Prospeção de Dados 2021/2022

First Home Assignment

Group 4
Catarina Canastra (57766) - 31h contribution
Daniel Dias (59056) - 27h contribution
João Raimundo (57454) - 28h contribution

March 20, 2022

Abstract

The naive bayes is a classification algorithm that follows the bayesian modelling. Key features include simplicity (based on count ratios), stability, fast to learn and make predictions, and updatable. In this research, gaussian naive bayes and categorical naive bayes are used as a learning methods to produce the best possible multiclass and binary classification models. The methodology includes the use of data processed in different ways, reserve a data set for final validation, determine the best set of features simultaneously with the best value of a hyperparameter, choose the best model of each algorithm and choose the best one based on model validation. The best model for multiclass classification and the chosen model for binary classification is equally CategoricalNB.

Introduction

The naive bayes (NB) is a classification algorithm based on bayes theorem, that is suitable for binary and multiclass classification. It works by firstly converting the data set into a frequency tabel. Afterwards, creating likelihood table by finding the probabilities and, finally, using the naive **bayesian equation**

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \tag{1}$$

to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction. Because it assumes that the value of a particular feature in a class is independent of the value of any other feature, it performs better in cases of categorical input features than numeric variables [2]. Gaussian naive bayes (GaussianNB) is a special type of NB algorithm. It is used when the features have continuous values and assumes that all the features are following a gaussian distribution [1]. Whereas categorical naive bayes (CategoricalNB) algorithm is suitable for classification with discrete features that are categorically distributed. The Laplace's Alpha is a additive smoothing parameter of CategoricalNB that handles the problem of zero probability in naive bayes. This happens when a internal probability is zero and, consequently, this class will never be chosen because the probability of a point belonging to that class will be zero [3].

1 Methodology

The following steps were performed to determine the best possible naive bayes classification model between the two previously documented models:

1. Starting with an understanding of the data in order to get to know the data set and determine the preprocessing required for classification tasks;
2. Preprocessing data:
   - Discretize numerical features (based on the distribution of quantiles);
   - Encode all the features (a new binary variable is added for each unique integer value).
3. Separate the data into training and test sets;
4. Determine the best alpha (hyparameter of CategoricalNB) simultaneously with the best feature set;
5. Choose the best model of each algorithm;
6. Train the models and check the performance with the independent validation set;
7. Choose one and compare with the results of the previous phases.

The code developed to implement the methodology described was written in *Python* programming language, in the *Jupyter Notebook* format, using the *Anaconda* package manager.

2 Exploratory Data Analysis (EDA)

One of the most important needs when modelling is to know about the data. Exploratory data analysis is the process of performing initial investigations on data so as to spot anomalies, discover patterns, test hypothesis and check assumptions with the help of summary statistics and graphical representations. Dataset comprises of 11813 observations and 15 characteristics. Out of which one is dependent variable and rest 14 are independent variables. Target variable is discrete and categorical in nature. No variable column has null/missing values. There are 26 duplicated values that need to delete. A descriptive table provides various summary statistics, where it can be noticed a large difference between 75th quantile and max values of predictors. This observation suggests that there are extreme values-outliers in data set, however, the propotion of outliers is not very significant in this context/problem. Class distribution showed that most values are concentrated in categories 10, 6 and 8, respectively. By visualize the correlation matrix using heatmap in seaborn, it is inferred that "danceability" has positive correlation with "valence", and "energy" has strong positive correlation with "loudness".

3 Results

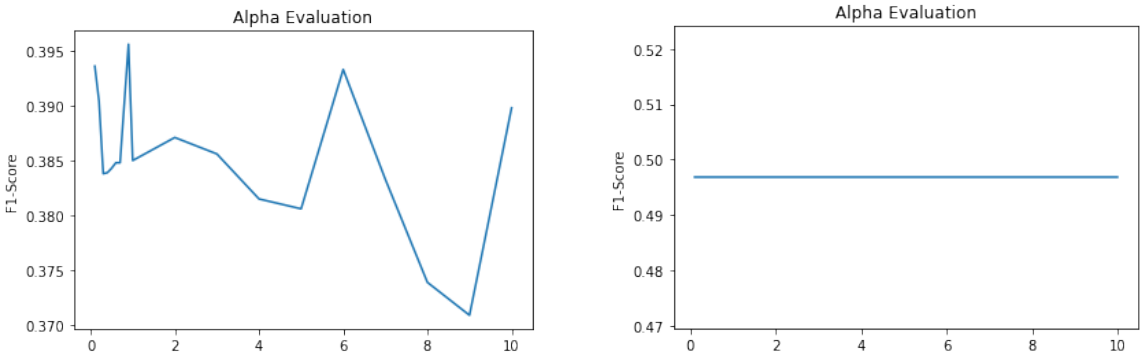
In the multiclass classification task:

- The best set of features for **GaussianNB** classifier is: ['instrumentalness_High', 'acousticness_High', 'duration_in min/ms_Low', 'valence_High', 'danceability_High', 'speechiness_Low', 'speechiness_Medium', 'duration_in min/ms_High', 'loudness_Low', 'energy_Medium', 'key_9', 'energy_Low', 'time_signature_3', 'valence_Medium'], with a final F1-score ~ 0.31;
- Whereas the best set of variables for **CategoricalNB** classifier is: ['instrumentalness_High', 'duration_in min/ms_Low', 'acousticness_High', 'valence_High', 'danceability_High', 'speechiness_High', 'speechiness_Low', 'time_signature_3', 'acousticness_Low', 'valence_Low', 'energy_Medium', 'instrumentalness_Low', 'duration_in min/ms_High', 'key_5', 'loudness_High', 'key_2', 'key_3', 'tempo_Medium', 'tempo_Low', 'key_4', 'key_9', 'mode_1', 'liveness_High', 'time_signature_4'], with alpha 0.9 reaching a score of ~ 0.40;
- GaussianNB was 32% accurate in classifying the data points in the validation phase, and scored ~ 0.30.
- The CategoricalNB had an accuracy of 42% and, in the model validation, it scored 0.38.
- So, the best multiclass classification model was CategoricalNB.

In the binary classification objective:

- In the **GaussianNB**, the best set of independent variables is: ['key_3', 'mode_1', 'acousticness_Low', 'mode_0', 'tempo_Medium'], with a final F1-score ~ 0.53;
- For the **CategoricalNB** the following features were chosen: ['valence_Medium'], with alpha 1 reaching a score of ~ 0.50;
- The GaussianNB had 0.97 precision and score ~ 0.53 with the independent validation set;
- Whereas CategoricalNB was 99% of correctly classified points and score ~ 0.5;
- Thus, CategoricalNB is the chosen binary classification model for this problem.

Figure 1 shows the model F1-Score according to the alpha and the best set of variables for that value, for the two classification tasks.



(a) Multiclass classification alpha evaluation.

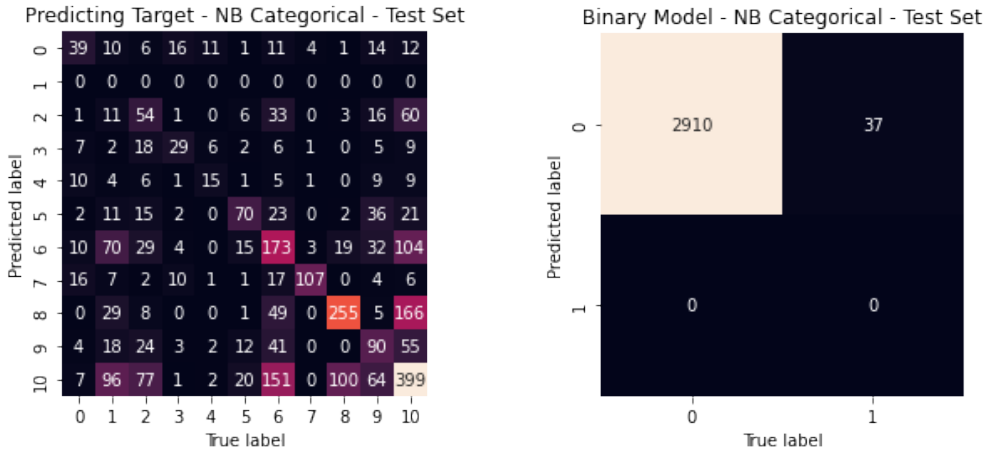
(b) Binary classification alpha evaluation.

Figure 1: Graphs on the correlation between alpha value and model performance.

4 Discussion

Discovering the best possible naive bayes multiclass classification model, we obtained a greater number of variables selected by the CategoricalNB algorithm than by the GaussianNB algorithm. This gives us a first indication that the GaussianNB algorithm is capable of producing a simpler model. The second point that reinforces this idea is the fact that it has no hyparameters. The relationship between performance and alpha values was not linear, and we were able to highlight four best alpha values (0.1, 0.9, 5 and 10). In the end, we chose the model that returned the best score. After selecting and validating the best model for each algorithm, CategoricalNB had the best performance in classifying the points among the existing classes. These results were expected since the algorithm is more suitable for classification with discrete features that are categorically distributed. And our dataset is treated this way. Therefore, the best model for multiclass classification was CategoricalNB. Comparing the results of the feature selection phase with the validation, the F1-Score result was slightly lower in the second task. The result is more realistic since in the validation new data are presented to the algorithm. That means you made a good generalization.

CategoricalNB needed fewer variables to classify a point between two distinct classes. This model is simpler compared to GaussianNB, and we justify this result with the characteristics of the algorithm. However, the score was slightly lower than the GaussianNB score in the selection of variables. The accuracy of the two best models with the independent validation set was very similar, and the score for each was equal to the feature selection phase. To choose a binary classification model, we took into account several points: accuracy, F1-Score and number of variables. Both models were viable options, however we chose CategoricalNB because it is more accurate, has fewer variables, despite having a lower score. Figure 2 presents the final statistics for both classification tasks.



(a) Multiclass classification confusion matrix.

(b) Binary classification confusion matrix.

Figure 2: Confusion matrices of the chosen classification models.

5 Conclusions

This document has presented the use of two types of naive bayes (NB) algorithms as learning methods, to determine the best possible classification model for multiclass and binary classification tasks.

Given the obtained and discussed results, these could be improved instead of discretizing the nominal features into "Low, Medium, High" categories. We should encode the categorical features with *OneHotEncoder()* method first, merge the resultant DataFrame with the nominal features, split the features into train and test set and lastly, perform a standardization with *StandardScaler()* method to the created train and test set. However, we tried to use some methodologies approached in practical classes (numerical features discretization) which might not have led to the best prediction model.

Additionally, another point that could be improved is the fact that we could train our NB classification model with other subsets of features, instead of only using the best features selected by our forward wrapper method. These subsets could be for example various subsets of features randomly selected and a subset without the features that have strong correlation with others.

On another hand, we agree that our data preprocessing and analysis steps could be further improved, even though we applied the steps we were aware of, through past experiences (e.g.: Exploratory Data Analysis). We hope to improve this and other aspects throughout the semester.

References

[1] Marlis Ontivero-Ortega, Agustin Lage-Castellanos, Giancarlo Valente, Rainer Goebel, and Mitchell Valdes-Sosa. Fast gaussian naïve bayes for searchlight classification analysis, 2017.

[2] Irina Rish. An empirical study of the naïve bayes classifier, 01 2001.

[3] Zhuoyuan Zheng, Yunpeng Cai, Yujie Yang, and Ye Li. Sparse weighted naive bayes classifier for efficient classification of categorical data, 2018.