

FEATURE SELECTION

Machine learning works on a simple rule - if you put garbage in, you will only get garbage to come out. By garbage here, I mean noise in data. This becomes even more important when the number of features are very large. You need not use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important. I have myself witnessed feature subsets giving better results than complete set of feature for the same algorithm.

Top reasons to use feature selection are:

It enables the machine learning algorithm to train faster.

It improves the accuracy of a model

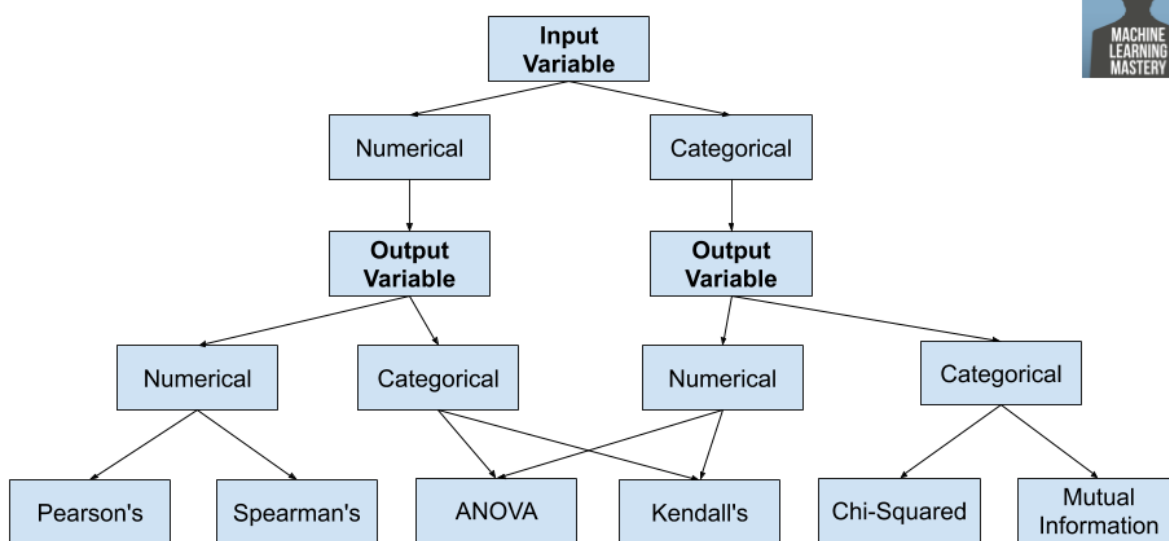
It reduces overfitting

FILTER METHODS

		Dependent Variable	
		Categorical	Continuous
Independent Variable	Categorical	Chi-squared test	ANOVA
	Continuous	Logistic Regression	Linear Regression

Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

How to Choose a Feature Selection Method



Copyright © MachineLearningMastery.com

Numerical Input, Numerical Output

Pearson's Correlation - Pearson's Correlation - Its value varies from -1 to +1. A value of 0 means no correlation. The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation

Pearson's correlation coefficient = $\text{covariance}(X, Y) / (\text{stdv}(X) * \text{stdv}(Y))$

where Covariance is one of the statistical measurement to know the relationship of the variance between the two variables. $\text{Cov}(x, y) = \text{SUM} [(x_i - x_m) * (y_i - y_m)] / (n - 1)$

Spearman's Rank Correlation - Two variables may be related by a nonlinear relationship, such that the relationship is stronger or weaker across the distribution of the variables. Further, the two variables being considered may have a non-Gaussian distribution.

In this case, the Spearman's correlation coefficient (named for Charles Spearman) can be used to summarize the strength between the two data samples. This test of relationship can

also be used if there is a linear relationship between the variables, but will have slightly less power (e.g. may result in lower coefficient scores).

Instead of calculating the coefficient using covariance and standard deviations on the samples themselves, these statistics are calculated from the relative rank of values on each sample. This is a common approach used in non-parametric statistics, e.g. statistical methods where we do not assume a distribution of the data such as Gaussian.

Spearman's correlation coefficient = $\text{covariance}(\text{rank}(X), \text{rank}(Y)) / (\text{stdv}(\text{rank}(X)) * \text{stdv}(\text{rank}(Y)))$

If you are unsure of the distribution and possible relationships between two variables, Spearman correlation coefficient is a good tool to use.

Categorical Input, Categorical Output

Pearson's Chi-Square - The Pearson's Chi-Square statistical hypothesis is a test for independence between categorical variables. The Chi-Squared test is a statistical hypothesis test that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable. The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same. The test returns test statistic value (chi-square value), p value, degree of freedom, table of expected frequencies.

After we have got test statistics value, now we need to find the critical value of chi-square using degree of freedom and significance factor from the critical values of the chi square distribution table. The formula for degree of freedom is $(\text{no. of rows} - 1) * (\text{no. of columns} - 1)$.

Critical values of the Chi-square distribution with d degrees of freedom							
Probability of exceeding the critical value							
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

The test can be interpreted as follows

Null Hypothesis: The features are independent (which means they are not associated).

Alternate Hypothesis: The features are not independent (which means they are associated)

In terms of statistics and critical values,

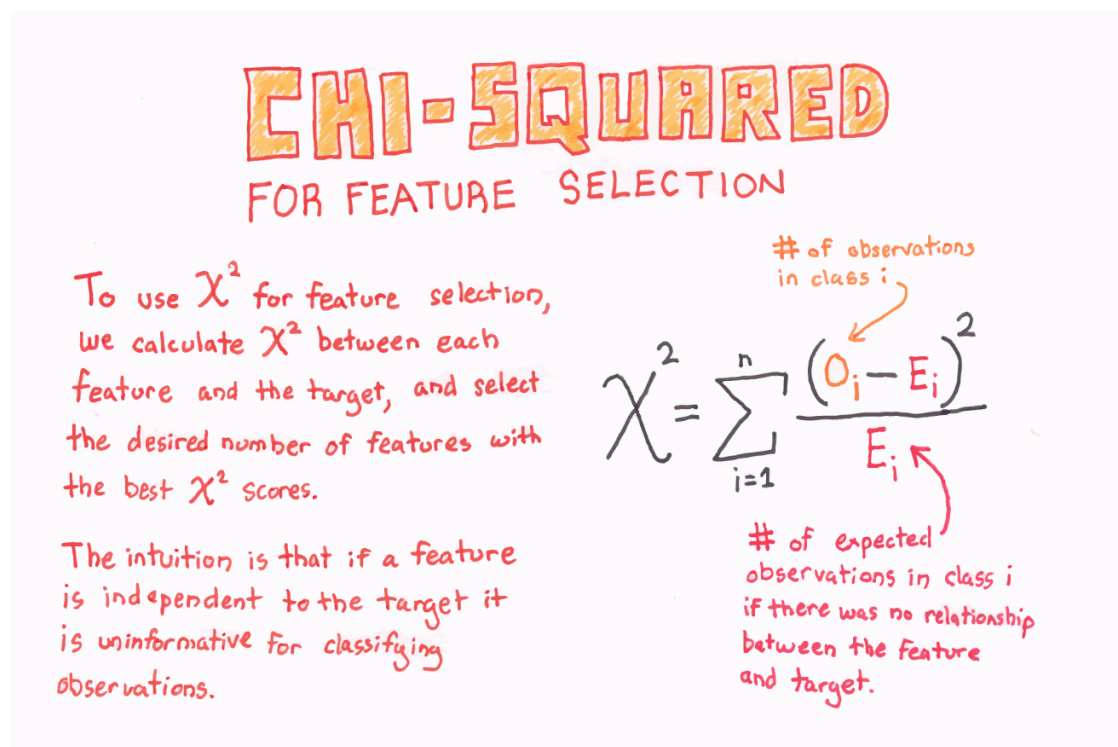
If test statistic value \geq critical value of chi square, then reject null hypothesis (H_0)

If test statistic value $<$ critical value of chi square then failed to reject null hypothesis (H_0)

In terms of a p value and chose significance level,

If p value \leq alpha then reject null hypothesis (H_0)

If p value $>$ alpha, then failed to reject null hypothesis (H_0)



CHI-SQUARED
FOR FEATURE SELECTION

To use χ^2 for feature selection, we calculate χ^2 between each feature and the target, and select the desired number of features with the best χ^2 scores.

The intuition is that if a feature is independent to the target it is uninformative for classifying observations.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

of observations in class i (pointing to O_i)

of expected observations in class i if there was no relationship between the feature and target. (pointing to E_i)

Numerical Input, Categorical Output

ANOVA - ANOVA stands for Analysis of variance. It provides a statistical test of whether the means of several groups are equal or not. An Analysis of Variance Test, or ANOVA, can be thought of as a generalization of the t-tests for more than 2 groups. The independent t-test is used to compare the means of a condition between two groups. ANOVA is used when we want to compare the means of a condition between more than two groups.

An F-statistic, or F-test, is a class of statistical tests that calculate the ratio between variances values, such as the variance from two different samples or the explained and unexplained

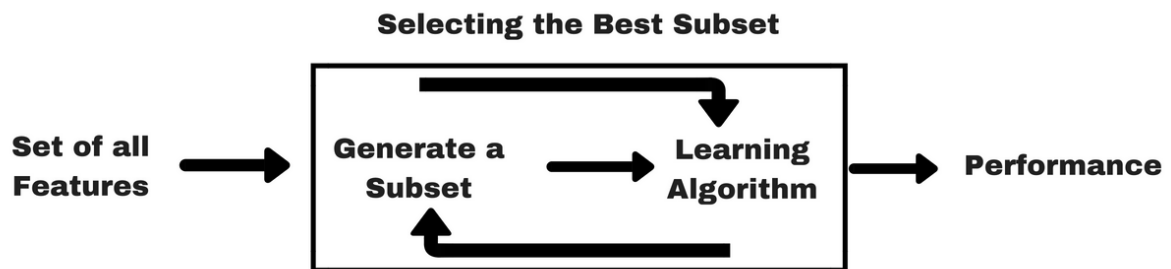
variance by a statistical test, like ANOVA. The ANOVA method is a type of F-statistic referred to here as an ANOVA f-test.

For checking of correlation between continuous input and categorical output we can use logistics regression as well.

Categorical Input, Numerical Output

ANOVA

WRAPPER METHODS



Recursive Feature elimination: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination. RFE is a wrapper-type feature selection algorithm. This means that a machine learning algorithm is given and a model is using the given machine learning algorithm, ranking features by importance, discarding the least important features and re-fitting the model and finally choosing best features. This process is repeated until a specified number of features remains

EMBEDDED METHODS

Embedded methods use algorithms that have built-in feature selection methods. For example: Decision Trees and RF have their own feature selection methods.