Conference Paper Title

Brewton Morais
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
brewtonlmorais@gmail.com

Cléber Lucas
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
seuemail@email.com Update

Lucas Abdalah
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
lucasabdalah@alu.ufc.br

***Abstract*—An analysis of classification models was done for a set of data with informations related to the request and approvals for research and academic activities. Using MATLAB 2022a, we trained different classification models, statistical and decision trees, such as Logistic Regression, KNN, LDA, CART, etc, making use of their respective confusion matrix as the results demonstrations. One of the goals these tests is to investigate if there are linear relationship between the provided predictors of the data base, in order to optimize the model that could be used by an university or a governamental instituion. Finally, we perform a comparison of the results and of the decision making about which model we understand to be the most efficient to the problem.**

***Index Terms*—classification model, logistic regression, grand application, KNN**

I. Introduction

This command is used for a strong suggestion.

This command is used for minor changes suggestion.

Classification models are a class of mathematical models constantly used in problems of assimilating observations of certain events to certain categories that define the problem. Nowadays, these models are considered tools of fundamental importance in the construction of Deep Learning and Machine Learning algorithms. To begin, it's important to evidence the main existing difference between this new class of models and the class of regression models which is the prediction of a qualitative variable instead of a quantitative one. This new class of tools present various practical applications, such as the development of a detection spam filter for emails based on the sender and on the content of the message, the development of classification techniques of a cell belonging to tumors, as beningn or malignant and on the development of a model of credit release for financing.

In addition to these pure classification models applications, there are mixed applications techniques that combine Data Mining techniques with some other types of models to perform a prediction. Some examples of models that use Data Mining techniques to improve their results are addressed by Sidropoulos Add reference, such as Web Mining/Search tensor models and Brain Data Analysis.

Meanwhile, in the work developed in this paper, some of the most used techniques in the development of classification models were approached, such as SVM, KNN and CART, to train and test a funding model that will separate observations into one of two available groups: Positive Founding and Negative Founding.

II. Method

A. Data set

The data used for the construction of the predictive models consists of 8708 samples of different requests for funding from universities around the world, to finance research, with the outcome being the success or failure of the request. The data set contains samples from the years 2005 to 2008, with a total of 1882 predictors (independent variables). 6633 samples from 2005 to 2007 and 1552 from 2008 were used for model training, and the remaining 518 samples from 2008 were used for testing the obtained model. Predictors can be separeted between continuous, such as the number of successes and failures passed by the "chief investigator", and categorical ones, such as the monetary value of the grant, divided into 17 groups of increasing amounts, and the month of application.

B. Pre-processing

Initially, the first step is verifying the data skewness, in case there is a strong tendence to the left or to the right, an adequate transformation would be applied in order to remove the skewness. The next step is to scale and center the data around the mean. It's done so since different predictors can have different scales, and if they're not normalized, models sensitive to the variance would be affected negatively, making it biased to those predictors with the highest values.

Then, what we should do is to verify which predictors have actual importance to the model construction, that is, which of them have a stronger say for the final prediction. We can study this by analysing their correlation. Those with a correlation larger than 0.99, with zero variance or sparse, that is, that have lots of zero values as data, were removed.

Then, the final approach is to verify the linearity of the predictors together with the output. This step is essential, and the reason for this is, once we have analysed this aspect, we can infer if using a linear model is the adequate way of resolving this problem. For example, if there are too much predictos with non-linear relationship with the final output, it makes no sense to insist in linear prediction models.

C. Cross-Validation

Cross-validation consists in a validation technique used to validate the model with the test set, usually taking into account the model flexibility and the mean squared error (MSE). Shortly, it divides the data set into k distinct subsets of size as equal as possible. From these groups, one of them is put aside to be used as validation set, while the model is trained based on the remaining $k-1$ subsets. Once the model is trained, the first removed subset is used as validation as previously stated. Then, the removed subset is restored to the principal set, and the following subset is put aside to perform the same procedure until all of the k subgroups are all used as validation set. This approach improves the model capability of generalization, once it's trained with all the data at dispose, it also makes the error estimation more robust.

This strategy generally serves to indicate which models have a better prediction capability on the test set, since it enables the comparison between the error levels and the variance generated.

It's important to state that if a k is chosen such as it's too small, e.g, $k = 2$ (two subgroups) or too large (k =sample size), we are going to have, respectively, a strongly biased model because we let a lot of data outside the training step, and a potential overfitting issue due the high model complexity, ocurring the model to have a high variance. Thus, to mitigate both of the effects, normally $k = 5, 10$ is employed, since they present an acceptable level of bias and variance.

III. Model validation performance

One of the most used metrics to measure the performance a classification model is the Receiver Operating Characteristic curve (ROC) and the Area Under the Curve (AUC) correct initials?. The ROC curve is traced in a graphic with the positive ratio in the y-axis and the negative ratio in the x-axis, and each point of the curve is computed varying the classification limiar. Decreasing this limiar makes the model classify more items as positive, increasing the true positive and false positive. Increasing this limiar, causes the inverse effect.

After calculating the ROC curve we can find the area under it. The area under the ROC curve represents how well the model divides the two classes, as close the AUC value is from 1, better the model. However, this kind of analysis shouldn't be applied alone to validate a model's performance, since there is an information loss in the construction of the graphic. Then, the dispersion table, which contains in its principal diagonal the number of true positives and true negatives, and in its secondary diagnoal the number of false positives and false negatives, becomes an interesting analysis complement to the ROC curve.

IV. Linear Methods

A. Logistic Regression (LR)

It is statistical model used to determine the probability of an event. Therefora, its values are probabilities, which belong to the $(0, 1)$ interval. The model is defined as:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}, \quad 0 \leq p(X) \leq 1 \tag{1}$$

This logistic model will aways produce a S-shaped curve, regardless of the value of X, getting a relatively precise prediction. After some mathematical astuces, it's possible to come up with the following equations that model the method.

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X} \tag{2}$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{3}$$

The right term of (2) is called Odds, then the right term of (3) is called log-odds or logistic. The Odds ratio represents the effects of predictor X, on the likelihood that an event will happen.

To adjust the model's parameters would be necessary to use the Maximum-Likelihood technique to perform an estimation of them. However, it's also possible to make use of the Least Squares for the adjust, as well as in the case of the coefficients in a linear regression.

Put Table 1 here

B. Linear Discriminant Analysis

Instead of directly estimating $P(Y|X)$, a model with the following characteristics will be developed:

- Modeling the distribution of predictors of X separately in each class Y.
- Baye's Theorem to estimate $P(Y = K|X = x)$.
- Normal distribution to describe each class.

Following from these informations, we initiate the model's development directly from the Baye's Theorem:

$$P(Y = k|X = x) = p_k(X) \tag{4}$$

$$p_k(X) = \frac{P(X = x|Y = k)P(Y = k)}{P(X = x)} \tag{5}$$

$$p_k(X) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{6}$$

Where $f_k(x)$ represents the probability density function (pdf) of the r.a X of an observation belonging to class K. Thus, instead of directly computing $p_k(X)$ it's possible to simply estimate $\pi_k(X)$ and $f_k(X)$. Then, assuming that the number of predictors is unitary, we can make some affirmations about the form of $f_k(x)$ in order to move on with the LDA method:

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{-(x-\mu_k)^2}{2\sigma_k^2}} \tag{7}$$

Besides, we assume $\sigma_1^2 = ... = \sigma_k^2$. Therefore, it's possible to write $p_k(X)$ as the following:

$$p_k(X) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{-(x-\mu_k)^2}{2\sigma_k^2}}}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi\sigma_l^2}} e^{\frac{-(x-\mu_l)^2}{2\sigma_l^2}}} \tag{8}$$

After some algebraic manipulations, it's possible to conclude that classifying one observation to a given class is equivalent to classify one observation to a given class such that the the linear discriminant function $\sigma_k(x)$ is larger:

$$\sigma_x = x\frac{\mu_x}{\sigma^2} - \frac{u_k^2}{2\sigma^2} + \log(\pi_k) \tag{9}$$

However, in practical situations, it's not always possible to know the parameter's values, then LDA approximates the Bayes classifier by the following expressions:

$$\hat{\mu_k} = \frac{1}{n_k}\sum_{i=k}^{K} x_i \tag{10}$$

$$\hat{\sigma^2} = \frac{1}{n-K}\sum_{i=k}^{K}\sum_{i=k}^{K}(x_i - \hat{\mu_k})^2 \tag{11}$$

$$\pi_k = \frac{n_k}{n} \tag{12}$$

Put Table 2 here

V. Non-Linear Methods

A. Quadratic Discriminant Analysis (QDA)

The QDA method is a non-linear variant of the LDA. The main difference between QDA and LDA lies in the fact that the covariance matrix of each class is different onr from another:

$$X \sim \mathcal{N}(u_k, \sum_k) \tag{13}$$

The discriminant function for this method is, after some algebraic astuces:

$$\sigma_k(x) = -\frac{1}{2}(x-\mu_k)^T \sum_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\sum_k| + \log(\pi_k) \tag{14}$$

$$\begin{aligned}\sigma_k(x) = & -\frac{1}{2}x^T\sum_k^{-1}x + x^T\sum_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\sum_k^{-1}\mu_k - \\ & -\frac{1}{2}\log|\sum_k| + \log(\pi_k)\end{aligned} \tag{15}$$

Thus, the QDA may be summarized as a way of computing $\sum_k$, μ_k and π_k in order to use the discriminant equation in the classification of a X observation into a class in which the discriminant has the larger absolute value.

Therefore, the QDA method is mainly recommended when there is a sufficiently large data set so the statements about the variance don't be a need or when it is not possible to sustain the statement abour the unicity of the covariance matrix. Hence, differently of the LDA method, the decision region of QDA is described by a non-linear curve.

B. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{16}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(16)", not "Eq. (16)" or "equation (16)", except at the beginning of a sentence: "Equation (16) is . . ."

C. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

D. Some Common Mistakes

- The word "data" is plural, not singular.

- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [1].

E. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

F. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

G. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [2]. The sentence punctuation follows the bracket [3]. Refer simply to the reference number, as in [4]—do not use "Ref. [4]" or "reference [4]" except at the beginning of a sentence: "Reference [4] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [5]. Papers that have been accepted for publication should be cited as "in press" [6]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [1].

References

[1] A. Author6 and C. CoAuthor6, *Title6*, 6th ed. publisher6, 2006.
[2] L. Wang, *Early Diagnosis of Breast Cancer*, 1st ed. sensors, 2017.
[3] A. Author2 and C. CoAuthor2, *Title2*, 2nd ed. publisher, 2002.
[4] A. Author3 and C. CoAuthor3, *Title3*, 3rd ed. publisher, 2003.
[5] A. Author4 and C. CoAuthor4, *Title4*, 4th ed. publisher, 2004.
[6] A. Author5 and C. CoAuthor5, *Title5*, 5th ed. publisher, 2005.