What is Best Subset Selection in Machine Learning and can you provide some examples?

Authored by

stats writer

April 22, 2024

Best Subset Selection is a method used in machine learning to select the most relevant features or variables for a given dataset. It involves evaluating all possible combinations of features and choosing the subset that produces the best performance for a given model. This method is commonly used to improve the accuracy and efficiency of predictive models. Examples of best subset selection can include choosing the most important genes for predicting a disease in bioinformatics or selecting the most relevant factors for predicting stock prices in finance.

Best Subset Selection in Machine Learning (Explanation & Examples)

In the field of machine learning, we're often interested in building models using a set of predictor variables and a response variable. Our goal is to build a model that can effectively use the predictor variables to predict the value of the response variable.

Given a set of *p* total predictor variables, there are many models that we could potentially build. One method that we can use to pick the *best* model is known as best subset selection and it works as follows:

1. Let M0 denote the null model, which contains no predictor variables.

2. For k = 1, 2, ... p:

Fit all pCk models that contain exactly *k* predictors.Pick the best among these pCk models and call it Mk. Define

"best" as the model with the highest R2 or equivalently the lowest RSS.

3. Select a single best model from among M0...Mp using cross-validation prediction error, Cp, BIC, AIC, or adjusted R2.

Note that for a set of *p* predictor variables, there are 2p possible models.

Example of Best Subset Selection

Suppose we have a dataset with p = 3 predictor variables and one response variable, y. To perform best subset selection with this dataset, we would fit the following 2p = 23 = 8 models:

A model with no predictorsA model with predictor x1A model with predictor x2A model with predictor x3A model with predictors x1, x2A model with predictors x1, x3A model with predictors x2, x3A model with predictors x1, x2, x3

Next, we'd choose the model with the highest R2 among each set of models with *k* predictors. For example, we might end up choosing:

A model with no predictorsA model with predictor x2A model with predictors x1, x2A model with predictors x1, x2, x3

Next, we'd perform cross-validation and choose the best model to be the one that results in the lowest prediction error, Cp, BIC, AIC, or adjusted R2.

For example, we might end up choosing the following model as the "best" model because it produced the lowest cross-validated prediction error:

A model with predictors x1, x2

Criteria for Choosing the "Best" Model

The last step of best subset selection involves choosing the model with the lowest prediction error, lowest Cp, lowest BIC, lowest AIC, or highest adjusted R2.

Cp: (RSS+2dσ?) / n

AIC: (RSS+2dσ?2) / (nσ?2)

BIC: (RSS+log(n)dσ?2) / n

Adjusted R2: 1 - ((RSS/(n-d-1)) / (TSS / (n-1)))

where:

d: The number of predictorsn: Total observationsσ?: Estimate of the variance of the error associate with each response measurement in a regression modelRSS: Residual sum of squares of the regression modelTSS: Total sum of squares of the regression model

Pros & Cons of Best Subset Selection

Best subset selection offers the following pros:

It's a straightforward approach to understand and interpret.It allows us to identify the best possible model since we consider all combinations of predictor variables.

However, this method comes with the following cons:

It can be computationally intense. For a set of *p* predictor variables, there are 2p possible models. For example, with 10 predictor variables there are 210 = 1,000 possible models to consider.Because it considers

such a large number of models, it could potentially find a model that performs well on training data but not on future data. This could result in**overfitting**.

Conclusion

While best subset selection is straightforward to implement and understand, it can be unfeasible if you're working with a dataset that has a large number of predictors and it could potentially lead to overfitting.

An alternative to this method is known as **stepwise selection**, which is more computationally efficient.