## Logistic Regression Model

Logistic regression is a statistical method that is used for binary classification problems, meaning cases where the dependent varible is categorical with two possible outcomes. Logistic Regression models the propability that a given input belongs to a particular category. Despite its name, logistic regression is a classification algorithm, not a regression algorithm.

Mathematical Representation

In logistic regression, the logistic function is used to model the relationship between the independent variables and the probability of a particular outcome. The logistic function is represented as shown below:

$$P(y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

Where:

P(y = 1) is the is the probability that the dependent variable y is equal to 1 (the positive class).

e is the base of the natural logarithm.

$\beta_0$, $\beta_1$, $\beta_2 \cdots \beta_n$ are the coefficients of the model, which are estimated from the training data.

$x_1$, $x_2 \cdots x_n$ are the input features.

Practical Example

For the practical examle let us consider an example of predicting if a student will pass or fail on their exam based on the number of hours they studied for it. Below is given a dataset with the information about the hours studied for each student (per week) and the information about wether each of them passed or failed on the exam. The data was put into a CSV file named exam.csv

The data has two columns column on the left is „Hours Studied per week“ and the one on the right is „Passed (1) / Failed (0)“

Data from the csv is as follows:

| Hours Studied per week | Passed (1) / Failed (0) |
|---|---|
| 8.5 | 0 |
| 9.2 | 0 |
| 10.1 | 0 |
| 10.5 | 0 |
| 10.8 | 0 |
| 11.2 | 0 |
| 11.5 | 0 |
| 12.3 | 0 |
| 12.8 | 0 |
| 13.2 | 0 |
| 13.5 | 0 |
| 14.1 | 0 |
| 14.5 | 0 |
| 15 | 1 |
| 15.3 | 1 |
| 16 | 1 |
| 16.5 | 0 |
| 17.2 | 1 |
| 17.8 | 1 |
| 18.5 | 1 |
| 19 | 1 |
| 19.5 | 1 |
| 20.2 | 1 |
| 20.8 | 1 |
| 21.5 | 1 |
| 22 | 1 |
| 22.5 | 1 |
| 23 | 1 |
| 23.5 | 1 |
| 24 | 1 |

Python Code

```python
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression

# Read data from CSV file
data = pd.read_csv("exam.csv")

# Separate independent and dependent variables
X = data.iloc[:, :-1]  # Features (hours studied)
y = data.iloc[:, -1]   # Target (pass or fail)

# Create a logistic regression model
model = LogisticRegression()

# Fit the model to the data
model.fit(X, y)

# Predict probabilities of passing for new data points
new_X = np.array([[8.3], [16], [17.2], [23.7]])
probabilities = model.predict_proba(new_X)

# Output the predicted probabilities
for i, prob in enumerate(probabilities):
    print(f"Probability of passing for {new_X[i][0]} hours: {prob[1]:.2f}")
```

We calculated probabilities of passing for students who studied following hours per week: 8.3, 16, 17.2, 23.7 .

This yielded the results:

```
Probability of passing for 8.3 hours: 0.00
Probability of passing for 16.0 hours: 0.68
Probability of passing for 17.2 hours: 0.89
Probability of passing for 23.7 hours: 1.00
```

In this example, we used logistic regression to predict whether a student will pass or fail an exam based on the number of hours they studied. The model learned the relationship between hours studied and the probability of passing from the training data and made predictions for new data points.