**Machine Learning: Logistic Regression Implementation**

Devin Schmidt

Department of Mathematics and Computer Science

COSC 320: ADVANCED DATA STRUCTURES AND ALGORITHM ANALYSIS

Dr. Lu

December 11, 2020

**Abstract**

Machine learning is an inevitable technological advancement for the progression of A.I. (Artificial Intelligence). Researching this topic to gain an understanding is critical for solving modern problems. Particularly, the problem to solve is classifying whether a person will buy a product or not. The data incorporated with the algorithm to solve this problem is collected information from the people who have been redirected from a social networking ads. An implementation of multivariable logistic regression in Python was used to predict a person purchasing the product with an average accuracy of 90 percent. With more analysis on the performance; the algorithm was able to predict ‘no purchase’ with 95 percent and ‘purchased’ with 85 percent accuracy. The predicted accuracy for purchasing the product could be improved upon. This could be done from various intuitive methods such as collecting more variables of information related to the people being redirected from the ads. The implementation of the project was successful in getting a rough prediction, however; the implementation could be further improved from ongoing research.

**Introduction to Machine Learning**

Algorithmic innovation encompassing machine learning is prudent to the advancement in this field and humanity. Currently, machine learning “use statistics to find patterns in massive\* amounts of data” (Hao, Karen, 2020). Patterns in the data would be distinguished as to what problem you are solving at the given time. Some problems include identifying pictures or predicting stock prices. Various types of machine learning are available to identify patterns which include supervised, unsupervised, and reinforcement learning. The culminating affects machine learning has is on A.I., enhancing the capabilities pushing the technology further.

The difference between A.I. and machine learning is mis conceptualized in the current climate of today. The terms are used inversely by many people and companies have diluted it for marketing purposes. A.I. resembles human intelligence such as playing game of chess and it makes moves that would resemble what a human would do. Machine learning are the algorithms used to create A.I. thus, machine learning is the tool used to enable the creation of A.I.

Machine learning algorithms can be grouped by their style of learning. Supervised, unsupervised and reinforcement are the main three styles. When analyzing these types of learning processes the main difference between them is what data the machine learning algorithm uses. This taxonomy or way of organizing machine learning algorithms is useful because it forces you to think about the roles of the input data and the model preparation process and select one that is the most appropriate for your problem in order to get the best result (Brownlee, Jason). The faster you can identify which algorithm the faster the problem can get solved.

Supervised learning uses data that has predefined outcomes the algorithm uses to learn enabling it to make predictions on something that does not. This algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables) (Sunil Ray, 2020). The independent variables are what it makes a prediction from such as previous stock prices. Multiple independent variables may be used in some implementations. The dependent variable is the predicted value in example the price of the stock in the future. These two variables are coherent with each other for supervised learning to work.

Unsupervised Learning resembles supervised, but the data does not have a result to learn from. This algorithm will learn based on features of the data provides it and then organize based on those features. Algorithms in this category will be used for organization and clustering in example grouping fruits based on their size and shape so that a robot can sort the when it sees the fruits. Unsupervised learning is still constricted by the variables it can use but it still learns without a dependent variable because of the type of problem it solves does not necessarily need it.

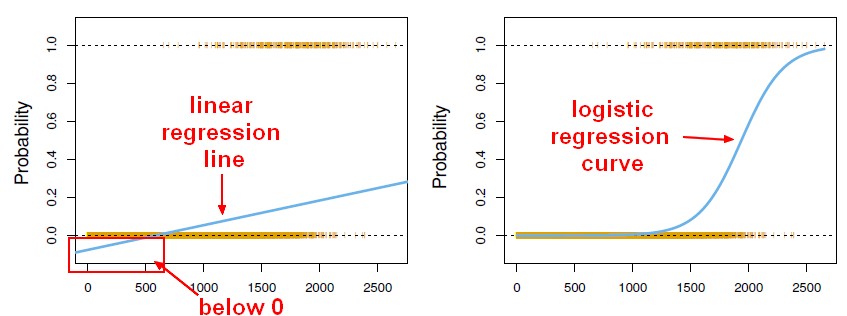
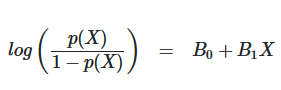
Reinforcement learning starts out with essentially no data. These algorithms learn from trial and error. It does a task and then collects the data from the task it has done. At the beginning the task does not get complete but after a while the algorithm is able to complete the task. The best way to understand is from example. If you were to have a racing game and want A.I. to race around the track, then you would use this algorithm. At first, they would not know how to get around the track, but eventually they would. Each time they run into a wall the data would be collected for how it got there and adjust to what it needs to do to get around the track. This type of learning starts form the ground up.

Picking the correct learning style will lead into the desired path to solve a problem with machine learning. Many aspects of our lives are using machine learning which in return uses one of these learning styles. Netflix recommendations is one of the things that we use. When searching Netflix there are shows and movies it recommends that the user might like. There is also a percentage of how closely it matches to you. I have shown this in Figure 1. A supervised learning algorithm could take in the what you watch and predict what you might like to watch.

**Figure 1**

**Logistic Regression**

Logistic regression is a supervised regression algorithm that outputs a classification. Since it is supervised the data will have independent and dependent variables. Independent variables are attributes that would define the classification. The dependent variable will be a one or zero. One dignifies a ‘yes’ while zero is ‘no.’ Multiple types of regression algorithms are used for various problems. Linear regression, Binomial Logistic regression, Multivariable logistic regression are different types of algorithms that are closely related.

Linear regression is related to logistic regression for a multitude of reasons, yet they are used for completely different problems. The outcome for logistic regression is categorical while the linear regression is continuous. This specific difference comes from the logit function (LOH -git). As Swaminathan states, “Data is fit into linear regression model, which then be acted upon by a logistic function predicting the target categorical dependent variable (Swaminathan, Saishruthi, 2019).” In Figure 2, the left graph denotes the continuous property with the blue line. While the blue line in the right graph denotes the constraint between 0 and 1. These ranges are important to conceptualize because any point on these lines are predicted values at that point. The formula shown in Figure 3 is what gives the s-curve so that the line of best fit can be plotted. The p value is the probability and the x are the probability at that specific position. The log of the function fits the probabilities as an s-curve. As for multivariable logistic regression the difference is it uses more than one independent variable to predict the outcome. Logistic regression is a close neighbor to other algorithms which makes it fundamental to learning machine learning algorithms.

**Figure 2**

**Figure 3**

**Implementation and Analysis**

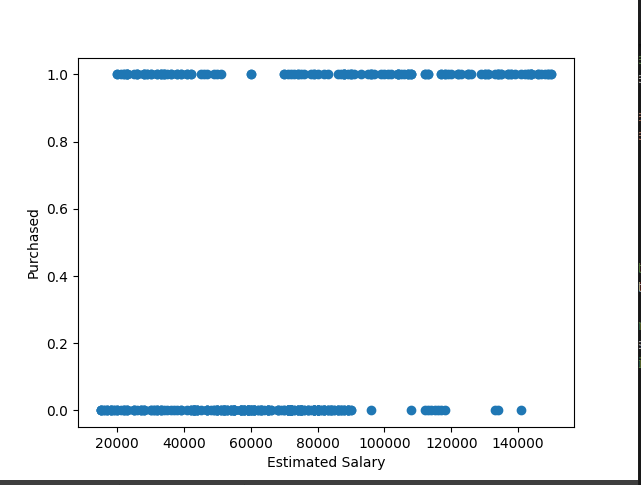
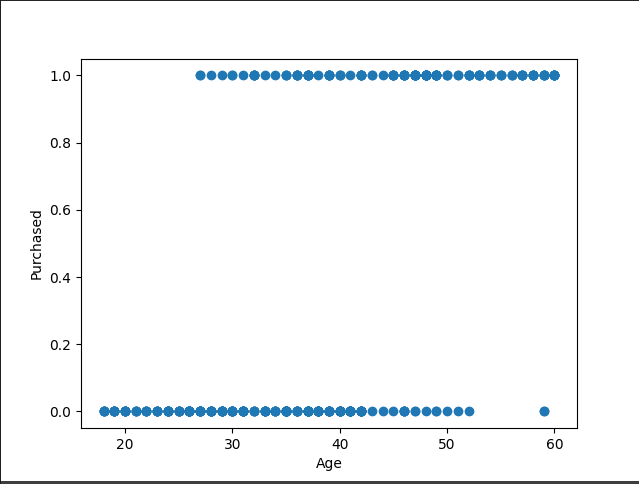
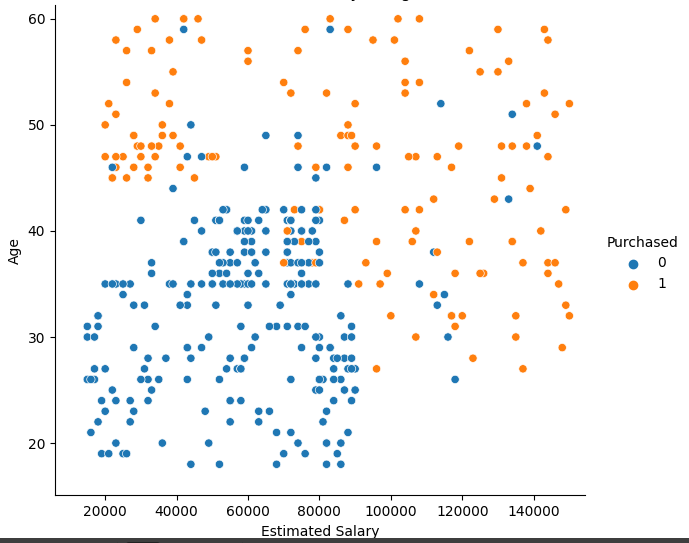
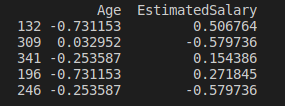
 Now that there is an established level of understanding for what machine learning is and the algorithm used to solve the problem, preparing to code it can begin. Things that are taken into consideration is the dataset set being used and the programming language. The dataset chosen is a .csv file. The reason for this is to be able to use the API (Application programming interface) for manipulating the dataset so that specific data needed can be extracted. The .csv file contained social networking ad information such as gender, age, estimated salary, and purchase. The age and estimated salary were extracted to be used as independent variables. While the purchase was used for the dependent variable. The API used to plot the data for analysis was matplotlib and seaborn. Breaking down the data set into smaller chunks will give you a sense of what variables are important. If a variable does not make an affect the logistic regression algorithm does not need it input. From Figure 4 and 5 you can get a better idea of what relationship the data has. Figure 4 shows that the more you make the more people purchase while Figure 5 shows the older people purchased more. In Figure 6 you can see the relationship between age, estimated salary, and purchased. People who bought younger the 40 and made less than 80,000 did not purchase the product from a social network ad while the opposite did.

Figure 4

Figure 5

Figure 6

Since the data being input in the algorithm has multiple predictor variables they need to be scaled. Scaling the variables will allow the algorithm to predict accurately. Estimated salary has a much larger magnitude compared to the age of people. This is going to cause it to affect the prediction because it will ‘think’ the estimated salary has a larger weight on the outcome. Multiplying both the variables by a coefficient will make them have similar weights for an accurate prediction. This is achieved from line 31 to 34 and scaled data is shown in Figure 7. The data then gets split up into two different groups, one for testing one for training. The training data allow the algorithm to learn what attributes makes a person purchase the product. Then this gets tested on the test data. Twenty percent was used for testing while eighty percent was used for training. This split occurs on line 37. After, the training data get put into the logistic regression function so that it can be able to predict the test data. Which occurs on line 42 and 43.



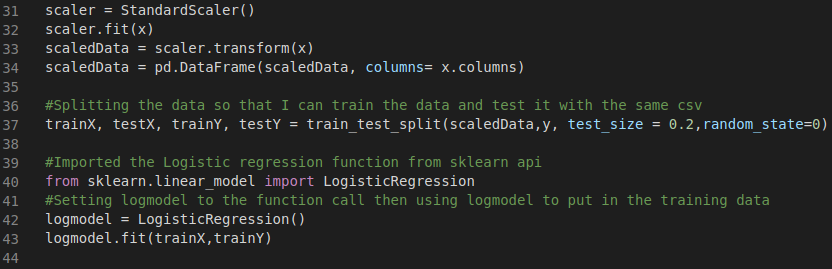
Figure 8

Figure 7

The performance of logistic regression algorithm is then printed to the console in Figure 9 line 52.



Figure 9

From the classification report for the algorithm it shows that it predicted 22 purchases and 58 people who did not. These results are the predictions for the test data then checked afterwards for correctness. The 58 no purchases were 95 percent correct while the prediction for purchases was 85 percent correct. The difference between the two could be that the logistic regression algorithm is overfitting when calculating the line of best fit. This could be influenced by the outliers from people who did not purchase even though the made above 80,000 and older than 40 from Figure 4. The overall/macro average was 90 percent. To improve this algorithm, we could account for the outliers and add more predictor variables from gathering more information on the people who purchase products from social networking ads.

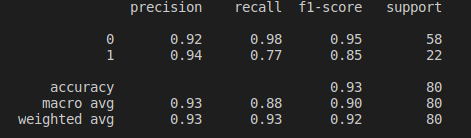


Figure 10

**Conclusion**

The implementation of this algorithm was successful, but as everything has room for improvement. Machine learning is a deep topic with many powerful algorithms to solve complex problems humans would not be able to do as efficiently. Many machine learning styles are taken into consideration trying to find the right machine learning algorithm. Then learning the complexities of the individual algorithm adds multitudes of depth as well but can also be related to others for better understanding. Implementing can be easily achieved with the use of APIs in the correct language. The result of the implementation on this dataset achieved an average of 90 percent correct predictions.

**References**

Alexopoulos, E C. *Introduction to Multivariate Regression Analysis*. Dec. 2010, www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/.

Brownlee, Jason. *A Tour of Machine Learning Algorithms*. 14 Aug. 2020, machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.

Hao, Karen. “What Is Machine Learning?” *MIT Technology Review*, MIT Technology Review, 2 Apr. 2020, www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/.

Hidalgo, Bertha, and Melody Goodman. *Multivariate or Multivariable Regression?* Jan. 2013, [www.ncbi.nlm.nih.gov/pmc/articles/PMC3518362/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518362/).

London guyLondon guy 24k3939 gold badges108108 silver badges162162 bronze badges, et al. *What Is the Difference between Linear Regression and Logistic Regression?* 1 Oct. 1961, stackoverflow.com/questions/12146914/what-is-the-difference-between-linear-regression-and-logistic-regression.

*Machine Learning - (Univariate: Simple) Logistic Regression*. datacadamia.com/data\_mining/simple\_logistic\_regression.

Sunil Ray I am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years. *Commonly Used Machine Learning Algorithms: Data Science*. 22 Nov. 2020, www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/.

Swaminathan, Saishruthi. *Logistic Regression - Detailed Overview*. 18 Jan. 2019, towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.