A multinomial logistic regression model for text in Albanian language

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

In this paper we present a multinomial logistic regression model for authorship identification in the Albanian language texts. In the model fitted the dependent variable is categorical which takes different values from 1 to 10 for each of the author and the independent variables are number of words, number of letters, number of vowels, number of consonants, number of punctuations and number of sentences for each text. The model was applied with success in the set of ten authors, each of them being represented by a set of one hundred texts they authored. As results first, second and the third authors have the higher correct predicted percentage and the highest overall correct predicted probability taken was 0.738. As conclusion adding in the model number of consonants, number of punctuations and number of sentences as independent variables the overall correct predicted percentage is increased.

Keywords

Multinomial logistic regression; classification.

Academic Discipline And Sub-Disciplines

Statistics;

SUBJECT CLASSIFICATION

Statistical Subject Classification;

TYPE (METHOD/APPROACH)

Application of multinomial logistic regression model.

1. INTRODUCTION

Different statistical methods are applied to identify the authorship of a text. To recognize many linguistic features of an individual, we are developing mathematical models of letters, vocabulary and syntax of a language. In our previous work we have developed different Markov chains models defined on n-gram of letters for Albanian linguistic elements estimating the texts entropy via Markov chains of high order and fitted logistic regression models for a set of 6 books where number of words, number of letters and number of vowels are considered as independent variables. For a text we must make a statistical decision whether this text belongs or not to a particular author. In this paper Authorship identification is considered as a statistical classification problem increasing the number of texts. We find in statistical classification theory different statistical models one of which is the logistic regression I and II. In this paper a multinomial logistic regression model is fitted as generalization of previous logistic regression models. Such models have been treated for English texts by Genki III. In the model fitted we use the Albanian linguistic elements of syntax treated in IV. Applying the models we define if a text belongs to the set of ten authors, each of them being represented by a set of one hundred texts they authored. The application gives good results for each of the author identification. First, second and the third authors have the higher correct predicted percentage. Adding in the model number of consonants, number of punctuations and number of sentences as independent variables the overall correct predicted percentage is increased.

2. MULTINOMIAL LOGISTIC REGRESSION

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). Logistic function is called the function $f\colon R \to R$ defined by the equation $f(X) = \frac{1}{1+e^{-X}} = \frac{e^{X}}{1+e^{X}}$.

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 $P(y_k|x)$ directly. 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. Logistic regression can be easily generalized to multiple classes. Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. For a depended variable with K categories, this requires the calculation of K-1 equations, one for each category relative to the reference category, to describe the relationship between the depended variable and the independent variables. In order to predict the authorship of a text by a set of known authors the model of multinomial logistic regression is given as follows. Let the feature vector of a text be denoted by $x = [x_1, \dots, x_j, \dots, x_d]^T$. We encode the fact that a text belongs to an author $k \in \{1, 2, 3, \dots, K\}$ by a vector $y = [y_1, \dots, y_K]^T$ where $y_k = 1$ and all other coordinates are 0. Multinomial logistic regression is a conditional probability model, defined via the softmax function

$P(y_k = 1 \mid x, B) = \frac{\exp\,(\beta_k^T x)}{\sum_{j=1}^{K} \exp\,(\beta_j^T x)}$ parameterized by the matrix $B = [\beta_1, \dots, \beta_K]^T$ is parameter model matrix where Each column of B is a parameter vector corresponding to one of the classes: $\beta_k = [\beta_{k1}, \dots, \beta_{kd}]^T$. Classification of a new observation is based on the vector of conditional probability estimates produced by the model. In this paper we simply assign the class with the highest conditional probability estimate: $\hat{y}(x) = arg\, \underbrace{\arg max}_{k} P(y_k = 1 \mid x)$. Consider a set of training examples $D = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$. Maximum likelihood estimation of the parameters B is equivalent to minimizing the negated log-likelihood:

$$l(B|D) = -\sum_i [\sum_k y_{ik} \beta_k^T x_i - ln \sum_k \exp\,(\beta_k^T x_i)].$$

3. THE MODELS FOR ALBANIAN TEXTS

There are 36 letters in Albanian alphabet from which 7 are vowels. We have applied the multinomial logistic regression model in one hundred texts of ten different authors. For each text formatted in .doc file we have calculated letters frequency, word frequency, vowels frequency, consonants frequency, punctuations frequency and number of sentences. The model is defined using the SPSS software for different case of study. In the first case we used as the independent variable all the above variables. The model fitted well the data and all the parameters are statistically significant but the number of vowels variable results redundant. So in the next case of study we had to exclude this variable. This result is because of the high correlation between number of letters and number of vowels. For the same reason we had to exclude and the number of consonants. Excluding the number of consonants we get improvement of parameters too. But the best model was taken excluding number of letters variable which consist both of number of vowels and number of consonants. The results are given below.

Table 1: The model fitting information

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	460.517			
Final	206.479	254.038	45	.000

As it is shown in table 1 the model gives adequate predictions compared to the intercept because the -2 Log Likelihood is decrease in the final model and the p-value is <0.05.

The results in table 2 conclude that the model fits the data due to p-value is > 0.05.

Table 2: The goodness-of-fit information of the data

Goodness-of-Fit

	Chi-Square	Df	Sig.
Pearson	656.324	846	1.000
Deviance	206.479	846	1.000

Using pseudo r square regarding to Cox and Snell coefficient the result presented in table 3 shows that about 92% of the variation in the authors are explained from the multinomial logistic model.

Table 3: Pseudo R-Square coefficients

Pseudo R-Square

Cox and Snell	.921
Nagelkerke	.930
McFadden	.552

According to the table 4 results, all the parameters are statistically significant due to p-value <0.05.

Table 4. Likelihood ratio tests of model

Likelihood Ratio Tests

	Model Fitting Criteria	**Likelihood Ratio Tests**		
Effect	**-2 Log Likelihood of Reduced Model**	**Chi-Square**	**Df**	**Sig.**
Intercept	259.230	52.751	9	.000
N_words	276.699	70.220	9	.000
N_consonant	262.179	55.699	9	.000
N_vowels	230.636	24.156	9	.004
N_punctuations	234.433	27.954	9	.001
N_sentences	249.452	42.973	9	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

In the classification table results presented in table 5 the overall correct predicted percentage is 59%. This result shows that 59% of texts are classified correctly in the fitted model. First second and the third author has the high overall correct predicted percentage.

Table 5

Classification

Observed	Predicted										
	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	Percent Correct
1.00	9	1	0	0	0	0	0	0	0	0	90.0%
2.00	1	7	0	0	0	0	0	2	0	0	70.0%
3.00	0	0	10	0	0	0	0	0	0	0	100.0%
4.00	0	0	0	6	2	1	0	0	0	1	60.0%
5.00	0	0	0	3	3	2	0	0	2	0	30.0%
6.00	0	1	0	2	1	3	0	0	2	1	30.0%
7.00	1	0	0	0	0	0	6	3	0	0	60.0%
8.00	1	1	0	0	0	0	3	4	0	1	40.0%
9.00	0	0	0	1	1	0	0	0	6	2	60.0%
10.00	0	0	0	1	0	2	1	0	1	5	50.0%
Overall Percentage	12.0%	10.0%	10.0%	13.0%	7.0%	8.0%	10.0%	9.0%	11.0%	10.0%	59.0%

In all the cases studied for the data as described above the overall correct predicted percentage is 59%. It decreases for the other cases when we exclude more variables. As conclusion we say that increasing number of independent variable increases the overall correct predicted percentage.

In our data we notice the presence of outliers which correspond to the same texts for the entire variables. So in the next we define the model without outliers which increase the overall predicted percentage to 64%. We notice in table 5 that the low predicted percentage is for the fifth and sixth authors. Making multiple comparisons among the means for all the authors' variables we note that all the authors are divided in two groups. The first group of authors 2, 3, 6, and 9 results they have no significance difference of means and the second group of authors 1,4,5,7,8 and 10 which results they have no significance difference of means too. Because the fifth and sixth authors belong to different groups are likely the same with the other authors and have the lowest predicted percentage we had to exclude their data from the dataset in the next model. So we define another model excluding data for these two authors where the overall predicted percentage increases to 73.8%. In all the models the third author was predicted correctly this is because he writes with long sentences and use less punctuations different from the other authors. The best multinomial logistic regression in Albanian texts is achieved increasing the number of independent variables and for authors that have individual text data features. The parameters used in these models maybe are not the best due to the small number of dependent variable taken for study.

4. CONCLUSIONS

The fitted multinomial logistic regression model allows analyzing the qualitative and quantitative variable like independent variable different from the Markov chains models which analysis only the appearance frequency of a variable after a defined variable. In our multinomial logistic model we used six depended variables drawn from 100 texts of ten different Albanian authors. As result only five variables fitted the best model. The highest correct predicted probability taken was 0.738 although the model best fitted in all the case of study. First, second and the third authors have the higher correct predicted percentage. In all the models the third author was predicted correctly this is because he writes with long sentences and use less punctuations different from the other authors. So the model gives good predictions for the authors with individual characteristics of writing. Adding in the model number of consonants, number of punctuations and number of sentences as independent variables the overall correct predicted percentage is increased.

REFERENCES

1. Alan Julian Izenman Modern Multivariate Statistical Techniques Regression, Classification,and Manifold Learning.
2. T. Zhang and F. Oles. Text categorization based on regularized linear classifiers. Information Retrieval, 4(1):5.31, April 2001.
3. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization., 2004
4. D. Salillari, L. Prifti, Sh. Kuka “Logistic regression for authorship attribution in albanian text ” Alb-shkenca Conference