Statistical Machine Learning

Kaggle Project

To develop model to classify data

Introduction

We have to develop a machine learning model which trains on sample data given in a csv file with some features and categories (which is called training) and finally predicts the category for new data stored in the csv file (which is called testing on new data for getting accuracy of the model trained).

Tools Used : -

1 > Numpy Module

2 > Pandas Module

3 > AccuracyScore from Sklearn.metrics module

4 > LogisticRegression from sklearn.linear_model module

5 > CrossValueScore from sklearn.model_selection module

6 > Kmeans from sklearn.cluster module

7 > LocalOutlierFactor from sklearn.neighbors module

Brief Description -

Firstly, I have converted the csv file into a dataframe using the pandas package and then made a list of categories from that dataframe and dropped that category column from the data frame as it is not going to help in the training model.

Secondly, I have done outlier detection(n_neigbors=10 and contamination =0.1) (Here, I have made a list of indices using numpy module which are outlier and it is removed from both dataframe and category list which is stored from dataframe as written above) so that model will be more precise and not have more variance in sample of data on which it is going to be feed/trained and such that outlier doesn't affect its accuracy as they are already very much alike from other sample of data. Then,Ii have used k means clustering with parameter of k=2 which is giving best silhouette score so that within class variance get reduced and we get similar data sample already in approximate together with each other so that while applying the next process that is algorithm or classifier, it get easier for them to feed on the sample data which are already clustered and have as much as possible same category(Also, the cluster's label is added as a column like an extra feature to dataframe using numpy and pandas module). Finally as written above i have applied the classifier which is in my case Logistic Regression (with max_iter=100000) and lastly i have checked the accuracy of my data using cross value score with parameter =5 and giving me 5 possible accuracy of my model and average of it.I also checked the accuracy of my model using Accuracy score function from module I have imported. In this way, my training of model got completed then, I again use pandas to convert test csv file into dataframe and then apply k mean clustering on it with same parameters and finally predicting the category for the test data and saved it in a list. At last, I again converted the test csv file into dataframe using pandas and removed all columns in it other than order id and added the column with category as header in it which is the predicted category list already saved as list as i wrote above. Finally, I have converted this dataframe using pandas again to a csv file and this way my work to predict the category of data based on training data got finished.

Reference : - Some websites along with lecture slides to read extra theory about algorithms (from lectures) to know more about them and increase the accuracy of my model.