Problem Statement –

Thousands of songs have to be classified into their respective high-level genres by research and analytics division of audio streaming and media services provider, 'MUZI', for seamless song discovery and a delightful user experience. Your task is to utilize the features of the songs to determine their respective genres out of 7 possible categories.

Data Set –

The data has been split into two groups:

1. training set (train.csv)
2. test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each entry in the dataset.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each data entry. It is your job to predict these outcomes.

Algorithm Used-

AdaBoost Classification Algorithm with Random Forest as the base classifier. The problem statement involves classification as we need to predict a label (genres). So we use a classifier instead of a regressor.

Random Forest Classifier –

Random forest is a supervised learning algorithm used for classification and regression. Random forest consists of large number of individual decision trees that operate as an ensemble. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. The concept behind random forest is wisdom of crowds which means ‘A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.’ The reason behind this is that the trees protect each other from their individual errors While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.

The algorithm works as follows-

- Random samples are selected from the training set.
- Decision tree for each sample is constructed to get a prediction result from each tree.
- A vote for each result is performed.
- The result with most votes is considered the final prediction.

The algorithm has following advantages –

- Highly accurate and robust method because of the number of decision trees participating in the process.
- Does not suffer from the overfitting problem.
- Can also handle missing values.

AdaBoost Algorithm –

AdaBoost algorithm is a boosting algorithm to increase the accuracy of the base classifier. It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations.

The algorithm works as follows-

- Select a training subset randomly.
- Iteratively train the AdaBoost machine learning model by selecting the training set based on the accurate prediction of the last training.
- Assign the higher weight to wrong classified observations so that in the next iteration these observations will get the high probability for classification.
- Iterate until the complete training data fits without any error or until reached to the specified maximum number of estimators.

Implementation-

- Import all the required python libraries like pandas, numpy, sklearn, etc.
- Read the training and testing datasets using pandas read_csv function.
- Define a list named 'parameters' containing the attributes used for classifying the data into genres.
- The large number of NaN values in the training and testing sets was a hindrance for the implementation of the classifier.
- Deleting the records containing the NaN values hampered the accuracy of the classifier.
- So I replaced the NaN values in a particular column with the median value of that attribute.
- Then four variables named x1, y1, x2, y2 are initialized. x1 and x2 contain the columns containing the attributes from the training and test set respectively. y1 contains the target column, that is, genre of the training set.
- Then the random forest classifier is called with the n_estimators parameter as 1100. N_estimators is the number of decision tress to be constructed. Selecting the value of n_estimators was very crucial as a lower value would result in underfitting and higher value would be overfitting of data. So I used trial and error to find the optimum value of n_estimators.
- To boost the classifier, Adaboost classifier is called on the random forest classifier.
- Then the classifier is fitted on x1 and y1 to create a model.

- The model is then used to predict the records present in x2.
- A csv file is generated for the predicted values.

Result –

First I used basic Decision Tree which gave me an accuracy of around 0.3. Then Decision Tree combined with Adaboost gave an accuracy of 0.36. For the given dataset, random forest classifier gives an accuracy of 0.36. To boost the accuracy, Adaboost classifier is used and the final accuracy is 0.42.

-Yash Rajendra Bhanushali