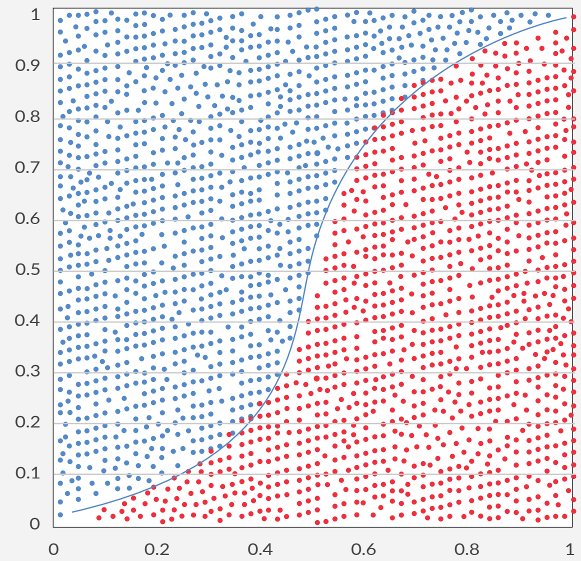
Logistic Regression

- Logistic regression is a statistical method that is primarily used for binary classification/
- It can be used for classification problems, such as spam detection, diabetes prediction and cancer detection problems.
- Instead of fitting a line to the data, logistic regression fits an S-shaped curve (called sigmoid function).
- Logistic regression computes the probability of a data point along with predictive classification.
- Simple logistic regression: Here, a single independent variable is used to predict the output.
- Multiple logistic regression: Here, multiple independent variables are used to predict the output.
- Logistic regression can be extended to solve multiclass classification problems.

Samples divided into two groups using logistic regression



Common Interview Questions:

- Why can linear regression be used for binary classification?
- Why is logistic regression called “regression” but still used for “classification” problems?
- List the advantages and disadvantages of using logistic regression for classification problems.
- What is the impact of outliers on logistic regression?
- Do residues exist in logistic regression? If not, why?
- How will you evaluate the performance of a logistic regression-based model?
- Explain the cost function used in logistic regression models.
- What is the one-vs-all method in logistic regression?
- How will you compare the performance of multiple logistic regression models?

Assumptions of the Logistic Regression Model

- The dependent variable must be linearly separable.
- Groups must be mutually exclusive and exhaustive.
- Absence of strong outliers

Types of Logistic Regression

Binary logistic regression: The target variable has only two possible outcomes: Gmail inbox message is spam or not spam.

Multinomial logistic regression: The target variable has three or more nominal categories (e.g., predicting the type of wine).

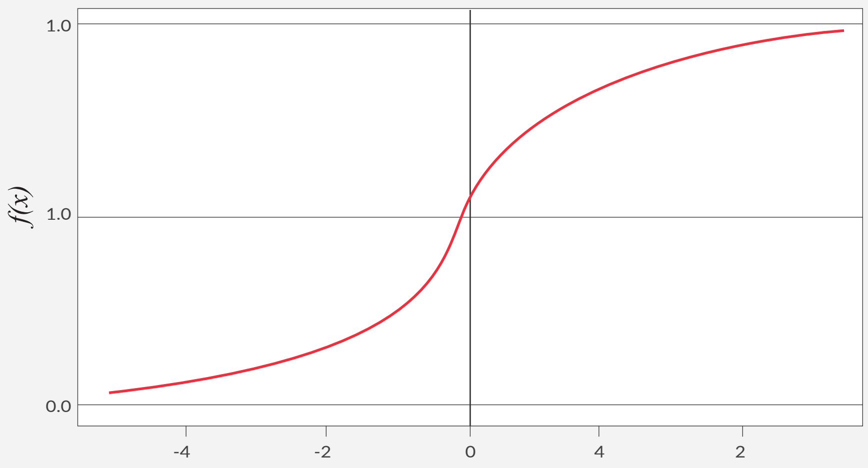
Ordinal logistic regression: The target variable has three or more ordinal categories, such as a rating of 1–5.

Logistic Regression

Sigmoid Function (logistic function)

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

- S-shaped curve
- Takes any real number as an input and gives a value between 0 and 1

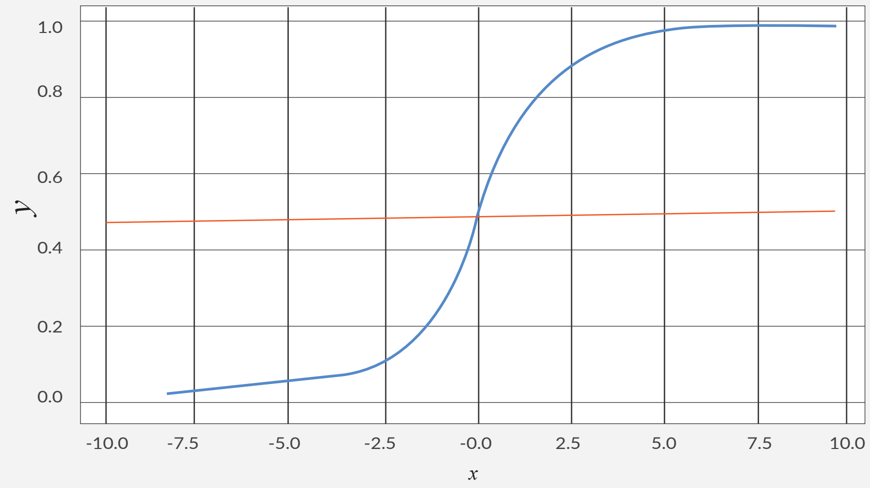


- As x tends toward ± infinity, the output becomes ± 1.
- These properties of the sigmoid function are used to estimate t.
- The sigmoid function is used as the "cost function" in logistic regression
- We can replace f(x) with p (probability) and x with a linear equation to get the following:

$$p = \frac{1}{1+e^{-(\beta_o+\beta_1 X_1+\beta_2 X_2+\ldots\ldots+\beta_n X_n)}}$$

Decision Boundary

If f(x) > 0.5, then "A;" otherwise, it will be "B." This will create a "decision boundary," which can classify samples into two different groups.



Best Fit

- Linear regression is estimated using ordinary least squares (OLS), whereas logistic regression is estimated using the maximum likelihood estimation (MLE) approach.
- The MLE approach is a "likelihood" maximisation method, whereas OLS is a distance-minimising approximation method.

Code

```
model = LogisticRegression(solver='liblinear')
model.fit(x,y)
```

note - max_iter parameter can be used in LogisticRegression to increase the iteration of model

Logistic Regression

Odd Logs

$$\text{ODDS} = \frac{\text{Probability of winning}}{\text{Probability of losing}} = \frac{p}{1-p}$$

- Taking the log of odds will fix the scale and help compare events.
- Log(odds) is also known as the "logit" function.

$$\text{Logit Function} = \log\left(\frac{p}{1-p}\right)$$

Multi-Variate Logistic Regression

- These have more than one input variable.
- The equation used in the case of univariate logistic regression is given below.

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

So, for multivariate logistic regression, the equation will be as follows.

$$p = \frac{1}{1+e^{-(\beta_o+\beta_1 X_1+\beta_2 X_2+\beta_3 X_{3+}\ldots+\beta_n X_n)}}$$

RFE

- RFE stands for Recursive Feature Elimination.
- It is used to eliminate features from a data set that might not be important for the model.
- Redundant features can be mostly eliminated.
- This saves processing power and memory space.
- RFE is used before applying the model to the data.

Here is the code for using RFE before logistic regression.

```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
model = Logistic Regression()
rfe = RFE (model, 4) # selecting 4 top features
fit = rfe.fit(X, Y) # Logistics regression model
```

Logistic Regression

Model Performance

The confusion matrix is used to define the performance of a classification algorithm

	Predicted Class	
True Class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

To evaluate this performance, we can use these measures:

Precision : Indicates the precision of predictions

Recall: Indicates the approximate percentage of classes captured by the model

Accuracy: Indicates the accuracy of predictions

F1 score: Is similar to the weighted average of the precision and recall values.

Here's how these can be calculated

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + F}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

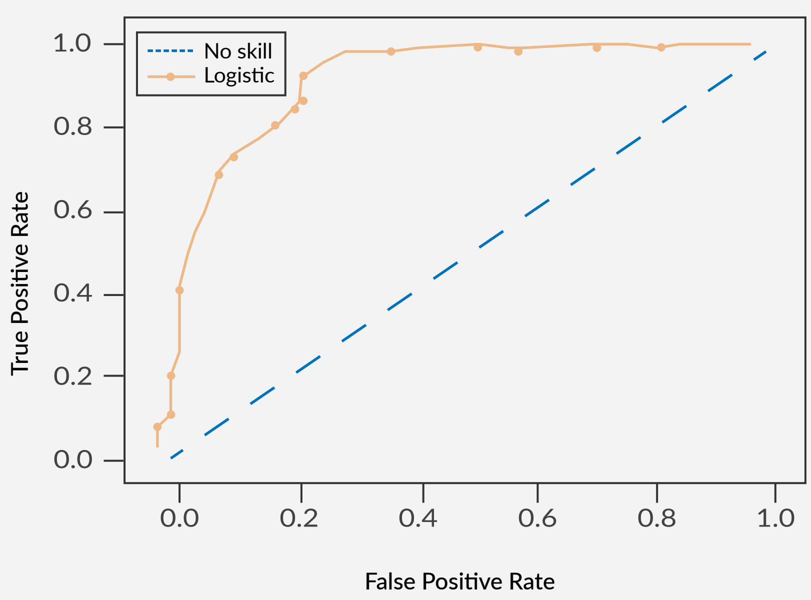
$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

ROC Curve

- This is used for determining the best cutoff value.
- The ROC curve gives information on the trade-off between sensitivity (or TPR) and specificity (1 – FPR).
- This indicates the performance of a classification model at all classification thresholds.
- The curve is plotted against two parameters: true positive rate and false positive rate.

Example of ROC curve



Here:

True positive rate (TPR) = True Positives / (True Positives + False Negatives)

False positive rate (FPR) = False Positives / (False Positives + True Negatives)

Note: TPR is also known as sensitivity and FPR = 1 - specificity.