Cheatsheet

15 May 2023 17:38

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.