Logistic Regression

Oommen P. Jose, Dibya Bharati Pradhan

National Institute of Science Education and Research (NISER)
Bhubaneswar

January 30, 2023

Part I: Demo Presentation Part

INTRODUCTION

Logistic Regression is an supervised regression algorithm that can be used for classification by setting a parameter called threshold. Logistic regression gives the probability that a data point of belongs to one of the two classes(let us take it as positive and negative class). If a threshold is set with the condition that if the output value of the regression is greater than the threshold it is classified as belonging to the positive class and if it is less than the threshold it is classified as negative.*Understanding logistic regression* 2023

INTRODUCTION

"Logistic" in logistic regression signifies that the standard logistic function which is the sigmoid function is used by this algorithm.*Logistic regression* 2023

$$f(x) = \frac{1}{1 + e^{x}}$$

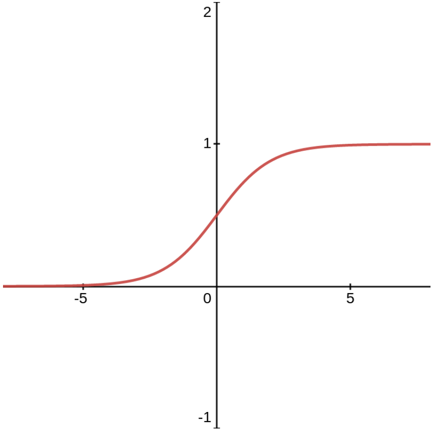

Figure. Sigmoid function

Model

The model uses the linear regression formula whose output is then mapped to the interval between 0 and 1 using the sigmoid function. The final function can be represented as follows

$$S_{W,b} = \frac{1}{1 + e^{wx+b}}$$

This function gives the probability that the given data point x belongs to positive class(class 1). The probability that the data point belongs to negative class(class 0) is given by $1 - S_{W,b}$

Expected Loss

The expected loss for logistic regression is represented using the concept of likelihood. Likelihood quantifies how likely it is for a data point to be assigned the correct label. The expression for likelihood is given below.

$$L_{(W,b)} \equiv L((W,b) \mid y; X) = \prod_{i=1}^{N} S_{W,b}\left(X_i\right)^{y_i} \left(1 - S_{W,b}\left(X_i\right)\right)^{(1-y_i)}$$

To minimize loss we need to maximise how likely a data point is assigned the correct label hence we try to maximise the likelihood function.[Shekhar n.d.]

Expected Loss

We see that the equations contain exponentials which are hard to calculate so we define log likelihood for easier calculation. The expression for log likelihood is given below.

$$\frac{1}{N}log(L_{(W,b)}) \equiv \frac{1}{N}log(L((W,b) \mid y; X)) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i * log(S_{W,b}(X_i)) + (1-y_i) * log\left(1 - S_{W,b}(X_i)\right)\right)$$

Sample Code

```
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

# Generate a random dataset
np.random.seed(0)
X = np.random.randn(100, 2)
y = (X[:, 0] + X[:, 1] > 0).astype(int)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Train a logistic regression model
clf = LogisticRegression()
clf.fit(X_train, y_train)

# Evaluate the model on the test set
accuracy = clf.score(X_test, y_test)
print("Accuracy: ", accuracy)
```

```
Accuracy:  0.9
```

This is a sample code which generates a random dataset of 100 samples and 2 features, splits the data into training and testing sets, trains a logistic regression model on the training data, and evaluates the model's accuracy on the test data.

Multinomial Regression

Logistic Regression function can be used to classify data to multiple classes. Say we have four classes i,j,k. First we evaluate whether the data point belongs to the class i or not using logistic regression. Then the data points that don't belong to class i are taken and evaluated using logistic regression to see if the data belongs to class j or k. This can be scaled up for n classes.

REFERENCES

Logistic regression (Jan. 2023). URL: https://en.wikipedia.org/wiki/Logistic_regression.

Shekhar, Danush (n.d.). *CS460 Lecture 9 - National Institute of Science Education and research*. URL: https://niser.ac.in/~smishra/teach/cs460/2020/lectures/lec9/.

Understanding logistic regression (Jan. 2023). URL: https://www.geeksforgeeks.org/understanding-logistic-regression/.