Machine Learning

Role of Log Odds in Logistic Regression

Satishkumar L. Varma

Department of Information Technology
SVKM's Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai.
ORCID | Scopus | Google Scholar | Google Site | Website

Outline

- Learning with Regression and Trees
  - Learning with Regression
    - Simple Linear Regression
    - Multiple Linear Regression
    - Logistic Regression
  - Learning with Trees
    - Decision Trees
    - Constructing Decision Trees using Gini Index
    - Classification and Regression Trees (CART)

Types of Regression

- Regression models used to find the relationship between a DV and IV.
- Simple linear regression
  - To models the relationship between a DV and a single IV.
- Multiple linear regression
  - If you have more than one independent variable.
- Multiple Regression vs. Multivariate Regression
- Multiple Regression:
  - The influence of several IVs on a DV is examined.
  - One DV is taken into account to analyzed.
- Multivariate Regression:
  - Several regression models are calculated to allow conclusions to be drawn about several DV.
  - Several dependent variables are analyzed.

Simple Linear Regression

$$\hat{y} = b \cdot x + a$$

Multiple Linear Regression

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

Logistic Regression

- Types of Logistic Regression
- Binomial Logistic Regression:
  - There can be only two possible types of the DVs, such as 0 or 1, Pass or Fail, etc.
- Multinomial Logistic regression:
  - There can be 3 or more possible unordered types of the DV, such as “cat”, “dogs”, or “sheep”
- Ordinal Logistic regression:
  - There can be 3 or more possible ordered types of DVs, such as “low”, “Medium”, or “High”.

Assumptions of Logistic Regression

- Assumptions of Logistic Regression
- Independent observations:
  - Each observation is independent of the other. meaning there is no correlation between any input variables.
- Binary dependent variables:
  - It takes the assumption that the DV must be binary or dichotomous, meaning it can take only two values.
  - For more than two categories SoftMax functions are used.
- Linearity relationship between independent variables and log odds:
  - The relationship between the IVs and the log odds of the DV should be linear.
- No outliers:
  - There should be no outliers in the dataset.
- Large sample size:
  - The sample size is sufficiently large.

Role of Log Odds in Logistic Regression

- Role of Log Odds in Logistic Regression
- Logistic Regression models the Log Odds of an event as a linear combination of one or more IVs.
- The function that converts Log Odds to probability is the logistic function, hence the name.
- The unit of measurement for the Log Odds scale is called a logit (logistic unit), hence the alternative names.
- There is need of linearity relationship between IVs and Log Odds;
  - That is the relationship between the IVs and the log odds of the DV should be linear.
- Log Odds is nothing but log of odds, i.e., log(odds).
- **Why Log odds?**
  - Log odds helps to convert the LR model from probability based to a likelihood based (log odds) model.
  - Both probability and log odds have their own set of properties,
    - however log odds makes interpreting the output easier.
    - Thus, using log odds is slightly more advantageous over probability.

Role of Log Odds in Logistic Regression

- Let us understand what odds are before understanding the logistic regression,

$$Odds\ Ratio = \frac{p}{1-p}$$

- **Odds:**
  - Simply put, odds are the chances of success divided by the chances of failure.
  - It is represented in the form of a ratio (as shown here in equation).
    - *Odds Ratio = p / (1-p)*; where, p -> success odds; 1-p -> failure odds
- In logistic regression, the odds of IV corresponding to a success is given by p =>

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

  - p -> odds of success;
  - β0, β1 -> assigned weights;
  - x -> independent variable;
- So, the odds of failure in this case will be given by (1- p) =>

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

- Therefore, the odds ratio is defined as *p / (1-p) =>*
- *Now, we take the log of the odds ratio to get symmetricity in the results.*
- *Therefore, taking log on both sides gives ln[p/(1-p] =>*

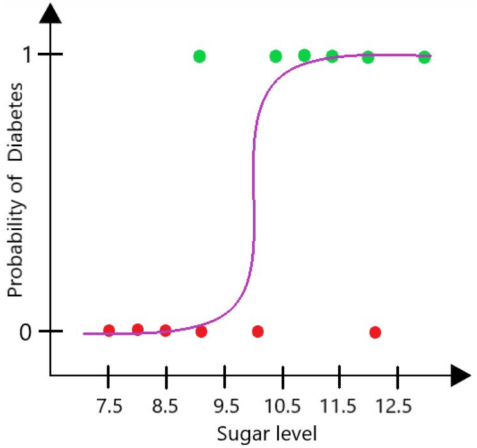
$$\frac{p}{1-p} = \frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\frac{1}{1+e^{(\beta_0+\beta_1 x)}}} = e^{(\beta_0+\beta_1 x)}$$

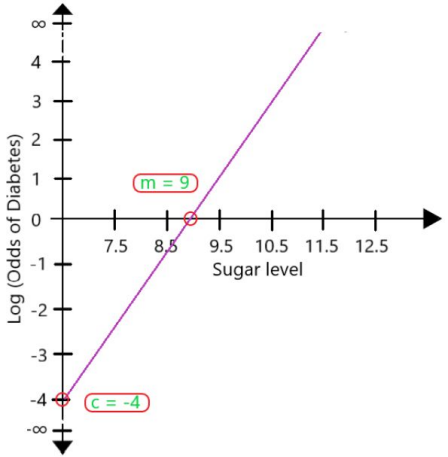
  - *which is the general equation of logistic regression.*
- *Now, in the logistic model, L.H.S contains the log of odds ratio*
  - *that is given by the R.H.S involving a linear combination of weights and IVs.*

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x$$

Role of Log Odds in Logistic Regression

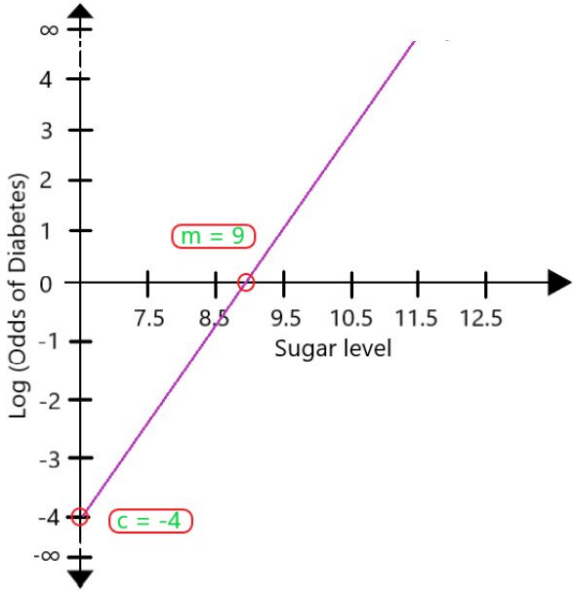
- **1. Example**: (Problem: with Probability based output in Logistic Regression.)
  - Build a LR model to determine
    - the prob. of a person suffering from diabetes based on their sugar level.
  - Observe the plot for this shown in the Figure:
  - The problem remains that o/p of model is only binary based on this plot.
  - To tackle this problem,
    - we use the concept of log odds present in logistic regression.
- 2. **Example** to solve the above problem: (Solution: Transforming Output)
  - We convert the probability-based output to log odds based output as:
    - P(diabetes) = log (odds of diabetes) =log_e [p/(1-p)]
    - Assume random values of p and see how the y-axis is transformed.





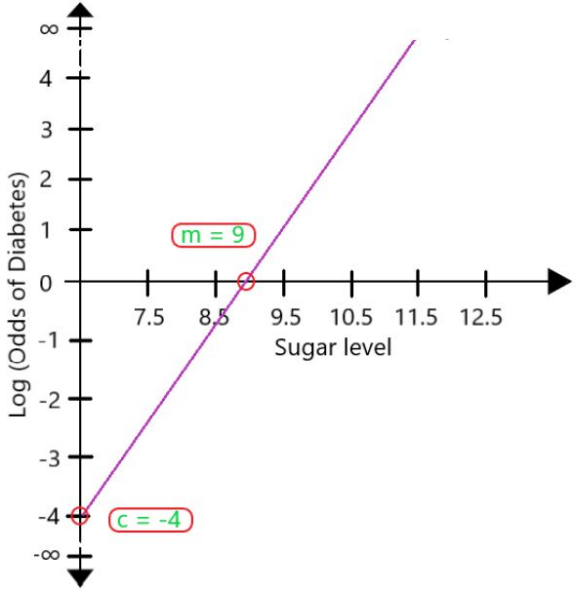
Role of Log Odds in Logistic Regression

- P(diabetes) = log (odds of diabetes) =$\log_e$ [p/(1-p)]
- Assume random values of p and see how the y-axis is transformed.
- 1. Boundary values;
  - at p=1 => $\log_e(p/(1-p)) = \log_e(1/0) = \log_e(1) - \log_e(0) = 0-(-) =$
  - at p=0 => $\log_e(p/(1-p)) = \log_e(0/1) = \log_e(0) - \log_e(1) = (-)-0 = -$
  - So, the domain of y axis is: (-∞, ∞)
- 2. Middle value;
  - at p=0.5 => $\log_e(p/(1-p)) = \log_e(0.5/0.5) = \log_e(1) = 0$
  - So, at p = 0.5 -> log (odds) = y = 0.
- 3. At random values;
  - at p=0.75 => $\log_e(p/(1-p)) = \log_e(0.75/0.25) = \log_e(3) = 1.09$
  - at p=0.3 => $\log_e(p/(1-p)) = \log_e(0.3/0.7) = \log_e(3) - \log_e(7) = -0.84$
  - So, at p > 0.5 -> we get value of log(odds) in range (0, ∞)
  - and at p < 0.5 -> we get value of log(odds) in range (-∞, 0)
- If we map these values onto a transformed plot,
  - it would look like as shown in this Figure.



Role of Log Odds in Logistic Regression

- Based on the value of slope (m) and intercept (c),
  - we can easily interpret the model and get non-binary deterministic output.
  - This is power of log odds in Logistic Regression.
- Log odds commonly known as Logit function is used in Logistic Regression models
  - when we are looking non-binary output.
- This is how logistic regression is able to work as both a regression as well as classification model.



References

Text books:

1. Ethem Alpaydin, “Introduction to Machine Learning , 4th Edition, The MIT Press, 2020.
2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012."
3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.
4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1ed, O'reilly, 2016.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012."

Reference Books:

6. Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd Edition, Shroff/O'Reilly, 2019.
7. Witten Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal., “Data Mining: Practical machine learning tools and techniques”, 1st Edition, Morgan Kaufmann, 2016.
8. Han, Kamber, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.
9. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, “Foundations of Machine Learning”, 1ed, MIT Press, 2012.
10. H. Dunham, “Data Mining: Introductory and Advanced Topics”, 1st Edition, Pearson Education, 2006.

Thank You.