Week 4 Assignment: Predictive Modeling Naïve Bayes Classification

Jonathan Ibifubara Pollyn

College of Science, Engineering and Technology, Grand Canyon University

DSC-530: Predictive Modeling

Filippo Posta

September 27, 2021

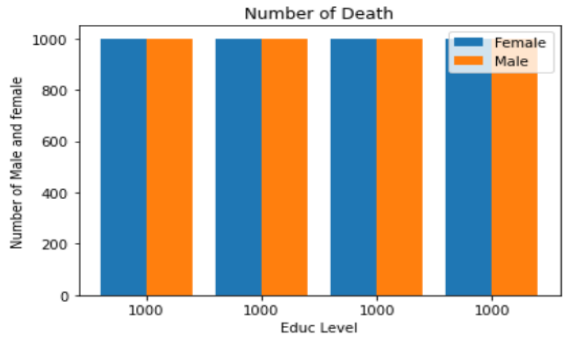
Part 1: Operational Tasks

Convert all variables (*Death*, *Sex*, and *Educ*) to factors:

It is a factor that can have any value as an element, if it comes from a specific set of values. When it comes to Sex, the only options are M or F. However, there are many options for what to call it. The level refers to the set of possible values for the vapor component when a new level is included in a segment.

Use the graphs from the previous exercise to answer the following questions:

If we know a person is dead, are they more likely to be male or female? The model shows that they can be both male or female. Also, if we know a person is alive, are they more likely to be male or female? The model shows that they are both likely to be both. Next, if we know a person is dead, what education level are the most likely to have? The model tells us that they are likely to have all educational levels, and if we know a person is alive, what education level are they most likely to have? We see from the model that they are likely to have all levels of education. Finally, which education levels are more prevalent for dead persons? For living persons? we see that they are of all levels of education. Using the contingency table, we found out that 525 are correctly classified as dead while 431 are classified as living people.



PART 2: Mathematical and Statistical Basis

In machine learning, "Naive Bayes" and "Logistic Regression" are two common models that are both similar in many ways and significantly distinct in other ways. A comparison of two well-known algorithms is presented in this blog post. Ismail ElarabIS(2014) (El Din Ahmed AB) When using 'Nave Baye,' the probability of a label being associated with the characteristics vector is calculated using Bayes' theorem. The Naive Bayes algorithm makes the erroneous assumption that each feature is always independent, even if it isn't. 'Logistic Regression' is a linear classification strategy that knows how probable a sample is to be included in a specific class. The goal of logical regressions is to identify the classroom decision-making boundaries that are appropriate. Classification problems are solved with the help of both techniques.

Both methods use categorization to determine whether samples belong to a particular class, such as spam or a ham email. This is where they have the most in common. Zhu H., Shang W., and T. DONG (2011) The mechanism for algorithmic learning Both strategies are used for categorization jobs. In categorization situations like these, logistic regression and Naive Bayes are employed to determine if samples fall into a given category like spam or ham email.

The process by which algorithm mechanisms learn While "Naive Bayes" is a generative model, the training procedure differs differently; Logistics Regression is a discriminatory model. Researchers Faisal KM, RC, Alamgir H, and Kesav D. (2011) found that a genetic model: the naive Bayes models use the joint distributions of X and Y to predict the chance. The approach directly models by P(y|x) by decreasing mistakes with mapping and learning the inputs and outputs with discriminatory models. When compared to other models, the Naive Bayes model performed the best. Assumptions used about the model. There are no qualities that are dependent on "Naive Bayes." The prognosis may be unfavorable if some characteristics are interdependent

(in large feature spaces). If a few variables are linked, "Logistic Regression" splits the space linearly. This method often works well. Researchers Faisal KM, RC, Alamgir H, and Kesav D. (2011) found that a method to improve the model's output. Prior probability information/data helps to improve the outcomes when the data training is small compared to the number of features. Overfitting can be reduced, and a more generic model can be developed using Logistic Regression if the data is little compared to a variety of characteristics, such as the regression of Lasso and Ridge. How did the model's sensitivity affect the model's validity? Model output uncertainty is linked to model input uncertainty, and the analysis sensitivity takes various techniques to quantify this link. To put it another way, the sensitivity analysis determines how "sensitive" the model is to changes in its input parameters and output data. The sensitivity analysis's results may significantly impact the modeling process, from identifying model defects to informing decision-making about model parameters. A key aspect of sensitivity analysis is that it allows researchers to detect model inadequacy and then uses a second experiment design result to quantify the total sensitivity of the model of interest. Due to the model's sensitivity will always outperform other models when tested on the original data packets (e.g., over-performance). Sensitivity analysis often includes subgroup analysis as a variation. Wang (2013) cites Jiang et al. (2013), who state that.

Explain how the Naïve Bayes Classifier outlined in Section 4.1(c) applies to the Internet of Things as evaluated in the article.

Researchers used the internet of things and cutting-edge computers to examine how public health environmental management affects health outcomes. Optimization of algorithms and architectural design based on experimental data substantially improved the level of public health assessment. However, due to the rapid expansion of the Internet of Things architectural

system, the system could not include all public health management of government organizations. The system's future maintenance should be centered on making it more functional. Wang (2013) cites Jiang et al. (2013), who states that. Malware detection methods today are reactive rather than proactive, relying on previously discovered malware structural expertise. There is no way to stop malicious software from doing its work if it has infected the device. This project aims to create a mobile app that uses "behavioral analysis" to find connections between different device parameters and decide when the data don't match these correlations to detect malware on mobile devices. Wan S, Jiang L, and Xie C. (2015) It's difficult to grasp why the other behavior is occurring without first understanding these two fundamental principles. It looks at the number of processes performed and the amount of data the device is generating. A Gaussian probability distribution calculates and determines the risk level for each possible value of a characteristic. max(P(T|Ci))]

References List

El Din Ahmed AB, ElarabIS (2014). *Data Mining: A prediction for Student's Performance using Classification Method*. World Journal of Computer Application and Technology.; 2(2):43–7.

Dong T, Shang W, Zhu H. (2011). *An improved algorithm of Bayesian text categorization. Journal of Software*.; 6(9):1837–43.

Faisal KM, Mofizur RC, Alamgir H, Kesav D. (2011). *Enhanced classification accuracy on naive Bayes data mining models.* International Journal of Computer Applications.; 28(3):9–16.

Jiang L, Cai Z, Zhang H, Wang D. (2013). *Naïve Bayes text classifiers: a locally weighted learning approach. Journal of Experimental and Theoretical Artificial Intelligence*.; 25(2):273–86.

Kohavi R. (1996). *Scaling up the accuracy of the Naïve Bayes Classifiers a Decision-Tree*

Hybrid.Proceedings of the second international conference on knowledge discovery and data mining. p. 202–7.

Wang S, Jiang L, Li C. (2015). *Adapting naive Bayes tree for text classification. Knowledge and Information Systems*. 44(1):77–89.