What is the concept of Multivariate Adaptive Regression Splines and how does it relate to regression analysis?

Authored by

stats writer

April 22, 2024

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression technique that combines the flexibility of polynomial regression with the adaptability of piecewise linear regression. It works by breaking the data into smaller subgroups and fitting separate linear regression models to each subgroup. These models are then combined to create a more accurate and flexible overall model. MARS is particularly useful for analyzing complex relationships between multiple variables and can handle both continuous and categorical data. It is often used in regression analysis to better understand the relationship between a dependent variable and multiple independent variables.

An Introduction to Multivariate Adaptive Regression Splines

When the relationship between a set of predictor variables and a response variable is linear, we can often use linear regression, which assumes that the relationship between a given predictor variable and a response variable takes the form:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

But in practice the relationship between the variables can actually be nonlinear and attempting to use linear regression can result in a poorly fit model.

One way to account for a nonlinear relationship between the predictor and response variable is to use polynomial regression, which takes the form:

$$Y = \beta 0 + \beta 1X + \beta 2X2 + ... + \beta hXh + \varepsilon$$

In this equation, *h* is referred to as the "degree" of the polynomial. As we increase the value for *h*, the model becomes more flexible and is able to fit nonlinear data.

However, polynomial regression has a couple drawbacks:

1. Polynomial regression can easily **overfit** a dataset if the degree, *h*, is chosen to be too large. In practice, *h* is rarely larger than 3 or 4 because beyond this point it simply fits the noise of a training set and does not generalize well to unseen data.

2. Polynomial regression imposes a global function on the entire dataset, which is not always accurate.

An alternative to polynomial regression is multivariate adaptive regression splines.

The Basic Idea

Multivariate adaptive regression splines work as follows:

1. Divide a dataset into *k* pieces.

First, we divide a dataset into *k* different pieces. The points where we divide the dataset are known as *knots*.

We identify the knots by assessing each point for each predictor as a potential knot and creating a linear regression model using the candidate features. The point that is able to reduce the most error in the model is deemed to be the knot.

Once we've identified the first knot, we then repeat the process to find additional knots. You can find as many knots as you think is reasonable to start.

2. Fit a regression function to each piece to form a hinge function.

For example, the hinge function for a model with one knot may be as follows:

$y = \beta 0 + \beta 1(4.3 - x)$ if $x < 4.3$ $y = \beta 0 + \beta 1(x - 4.3)$ if $x > 4.3$

In this case, it was determined that choosing 4.3 to be the cutpoint value was able to reduce the error the most out of all possible cutpoints values. We then fit a different regression model to the values less than 4.3 compared to values greater than 4.3.

A hinge function with two knots may be as follows:

$y = \beta 0 + \beta 1(4.3 - x)$ if $x < 4.3$ $y = \beta 0 + \beta 1(x - 4.3)$ if $x > 4.3$ & $x < 6.7$ $y = \beta 0 + \beta 1(6.7 - x)$ if $x > 6.7$

In this case, it was determined that choosing 4.3 and 6.7 as the cutpoint values was able to reduce the error the most out of all possible cutpoint values. We then fit one regression model to the values less than 4.3, another regression model to values between 4.3 and 6.7, and another regression model to the values greater than 4.3.

3. Choose *k* based on k-fold cross-validation.

Lastly, once we've fit several different models using a different number of knots for each model, we can perform k-fold cross-validation to identify the model that produces the lowest test mean squared error (MSE).

The model with the lowest test MSE is chosen to be the model that generalizes best to new data.

Pros & Cons

Multivariate adaptive regression splines come with the

following pros and cons:

Pros:

It can be used for both regression and classification problems.It works well on large datasets.It offers quick computation.It does not require you to standardize the predictor variables.

Cons:

It tends to not perform as well as non-linear methods like random forests and gradient boosting machines.

How to Fit MARS Models in R & Python

The following tutorials provide step-by-step examples of how to fit multivariate adaptive regression splines (MARS) in both R and Python:

Multivariate Adaptive Regression Splines in R
Multivariate Adaptive Regression Splines in Python