

The main difference between them is that the output variable in regression is numerical (or continuous) while that for classification is categorical (or discrete).

**Classification Predictive Modeling**

Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to **discrete** output variables (y).

The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.

For example, an email of text can be classified as belonging to one of two classes: “spam*“* and “*not spam*“.

* A classification problem requires that examples be classified into one of two or more classes.
* A classification can have real-valued or discrete input variables.
* A problem with two classes is often called a two-class or binary classification problem.
* A problem with more than two classes is often called a multi-class classification problem.
* A problem where an example is assigned multiple classes is called a multi-label classification problem.

It is common for classification models to predict a continuous value as the probability of a given example belonging to each output class. The probabilities can be interpreted as the likelihood or confidence of a given example belonging to each class. A predicted probability can be converted into a class value by selecting the class label that has the highest probability.

**Regression Predictive Modeling**

Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a **continuous** output variable (y).

A continuous output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes.

For example, a house may be predicted to sell for a specific dollar value, perhaps in the range of $100,000 to $200,000.

* A regression problem requires the prediction of a quantity.
* A regression can have real valued or discrete input variables.
* A problem with multiple input variables is often called a multivariate regression problem.
* A regression problem where input variables are ordered by time is called a time series forecasting problem.

Because a regression predictive model predicts a quantity, the skill of the model must be reported as an error in those predictions.

**Classification vs Regression**

Classification predictive modeling problems are different from regression predictive modeling problems.

* Classification is the task of predicting a discrete class label.
* Regression is the task of predicting a continuous quantity.

There is some overlap between the algorithms for classification and regression; for example:

* A classification algorithm may predict a continuous value, but the continuous value is in the form of a probability for a class label.
* A regression algorithm may predict a discrete value, but the discrete value in the form of an integer quantity.

Some algorithms can be used for both classification and regression with small modifications, such as decision trees and artificial neural networks. Some algorithms cannot, or cannot easily be used for both problem types, such as linear regression for regression predictive modeling and logistic regression for classification predictive modeling.

Importantly, the way that we evaluate classification and regression predictions varies and does not overlap, for example:

* Classification predictions can be evaluated using accuracy, whereas regression predictions cannot.
* Regression predictions can be evaluated using root mean squared error, whereas classification predictions cannot.