Machine Learning with R: Logistic Regression

Humayun Kabir, BScN, MPH, MSc (Student)
MSc in Health Research Methodology

Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada

Date: March 22, 2024

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University | Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information:
scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Attendance that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

Verify your participation at a session: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events:

March 28: Intermediate Python Programming – Seyed Amirreza Mousavi

April 30: Survival Analysis with R – Humayun Kabir

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- □ Creating data visualizations, including charts, graphs, and scatter plots
- □ Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- □ Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- □ Choosing which software package to use, including free and open-source software
- □ Troubleshooting problems related to file formats, data retrieval, and download
- □ Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: https://library.mcmaster.ca/services/dash

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Machine Learning with R: Logistic Regression

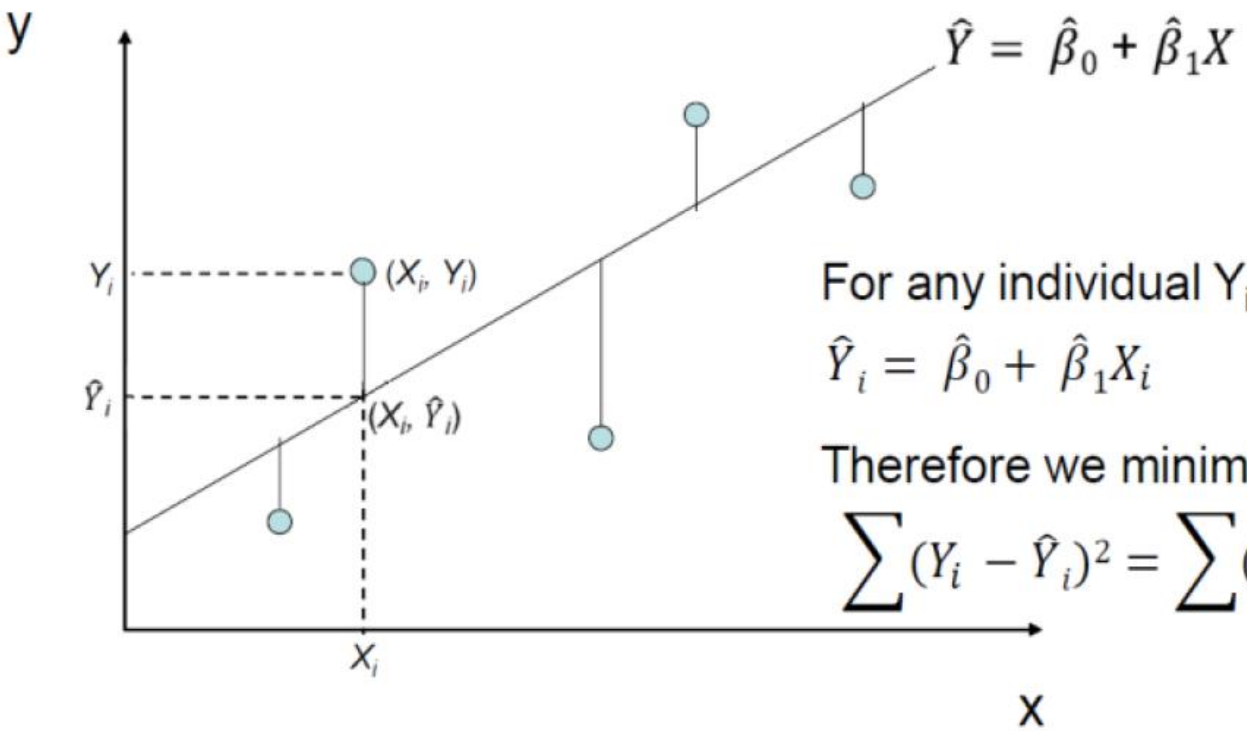
Objective

LEARNING THE BASICS OF LOGISTIC REGRESSION

MACHINE LEARNING CODE WITH R

Linear regression

Geometry of Least Square



For any individual Y_i,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Therefore we minimize:

$$\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Source: Dr. Shofiqul Islam, McMaster University

What is Logistic Regression?

- Many situations where the dependent variable is binary or categorical.

✓ Dead vs. Aliv

✓ CVD vs No CVD

✓ Employed vs. Unemployed

✓ Guilty vs. Not guilty

- Requires to consider a special type of model called logistic regression

Source: Dr. Shofiqul Islam, McMaster University

A new Regression Model

- Need to modify our regression equation so that:
  1. Predictions lie between 0 and 1
  2. Effects of covariates can be interpreted on a relative (multiplicative) scale.

- Solution
  - Use the log odds or logit of p to represent the
  - Outcome: $\text{logit } p = \ln\left(\frac{p}{1-p}\right)$
- Where p is the probability of having the outcome
  - When $p \rightarrow 0$, $\text{logit } p \rightarrow -\infty$
  - When $p \rightarrow 1$, $\text{logit } p \rightarrow \infty$

Source: Dr. Shofiqul Islam, McMaster University

A new Regression Model

- Logit transformation of the binary Outcome leads to Logistic Regression
- The logistic function: $\frac{1}{1+e^{-x}}$
- Inverse is called the logit: $\ln(\frac{x}{1-x})$
- Logistic regression models:
  - The log-odds (or logit) of a binary outcome as a straight-line function of covariates:

$$\mathrm{E}[\ln(\frac{p}{1-p})|\mathrm{X}] = \beta_0 + \beta_1 * \mathrm{X}_1 + \; ... + \beta_k * \mathrm{X}_k$$

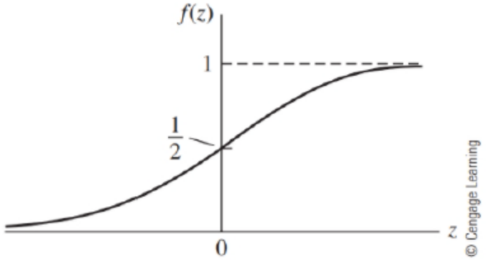


FIGURE 22.1 The logistic function $f(z) = \frac{1}{1+e^{-z}}$

Source: Dr. Shofiqul Islam, McMaster University

Logistic Function

- How do we transform the logit back to p ?

$$\text{Logit(p)} = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$e^{ln\left(\frac{p}{1-p}\right)} = e^{(\beta_0+\beta_1 X)}$$

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

$$p = (1-p)\, e^{(\beta_0+\beta_1 X)}$$

$$p = e^{(\beta_0+\beta_1 X)} - pe^{(\beta_0+\beta_1 X)}$$

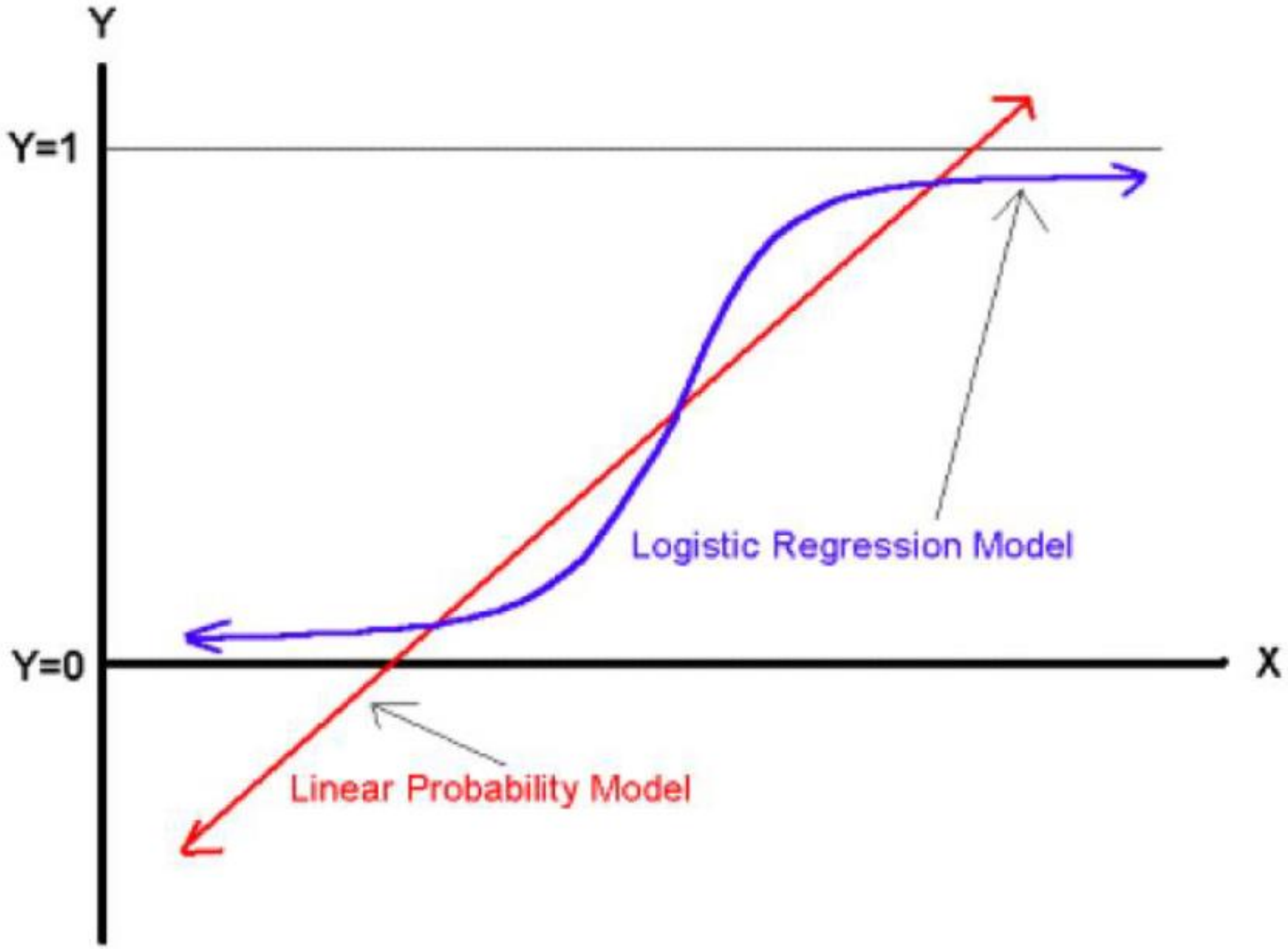
$$p + pe^{(\beta_0+\beta_1 X)} = e^{(\beta_0+\beta_1 X)}$$

$$p\left(1 + e^{(\beta_0+\beta_1 X)}\right) = e^{(\beta_0+\beta_1 X)}$$

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

Source: Dr. Shofiqul Islam, McMaster University

Comparing the Linear vs Logistic fit



Source: Dr. Shofiqul Islam, McMaster University

Assumptions of Logistic Regression

- The logistic regression model assumes:
  - Outcome is a binary or dichotomous variable
  - There is a linear relationship between the logit of the outcome and each predictor variables
  - There is no influential values (extreme values or outliers) in the continuous predictors
  - There are no high correlations (multi-collinearity) among the predictors

Source: Dr. Shofiqul Islam, McMaster University

Logistic Regression

- Goodness of fit is examined using
  - Measures of predictive ability:
    - Pseudo R^2 by McFadden (1974)
    - Generalized R^2 by Cox-Snell (1989)
    - Tjur (2009) coefficient of discrimination
    - Diagnostic test criteria - sensitivity/specificity, area under the ROC curve
- Goodness of fit statistics –
  - Deviance and Pearson chi-squared statistics
  - Hosmer-Lemeshow (1980) test
- Information criteria – Akaike (AIC) & Bayesian (BIC)

Source: Dr. Shofiqul Islam, McMaster University

Machine learning

Machine learning involves showing a large volume of data to a machine/model so that it can learn and make predictions, find patterns, or classify data.

Source: coursera.org

Types of machine learning

Basically, machine learning are three types.

Supervised

Unsupervised

Reinforcement learning.

Source: coursera.org

Supervised learning

Machine learning ***feeds historical input and output data*** *in machine learning algorithms, with processing in between each input/output pair that allows the algorithm to shift the model* ***to create outputs as closely aligned with the desired result as possible****.*

Common algorithms used during supervised learning include linear regression, neural networks, decision trees, and support vector machines.

Source: coursera.org

Unsupervised learning

While supervised learning requires users to help the machine learn, unsupervised ***learning algorithms do not use the same labeled training sets and data****. Instead, the machine looks for less obvious patterns in the data.*

Unsupervised machine learning is very helpful when you need ***to identify patterns*** *and use data to make decisions.*

Common algorithms used in unsupervised learning include k-means clustering, and Gaussian mixture models.

Source: coursera.org

Reinforcement learning

Reinforcement learning is the closest machine learning type to how humans learn.

*The algorithm used **learns by interacting with its environment and getting a positive or negative reward**.*

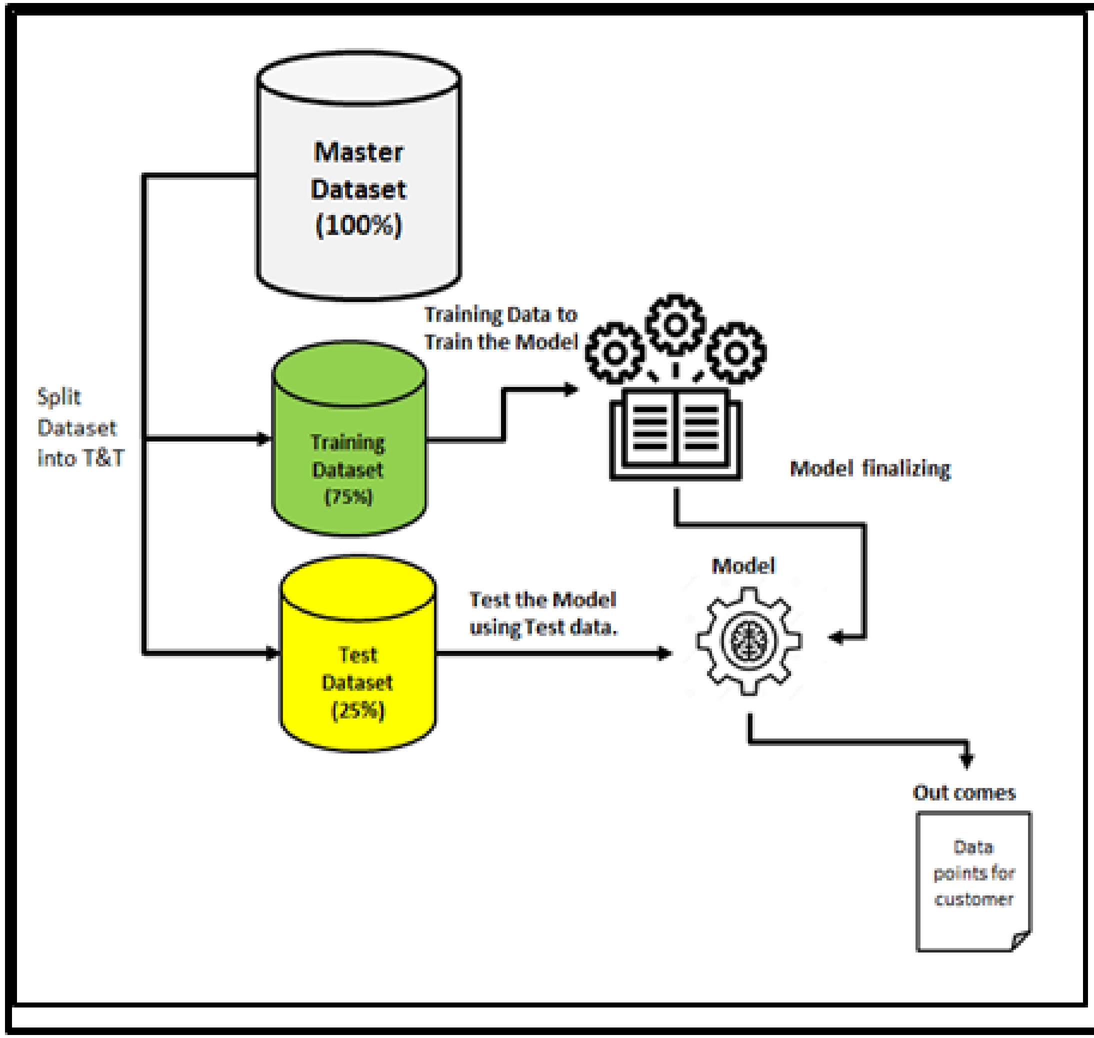
Common algorithms include temporal difference, and Q-learning.

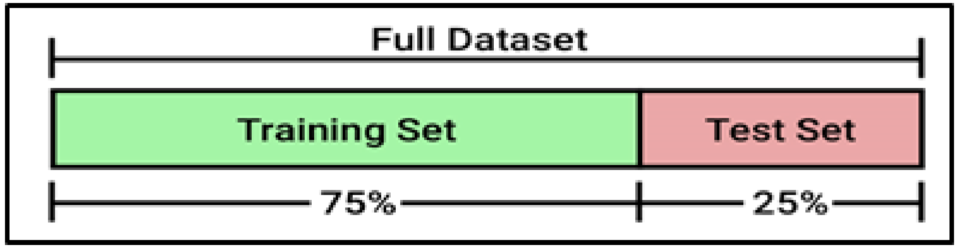
Source: coursera.org

Logistic regression

- A supervised machine learning
- Learns from labeled data
- Make predictions on unseen data

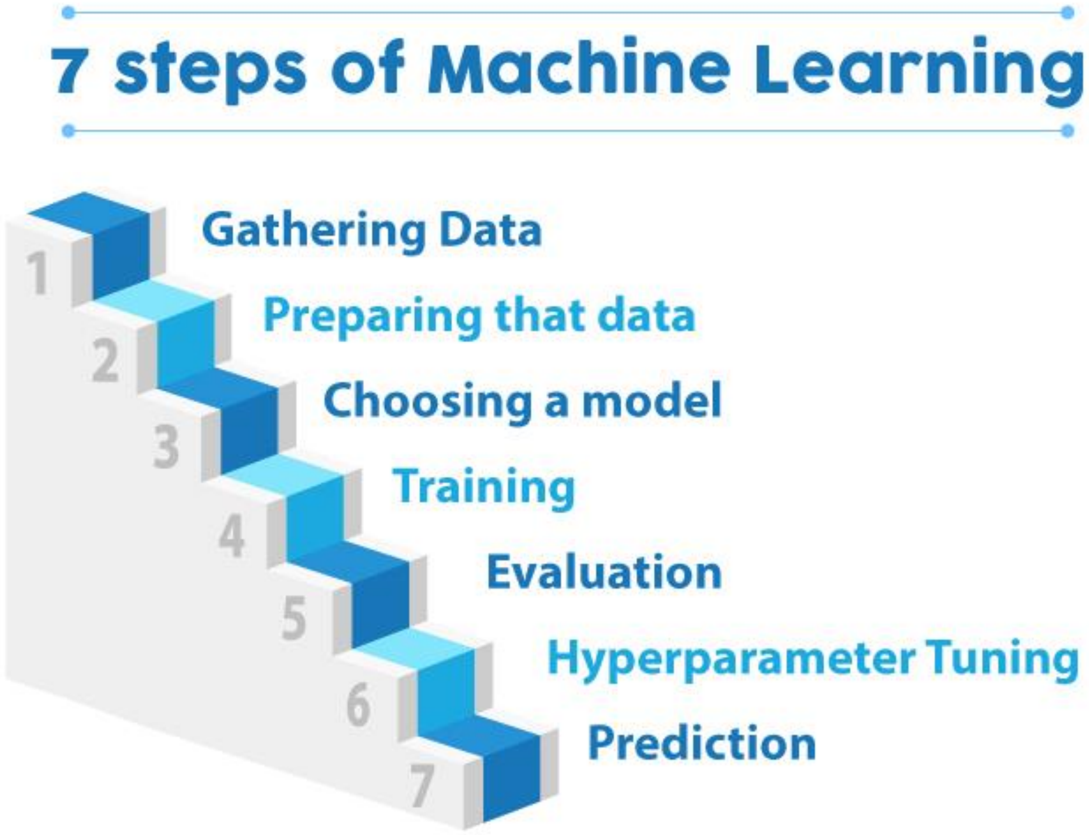
Steps of ML





https://www.analyticsvidhya.com/

Steps of ML including tuning



7 steps of Machine Learning

1. Gathering Data
2. Preparing that data
3. Choosing a model
4. Training
5. Evaluation
6. Hyperparameter Tuning
7. Prediction

www.mygreatlearning.com

Contact

Book an appointment with DASH: https://library.mcmaster.ca/services/dash

Contact DASH: Data Analysis Support Hub: libdash@mcmaster.ca

Let move to the coding part

https://colab.research.google.com/drive/1iOvC52mkSdQ-EkNQZFSi-O9y05YSRSMs#scrollTo=yi7vfdCrHQaU