Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Website: Kjar.spu.edu.iq | Email: kjar@spu.edu.iq

Utilizing Statistical Tests for Comparing Machine Learning Algorithms

Hozan Khalid Hamarashid
Information Technology Department
Computer Science Institute
Sulaimani Polytechnic University
Sulaimani, Iraq
hozan.khalid@spu.edu.iq

Article Info

Volume 6 – Issue 1- June 2021

DOI: science.2021.1.8

Article history:

Received 1/6/2021
Accepted 28/6/2021

Keywords:

machine learning, machine learning assessment, statistical tests, machine learning algorithm, machine learning comparison.

ABSTRACT

The mean result of machine learning models is determined by utilizing k-fold cross-validation. The algorithm with the best average performance should surpass those with the poorest. But what if the difference in average outcomes is the consequence of a statistical anomaly? To conduct whether or not the mean result differences between two algorithms is genuine then statistical hypothesis test is utilized. Using statistical hypothesis testing, this study will demonstrate how to compare machine learning algorithms. The output of several machine learning algorithms or simulation pipelines is compared during model selection. The model that performs the best based on your performance measure becomes the last model, which can be utilized to make predictions on new data. With classification and regression prediction models it can be conducted by utilizing traditional machine learning and deep learning methods. The difficulty is to identify whether or not the difference between two models is accurate.

1. INTRODUCTION

Data may be understood by using a determined structure for the result and then utilizing statistical techniques to validate or invalidate the estimation. The estimation is called hypothesis, and the validation that conducted by using statistical tests are called statistical hypothesis tests. If it is wanted to build statements on the data distribution or the grouped results are different in applied machine learning, statistical hypothesis testing has to be performed. [1][2].

Data is uninteresting on its own. What interests the most is how data is interpreted. When the time is started to request the inquiry about the data and comprehend the discovery, then statistical methodologies will be utilized that provide a certainty or probability regarding the replies. This sort of procedure is called testing for significance or hypothesis test [3]. The theory term might bring up the concept of scientific investigations in which a hypothesis is tested. This is a good step in the right direction. Hypothesis test, statistically, computes a number for a given estimation. The test results enable the researcher to see whether the estimation is valid or whether it is falsified. Particularly, two examples that are widely utilized in machine learning are as follows:

- An assumption test conducted for the data that followed a normal distribution.
- An assumption test conducted that two samples are picked from the same population distribution.

The null hypothesis, commonly known as hypothesis 0, is a statistical test assumption (H0 for short). The default assumption, often known as the "nothing has changed" assumption, is extensively employed. Because all available information indicates that the evidence shows that the H0 may be rejected, the first hypothesis, often known as hypothesis 1 or H1, is a violation of the test's premise. H1 is essentially only a shorthand for some other hypothesis.

Hypothesis 0 (H0): The test assumption is correct and is rejected at a significant level.

Hypothesis 1 (H1): At a given level of significance, the test's assumption fails and is rejected. Before it may reject or fail to reject the null hypothesis, the test findings have to be evaluated [3]. Disregarding to the level of significance, testing of hypothesis results might contain errors. By estimating a determined structure for the result and then utilizing statistical techniques to validate or invalidate the estimation, data may be understood. The assumption is called hypothesis, and the statistical tests utilized to validate it is called statistical hypothesis tests. Whenever it is needed to investigate data distribution or when the results of the groups are different [2-4].

1.1 P-Value Interpretation

What is the significance of the p-value? It is possible to establish if a result is significant statistically by calculating p-value. It may, for instance, conduct a test for normality on a sample data and discover that the data sample seldom different from a Gaussian distribution, rejecting the null hypothesis [5]. The hypothesis test in statistics returns the p-value as a result. This is a statistic that may be used to evaluate or quantify the test's results and decide whether or not the null hypothesis should be rejected. This is accomplished by comparing the p-value to a significance level, which is a predetermined value [5]. Alpha is widely utilized to indicate the degree of significance. The most often used alpha value is 0.05. A smaller value of alpha, for instance 0.01 percent, implies a more robust null hypothesis interpretation [6]. Previously obtained the value of alpha will be compared to the p-value. The statistical significance of the result is obtained when the alpha is greater than the p-value. This implies that there has been a change: The default hypothesis is now ruled out [7] [8]. The null hypothesis is not rejected if the alpha is smaller than the p-value which means it is not a significant result. The null hypothesis must be rejected if the alpha is equal to the p-value which means it is a significant result [9]. For instance, if a test was run to see whether a sample of data was normal and the p-value was 0.07, the following might be told: The test failed to reject the null hypothesis at a 0.05 significance level, suggesting that the data sample was normal. You may generate a confidence level for the hypothesis based on the observed sample data by subtracting 1 from the significance level [10].

2. METHODOLOGY

One approach is to score and divide the data using the same k-fold cross-validation split for example, in each case, utilizing the same number of seed randomly. This will provide a sample of ten ratings for 10-fold cross validation. The values might be compared by utilizing paired statistical hypothesis test since each algorithm utilized the identical treatment which means rows of data to calculate every value. The paired student's t-test might be utilized. Thus, utilizing the paired Student t-test in this place of activity is that each model evaluation might not independent due to with the exception of the hold-out test fold, which uses a new row of data each time, to train the data several times in reality the same rows of data are utilized. The paired student t-test is biased in the perspective of optimistically due to the lack of evaluation flexibility. To account for the lack of flexibility, this statistical test may be adjusted. Furthermore, the number of folds and repetitions in the technique may be changed to produce a decent sample of model output that can be used to a broad variety of problems and methods. 52-fold cross-validation is a two-fold cross-validation method with five repetitions.

2.1 *Machine Learning xtend 5x2 method*

The paired t-test 5x2cv() function in Sebastian Raschka's MLxtend package offers an implementation. To utilize the evaluation, first import the dataset, then choose the two methods for comparison. The data and methods might then be utilized with the paired t-test 5x2cv() function, which will provide the t-test and p-value showing when the performance difference between the methods is significant.

```
# algorithms comparison
ttest, pvalue = paired_t_test_5x2cv(first_estimator=First_Model,
second_estimator=Second_Model, X=X, y=y)
```

To comprehend the p-value, the alpha value, which is the intended degree of significance, must be employed. When the same mean for performance is obtained for the methods, the null hypothesis is rejected if the given alpha if greater than or equal to the p-value, suggesting that the difference is most likely true. The null hypothesis of that methods that have the same mean of performance is not rejected if the p-value is larger than alpha, and any observed variance in mean accuracies is most likely attributable to chance. The lower the alpha value, the better, and 0.05 is commonly utilized.

```
# results interpretation
if pvalue <= 0.05:
    print('mean performance differences might be real')
else:
    print('Algorithms may perform in the same way')
```

3. CLASSIFIER ALGORITHMS COMPARISON

Now, on a binary classification problem, the results of the two machine learning algorithms can be compared, and see whether there is a statistically significant difference. To begin, the make classification() method will be utilized to build synthetic dataset with a 1,000 samples and 20 input variables. The following example makes a dataset and explains its structure.

```
# build dataset and definition
X, y = build_classification(m_sample=1000, m_feature=10, m_informative=10,
m_redundant=0, random_state=1)
# summary
print(X.shape, y.shape)
```

The dataset is built, and the rows and columns number is calculated, validating the assumptions. This data may be used to compare two different algorithms.

```
(1000, 10)
(1000,)
```

A comparison of the output of the two linear algorithms will be performed on the created dataset. To be particular, a Logistic Regression LR method and a Linear Discriminant Analysis LDA methodology. The most recommended approach is repeated stratified k-fold cross-validation with 10 folds and three repetitions. This approach will be used to evaluate each algorithm and calculate the average classification accuracy. The whole example is shown below.

```
# LR and LDA binary classification comparison
X, y = make_classification(m_sample=1000, m_feature=10, m_informative=10,
m_redundant=0, random_state=1)
# first model assessment
First_model = LogisticRegression()
cv1 = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
first_score = cross_val_score(first_model, X, y, scoring='accuracy', cv=cv1, n_jobs=-1)
print('LogisticRegression Mean Accuracy: %.3f (%.3f)' % (mean(first_score),
std(first_score)))
# second model assessment
Second_model = LinearDiscriminantAnalysis()
cv2 = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
second_score = cross_val_score(second_model, X, y, scoring='accuracy', cv=cv2,
n_jobs=-1)
print('LinearDiscriminantAnalysis Mean Accuracy: %.3f (%.3f)' % (mean(second_score),
std(second_score)))
# consequences plot
pyplot.boxplot([first_score, second_score], labels=['LR', 'LDA'], showmeans=True)
pyplot.show()
```

When the example is performed, the mean classification accuracy for each method is presented first. When the mean ratings are examined, it is clear that LDA beats logistic regression: logistic regression scores 89.2 percent, whereas LDA scores 89.3 percent.

```
LR Mean Accuracy: 0.892 (0.036)
LDA Mean Accuracy: 0.893 (0.033)
```

Because of nature of the stochastic or assessment technique of the algorithm, as well as changes in numerical precision. You can find that the outcomes are different if you repeat the example and compare the average results. If you simply looked at the averages, they seemed to be fairly excellent. As can be seen, LDA beats logistic regression: logistic regression scores 89.2 percent, whereas LDA scores 89.3 percent.

```
LR Mean Accuracy: 0.892 (0.036)
LDA Mean Accuracy: 0.893 (0.033)
p-value: 0.328, t-Statistic: 1.085
'Algorithms may perform in the same way'
```

To evaluate if the observed findings are statistically significant, a hypothesis test may now be used. It will begin by reviewing the algorithms and utilizing the 52 procedure to compute the p-value and t-test value.

```
# Check to see if the difference in algorithms is real.
ttest, pvalue = paired_ttest_5x2cv(first_estimator=first_model,
second_estimator=second_model, X=X, y=y, scoring='accuracy', random_seed=1)
# summarize
print('P-value: %.3f, t-Statistic: %.3f' % (pvalue, ttest))
# Check to see if the difference in algorithms is real.
ttest, pvalue = paired_ttest_5x2cv(first_estimator=first_model,
second_estimator=second_model, X=X, y=y, scoring='accuracy', random_seed=1)
# summary
print('P-value: %.3f, t-Statistic: %.3f' % (pvalue, ttest))
```

The outcomes might vary due to the nature of stochastic of the algorithm or assessment approach, as well as numerical precision variations. Consider re-enacting the scenario and comparing the mean of results. Thus, the p-value is roughly 0.3, which is much more than 0.05. As a consequence, it is incapable to reject the null hypothesis, meaning that any discernible difference between the algorithms is most likely unreal. It could just as well apply LR or LDA, and the results could be the same. This demonstrates that choosing a model just on the basis of average of performance night not be adequate.

```
LR Mean Accuracy: 0.894 (0.012)
LDA Mean Accuracy: 0.890 (0.013)
P-value: 0.328, t-Statistic: 1.085
'Algorithms may perform in the same way'
```

As a consequence, figure 1 shows means and accuracy of LR and LDA algorithms:

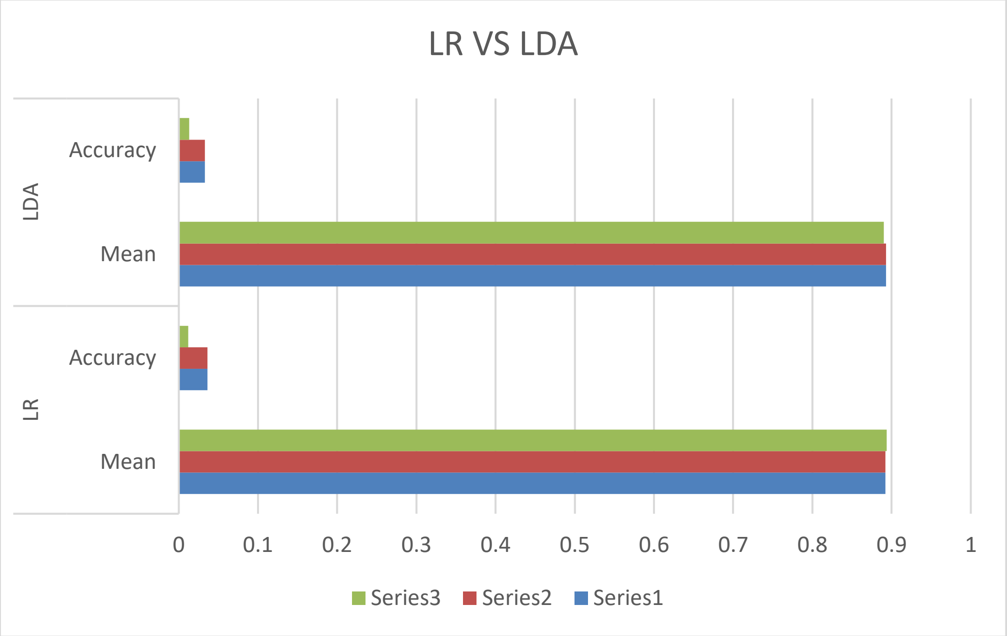

Figure 1: shows means and accuracy of LR and LDA

4. CONCLUSION

This study demonstrates the use of statistical hypothesis tests to evaluate machine learning algorithms. Furthermore, it instructs researchers on how to choose models based on model performance averages, which might be misleading. An appropriate methodology for comparing machine learning algorithms is five rounds of two folds of cross-validation using a adapted student t-test. Using MLxtend machine learning and a statistical hypothesis test, compare algorithms.

REFERENCE

[1] W. Daniel, C. Cross. 2013, “Biostatistics: Basic Concepts and Methodology for the Health Sciences”, 10th Edition International Student Version, chapter 6, ISBN: 978-1-118-65291-6, 2013.

[2] H. K. Hamarashid, S. A. Saeed, and T. A. Rashid, “Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji”. Neural Computing and Applications, 33(9), 4547-4566. 2021

[3] F. Emmert-Streib and M. Dehmer. “Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference”, https://doi.org/10.3390/make1030054, 2019.

[4] S. Hartshorn. “Hypothesis Testing: A Visual Introduction To Statistical Significance. USA, ASIN : B019N212NE. 2015.

[5] L. Surhone, M. Timpledon, and S. Marseken. “P-Value”, ISBN 6130502370, 9786130502379, VDM publishing. 2010.

[6] S. McLeod. “What a p-value tells you about statistical significance, 2019, [online] available at: https://www.simplypsychology.org/p-value.html. Accessed on (02/04/2021).

[7] T. Dahiru. “P - value, a true test of statistical significance? A cautionary note. Annals of Ibadan postgraduate medicine”, 6(1), 21–26. https://doi.org/10.4314/aipm.v6i1.64038, 2008.

[8] V. Johnson. “Revised standards for statistical evidence”. Proc Natl Acad Sci 110(48):19313–19317, 2013.

[9] R. Nuzzo. “Statistical errors: P values, the ‘gold standard’of statistical validity, are not as reliable as many scientists assume”. Nature 506:150–152. 2014.

[10] R. Wasserstein, N. Lazar. “The ASA’s statement on p-values: context, process, and purpose”. Am Stat 70(2):129–133. 2016.