A close up of a sign

Description generated with high confidence

# Feature Selection Algorithm

|  |  |
| --- | --- |
| Version | 1.0 |
| Document Title | One Page Report of Understanding of the Problem |
| Group Members | Tanuj Gyan, Sayanti Ray, Raviprakash Shreeraksha, Babak Shabani |
| Date of Submission | 26-March-2018 |
| Instructor Name | Dr. Johnson P Thomas |
| TA Name | Vasanthi Mudunuri |

## Objective:

Extract the feature subset from the given dataset using wrapper method (Brute force, forward and backward method) to calculate the accuracy using the machine learning algorithms (KNN, Logistic Regression, Decision Trees and Support Vector Machines) and to get the most accurate subset.

## Description:

Supervised Learning is a technique where we have two variables – X (input) and Y(output) and the machine is trained to derive a function f which when applied to X will give Y

f(X) = Y. It is implicit that we will have multiple inputs i.e. more than one X and one output Y

We are provided with 4 files - test\_data, test\_labels, train\_data, train\_labels. Train\_data and Train\_labels are our training data and will be used to train our model based on the machine learning algorithms as mentioned above (in Objective). Test\_data will be fed to this model to get labels for this test\_data which will be compared with test\_labels file to estimate the accuracy using the formula.

*𝑎𝑐𝑐𝑢𝑟𝑎𝑐𝑦=number of correct predictions / total number of predictions*

There are two parts of the problem. The first one will calculate the accuracies of all subsets from test\_data. We have 760 features in the dataset provided. All 760 features will be fed to the machine learning algorithm to calculate accuracy. The dataset will be reduced using Brute force wrapper algorithm and the new subset will be fed to machine learning algorithm again to calculate the accuracy. The process will continue for all subsets.

In the second part of the problem the subset accuracies calculated in part 1 will be compared with accuracy obtained from 760 features to check if the accuracy is within threshold (empirically set to 0.03 for our project). If the difference is more than 0.03 the subset will be discarded. For all the subsets within the range the one with least accuracy will be considered. If the accuracy of two subsets will be same the one with smallest number of features will be considered. Also the time taken to run the wrapper algorithm with machine learning algorithm is considered.For one subset , one machine learning algorithm may take less time and for another subset some other machine learning algorithm may run fast. So it is not possible to decide by considering one subset and the machine learning algorithm that took less time to run.

Here we are using a set of supervised machine learning algorithms as suggested for classification and analysis of the dataset.

* K-Nearest Neighbor: This is a simple machine learning algorithm used for classification and regression purposes where the value of K denotes the number of nearest neighbors to be considered. Depending on the value of K the classification results are expected to vary. Here we are asked to consider K = 1, 3, 5, 7, 9.
* Logistic regression : As the name suggests Logistic regression is a regression method used for predictive analysis of dataset. We used this method to develop an idea of the dataset having one or more independent variables.
* Decision trees : Decision trees are graph based model used for classification and regression purposes which uses graph like root to leaf traversal to find out the appropriate strategy to analyze the dataset.
* Support Vector Machines : Another set of algorithms used for classification and regression which uses the concept of boundaries to classify the dataset and is used for high dimensional spaces.

The objective function of Support vector Machines is sum of Regularization function and loss function.In Regularization function the trained system builds its own model and in the loss function accuracy is determined .Here loss is to be minimized.