What is logistic regression?

Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with COVID-19 or not. Since we have two possible outcomes to this question - yes they are infected, or no they are not infected - this is called binary classification.

In this imaginary example, the probability of a person being infected with COVID-19 could be based on the viral load and the symptoms and the presence of antibodies, etc. Viral load, symptoms, and antibodies would be our factors (Independent Variables), which would influence our outcome (Dependent Variable).

How is logistic regression different from linear regression?

In linear regression, the outcome is continuous and can be any possible value. However, in the case of logistic regression, the predicted outcome is discrete and restricted to a limited number of values.

For example, say we are trying to apply machine learning to the sale of a house. If we are trying to predict the sale price based on the size, year built, and number of stories we would use linear regression, as linear regression can predict a sale price of any possible value. If we are using those same factors to predict if the house sells or not, we would logistic regression as the possible outcomes here are restricted to yes or no.

Hence, linear regression is an example of a regression model and logistic regression is an example of a classification model.

Where to use logistic regression

Logistic regression is used to solve classification problems, and the most common use case is binary logistic regression, where the outcome is binary (yes or no). In the real world, you can see logistic regression applied across multiple areas and fields.

• In health care, logistic regression can be used to predict if a tumor is likely to be benign or malignant.

• In the financial industry, logistic regression can be used to predict if a transaction is fraudulent or not.

• In marketing, logistic regression can be used to predict if a targeted audience will respond or not.

The three types of logistic regression

1. Binary logistic regression - When we have two possible outcomes, like our original example of whether a person is likely to be infected with COVID-19 or not.

2. Multinomial logistic regression - When we have multiple outcomes, say if we build out our original example to predict whether someone may have the flu, an allergy, a cold, or COVID-19.

3. Ordinal logistic regression - When the outcome is ordered, like if we build out our original example to also help determine the severity of a COVID-19 infection, sorting it into mild, moderate, and severe cases.

Mathematics behind logistic regression

Probability always ranges between 0 (does not happen) and 1 (happens). Using our Covid-19 example, in the case of binary classification, the probability of testing positive and not testing positive will sum up to 1. We use logistic function or sigmoid function to calculate probability in logistic regression. The logistic function is a simple S-shaped curve used to convert data into a value between 0 and 1.

$$h\Theta(x) = 1/1 + e - (\beta o + \beta 1X)$$

'hΘ(x)' *is output of logistic function , where 0 ≤ hΘ(x) ≥ 1*

*'**β1**' is the slope*
*'**βo**' is the y-intercept*
*'**X**' is the independent variable*

*(βo + β1*x) - derived from equation of a line Y(predicted) = (βo + β1*x) + Error value*

What is Classification?

Classification is defined as the process of recognition, understanding, and grouping of objects and ideas into preset categories a.k.a "sub-populations." With the help of these pre-categorized training datasets, classification in machine learning programs leverage a wide range of algorithms to classify future datasets into respective and relevant categories.

Classification algorithms used in machine learning utilize input training data for the purpose of predicting the likelihood or probability that the data that follows will fall into one of the predetermined categories. One of the most common applications of classification is for filtering emails into "spam" or "non-spam", as used by today's top email service providers.

In short, classification is a form of "pattern recognition,". Here, classification algorithms applied to the training data find the same pattern (similar number sequences, words or sentiments, and the like) in future data sets.

We will explore classification algorithms in detail, and discover how a text analysis software can perform actions like sentiment analysis - used for categorizing unstructured text by opinion polarity (positive, negative, neutral, and the like).

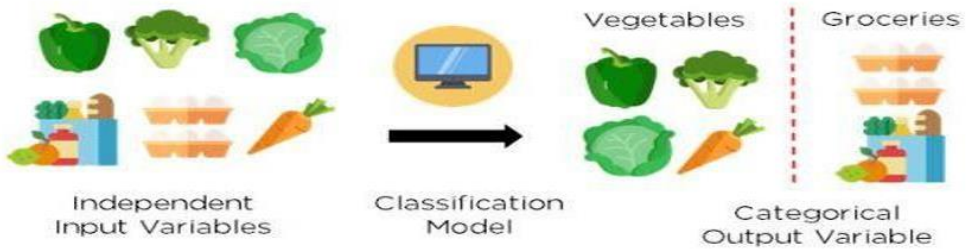

Figure 2: Classification of vegetables and groceries

What is Classification Algorithm?

Based on training data, the Classification algorithm is a Supervised Learning technique used to categorize new observations. In classification, a program uses the dataset or observations provided to learn how to categorize new observations into various classes or groups. For instance, 0 or 1, red or blue, yes or no, spam or not spam, etc. Targets, labels, or categories can all be used to describe classes. The Classification algorithm uses labeled input data because it is a supervised learning technique and comprises input and output information. A discrete output function (y) is transferred to an input variable in the classification process (x).

In simple words, classification is a type of pattern recognition in which classification algorithms are performed on training data to discover the same pattern in new data sets.

There are four different types of Classification Tasks in Machine Learning and they are following -

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

Binary Classification

Those classification jobs with only two class labels are referred to as binary classification.

Examples comprise -

- Prediction of conversion (buy or not).
- Churn forecast (churn or not).
- Detection of spam email (spam or not).

Binary classification problems often require two classes, one representing the normal state and the other representing the aberrant state.

Multi-Class Classification

- Multi-class labels are used in classification tasks referred to as multi-class classification.

Examples comprise –

- Categorization of faces.
- Classifying plant species.
- Character recognition using optical.

The multi-class classification does not have the idea of normal and abnormal outcomes, in contrast to binary classification. Instead, instances are grouped into one of several well-known classes.

Multi-Label Classification

- Multi-label classification problems are those that feature two or more class labels and allow for the prediction of one or more class labels for each example.

Imbalanced Classification

- The term "imbalanced classification" describes classification jobs where the distribution of examples within each class is not equal.
- A majority of the training dataset's instances belong to the normal class, while a minority belong to the abnormal class, making imbalanced classification tasks binary classification tasks in general.

Examples comprise –

• Clinical diagnostic procedures

• Detection of outliers

• Fraud investigation