. (2017) Probabilistic Machine Learning: Models, Algorithms, and a Programming Library. *Proceedings of the Twenty-seventeen International Joint Conference on Artificial intelligence (IJCAI-18)*

Deborah, J. R. (2020) *What the Distribution Tells You About a Statistical Datasets*. http://www.dummies.com/education/math/statistics/what-the-distribution-tells-you-about-a-statistical-data-set/

Britannica, (2020) *Distribution Function.* https://www.britannica.com/science/distribution-, function

Scientia Africana, Vol. 21 (No. 3), December, 2022. Pp 49-56 https://dx.doi.org/10.4314/sa.v21i3.4
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

Statisticshowto, (2020) *Binomial Distribution Formula*. https://www.statisticshowto.com/ probability-and-statistics/binomial-theorem/binomial-distribution-formular/

Tan, S. K. (2004) *Classification: Basic Concepts, Decision Trees, and Model Evaluation*. https://slideplayer.com/slide/7696713/

Jason, B. (2018) *Confidence Interval for Machine Learning*. https://machinelearningmastery.com/confidence-intervals-for-machine-learning/

Mathisfun, (2019) *Normal Distribution*. https://www.mathsisfun.com/data/standard-normal-distribution.html

Yi-Xin, (2018) *Understanding Generalization Error in Machine Learning*. https://medium.com/@yixinsun_56102/understanding-generalization-error-in-machine-learning-e6c03b203036

Jim, F. (2020) *Normal Distribution in Statistics*. https://statisticsbyjim.com/basics/normal-distribution/

Shona, M. (2020) *How to Write a Hypotheses*. https://www.scribbr.com/ research-process/hypotheses/

Statisticshowto (2020) *Hypothesis Testing*. https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/

Statisticshowto (2020) *Confident Interval*: How to Find a Confident Interval. https://www.statisticshowto.com/probability-and-statistics/confident-interval/

Jason, B. (2019) *What is Hypothesis in Machine Learning?* https://machinelearningmastery.com/what-is-a-hypothesis-in-machine-learning/