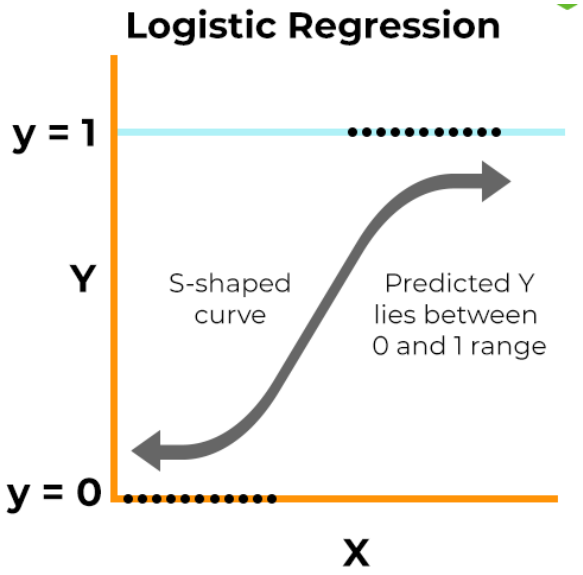
Logistic regression

Logistic regression is a well-known statistical technique that is used for modeling binary outcomes.

Logistic regression is a type of regression analysis used for predicting the probability of a categorical outcome. It's widely used in various domains such as healthcare (disease diagnosis), finance (credit risk assessment), marketing (customer churn prediction), and more. Unlike linear regression, which predicts continuous outcomes, logistic regression models the probability of an event occurring based on one or more predictor variables. It's a powerful and interpretable method for binary classification tasks, providing insights into the relationship between predictors and the likelihood of the outcome.



There are various implementations of logistic regression in statistics research, using different learning techniques. The Microsoft Logistic Regression algorithm has been implemented by using a variation of the Microsoft Neural Network

algorithm. This algorithm shares many of the qualities of neural networks but is easier to train.

One advantage of logistic regression is that the algorithm is highly flexible, taking any kind of input, and supports several different analytical tasks:

- Use demographics to make predictions about outcomes, such as risk for a certain disease.
- Explore and weight the factors that contribute to a result. For example, find the factors that influence customers to make a repeat visit to a store.
- Classify documents, e-mail, or other objects that have many attributes.

Example

Consider a group of people who share similar demographic information and who buy products from the Adventure Works company. By modeling the data to relate to a specific outcome, such as purchase of a target product, you can see how the demographic information contributes to someone's likelihood of buying the target product.

How the Algorithm Works

Logistic regression is a well-known statistical method for determining the contribution of multiple factors to a pair of outcomes. The Microsoft implementation uses a modified neural network to model the relationships between inputs and outputs. The effect of each input on the output is measured, and the various inputs are weighted in the finished model. The name logistic regression comes from the fact that the data curve is compressed by using a logistic transformation, to minimize the effect of extreme values.

Data Required for Logistic Regression Models

When you prepare data for use in training a logistic regression model, you should understand the requirements for the particular algorithm, including how much data is needed, and how the data is used.

The requirements for a logistic regression model are as follows:

A single key column Each model must contain one numeric or text column that uniquely identifies each record. Compound keys are not allowed.

Input columns Each model must contain at least one input column that contains the values that are used as factors in analysis. You can have as many input columns as you want, but depending on the number of values in each column, the addition of extra columns can increase the time it takes to train the model.

At least one predictable column The model must contain at least one predictable column of any data type, including continuous numeric data. The values of the predictable column can also be treated as inputs to the model, or you can specify that it be used for prediction only. Nested tables are not allowed for predictable columns, but can be used as inputs.

Viewing a Logistic Regression Model

To explore the model, you can use the Microsoft Neural Network Viewer, or the Microsoft Generic Content Tree Viewer.

When you view the model by using the Microsoft Neural Network Viewer, Analysis Services shows you the factors that contribute to a particular outcome, ranked by their importance. You can choose an attribute and values to compare.

If you want to know more, you can browse the model details by using the Microsoft Generic Content Tree Viewer. The model content for a logistic

regression model includes a marginal node that shows you the all the inputs used for the model, and subnetworks for the predictable attributes.

Creating Predictions

After the model has been trained, you can create queries against the model content to get the regression coefficients and other details, or you can use the model to make predictions.

- For general information about how to create queries against a data mining model.
- For examples of queries on a logistic regression model.

Explanation:

1. **Model Equation:**
    - In logistic regression, the relationship between the independent variables and the log-odds of the dependent variable being in a particular category is modeled using the logistic function (sigmoid function). The logistic function maps any real-valued number to a value between 0 and 1, making it suitable for modeling probabilities.
    - The logistic regression model equation for binary classification is:

$$P(Y=1|X) = 1 / (1 + \exp(-z))$$

Where:

- P(Y=1|X) is the probability of the dependent variable (Y) being in category 1 given the independent variables (X).
- exp() denotes the exponential function.
- z is the linear combination of the independent variables and their coefficients.

2. **Coefficients Interpretation:**
    - In logistic regression, coefficients represent the change in the log-odds of the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant.
    - The odds ratio (exponentiated coefficient) indicates the multiplicative change in the odds of the event (Y=1) for a one-unit change in the predictor variable.
3. **Model Training:**

- Logistic regression models are trained using maximum likelihood estimation, aiming to maximize the likelihood of observing the actual outcomes given the model predictions.
- Optimization algorithms such as gradient descent or Newton's method are commonly used to find the coefficients that maximize the likelihood.

4. **Prediction:**

- Once the model is trained, it can predict the probability of the binary outcome for new observations.
- A threshold (usually 0.5) is applied to the predicted probabilities to classify observations into the respective categories (e.g., class 1 if $P(Y=1|X) >= 0.5$, otherwise class 0).

5. **Evaluation:**

- Common evaluation metrics for logistic regression include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
- Confusion matrix and ROC curves are often used to assess model performance and determine the trade-off between true positive rate and false positive rate.

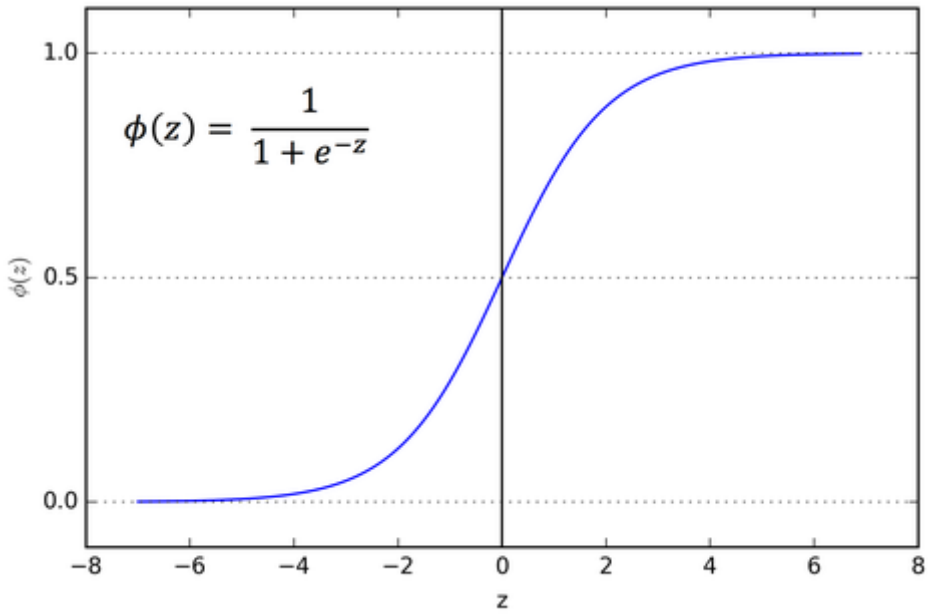
Sigmoid Function (Logistic Function):

The sigmoid function, also known as the logistic function, is a key component of logistic regression. It maps any real-valued number to a value between 0 and 1, making it suitable for modeling probabilities. The sigmoid function is defined as:

$\sigma(z) = 1 / (1 + \exp(-z))$

Where:

- **z** is the linear combination of the independent variables and their coefficients ($\beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_r x_r$).
- **exp()** denotes the exponential function.
- σ(z) represents the predicted probability of the dependent variable being in category 1 (Y=1) given the values of the independent variables.



The sigmoid function ensures that the predicted probabilities fall within the range [0, 1], allowing logistic regression to model binary outcomes effectively. The decision boundary (separating the two classes) is typically set at 0.5, with

probabilities greater than 0.5 classified as class 1 and probabilities less than or equal to 0.5 classified as class 0.

Assumptions of Logistic Regression:

1. **Binary Outcome**:
    - Logistic regression assumes that the dependent variable is binary (e.g., 0 or 1, Yes or No).
2. **Linearity of Log-Odds**:
    - The log-odds of the dependent variable being in a particular category should be a linear combination of the independent variables.
3. **Independence of Observations**:
    - Observations are assumed to be independent of each other, similar to linear regression.

Types of Logistic Regression:

1. **Binary Logistic Regression**:
    - The most common type, used when the dependent variable has two categories.
2. **Multinomial Logistic Regression**:
    - Used when the dependent variable has more than two unordered categories.
3. **Ordinal Logistic Regression**:
    - Suitable when the dependent variable has more than two ordered categories.

Limitations of Logistic Regression:

1. **Assumption of Linearity**:
    - Logistic regression assumes a linear relationship between the log-odds of the dependent variable and the independent variables. If the relationship is non-linear, logistic regression may provide poor predictions.
2. **Feature Engineering**:
    - Logistic regression relies on meaningful and relevant features for accurate predictions. In cases where feature engineering is challenging, the model's performance may suffer.
3. **Sensitivity to Outliers**:
    - Outliers in the dataset can significantly influence logistic regression coefficients and predictions, potentially leading to biased results.
4. **Assumption of Independence**:
    - Logistic regression assumes independence of observations. Violations of this assumption, such as in time series data or spatial data, can affect model performance.
5. **Limited to Linear Decision Boundary**:
    - Logistic regression can only capture linear decision boundaries between classes. For non-linear relationships, more complex models like decision trees or support vector machines may be more appropriate.