ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.5, 2809–2815
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.5.42

A COMPARATIVE STUDY ON THE LOGISTIC REGRESSION AND NAÏVE BAYES MODELS UPON MEDICAL DATA THROUGH A MACHINE LEARNING APPROACH

S. PARTHASARATHY[1] AND V. MADHU

ABSTRACT. Machine learning models plays a vital role in medical data analysis. This article deals with the comparison of two machine learning models for medical data. Based on the precision, recall, f-score we estimated model accuracy and identified best model among Logistic Regression and Naive Bayes.

1. INTRODUCTION

Machine Learning in medicine has been reliably hopeful, backed by constantly available and ever-flourishing data. "Disease identification and diagnosis of ailments are at the forefront of ML research in medicine, [1]. It is no surprise that large players were some of the first to jump on the bandwagon, particularly in high-need areas like cancer identification and treatment. In Gregorian calendar month 2016, IBM Watson Health announced IBM Watson Genomics which aims to make strides in precision medicine by integrating cognitive computing and genomic tumor sequencing. Current analysis comes current include dosage trials for blood vessel tumour treatment and detection and management of prostate cancer"- says Daniel Faggella, a sought-after expert on the competitive strategy implications of AI for business and government leaders.

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 62J02.

Key words and phrases. Logistic regression, naive Bayes, machine learning, medical data.

2. MODELS

Data analysis ways is also delineated by their areas of applications, except for this text, i.e. exploitation definitions that area unit strictly methods-oriented, [8].

A Statistical Model (SM) may be a data model that comes with possibilities for the information generating mechanism and has identified unknown parameters that are sometimes explainable and of interest, e.g., effects of predictor variables and spacing parameters regarding the result variable, [1]. The foremost normally used SMs are regression models, that probably leave a separation of the results of competitor predictor variables, [7].

SMs embrace standard regression, Bayesian regression, semiparametric models, generalized additive models, longitudinal models, time-to-event models, penalised regression, and others. penalised regression includes ridge regression, lasso, and elastic net. Contrary to what some Machine Learning (ML) researchers believe, SMs simply leave quality (nonlinearity and second-order interactions) and a limitless range of candidate options, [4]. It's particularly simple, exploitation regression splines, to permit each continuous predictor to possess a sleek nonlinear result, [5]. It assumes that every one, predictor have a linear result on the result, which the model is absolutely additive. This can be as SM in concert can get, [8, 9].

ML sometimes doesn't decide to isolate the result of any single variable. ML doesn't model the information generating method however rather makes an attempt to be told from the dataset at hand. ML is additional a neighbourhood of technology than it's a part of statistics. Maybe the only thanks to distinguish ML kind SMs is that SMs favour additivity of predictor effects whereas ML sometimes doesn't, [3].

In this article we are using the Melanoma data to compare two statistical model Logistic regression and Naive Bayes.

2.1. **Logistic Regression.**

In statistics, multinomial logistic regression is a classification approach that generalizes logistic regression to multiclass problems, that is with more than two viable discrete results. That is, it is a model which is used to predict the possibilities of the different possible results of a categorically distributed dependent variable, given a fixed of unbiased variables, [2].

The multinomial logistic model assumes that information are case specific; that is, every impartial variable has a single price for every case. The multinomial logistic version additionally assumes that the established variable cannot be perfectly predicted from the impartial variables for any case. As with other kinds of regression, there is no need for the independent variables to be statistically independent from each other (unlike, for example, in a naive Bayes classifier); however, collinearity is assumed to be rather low, as it becomes tough to distinguish between the effect of numerous variables if this is not the case.

If the multinomial logit is used to version choices, it relies on the idea of independence of irrelevant alternatives, which isn't continually desirable. This assumption states that the percentages of who prefer one class over some other do not depend upon the presence or absence of other "irrelevant" alternatives.

When using multinomial logistic regression, one category of the structured variable is chosen because the reference category. Separate odds ratios are determined for all impartial variables for each class of the established variable excluding the reference class, which is omitted from the analysis. The exponential beta coefficient represents the alternate within the odds of the based variable being in a particular class vis-a-vis the reference category, related to a one unit alternate of the corresponding unbiased variable.

As in other forms of linear regression, multinomial logistic regression uses a linear predictor function $C(n, i)$ to predict the probability that observation i has outcome $'n'$, of the form

$$C(n, i) = \delta_{0,n} + \delta_{1,n}x_{1,i} + \ldots \delta_{M,n}x_{M,i}$$

where $\delta_{M,n}$ is a regression coefficient associated with the M^{th} explanatory variable and the n^{th} outcome.

$$Z = \sum_{n=1}^{N} e^{\delta_n X_i}$$

where z represents the output of the linear model, where X_i is the vector of explanatory variables describing observation i, δ_n is a vector of weights corresponding to outcome n, and score(X_i, n) is the score associated with assigning observation i to category n.

2.2. **Naive Bayes.** The Naive Bayes (NB) model - is a simple but surprisingly powerful algorithm for predictive modeling. In machine learning, we are often

interested in selecting the best hypothesis (h) given data (d).
In a classification drawback, our hypothesis (h) may be the class to assign for a new data instance (d).
Bayes Theorem is stated as:

$$P\left(h|d\right) = \quad \frac{\left(P\left(d|h\right) * \ P\left(h\right)\right)}{P\left(d\right)}$$

Naive Thomas Bayes could be a classification rule for binary (two-class) and multi-class classification issues, [4]. The technique is best to grasp once represented exploitation binary or categorical input values.

The representation of naive Bayes is probabilities. A list of probabilities is stored to file for a learned Naive Bayes model.
This includes: class probabilities: the possibilities of every category within the training dataset; conditional probabilities: the conditional probabilities of each input value given each class value. On account of discrete data sources, Naive Bayes classifiers structure a generative-discriminative pair with Logistic regression classifiers: each Naive Bayes classifier can be viewed as a method for fitting a likelihood $p(H, d)$ model that upgrades the joint probability while calculated relapse fits a similar likelihood model to enhance the contingent $P(H|d)$.

The connection between the two can be seen by seeing that the choice capacity for Naive Bayes can be modified as "anticipate class H_1 if the chances of $p(H_1, d)$ surpass those of $p(H_2, d)$. Communicating this in log-space gives:

$$log\frac{p(H_1/d)}{p(H_2/d)} = logp\left(H_1/d\right) - log(H_2/d) > 0$$

The left-hand side of this condition is the log-chances, or logit, the amount anticipated by the linear model that underlies calculated relapse. Since Naive Bayes is likewise a direct model for the two "discrete" event models, it very well may be reparametrized as a linear function $a + B_k^T x > 0$, where $a = \log p\left(H_k\right)$ and $B_k = logp_{ki}$.

Discriminative classifiers have lower asymptotic mistake than generative ones; nonetheless, look into by Ng and Jordan has indicated that in some down to earth cases Naive Bayes can outperform Logistic regression since it arrives at its asymptotic error faster, [6].

3. Data Base

In this section we have considered the 'Survival from Malignant Melanoma' data for comparison of Logistic Regression and Naive Bayes model. We have analyzed the data using Python software for our calculations'.

The Melanoma data involves 205 rows and 7 columns. The following variables are considered for modelling whose descriptions are given below in the table 3.0

Table 3.0: List of Variable names

Time	Survival time in days since the operation, possibly censored
Status	1. Indicates that they had died from melanoma 2. Indicates that they were still alive 3. Indicates that they had died from causes unrelated to their melanoma.
Sex	1- Male, 0- Female
Age	Age in years at the time of the operation
Year	Year of operation
Thickness	Tumour thickness in millimeter
Ulcer	1-Present, 0- Absent

4. Discussion and conclusion

This comparative study on the accuracy of the models in the analysis of medical data would let us conclude that the LR model is more accurate or the NB model.

For this, we consider medical data on people affected by melanoma - a tumor of melanin-forming cells, especially a malignant tumor associated with skin cancer, considering different parameters like the patient's age, gender, the state of their ulcer, the thickness to classify the status of the patient under the distinct classifications dead, alive and dead but not out of melanoma. The results of this analysis their accuracy decides which model from among the LR and NB, is the better one.

From table 3.0, we used the Melanoma data frame, it shows that, it has 205 rows and 7 columns. The data consist of measurements made on patients with malignant melanoma. Each patient had their tumour removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during

the period 1962 to 1977, [1]. When we analyzed under both the models, with 25% and 40% of the point from the data set being considered, to obtain the accuracy of the model, [3, 4].

Courtesy: Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) Statistical Models Based on Counting Processes. Springer-Verlag.

Table 3.1: Comparison of Logistic and Naive Bayes Model

Model	Logistic Regression						Naïve Bayes					
	At 25% testing set			At 40% testing set			At 25% testing set			At 40% testing set		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
1	.85	.85	.85	.82	.90	.86	.75	.69	.72	.64	.70	.67
2	.95	1	.97	.97	1	..98	.9	.97	.94	.88	.91	.90
3	0	0	0	0	0	4	0	0	0	0	0	4
Weighted Average	.6	.62	.61	.59	.93	.90	.83	.87	.85	.78	.82	.80
Model Accuracy	92.30			92.68			86.53			81.70		

From table 3.1 LR Model and NB models exhibited accuracy at 25% testing set respectively 92.3 and 86.53. When compared to NB model LR model gives best accuracy. Similarly increase the testing set size from 25% to 40%, we can see the LR model gives more accuracy than NB model. This concludes that when we increase the testing set size LR models perform well than the NB model. And thus, the LR model could be used to deliver comparatively more accurate analysis results on the statuses of patients suffering from melanoma.

REFERENCES

[1] P. K. ANDERSEN, O. BORGAN, R. D. GILL, N. KEIDING: *Statistical Models Based on Counting Processes*, Springer-Verlag, 1993.

[2] D. BELSLEY: *Conditioning diagnostics: collinearity and weak data in regression*, Wiley, New York, ISBN 9780471528890.

[3] V. J. CAREY: *Machine learning concepts and tools for statistical genomics*, Bioinformatics and computational biology solutions using R and Bioconductor, Springer, New York, 2005, 273–292.

[4] S. DUDOIT, J. FRIDLYAND, T. SPEED: *Comparison of discrimination methods for the classification of tumors using gene expression data*, J. Am. Stat. Assoc., **97**(457) (2002), 77–87.

[5] T. HASTIE, R. TIBSHIRANI, J. FRIEDMAN: *The elements of statistical learning: Data mining, inference, and prediction*, Springer, New York, 2001, 533.

[6] A. U. NG, M. I. JORDAN: *On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes*, Proceedings of 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, 2001, 841–848.

[7] F. ROSENBLATT: *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychol. Rev., **65**(6) (1958), 386–408.

[8] W. N. VENABLES, B. D. RIPLEY: *Modern Applied Statistics with S-Plus*, Springer-Verlag, 1994.

[9] J. P. WILLEMS, J. T. SAUNDERS, D. E. HUNT, J. B. SCHORLING: *Prevalence of coronary heart disease risk factors among rural blacks: A community-based study*, Southern Medical Journal, **90**(8) (1997), 814–820.

DEPARTMENT OF MATHEMATICS
SRM IST, RAMAPURAM
CHENNAI-89, INDIA
E-mail address: parthass@srmist.edu.in

DEPARTMENT OF MATHEMATICS
VIT, VELLORE CAMPUS INDIA
E-mail address: madhu.riasm@gmail.com