Feature Selection

Feature Selection is the process of identifying and choosing the most relevant features from the dataset to use when training and building a machine learning model.

The main goal of feature selection is to streamline the dataset by removing irrelevant or redundant features, leaving only those that contribute the most significantly.

Feature selection is needed because:

- Improved model performance
- Reduced overfitting
- Enhanced Interpretability
- lower Computational cost
- Simplified Data visualization
- Curse of Dimensionality

Types of feature Selection Techniques:

i. Filter based technique

ii. Wrapper Method

iii. Embedded Method

iv. Hybrid Technique.

i) Filter based Technique

- Filter based feature selection techniques are methods that use statistical measures to score each feature independently, and then select a subset of features based on these scores. These methods are called "filter" methods because they essentially filter out the features that do not meet some criterion.

Statistical techniques used in filter based selection:

- Variance Threshold
- Correlation
- ANOVA
- Chi-Square
- Mutual information

First step is always to remove duplicate features.

i) Variance Threshold

A variance selection method that removes all features from a dataset whose variance does not meet specified threshold, usually to eliminate that carry little information because their values rarely change.

There are two methods.

- Constant feature: A feature where all values are the same across all samples (variance = 0). These features are uninformative and are removed by default with variance thresholding.
- Quasi-Constant feature:

  A feature where almost all values are the same, with very little variation (var close to 0). These features are also generally uninformative and can be removed by setting a small variance threshold.

Points to Consider

1. **Ignores Target Variable:** Variance Threshold is a univariate method, meaning it evaluates each feature independently and doesn't consider the relationship between each feature and the target variable. This means it may keep irrelevant features that have a high variance but no relationship with the target, or discard potentially useful features that have a low variance but a strong relationship with the target.

2. **Ignores Feature Interactions:** Variance Threshold doesn't account for interactions between features. A feature with a low variance may become very informative when combined with another feature.

3. **Sensitive to Data Scaling:** Variance Threshold is sensitive to the scale of the data. If features are not on the same scale, the variance will naturally be higher for features with larger values. Therefore, it is important to standardize the features before applying Variance Threshold.

4. **Arbitrary Threshold Value:** It's up to the user to define what constitutes a "low" variance. The threshold is not always easy to define and the optimal value can vary between datasets.

ii) Correlation

Pearson Correlation Method

- Definition:
  The Pearson correlation coefficient (r) measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 (perfect negative) to +1 (perfect positive), with 0 meaning no linear relationship.
  - Interpretation:
  - $r > 0$: Positive correlation (both variables increase together).
  - $r < 0$: Negative correlation (one variable increases, the other decreases).
  - $r = 0$: No linear correlation

Disadvantages

1. **Linearity Assumption:** Correlation measures the linear relationship between two variables. It does not capture non-linear relationships well. If a relationship is nonlinear, the correlation coefficient can be misleading.

2. **Doesn't Capture Complex Relationships:** Correlation only measures the relationship between two variables at a time. It may not capture complex relationships involving more than two variables.

3. **Threshold Determination:** Just like variance threshold, defining what level of correlation is considered "high" can be subjective and may vary depending on the specific problem or dataset.

4. **Sensitive to Outliers:** Correlation is sensitive to outliers. A few extreme values can significantly skew the correlation coefficient.

iii) ANOVA

Already done before ....

Disadvantages

1. **Assumption of Normality:** ANOVA assumes that the data for each group follow a normal distribution. This assumption may not hold true for all datasets, especially those with skewed distributions.
2. **Assumption of Homogeneity of Variance:** ANOVA assumes that the variances of the different groups are equal. This is the assumption of homogeneity of variance (also known as homoscedasticity). If this assumption is violated, it may lead to incorrect results.
3. **Independence of Observations:** ANOVA assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).
4. **Effect of Outliers:** ANOVA is sensitive to outliers. A single outlier can significantly affect the F-statistic leading to a potentially erroneous conclusion.
5. **Doesn't Account for Interactions:** Just like other univariate feature selection methods, ANOVA does not consider interactions between features.

iv) Chi-Square

Disadvantages

1. **Categorical Data Only:** The chi-square test can only be used with categorical variables. It is not suitable for continuous variables unless they have been discretized into categories, which can lead to loss of information.
2. **Independence of Observations:** The chi-square test assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).
3. **Sufficient Sample Size:** Chi-square test requires a sufficiently large sample size. The results may not be reliable if the sample size is too small or if the frequency count in any category is too low (typically less than 5).
4. **No Variable Interactions:** Chi-square test, like other univariate feature selection methods, does not consider interactions between features. It might miss out on identifying important features that are significant in combination with other features.

Advantage and Disadvantage of Filter based Selection

Advantages

1. **Simplicity:** Filter methods are generally straightforward and easy to understand. They involve calculating a statistic that measures the relevance of each feature, and selecting the top features based on this statistic.

2. **Speed:** These methods are usually computationally efficient. Because they evaluate each feature independently, they can be much faster than wrapper methods or embedded methods, which need to train a model to evaluate feature importance.

3. **Scalability:** Filter methods can handle a large number of features effectively because they don't involve any learning methods. This makes them suitable for high-dimensional datasets.

4. **Pre-processing Step:** They can serve as a pre-processing step for other feature selection methods. For instance, you could use a filter method to remove irrelevant features before applying a more computationally expensive method, such as a wrapper method.

Disadvantages

1. **Lack of Feature Interaction:** Filter methods treat each feature individually and hence do not consider the interactions between features. They might miss out on identifying important features that don't appear significant individually but are significant in combination with other features.

2. **Model Agnostic:** Filter methods are agnostic to the machine learning model that will be used for the prediction. This means that the selected features might not necessarily contribute to the accuracy of the specific model you want to use.

3. **Statistical Measures Limitation:** The statistical measures used in these methods have their own limitations. For example, correlation is a measure of linear relationship and might not capture non-linear relationships effectively. Similarly, variance-based methods might keep features with high variance but low predictive power.

4. **Threshold Determination:** For some methods, determining the threshold to select features can be a bit subjective. For example, what constitutes "low" variance or "high" correlation might differ depending on the context or the specific dataset.

V. Mutual Information

Mutual Information (MI) is a measure of the dependency between two variables. It qualifies the amount of information obtained about one random variable through observing the other random variable. It is fundamental quantity in information theory.

$$MI = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \left[\frac{P(x,y)}{P(x)\,P(y)} \right]$$

where,

$P(x,y) \longrightarrow$ Joint Probability of x and y

$P(x) \longrightarrow$ Marginal Probability of x

$P(y) \longrightarrow$ Marginal Probability of y

1. Joint Probability: This is the probability of two (or more) simultaneous events. For example, if we have two random variables, X and Y, the joint probability of X and Y is denoted as P(X, Y), and it represents the probability that X takes on a specific value and Y takes on a specific value at the same time. In other words, it represents the probability of both events happening at the same time.

2. Marginal Probability: This is the probability of an event occurring regardless of the outcome of another event. If we have two random variables, X and Y, the marginal probability of X is simply denoted as P(X), and it represents the probability that X takes on a specific value irrespective of the values of Y. The term "marginal" refers to the process of summing or integrating over the distribution of the other variable(s) to obtain the distribution of the variable of interest.

Mutual Information has several properties that make it useful for feature selection:

1. It is non-negative: MI is always zero or positive, with zero indicating that the variables are independent (i.e., no information about one variable can be obtained by observing the other variable).

2. It is symmetric: MI(X, Y) = MI(Y, X). The mutual information from X to Y is the same as from Y to X.

3. It can capture any kind of statistical dependency: Unlike correlation, which only captures linear relationships, mutual information can capture any kind of relationship, including nonlinear ones.

Wrapper Method

Wrapper methods for feature selection are a type of feature selection methods that involve using a predictive model to score the combination of features. They are called "wrapper" methods because they "wrap" this type of model-based evaluation around the feature selection process.

Here's how wrapper methods work in general:

1. **Subset Generation**: First, a subset of features is generated. This can be done in a variety of ways. For example, you might start with one feature and gradually add more, or start with all features and gradually remove them, or generate subsets of features randomly. The subset generation method depends on the specific type of wrapper method being used.
2. **Subset Evaluation**: After a subset of features has been generated, a model is trained on this subset of features, and the model's performance is evaluated, usually through cross-validation. The performance of the model gives an estimate of the quality of the features in the subset.
3. **Stopping Criterion**: This process is repeated, generating and evaluating different subsets of features, until some stopping criterion is met. This could be a certain number of subsets evaluated, a certain amount of time elapsed, or no improvement in model performance after a certain number of iterations.

Types of Wrapper Selection Method

1. Exhaustive feature Selection
2. forward Selection
3. Backward Elimination
4. Recursive feature Elimination