ISSN 2347-1921
Volume 12 Number 09
Journal of Advances in Mathematics

Comparison Study of Logistic Regression Model for Albanian Texts.

Denisa Salillari[1], Luela Prifti[2]
[1]Department of Mathematical Engineering, Polytechnic University of Tirana, Sheshi "Nene Tereza", nr. 1, Tirana, Albania
salillaridenisa@yahoo.com
[2]Department of Mathematical Engineering, Polytechnic University of Tirana, Sheshi "Nene Tereza", nr. 1, Tirana, Albania
luela_p@yahoo.com

ABSTRACT

Considering authorship attribution as a classification problem we attempt to estimate the probability to find the right author for each text under study. In this paper using R we first improve the simple model for six Albanian texts, (I) increasing number of texts and number of independent variables and then compare the results taken with them of the multinomial logistic regression (II). The model was applied on a set of one hundred texts of ten different authors. For all the authors under study the average correct predicted probability is 0.918. Analyzing data from different Albanian texts, results that about 40% of their letters consist of vowels. As conclusion comparing results taken with them of (II) multinomial logistic regression model for Albanian texts has more advantages than logistic regression model.

Keywords

Logistic regression; classification, R

Academic Discipline And Sub-Disciplines

Statistics;

SUBJECT CLASSIFICATION

Statistical Subject Classification;

TYPE (METHOD/APPROACH)

Application of logistic regression model.

1. INTRODUCTION

Logistic regression refers to a classifier that classifies an observation into one of two classes. We find in statistical classification theory different statistical models, the logistic regression is considered as a linear method for classification (III) which is implemented in many different softwares, one of them is the popular statistical software R (IV). Considering authorship attribution as a classification problem we attempt to estimate the probability to find the right author for each text under study. Logistic regression forms a best fitting model using the maximum likelihood method, which maximizes the probability of classifying the observed data into the appropriate category. In our previous paper (I) we defined a model for six Albanian texts using logistic regression, as a classification statistical method considering the authorship of a single author for a given text. As result the parameters used wasn't the best due to the small number of text taken for study. In this paper using R we improve the logistic regression model increasing the number of texts and number of independent variables. The application is realized with data from one hundred texts of ten different authors, considering as the independent variables in the model, number of letters, number of words, number of vowels, number of consonants, number of punctuations and number of sentences. Analyzing these Albanian texts data it results that about 40% of their letters consist of vowels with a 95% confidence interval of]0.3985;0.4032[that explains the high correlation between number of letters and number of vowels. By reviewing different cases of model we defined the most significant independent variables, as result for all the authors under study the average correct predicted probability was 0.918.

2. METHOD AND RESULTS

Logistic regression analysis belongs to the class of generalized linear models. These models are characterized by their response distribution and a link function, which transfers the mean value to a scale in which the relation to background variables is described as linear and additive. Logistic regression is a mathematical model through logistic function which is used to indicate the relationship between independent random variables with a qualitative dependent variable with two values 0/1 (dichotomous, dummy).
In a logistic regression analysis, the link function is logistic function $logit\, p = log[\, p/(1-p)]$. A binary classifier based on a logistic regression model learns the mapping of a feature vector x to a category label assignment y_k for the $k-th$ category label by modeling conditional probability $P(y_k|\boldsymbol{x})$ directly.

ISSN 2347-1921
Volume 12 Number 09
Journal of Advances in Mathematics

The conditional probability is modeled as $P(Y=1|X)=\frac{e^X}{1+e^X}=\frac{e^{(\alpha+\sum_i \beta_i x_i)}}{1+e^{(\alpha+\sum_i \beta_i x_i)}}$ considering X as a linear combination of x_i. For a text we must make a statistical decision whether this text belongs or not to a particular author. We have applied the logistic regression model in one hundred texts of ten different authors. All formatting are removed from the texts. For each text formatted in .doc file we have calculated number of letters, number of words, number of vowels, number of consonants, number of punctuations and number of sentences. The model is defined using R software for different case of study. Considering as the independent variable all the above variables in the model not all the parameters are statistically significant. Reviewing different cases of model it results that number of letters variable is correlated with both number of vowels and number of consonants variables. Referring to the above mentioned data we have obtained a simple logistic regression model for author 1 using glm() function in R:

```
model<-glm(Author1~N_words+N_letters+N_vowels+N_consonant+N_sentences+N_punctuations,
data = paper,family= binomial())

> summary(model1,corr=T)

Call:
glm(formula = Author1 ~ N_words + N_letters + N_vowels + N_consonant +
    N_sentences + N_punctuations, family = binomial(), data = paper)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4186  -0.2095  -0.1057  -0.0315   3.5414

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.102289   1.761839  -0.058  0.95370
N_words        -0.011986   0.011585  -1.035  0.30085
N_letters       0.006257   0.003962   1.579  0.11428
N_vowels       -0.004237   0.007422  -0.571  0.56810
N_consonant           NA         NA      NA       NA
N_sentences     0.145752   0.045147   3.228  0.00124 **
N_punctuations -0.162807   0.054598  -2.982  0.00286 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 29.031  on 94  degrees of freedom
AIC: 41.031

Number of Fisher Scoring iterations: 7

Correlation of Coefficients:
              (Intercept) N_words N_letters N_vowels N_sentences
N_words        -0.28
N_letters       0.06       -0.57
N_vowels        0.07        0.10    -0.85
N_sentences    -0.17        0.15     0.09     -0.01
N_punctuations  0.00       -0.26    -0.14      0.10    -0.85
```

From the results only two of the parameters of the model are significant and there is high correlation between number of letters and number of vowels and between number of punctuations and number of sentences. There are 36 letters in Albanian alphabet from which 7 are vowels. Analyzing data from different Albanian texts, results that about 40% of their letters consist of vowels with a 95% confidence interval of]0.3985;0.4032[and 60% are consonants. This explains the correlation between number of letters and number of vowels or number of consonants in the Albanian texts. In the number of punctuations is calculated and number of dots which indicate the number of sentences, this explains the correlation between number of punctuactions and number of sentences.

```
> pvowel<-paper$N_vowels/paper$N_letters
> summary(pvowel)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3236  0.3986  0.4029  0.4008  0.4070  0.4131
> t.test(pvowel,mu=0.4)

        One Sample t-test

data:  pvowel
t = 0.69933, df = 99, p-value = 0.486
alternative hypothesis: true mean is not equal to 0.4
95 percent confidence interval:
 0.3984701 0.4031952
```

```
sample estimates:
mean of x
0.4008327
```

Analyzing different models one of the best fitted is achieved considering as the independent variable number of word, number of letters, number of sentences and number of punctuations.

```
> Model1 <- glm(Author1 ~ N_words+N_letters+N_sentences+N_punctuations, data =
paper,family = binomial())
> summary(Model1,corr=T)

Call:
glm(formula = Author1 ~ N_words + N_letters + N_sentences + N_punctuations,
    family = binomial(), data = paper)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4454  -0.2173  -0.1020  -0.0303   3.5532

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.062957   1.725130   0.036  0.97089
N_words        -0.011299   0.011095  -1.018  0.30847
N_letters       0.004416   0.002038   2.166  0.03028 *
N_sentences     0.149242   0.045383   3.288  0.00101 **
N_punctuations -0.165346   0.054129  -3.055  0.00225 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 29.402  on 95  degrees of freedom
AIC: 39.402

Number of Fisher Scoring iterations: 7

Correlation of Coefficients:
               (Intercept) N_words N_letters N_sentences
N_words        -0.30
N_letters       0.23       -0.91
N_sentences    -0.18        0.14    0.18
N_punctuations  0.01       -0.26   -0.14      -0.86
```

According to the output, the model is:

$$\text{logitP(y)} = \ln\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = 0.062957 - 0.011299\text{N}_{\text{words}} + 0.004416\text{N}_{\text{letters}} - 0.165346\text{N}_{\text{punctuations}} + 0.149242\text{N}_{\text{sentences}}$$

Testing if it is true that the fifth text belonging to the first author:

$$P(Y = 1|X_5) = 9.660309e - 01 \quad , \qquad \text{P(Y} = 0|\text{X}_5) = 1 - 9.660309e - 01 = 0.339698\text{e} - 01$$

we note that

$$P(Y = 1|X_5) > \text{P(Y} = 0|\text{X}_5)$$

This shows that the text is written by the first author. These probabilities are calculated using the function predict(). As result the corrected predicted probability for this model is 0.96. This result shows that 96% of texts are classified correctly in the fitted model. For all the authors under study the average correct predicted probability is 0.918. In our previous work (II) we had defined a multinomial logistic regression model which determines as the most significant independent variables, number of words, number of vowels, number of consonants, number of punctuation and number of sentences, with the highest overall correct predicted probability 0.738. The logistic regression model determines as the most significant independent variable, number of letters, number of sentences, number of punctuations. Recognizing as many linguistic features of an author should give good mathematical models for authorship attribution of texts. In logistic regression models for Albanian texts results that not all the parameters are statistically significant, from six independent variables only three of them defined most significant while in the multinomial logistic regression model five variables results significant. Logistic regression model gives higher predicted probability than multinomial logistic regression model but we had to define as logistic regression models as the number of the authors. The problem is that it will take a lot of time to find the right model due to the number of authors while we define the authorship of a text, with one only model of

multinomial logistic regression. As conclusion multinomial logistic regression model for Albanian texts has more advantages than logistic regression model.

3. CONCLUSIONS

In our logistic model we used six independed variables drawn from 100 texts of ten different Albanian authors. Analyzing these Albanian texts data it results that about 40% of their letters consist of vowels with a 95% confidence interval of]0.3985;0.4032[that explains the high correlation between number of letters and number of vowels. By reviewing different cases of model we defined as most significant independent variables, number of letters, number of sentences, number of punctuations. As result for all the authors under study the average correct predicted probability is 0.918. Comparing the results taken in this paper with them taken in multinomial logistic regression model (II) we conclude that multinomial logistic regression model for Albanian texts has more advantages than logistic regression model.

REFERENCES

I. D. Salillari, L. Prifti, Sh. Kuka “Logistic regression for authorship attribution in albanian text ” 7th Annual Meeting of Institute Alb-Shkenca, Conference Of Natural Sciences.

II. D. Salillari, L.Prifti A multinomial logistic regression model for text in Albanian language, Journal of Advances in Mathematics, Volume 12 Number 07.

III. T.Hastie, R.Tibshirani, J.Friedman “The Elements of Statistical Learning” Data Mining, Inference, and Prediction Second Edition.

IV. G. James, D. Witten, T. Hastie, R. Tibshirani “An Introduction to Statistical Learning” with applications in R. Springer 2013