Feature Selection Methods

Comprehensive Study Guide for Machine Learning

Filter Methods

Statistical techniques that evaluate features independently

Variance Threshold

- **Purpose:** Eliminate low-variance features
- **Best for:** Datasets with constant/near-constant features
- **Key insight:** Low variation = minimal predictive value

Correlation Coefficient

- **Purpose:** Identify redundant features
- **Best for:** Removing highly correlated features
- **Limitation:** Only captures linear relationships

Chi-Square Test

- **Purpose:** Measure categorical variable association
- **Best for:** Categorical features with categorical targets
- **Application:** Feature-target dependency strength

Mutual Information

- **Purpose:** Capture linear and non-linear dependencies
- **Best for:** Complex relationships
- **Advantage:** More comprehensive than correlation

ANOVA

- **Purpose:** Compare means across groups
- **Best for:** Categorical features with continuous targets
- **Use case:** Factor impact through group comparison

Wrapper Methods

Model-based approaches that evaluate feature subsets

Recursive Feature Elimination (RFE)

- **Process:** Train model → Remove least important → Repeat
- **Best for:** Leveraging model performance to guide selection
- **Advantage:** Uses actual model feedback

Sequential Feature Selection

- **Process:** Add/remove one feature at a time
- **Best for:** Finding optimal subset incrementally
- **Types:** Forward selection or backward elimination

Exhaustive Selection

- **Process:** Tests all possible combinations
- **Best for:** Small feature sets only
- **Warning:** Computationally expensive

Embedded Methods

Built-in feature selection within algorithms

Lasso Regression

Ridge Regression

- **Mechanism:** L1 penalty shrinks coefficients to zero
- **Best for:** Simple, interpretable linear models
- **Advantage:** Automatic feature selection

- **Mechanism:** L2 penalty shrinks coefficients
- **Best for:** Handling multicollinearity
- **Note:** Doesn't eliminate features

Elastic Net

- **Mechanism:** Combines L1 and L2 penalties
- **Best for:** Multiple correlated features
- **Advantage:** More stable than Lasso

Random Forest Importance

- **Mechanism:** Mean decrease in impurity (MDI)
- **Best for:** Non-linear relationships
- **Application:** Ensemble method insights

Method Selection Guidelines

Choose Filter Methods When:

- Large number of features need preprocessing
- Want computationally efficient approach
- Need basic feature relationship understanding

Choose Wrapper Methods When:

- Model performance is primary concern
- Have moderate number of features
- Computational resources available

Choose Embedded Methods When:

- Want integrated approach
- Working with specific algorithms
- Need performance-efficiency balance

Key Considerations

Computational Complexity:

Filter: Fastest (O(n)) **Embedded:** Moderate **Wrapper:** Slowest

Model Dependency:

Filter: Model-agnostic **Embedded:** Algorithm-specific **Wrapper:** Uses target model

Interaction Detection:

Filter: Limited (univariate) **Embedded:** Good (tree-based) **Wrapper:** Excellent

Detailed Version - Original Reference

Complete definitions and explanations from original cheatsheet

1. Filter Methods:

- **Variance Threshold:** Removes all features whose variance doesn't meet a certain threshold. Use this when you have many features and you want to remove those that are constants or near constants.
- **Correlation Coefficient:** Finds the correlation between each pair of features. Highly correlated features can be removed since they contain similar information. Use this when you suspect that some features are highly correlated.
- **Chi-Square Test:** This statistical test is used to determine if there's a significant association between two variables. It's commonly used for categorical variables. Use this when you have categorical features and you want to find their dependency with the target

variable.

- **Mutual Information:** Measures the dependency between two variables. It's a more general form of the correlation coefficient and can capture non-linear dependencies. Use this when you want to measure both linear and non-linear dependencies between features and the target variable.
- **ANOVA (Analysis of Variance):** ANOVA is a statistical test that stands for "Analysis of Variance". ANOVA tests the impact of one or more factors by comparing the means of different samples. Use this when you have one or more categorical independent variables and a continuous dependent variable.

2. Wrapper Methods:

- **Recursive Feature Elimination (RFE):** Recursively removes features, builds a model using the remaining attributes, and calculates model accuracy. It uses model accuracy to identify which attributes contribute the most. Use this when you want to leverage the model to identify the best features.
- **Sequential Feature Selection (SFS):** Adds or removes one feature at the time based on the classifier performance until a feature subset of the desired size k is reached. Use this when computational cost is not an issue and you want to find the optimal feature subset.
- **Exhaustive Feature Selection:** This is a brute-force evaluation of each feature subset. This method, as the name suggests, tries out all possible combinations of variables and returns the best subset. Use this when the number of features is small, as it can be computationally expensive.

3. Embedded Methods:

- **Lasso Regression:** Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization. Use this when you want to create a simple and interpretable model.
- **Ridge Regression:** Ridge regression is a method used to analyze multiple regression data that suffer from multicollinearity. Unlike Lasso, it doesn't lead to feature selection but rather minimizes the complexity of the model.
- **Elastic Net:** This method is a combination of Lasso and Ridge. It incorporates penalties from both methods and is particularly useful when there are multiple correlated features.
- **Random Forest Importance:** Random forests provide a straightforward method for feature selection, namely mean decrease impurity (MDI). Use this when you want to leverage the power of random forests for feature selection.