

What is the difference between regression and classification?

Authored by
stats writer

April 21, 2024

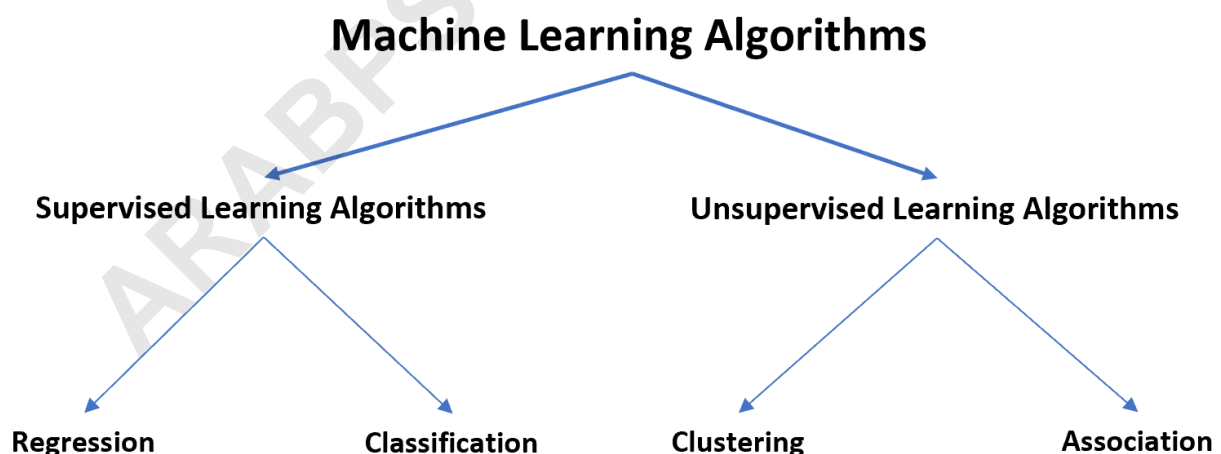
RECOMMENDED CITATION

stats writer (2024). *What is the difference between regression and classification?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=137705>

Regression and classification are two popular types of supervised machine learning algorithms used for predictive modeling. The main difference between these two techniques lies in their output variable. In regression, the output variable is continuous, meaning it can take on any numerical value within a given range. This makes regression suitable for predicting a quantity, such as the price of a house or the stock market performance. On the other hand, classification has a categorical or discrete output variable, meaning it can only take on a limited number of values. This makes classification suitable for predicting classes or categories, such as predicting whether a customer will churn or not. Additionally, the algorithms used for regression and classification differ, with regression using techniques such as linear regression, while classification uses algorithms such as decision trees or logistic regression. In summary, the main difference between regression and classification lies in the type of output variable they predict and the techniques used to achieve the prediction.

Regression vs. Classification: What's the Difference?

Machine learning algorithms can be broken down into two distinct types: supervised and unsupervised learning algorithms.



Supervised learning algorithms can be further classified into two types:

1. Regression: The response variable is continuous.

For example, the response variable could be:

Weight Height Price Time Total units

In each case, a regression model seeks to predict a continuous quantity.

Regression Example:

Suppose we have a dataset that contains three variables for 100 different houses: square footage, number of bathrooms, and selling price.

We could fit a regression model that uses square footage and number of bathrooms as explanatory variables and selling price as the response variable.

We could then use this model to predict the selling price of a house, based on its square footage and number of bathrooms.

This is an example of a regression model because the response variable (selling price) is continuous.

The most common way to measure the accuracy of a regression model is by calculating the root mean square error (RMSE), a metric that tells us how far apart our predicted values are from our observed values in a model, on average. It is calculated as:

$$\text{RMSE} = \sqrt{\sum(P_i - O_i)^2 / n}$$

where:

Σ is a fancy symbol that means "sum"
 P_i is the predicted value for the i th observation
 O_i is the observed value for the i th observation
 n is the sample size

The smaller the RMSE, the better a regression model is able to fit the data.

2. Classification: The response variable is categorical.

For example, the response variable could take on the following values:

Male or female
Pass or fail
Low, medium, or high

In each case, a classification model seeks to predict some class label.

Classification Example:

Suppose we have a dataset that contains three variables for 100 different college basketball players: average points per game, division level, and whether or not they got drafted into the NBA.

We could fit a classification model that uses average points per game and division level as explanatory variables and "drafted" as the response variable.

We could then use this model to predict whether or not a given player will get drafted into the NBA based on their average points per game and division level.

This is an example of a classification model because the response variable ("drafted") is categorical. That is, it can only take on values in two different categories: "Drafted" or "Not drafted."

The most common way to measure the accuracy of a classification model is by simply calculating the percentage of correct classifications the model makes:

Accuracy = correct classifications / total attempted classifications * 100%

For example, if a model correctly identifies whether or not a player will get drafted into the NBA 88 times out of 100 possible times then the accuracy of the model is:

Accuracy = (88/100) * 100% = 88%

The higher the accuracy, the better a classification model is able to predict outcomes.

Similarities Between Regression and Classification

Regression and classification algorithms are similar in the following ways:

Both are supervised learning algorithms, i.e. they both involve a response variable. Both use one or more explanatory variables to build models to predict some response. Both can be used to understand how changes in the values of explanatory variables affect the values of a response variable.

Differences Between Regression and Classification

Regression and classification algorithms are different in

the following ways:

Regression algorithms seek to predict a continuous quantity and classification algorithms seek to predict a class label. The way we measure the accuracy of regression and classification models differs.

Converting Regression into Classification

It's worth noting that a regression problem can be converted into a classification problem by simply discretizing the response variable into buckets.

For example, suppose we have a dataset that contains three variables: square footage, number of bathrooms, and selling price.

We could build a regression model using square footage and number of bathrooms to predict selling price.

However, we could discretize selling price into three different classes:

\$80k - \$160k: "Low selling price" \$161k - \$240k: "Medium selling price" \$241k - \$320k: "High selling

price"

We could then use square footage and number of bathrooms as explanatory variables to predict which class (low, medium or high) that a given house selling price will fall in.

This would be an example of a classification model since we're attempting to place each house in a class.

Summary

The following table summarizes the similarities and differences between regression and classification algorithms:

	Regression	Classification
Description	A regression model seeks to predict a continuous quantity.	A classification model seeks to predict some class label.
Type of algorithm	Supervised learning algorithm	Supervised learning algorithm
Type of response variable	Continuous	Categorical
How to assess model fit	Root mean squared error	Percentage of correct classifications