Best Subset Selection: A Comprehensive Guide to Feature Selection in Machine Learning

Authored by

Mohammed looti

November 6, 2025

In the expansive field of machine learning and statistical modeling, a common and critical task is determining the optimal set of predictor variables--also known as features--to build the most effective model. We are fundamentally concerned with accurately predicting a response variable based on available data. When faced with numerous potential predictors, choosing the right combination is essential for achieving high predictive accuracy while maintaining interpretability.

The challenge arises because incorporating irrelevant or redundant predictors can introduce noise, complicate the model, and potentially reduce its generalization ability. Conversely, omitting crucial predictors leads to a poorly specified model. Therefore, systematic feature selection is paramount. Given a total of *p* predictor variables, the number of possible models we could construct is vast, totaling 2p combinations. The technique known as **best subset selection** offers a rigorous, albeit computationally intensive, strategy for navigating this model space to identify the truly optimal subset of features.

The Process of Best Subset Selection

Best subset selection is a comprehensive model selection procedure designed to find the subset of predictors that results in the best-performing model for every possible subset size (k). This method systematically explores the entire model space, ensuring that no potential combination is overlooked. The procedure operates through three distinct phases, starting from the simplest model and expanding outward.

The core objective in the first phase is to fit every single possible model variation. For a system with *p* predictors, the process follows these steps:

Let M0 denote the **null model**. This foundational model contains zero predictor variables and serves as the baseline against which all other models are compared.

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

Fit all pCk models that contain exactly *k* predictors. This involves calculating every unique combination of *k* features from the total set *p*.

From this set of pCk models, select the single best model, denoted Mk. The definition of "best" at this stage typically involves metrics that measure fit quality, such as achieving the highest R-squared value or, equivalently, the lowest Residual Sum of Squares (RSS).

Finally, select the single overall best model from the collection M0, M1, ... Mp. This ultimate selection is made using predictive performance criteria, such as cross-validation prediction error, Mallows' Cp, the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), or the adjusted R-squared.

Illustrative Example of Subset Selection

To demonstrate the scope of this technique, consider a relatively small dataset where we have p = 3 predictor variables (x1, x2, x3) and one response variable (y). Applying **best subset selection** requires us to evaluate all 2p possible models, which in this case is 23 = 8 models.

The complete list of models fitted by this process is as follows:

A model with no predictors (M0)
A model with predictor x1 (k=1)
A model with predictor x2 (k=1)
A model with predictor x3 (k=1)
A model with predictors x1, x2 (k=2)
A model with predictors x1, x3 (k=2)
A model with predictors x2, x3 (k=2)
A model with predictors x1, x2, x3 (k=3)

Following the initial fitting phase, we proceed to select the best model for each subset size (k), typically based on the highest achieved R-squared value within that size constraint. For instance, the intermediate selection might yield the following four candidates:

M0: The null model (no predictors)
M1: The model containing only predictor x2
M2: The model containing predictors x1 and x2
M3: The model containing predictors x1, x2, and x3

The final step requires applying a criterion that balances model fit and complexity--such as cross-validation prediction error or one of the information criteria. The goal is to choose the simplest model among M0 through Mp that provides adequate predictive power. For example, if the model containing predictors x1 and x2 (M2) produced the lowest cross-validated prediction error, it would be designated as the overall best model.

Key Criteria for Optimal Model Selection

The final decision in the **best subset selection** procedure hinges on using sophisticated metrics that penalize model complexity. Simply choosing the model with the highest R-squared is inadequate, as R-squared always increases as more predictors are added, potentially leading to overfitting. The criteria below help to strike a necessary balance between minimizing the Residual Sum of Squares (RSS) and the number of predictors (d).

The following are the standard formulas used for model comparison, where a lower value generally

indicates a better model (except for adjusted R-squared, where a higher value is preferred):

Cp (Mallows' Cp): (RSS + 2dσ?) / n

AIC (Akaike Information Criterion): (RSS + 2dσ?2) / (nσ?2)

BIC (Bayesian Information Criterion): (RSS + log(n)dσ?2) / n

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

Understanding the components of these formulas is crucial for interpreting model selection results:

d: Represents the number of predictor variables included in the model, acting as the penalty for complexity.

n: Denotes the total number of observations or data points in the dataset.

σ**?:** An estimate of the variance of the error associated with each response variable measurement in the regression model.

RSS: The Residual Sum of Squares, measuring the discrepancy between the data and the estimation model.

TSS: The Total Sum of Squares of the regression model, representing the total variation in the response variable.

Advantages and Limitations of Best Subset Selection

While **best subset selection** is theoretically ideal for guaranteeing the discovery of the absolute optimal model, its practical application is often constrained by computational resources and the risk of data contamination.

This method offers several key **advantages**:

It is conceptually straightforward to understand and the resulting model structure is highly interpretable, as irrelevant features are explicitly excluded.

It provides the most exhaustive approach to model selection, ensuring that since every combination of predictor variables is considered, the mathematically best possible subset is identified according to the chosen performance metric.

However, the method suffers from significant **disadvantages**:

The computational intensity is immense. Since 2p models must be fitted, the time required grows

exponentially with the number of predictors (p). For example, a dataset with only p=20 predictors requires fitting over a million (220) distinct models, rendering the method infeasible for datasets common in modern machine learning.

Due to the sheer volume of models examined, there is an elevated risk of finding a model that performs exceptionally well on the training data purely by chance, a phenomenon known as overfitting. This results in a model with poor generalization ability on unseen, future data.

Conclusion and Alternatives

Best subset selection represents the gold standard for model selection when the number of predictor variables is small (typically $p < 15$). It guarantees the identification of the truly optimal feature subset by exhaustively searching the entire model space.

Nevertheless, the exponential increase in computational burden and the heightened risk of overfitting make this technique impractical for high-dimensional data. When working with a large number of predictors, researchers and practitioners often turn to more computationally efficient alternatives, such as stepwise selection (including forward or backward stepwise regression), or regularization methods like Lasso, which provide a practical balance between model complexity and predictive power.