Feature Selection Techniques in Machine Learning

Feature selection:

Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.

Feature selection is a critical step in the feature construction process. In text categorization problems, some words simply do not appear very often. Perhaps the word “groovy” appears in exactly one training document, which is positive. Is it really worth keeping this word around as a feature ? It’s a dangerous endeavor because it’s hard to tell with just one training example if it is really correlated with the positive class or is it just noise. You could hope that your learning algorithm is smart enough to figure it out. Or you could just remove it.

There are three general classes of feature selection algorithms: **Filter methods, wrapper methods and embedded methods**.

The role of feature selection in machine learning is,

1. To reduce the dimensionality of feature space.

2. To speed up a learning algorithm.

3. To improve the predictive accuracy of a classification algorithm.

4. To improve the comprehensibility of the learning results.

Features Selection Algorithms are as follows:

1. **Instance based approaches:** There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.

2. Nondeterministic approaches: Genetic algorithms and simulated annealing are also used in feature selection.

3. Exhaustive complete approaches: Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.

While building a machine learning model for real-life dataset, we come across a lot of features in the dataset and not all these features are important every time. Adding unnecessary features while training the model leads us to reduce the overall accuracy of the model, increase the complexity of the model and decrease the generalization capability of the model and makes the model biased. Even the saying "Sometimes less is better" goes as well for the machine learning model. Hence, **feature selection** is one of the important steps while building a machine learning model. Its goal is to find the best possible set of features for building a machine learning model.

Some popular techniques of feature selection in machine learning are:

- Filter methods
- Wrapper methods
- Embedded methods

Filter Methods

These methods are generally used while doing the pre-processing step. These methods select features from the dataset irrespective of the use of any machine learning algorithm. In terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity. Selection of feature is evaluated individually which can sometimes help when features are in isolation (don't have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model.

Set of all features → Selecting the best subset → Learning algorithm → Performance

Filter Methods Implementation

Some techniques used are:

- **Information Gain –** It is defined as the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Information gain of each attribute is calculated considering the target values for feature selection.
- **Chi-square test —** Chi-square method (X2) is generally used to test the relationship between categorical variables. It compares the observed values from different attributes of the dataset to its expected value.

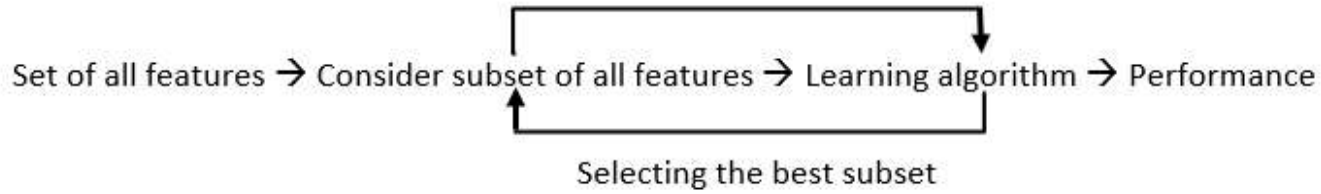
$$X^2 = \sum \frac{(Observed\ value - Expected\ value)^2}{Expected\ value}$$

Chi-square Formula

- **Fisher's Score –** Fisher's Score selects each feature independently according to their scores under Fisher criterion leading to a suboptimal set of features. The larger the Fisher's score is, the better is the selected feature.
- **Correlation Coefficient –** Pearson's Correlation Coefficient is a measure of quantifying the association between the two continuous variables and the direction of the relationship with its values ranging from *-1 to 1*.
- **Variance Threshold –** It is an approach where all features are removed whose variance doesn't meet the specific threshold. By default, this method removes features having zero variance. The assumption made using this method is higher variance features are likely to contain more information.
- **Mean Absolute Difference (MAD) –** This method is similar to variance threshold method but the difference is there is no square in MAD. This method calculates the mean absolute difference from the mean value.
- **Dispersion Ratio –** Dispersion ratio is defined as the ratio of the Arithmetic mean (AM) to that of Geometric mean (GM) for a given feature. Its value ranges from $+1$ *to* ∞ *as* $AM \geq GM$ for a given feature. Higher dispersion ratio implies a more relevant feature.
- **Mutual Dependence –** This method measures if two variables are mutually dependent, and thus provides the amount of information obtained for one variable on observing the other variable. Depending on the presence/absence of a feature, it measures the amount of information that feature contributes to making the target prediction.
- **Relief –** This method measures the quality of attributes by randomly sampling an instance from the dataset and updating each feature and distinguishing between instances that are near to each other based on the difference between the selected instance and two nearest instances of same and opposite classes.

Wrapper methods:

Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner. Based on the conclusions made from training in prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved. The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive.



Wrapper Methods Implementation

Some techniques used are:

- **Forward selection –** This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
- **Backward elimination –** This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.
- **Bi-directional elimination –** This method uses both forward selection and backward elimination technique simultaneously to reach one unique solution.
- **Exhaustive selection –** This technique is considered as the brute force approach for the evaluation of feature subsets. It creates all possible subsets and builds a learning algorithm for each subset and selects the subset whose model's performance is best.
- **Recursive elimination –** This greedy optimization method selects features by recursively considering the smaller and smaller set of features. The estimator is trained on an initial set of features and their importance is obtained using feature_importance_attribute. The least important features are then removed from the current set of features till we are left with the required number of features.

Embedded methods:

In embedded methods, the feature selection algorithm is blended as part of the learning algorithm, thus having its own built-in feature selection methods. Embedded methods encounter the drawbacks of filter and wrapper methods and merge their advantages. These methods are faster like those of filter methods and more accurate than the filter methods and take into consideration a combination of features as well.

Embedded Methods Implementation

Some techniques used are:

- **Regularization –** This method adds a penalty to different parameters of the machine learning model to avoid over-fitting of the model. This approach of feature selection uses Lasso (L1 regularization) and Elastic nets (L1 and L2 regularization). The penalty is applied over the coefficients, thus bringing down some coefficients to zero. The features having zero coefficient can be removed from the dataset.
- **Tree-based methods –** These methods such as Random Forest, Gradient Boosting provides us feature importance as a way to select features as well. Feature importance tells us which features are more important in making an impact on the target feature.

Conclusion:

Apart from the methods discussed above, there are many other methods of feature selection. Using hybrid methods for feature selection can offer a selection of best advantages from other methods, leading to reduce in the disadvantages of the algorithms. These models can provide greater accuracy and performance when compared to other methods. Dimensionality reduction techniques such as Principal Component Analysis (PCA), Heuristic Search Algorithms, etc. don't work in the way as to feature selection techniques but can help us to reduce the number of features.

Feature selection is a wide, complicated field and a lot of studies has already been made to figure out the best methods. It depends on the machine learning engineer to combine and innovate approaches, test them and then see what works best for the given problem.