Logistic Regression

Pena Benafa
Electronic Engineering
University of Applied Science Hamm-Lippstadt
Lippstadt, Germany
pena.benafa@stud.hshl.de

***Abstract*—This seminar paper is an overview of logistic regression, one of the most used tools for predicting categorical outcomes. It uses a type of curve to fit data, and it describes the data and relationship between dependent variables and independent variables [1]. The main advantage of logistic regression is that it can predict the associated probability (this can be seen from the example given in this paper).**

***Index Terms*—logistic regression, predicting, dependent variables, independent variables, data**

I. Introduction

The pioneer of machine learning (ML) defined ML as a "field of study that gives the computer the ability to learn without being explicitly programmed" [2]. Thus, ML is a subset of Artificial Intelligence with roots in computational statistics, focusing on using computers to make predictions. Like humans that learn from experiences, the objective in ML is to create mathematical models that will train to make useful outputs (predictions) when fed with input or training data experiences). From these experiences, the ML models are tuned by an optimization algorithm to produce accurate predictions. In addition, the ML models can generalize what they have learned from the training data (previous experience) so that they can make accurate predictions for new and unseen data [3]. An in-depth explanation and all the techniques in determining logistic regression percentage accuracy are discussed in section 2. Finally, section 3 concludes this seminar paper.

A. Types of Machine Learning Algorithms

Different types of ML algorithms exist and are used to solve a variety of problems. ML algorithms can be grouped into three types of problems being solved: supervised learning, unsupervised learning, and reinforcement learning.

1) Supervised Learning: Consider the equation of straight line (Eq 1), in ML studies commonly referred to as regression equation.

$$y = mx + b \tag{1}$$

Where :
y = dependent variable
m = slope of the regression equation
x = independent variable
b = constant of the equation

In supervised learning, the ML model is given training data containing y and x. The model learns the relationship between these, i.e., m and b. Afterwards, the model is given test data with the only y; from its experience discovering m and b, it can predict the y's for each x. The classification of skin lesions according to malignancy is an example of this [4]. Based on the value of the predicted label or dependent variable, supervised learning algorithms are further grouped into two, regression and classification. And logistic regression algorithm is an example of a classification supervised ML algorithm.

2) Unsupervised Learning: Unlike supervised learning, in which the training dataset of the ML model contains both the independent variables and dependent variable(s), in unsupervised learning, the training dataset contains only independent variables — x's. The ML model learns the relationship between all the independent variables and then groups them accordingly. An example of this is identifying and grouping customers who are high profit, high value or low-risk buyers [5]. Unsupervised ML algorithms are grouped into the following: association, clustering and dimension reduction.

3) Reinforcement learning: Reinforcement learning is different from supervised and unsupervised learning. The problem dataset will contain at least independent variable(s); instead of reinforcement learning, an agent and an environment are constructed. Through trial and error, the agent learns from its environment while optimizing some set objective function. After every performed action, depending on the set goal, the agent gets a reward, punishment, or nothing. The agent uses these to determine the optimal path to the goal. The AlphaGo and AlphaZero are good examples of this [4], and [6].

B. Description of the Classification Problem

One of the problems in supervised learning is the classification problem. In this type of problem, the ML model built using ML classification algorithms learns from training datasets containing features or independent variables, i.e. x's, dependent variables or labels or classes (as they referred to in this case), i.e. y's. After training, the test dataset containing only the features are 'classified' or grouped according to what the model has learned from the training dataset. The classes can be binary, true or false, spam or not spam, 0 or 1, dog or cat, male or female e.t.c. Or the dependent variables can be multi-class; that is, instead of being 0 or 1, spam or not, there can be third, fourth....nth class. For example, classification of Apps in the App Store, an App can be classified as one of

these, social media, game, news, weather, media (audio, photo, video), educational, health and fitness, entertainment e.t.c. In this seminar paper, the focus shall be on binary classification problems.

Here are some examples of ML algorithms used for solving classification problems:

1) Logistic Regression
2) Support Vector Machine
3) k-Nearest Neighbours
4) Decision Trees
5) Naive Bayes
6) Random Forest etc.

II. Logistic Regression

Logistic regression can estimates probabilities using a logistic function to measure the relationship between independent variables and categorical (binary or multi-class) dependent variables. During the training phase, the ML model learns the relationship between categorical dependent variables and the independent variable(s). Then, in the test phase, for a given x, the model makes a prediction based on the probability that the corresponding y value for that x is ”false” or ”true”. The sum of the probability for these two outcomes is 1. However, the ’y’ value would be ”false” if the probability value for ”false” is greater than ”true”. As mentioned earlier, the probability of the outcome is governed by a logistic function.

A. Logistic Function

Visually the shape of the logistic curve is shaped like an ”S”, and it is also called a sigmoid curve. A sigmoid curve that shows a sufficient degree of smoothness is a bounded differentiable real function defined for all the real input values that have a positive derivative [7]. The curve starts low, has a period of acceleration, and then approaches a straight line that defines the limit of a curve (asymptote).

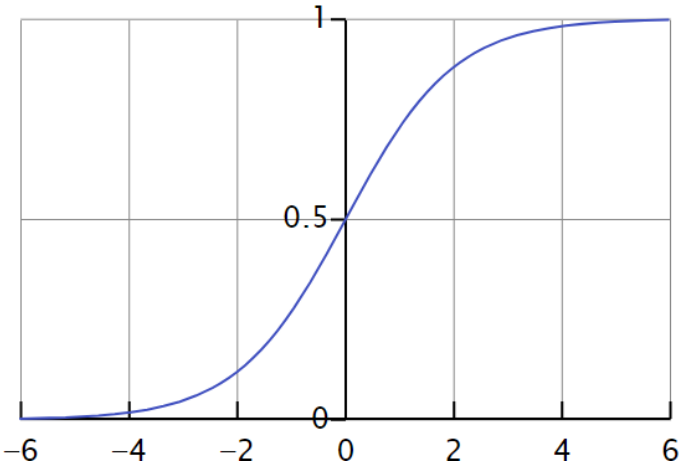

Fig. 1. Graph of the standard logistic sigmoid function [8]

$$f(x) = \frac{L}{1 + e^{-k(x-x0)}} \tag{2}$$

Where:
L = the curves maximum value or the carrying capacity
k = the curves steepness or logistic growth rate
x = the independent variable
x0 = value of x at sigmoid’s midpoint

The maximum value of f(x) is obtained when it approaches L, that is when x approaches $+\infty$. And the minimum value of f(x) is obtained when it approaches zero, that is when x approaches $-\infty$.

B. Making Predictions (Example on Binary Logistic Regression)

The dataset (social network ad [9]) used for this coding example was gotten from kaggle website. It is simple social network advert datasets, that contains 5 columns: UserID, Gender, Age, Estimated Salary and Purchase.There are 401 samples with the aim to train and test set the prediction.

UserID —: the social network’s user’s ID
Gender —: gender of the user
Age —: age of the user
Estimated-Salary —: estimated salary of the user and
Purchase —: purchase information about the user whether he/she purchased the advertised product through the social network.

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	26	24000	1
1	15810944	Male	45	30000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	1
4	15804002	Female	29	89000	0

Fig. 2. First five rows of the social network ad dataset [9]

1) Independent Variables (x): The independent variable shown in the fig 2, are the Age and Estimated Salary. These variables do not depend on other variables.

2) Dependent Variables (y): The y variable is completely depended on x variables and from fig 2 above, the purchased column is the dependent variable, as it has two classes 0 and 1.

Where :
1 —: indicates that the user purchased the the advertised product through the social network.
0 —: indicates that the user DID NOT purchase the the advertised product through the social network.

3) Data manipulation: The independent variables and dependent variables were separated from the main dataset, X and y, respectively. After that, X and y were randomly split into a training set and test set. The training set had 75% of the total

dataset, while the test set had 25%. Thus, both the training set and test set had X and the corresponding y values. The training set was used to train the ML model, while the test set was used to evaluate the model, to check how good or bad it was.

Furthermore, independent variables for both the training and set were feature scaled. This was done to bring the variables to the same scale. Since the maximum value for Age is less than 100, the maximum value for the EstimatedSalary is more than 200,000. Using these values might skew the results and thus produced a skewed and/or biased ML model. The standardised value for a sample of x can be calculated by using this equation.

$$z = \frac{(x - u)}{s} \tag{3}$$

Where :

z = standardised value for x

u = the mean of the training samples

s = standard deviation of the training samples

The figures below show the Age and EstimatedSalary columns before feature scaling and after applying the formula (3) above, it shows the Age and EstimatedSalary columns after feature scaling.

```
array([[    30,  87000],
       [    38,  50000],
       [    35,  75000],
       [    30,  79000],
       [    35,  50000],
       [    27,  20000],
       [    31,  15000],
       [    36, 144000],
       [    18,  68000],
       [    47,  43000],
```

Fig. 3. Age and EstimatedSalary columns before feature scaling

4) Training: After manipulation of the data, the training set was used to train the model. The figure below shows how the model classified the training dataset.

5) Evaluation Metrics: Specific evaluation metrics are used to determine how good or bad a model is. For classification problems, the most common evaluation metrics are accuracy and confusion matrix. The accuracy is derivative of the four basic cardinalities of confusion matrix. Which are:

1. True Positive (TP): predicted value is positive, and the actual value is positive, hence true.
2. True Negative (TN): predicted value is negative, and the actual value is negative, hence true.
3. False Positive (FP): predicted value is positive, but the actual value is negative, hence false.

```
array([[-0.81050083,  0.50499194],
       [-0.01556745, -0.56801926],
       [-0.31366747,  0.15698831],
       [-0.81050083,  0.27298952],
       [-0.31366747, -0.56801926],
       [-1.10860085, -1.43802834],
       [-0.71113416, -1.58302986],
       [-0.21430079,  2.15800919],
       [-2.00290091, -0.04601381],
       [ 0.87873261, -0.77102138],
```

Fig. 4. Age and EstimatedSalary columns after feature scaling

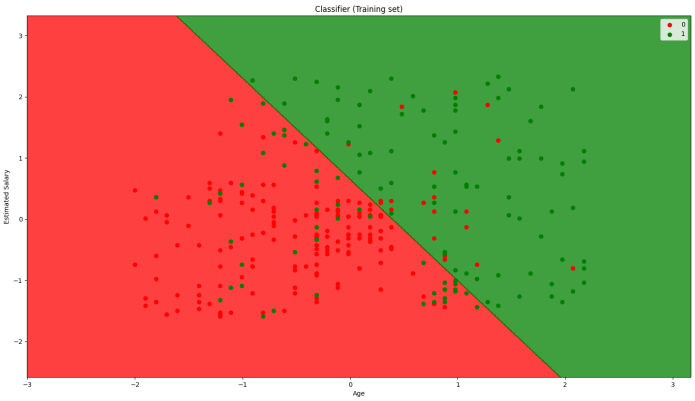

Fig. 5. Classification of the training set

4. False Negative (FN): predicted value is negative, but the actual value is positive, hence false.

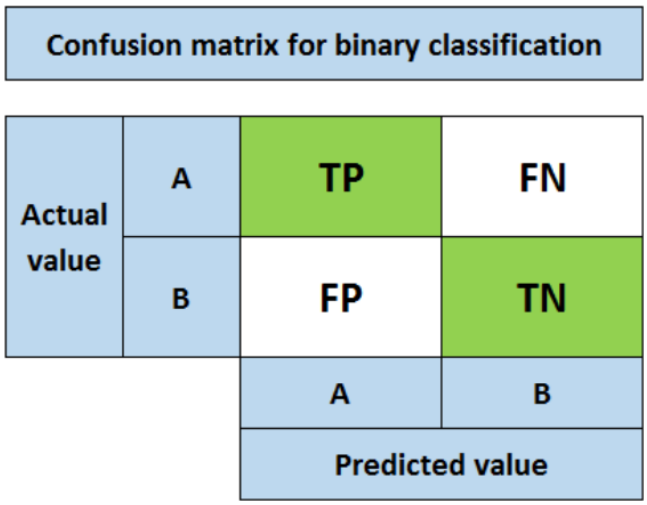

Fig. 6. Confusion Matrix

$$Accuracy = \frac{Totalnumberofcorrectpredictions}{Totalnumberofallprediction} \tag{4}$$

$$Accuracy == \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

For confusion matrix, the lower FN and FP, and higher TP and TN the better the model. The closer to 1.0, the accuracy is the better it is.

6) Test set Prediction and Model Evaluation: After training the model, the test set was predicted. As mentioned earlier, the test set contains both the independent variables and the dependent variable. Therefore, to know how good or bad the newly trained model was, the model was used to predict the test result. This means the independent variables were fed to the model, from which the model predicted the corresponding dependent variables. After this, the predicted values of the dependent variables were compared with the actual values, using the evaluation metrics explained earlier. Fig (7) shows how the model classified the test dataset.

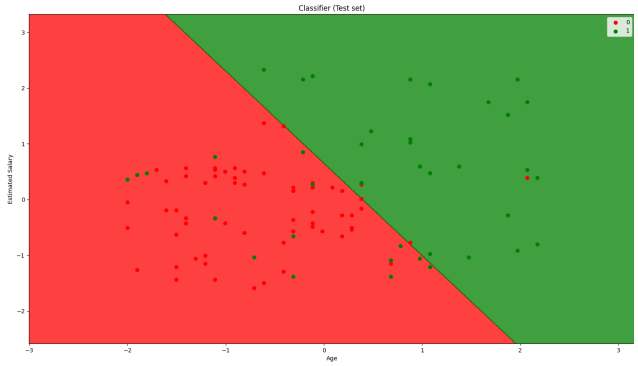

Fig. 7. Classification of the test set

The accuracy for the test result was 0.82 which is very good. Using the confusion matrix, we can explain the accuracy when Logistic Regression Classification is applied to the supplied dataset.
The confusion matrix gives the following:

```
[[58  3]
[15 24]]
```

Fig. 8. Confusion Matrix

This is entered in the diagram below and explained thus:

1. At the top left, it shows that the system correctly predicted 58 no purchase decisions.
2. At the bottom right, the system correctly predicted 24 purchase decisions.
3. At the top right, the system wrongly predicted 3 no purchase decisions as purchase decisions.
4. At the bottom left, the system wrongly predicted 15 purchase decisions as no purchase decisions.

The confusion matrix gives the following:

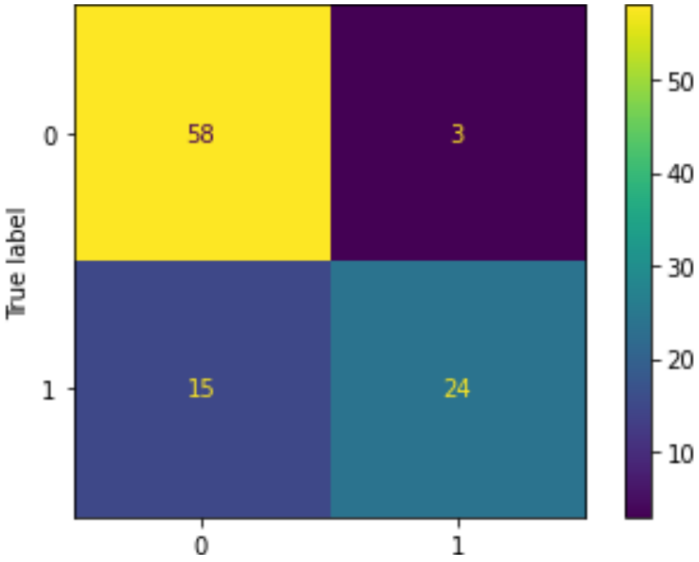

Fig. 9. Confusion Matrix Table for Logistic Regression

III. Advantages and Disadvantages of Logistic Regression

A. Advantages

1) Easier to Implement: Logistic regression requires less computation power because of its mathematical foundation. Therefore, it is easy to train and implement as compared to other methods.

2) Probability Prediction: Depending on the required output, logistic regression can be used to classify labels and predict the probability of an outcome. An example of such an application usage is Google reCAPTCHA V3 (uses advanced risk analysis techniques to predict humans from bots [10]).

3) Unlikely to Overfit: Unless high dimensional datasets, logistic regression is less inclined to over-fitting. Also, in high dimensional datasets, with the use of regularization techniques, over-fitting can be avoided [11].

4) Large Dataset: Logistic Regression can be used to classify large dataset very fast and reliable, unlike other ML algorithms such as SVMs, kNNs etc.

5) Classification of Unknown Records: Logistic regression can be used to classify unknown record, and the accuracy is perfect, especially when the dataset is low dimensional datasets (dataset features).

B. Disadvantages

1) Possibility of Over-fitting: On high dimensional datasets, there is a high possibility that over-fitting may occur. This will result in the model not predicting accurate results on the test set due to overstating the accuracy of predictions on the training set [12]. This usually occurs when there are

many features in the dataset, and the model is trained on little training data. Regularization techniques are used to avoid over-fitting, but this can lead to an under-fit on the training data.

2) Predict Only Discrete Functions: Logistic regression can only be used to predict discrete functions. In order words, the Logistic Regression dependent variable is bound to the discrete number set [11].

3) Limited Use Case: Unlike other algorithms methods such as SVMs, Naive Bayes, Random Forests, kNN etc., logistic regression is limited to classification methods, and it requires the independent variables to be linearly related to the log odds (log(p/(1-p)) [11], and [12]. Logistic regression can not be used to predict the continuous outcome. An example is the prediction of an increase in a patient's body temperature.

IV. Conclusion

There are countless use cases of logistic regression and other classification machine learning algorithms across multiple industries and sectors. In the example, the model I used could predict whether social network users will buy a particular product via network given their age and estimated salary. The example used is an oversimplified version of what is obtainable in a real social network like Instagram, Twitter or Facebook. In the real-life situation, there might not be an estimated salary; however, there would be a lot more independent variables that might include: race, employment histories, political interests, religion, sports interests and affiliations, time spent on the network per day, likes given to different posts, number of ads viewed, number of friends and family, and other information from friends and families e.t.c. Although in the coding example, the obtained accuracy was 0.82, which is good. It might not be the best we can get for this dataset. Changing the training set and test set split ratio to 80/20 might improve the result or using another classification machine learning algorithms as kNN, SVC, Decision Trees, Random Forest e.t.c. It is known that no single algorithm is better than all the others on all the problems; however, certain algorithms stand up tall in certain problem [13]. Thus, it is possible that logistic regression might not be the best for the problem in solved in the coding example.

The code and dataset used for the coding example are available in: https://github.com/Justblaise/Logistic-Regression

Affidavit

I Pena Benafa herewith declare that I have composed the present paper and work by myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The paper and work in the same or similar form has not been submitted to any examination body and has not been published. This paper was not yet, even in part, used in another examination or as a course performance.

References

[1] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011.

[2] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *Ieee access*, vol. 6, pp. 35 365–35 381, 2018.

[3] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[5] D. Rajagopal *et al.*, "Customer data clustering using data mining technique," *arXiv preprint arXiv:1112.2663*, 2011.

[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[7] J. M. F. Sandoval(Eds.), "From natural to artificial neural computation : International workshop on artificial neural networks, malaga-torremolinos, spain, june 7-9, 1995 : proceedings : International workshop on artificial neural networks (1995 : Torremolinos, spain) : Free download, borrow, and streaming : Internet archive," https://archive.org/details/fromnaturaltoart1995inte/page/194/mode/2up, (Accessed).

[8] Wikipedia, "Sigmoid function," https://en.wikipedia.org/wiki/Sigmoid_function, (Accessed on 05/16/2021).

[9] R. Raushan, "Social network ads — kaggle," https://www.kaggle.com/rakeshrau/social-network-ads, 2018, (Accessed on 05/10/2021).

[10] G. Developers, "recaptcha," https://developers.google.com/recaptcha, (Accessed on 05/16/2021).

[11] A. RanjanRout, "Advantages and disadvantages of logistic regression - geeksforgeeks," https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/, September 2020, (Accessed on 05/16/2021).

[12] K. Grover, "Advantages and disadvantages of logistic regression," https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/, (Accessed on 05/16/2021).

[13] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural computation*, vol. 8, no. 7, pp. 1341–1390, 1996.