Machine Learning

Learning with Regression and Trees

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”.

Logistic Regression

- Logistic Regression
- Logistic regression is a supervised machine learning algorithm.
- It is extensively used in predictive modeling.
- It helps to predict the probability of an outcome, event, or observation.
- It is to predict the probability that an instance belongs to a given class or not.
- The model delivers a binary outcome limited to two possible outcomes: yes/no, 0/1, or true/false.
- It analyzes the relationship between one or more IVs and classifies data into discrete classes.
- It is used for binary classification tasks using sigmoid function.
- Sigmoid function takes input as IVs and produces a probability value between 0 and 1.
- Example
  - For two classes Class 0 and Class 1
  - if the value of the logistic function for an input > 0.5 (threshold value)
    - then it belongs to Class 1 otherwise it belongs to Class 0.
  - It is referred to as regression because it is the extension of linear regression
    - but is mainly used for classification problems.

Logistic Regression

- Logistic regression is commonly used in binary classification.
- In statistics, the logistic model (or logit model) is a statistical model.
- It models the log odds of an event as a linear combination of one or more IVs.
- In regression analysis,
  - logistic regression (or logit regression) estimates
    - the parameters of a logistic model (the coefficients in the linear or non linear combinations).
- In binary logistic regression
  - there is a single binary DV, coded by an indicator variable,
  - where the two values are labeled "0" and "1",
  - while the IVs can each be a binary variable (two classes) or a continuous variable (any real value).
- 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.

Logistic Regression

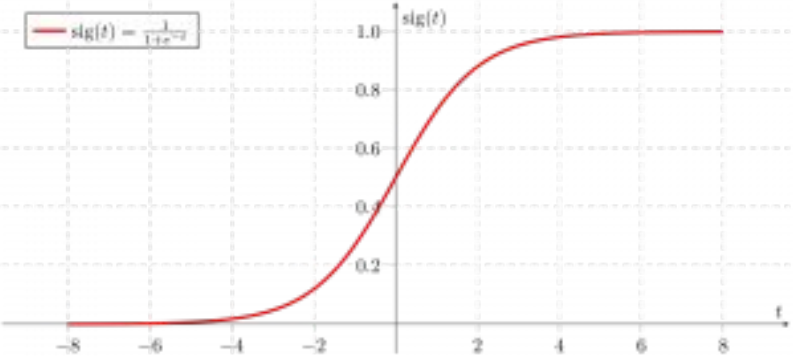
- Logistic regression predicts the output of a categorical dependent variable.
- Therefore, the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc.
  - But instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie bet. 0 and 1.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function,
  - which predicts two maximum values (0 or 1).

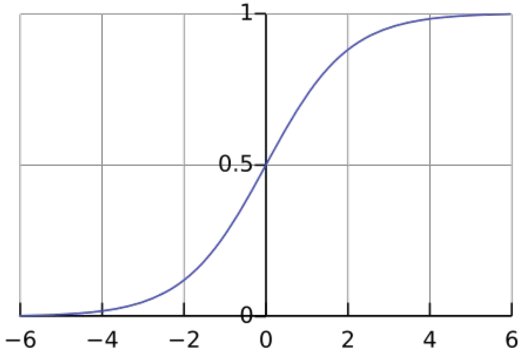
Logistic Regression

- 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.

Logistic Regression

- Understanding Sigmoid Function (the core of logistic regression)
- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1,
  - which cannot go beyond this limit, so it forms a curve like the “S” form.
  - The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value,
  - which defines the probability of either 0 or 1.
  - Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.
- Standard logistic function where; L = 1; k = 1; x0 = 0

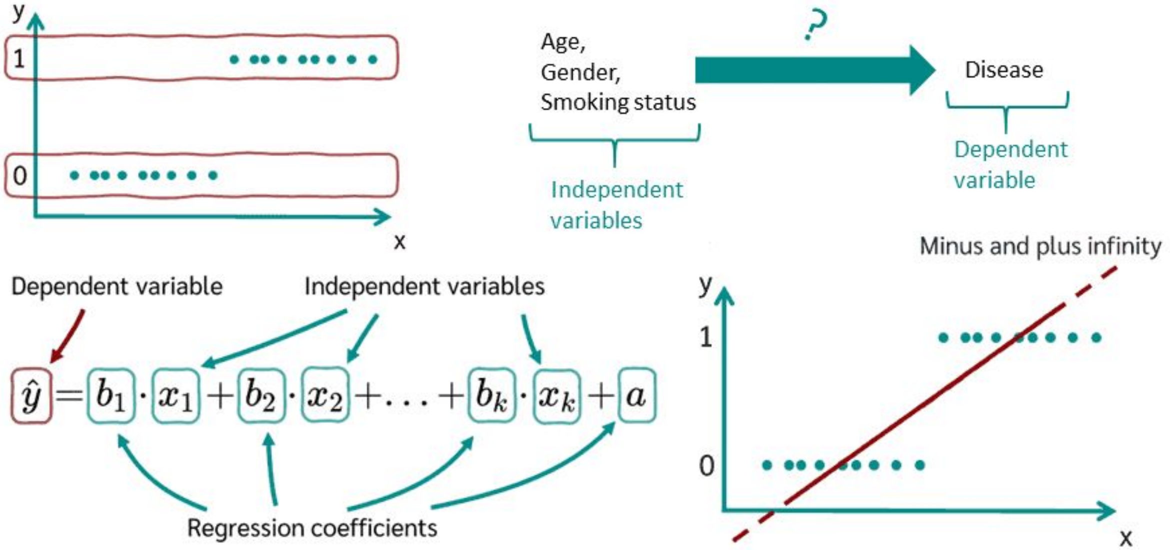




Logistic Regression

- Understanding Sigmoid Function (the core of logistic regression)
- The Figure sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.

$$\sigma = \frac{1}{1+e^{-x}}$$



Logistic Regression

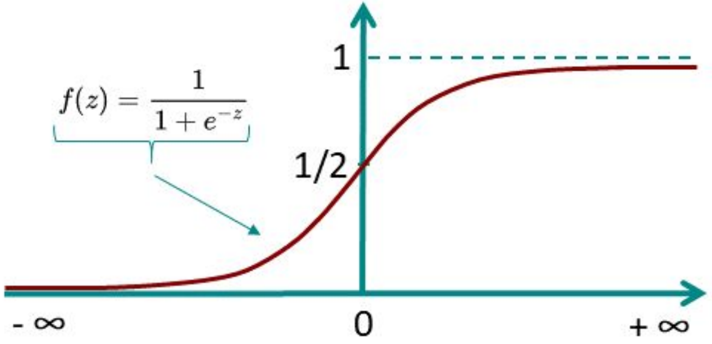
$$\sigma = \frac{1}{1+e^{-x}}$$

- In linear regression
  - IVs (e.g., age and gender) are used to estimate the specific value of the DV (e.g., body weight).
- In logistic regression
  - DV is dichotomous variables (0 or 1) and
  - the probability of the occurrence of value 1 (characteristic present) is estimated.
- Perform logistic regression:
  - To find out which variables have an influence on a disease.
  - To examine effect of age, gender and smoking or not for detecting a particular disease.
    - In this case, 0 could stand for not diseased and 1 for diseased.
  - To estimate the probability of occurrence and not the value of the variable itself.
  - To do this, it is necessary to restrict the value range for the prediction to the range bet. 0 and 1 (see figure).
  - To ensure that only values between 0 and 1 are possible, the logistic function f is used:
    - $f(x) = 1 / (1 + e^{-x})$

Logistic Regression

$\sigma = \frac{1}{1+e^{-x}}$

- Logistic function
- The logistic model is based on the logical function.
- The special thing about the logistic function is that for values between minus and plus infinity,
  - it always assumes only values between 0 and 1.
- To ensure that only values between 0 and 1 are possible, the logistic function f is used: f(z) = 1 / (1 + e^{-z})



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

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\boxed{(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}}$$

Logistic Regression

- To ensure that only values between 0 and 1 are possible, the logistic function f is used: f(z) = 1 / (1 + e^{-z})
- Diseased = Yes = 1 i.e P(y=1)
- Diseased = No = 0 i.e P(y=0)
- The probability that for given values of the independent variable the dichotomous dependent variable y is 0 or 1 is given by:

$$P(y=1|x_1,\ldots,x_n)=\frac{1}{1+e^{-(b_1\cdot x_1+\ldots+b_k\cdot x_k+a)}}$$

$$P(y=0|x_1,\ldots,x_n)=1-\frac{1}{1+e^{-(b_1\cdot x_1+\ldots+b_k\cdot x_k+a)}}$$

- To calculate the probability of a person being sick or not using the logistic regression:
  - The model parameters b1, b2, b3 and a must first be determined.
  - Once these have been determined, the equation for the example above is:

$$P(\text{Diseased})=\frac{1}{1+e^{-(b_1\text{Age}+b_2\text{Gender}+b_3\text{Smoking status}+a)}}$$

$$\sigma=\frac{1}{1+e^{-x}}$$

Logistic Regression

- Example 1
  - Use a logistic regression with one explanatory variable and two categories to answer the question:
  - Dataset: A group of 5 person with disease yes/no for age, gender and smoking status.
  - Question: How does the age, gender and smoking status affect the probability of having a particular disease?

Sr No	Age	Gender	Smoking	Disease
1	25	0	1	1
2	37	1	0	0
3	40	0	0	0
4	49	0	1	1
5	55	1	1	1

  - Answer:
    - Disease = Yes = 1 i.e P(y=1) and Disease = No = 0 i.e P(y=0)
    - $f(z) = P(Disease) = 1 / (1 + e^{-(b0+b1*Age+b2*Gender+b3*Smoking)})$
    - Logistic Regression Solved Example 1 [PDF]

Logistic Regression

- Logistic Regression Solved Example 1 [PDF]

Answer: IV is Independent variable and DV is Dependent variable

	x1	x2	x3	y	e =	2.71828
Sr No	**Age (x1)**	**Gender (x2)**	**Smoking (x3)**	**Disease (y)**	**Calcuated Y (Fitted Value)**	**Prediction**
1	25	0	1	1	0.9947	1
2	37	1	0	0	0.0034	0
3	40	0	0	0	0.0184	0
4	49	0	1	1	0.9263	1
5	55	1	1	1	0.4518	0

Determine Predicted value of y:

Disease = Yes = 1 i.e P(y=1) and Disease = No = 0 i.e P(y=0)

$$f(z) = P(Disease) = 1 / (1 + e^{-(b0+b1*Age+b2*Gender+b3*Smoking)})$$

C' Matrix	Predicted 1	Predicted 0
Actual 1	TP	FN
Actual 0	FP	TN

C' Matrix	**Predicted 1**	**Predicted 0**	**Total (N)**
Actual 1	2	1	3
Actual 0	0	2	2
Total (N)	2	3	**5**

P, Precision or PPV =	TP / (TP+FP)=	1.000
NPV =	TN / (TN+FN)=	0.667
False Omission Rate (FOR) =	1 - NPV=	0.333
R, Recall (Sensitivity) or TPR =	TP / (TP+FN)=	0.667
Specificity or NPV or TNR =	TN / (TN+FP)=	1.000
False Positive Rate (FPR) =	FP / (FP+TN)=	0.000
False Negative Rate (FNR) =	FN / (FN+TP)=	0.333
Accuracy =	(TP+TN)/(TP+TN+FP+FN)=	0.800
F1 score =	2 * P * R / (P + R)=	0.8

Logistic Regression

- Example 2
  - Use a logistic regression with one explanatory variable and two categories to answer the question:
  - Dataset: A group of 20 students spends between 0 and 6 hours studying for an exam.
  - Question: How does the # hours spent studying affect the probability of the student passing the exam?
  - Answer:
    - Result = Pass = 1 i.e P(y=1) and Result = Fail = 0 i.e P(y=0)
    - f(z) = P(Result) = 1 / (1 + $e^{-(b0+b1*Slept+b2*Study)}$)
    - Logistic Regression Solved Example 2 [PDF]

Hours (x_k)	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass (y_k)	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	1	1	1	1

Logistic Regression

- Logistic Regression Solved Example 2 [PDF]

Answer: IV is Independent variable and DV is Dependent variable

e = 2.71828

Sr No	Slept (x1) (IV)	Study (x2) (IV)	Result (y) (DV)	Calcuated Y (Fitted Value)	Prediction
1	7.23	5.55	0	0.0101	0
2	8.12	5.12	1	0.0023	0
3	9.23	6.23	0	0.0018	0
4	5.12	8.12	1	0.6342	1
5	9.23	6.23	0	0.0018	0
6	3.55	9.55	1	0.9798	1
7	4.32	8.32	1	0.8472	1

Determine Predicted value of y: Result = Pass = 1 i.e P(y=1) and Result = Fail = 0 i.e P(y=0)

$f(z) = P(Result) = 1 / (1 + e^{-(b0+b1*Slept+b2*Study)})$

C' Matrix	Predicted 1	Predicted 0
Actual 1	TP	FN
Actual 0	FP	TN

C' Matrix	Predicted 1	Predicted 0	Total (N)
Actual 1	3	1	4
Actual 0	0	3	3
Total (N)	3	4	7

P, Precision or PPV =	TP / (TP+FP)=	1.000
NPV =	TN / (TN+FN)=	0.750
False Omission Rate (FOR) =	1 - NPV=	0.250
R, Recall (Sensitivity) or TPR =	TP / (TP+FN)=	0.750
Specificity or NPV or TNR =	TN / (TN+FP)=	1.000
False Positive Rate (FPR) =	FP / (FP+TN)=	0.000
False Negative Rate (FNR) =	FN / (FN+TP)=	0.250
Accuracy =	(TP+TN)/(TP+TN+FP+FN)=	0.857
F1 score =	2 * P * R / (P + R)=	0.8571428571

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.