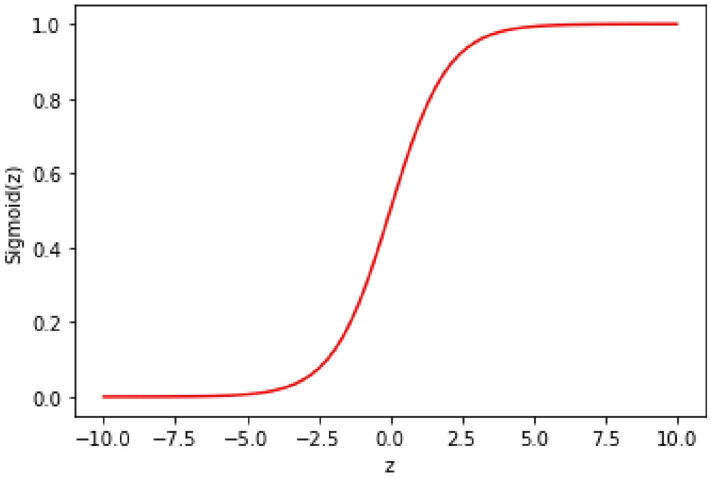
LOGISTIC REGRESSION

Logistic Regression Theory

Logistic Regression solves the classification problem, where the target variable is categorical in nature. It models the data using sigmoid function that takes in any value and outputs it to be between 0 and 1.

Sigmoid function, f(z)= 1/(1+e^(-z))



Model Evaluation

Confusion matrix is used to evaluate classification models performance on a set of test data for which the true values are already known.

		Predicted	
		NO	YES
Actual	NO	TN	FP
	YES	FN	TP

True Positive (TP): Actual observation is positive and is predicted to be positive.
True Negative (TN): Actual observation is negative and is predicted to be negative.
False Positive (FP): Actual observation is negative but is predicted positive.
False Negative (FN): Actual observation is positive but is predicted negative.

Evaluation Parameters

Accuracy: Ratio of correctly predicted classes to all the classes.
Accuracy= (TP+TN) / (TP+TN+FP+FN)

Misclassification Rate (Error rate): Ratio of wrong predictions to the total number of classes.
Error rate = (FP+FN) / (TP+TN+FP+FN)

Recall: Ratio of correctly predicted positive classes to all the actual positive classes.
Recall= TP / (TP+FN)

Precision: Ratio of correctly predicted positive classes to all positive predicted classes.
Precision= TP / (TP+FP)

F-measure: It is the harmonic mean of Recall and Precision.
*F-measure = (2 * Recall * Precision) / (Recall + Precision)*

Logistic Regression with Python

Project Background

Data Source:
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

The dataset has been taken from UCI machine learning repository. The main objective of the analysis is to perform classification of tumors i.e., benign(B) or malignant(M). A **benign tumor** is a tumor that does not invade its surrounding tissue or spread around the body. A **malignant tumor** is a tumor that may invade its surrounding tissue or spread around the body. This dataset consists of 569 rows and 33 columns.

Attribute Information:

id: ID number

diagnosis: The diagnosis of breast tissues (M = malignant, B = benign)

radius_mean: mean of distances from center to points on the perimeter

texture_mean: standard deviation of gray-scale values

perimeter_mean: mean size of the core tumor

area_mean

smoothness_mean: mean of local variation in radius lengths

compactness_mean: mean of perimeter^2 / area — 1.0

concavity_mean: mean of severity of concave portions of the contour

concave points_mean: mean for number of concave portions of the contour

symmetry_mean

fractal_dimension_mean: mean for “coastline approximation” — 1

radius_se: standard error for the mean of distances from center to points on the perimeter

texture_se: standard error for standard deviation of gray-scale values

perimeter_se

area_se

smoothness_se: standard error for local variation in radius lengths

compactness_se: standard error for perimeter^2 / area — 1.0

concavity_se: standard error for severity of concave portions of the contour

concave points_se: standard error for number of concave portions of the contour

symmetry_se

fractal_dimension_se: standard error for “coastline approximation” — 1

radius_worst: “worst” or largest mean value for mean of distances from center to points on the perimeter

texture_worst: “worst” or largest mean value for standard deviation of gray-scale values

perimeter_worst

area_worst

smoothness_worst: “worst” or largest mean value for local variation in radius lengths

compactness_worst: “worst” or largest mean value for perimeter^2 / area — 1.0

concavity_worst: “worst” or largest mean value for severity of concave portions of the contour

concave points_worst: “worst” or largest mean value for number of concave portions of the contour

symmetry_worst

fractal_dimension_worst: “worst” or largest mean value for “coastline approximation” — 1