UE19CS322 Big Data Project 2

Machine Learning with Spark MLlib

SPAM HAM DETECTION

Aryan Raj Chhetri CSE
PES1UG19CS093
aryanchhetri123@gmail.com

Kshitiz Kumar CSE
PES1UG19CS235
kshitizkumar80@gmail.com

Shreyas M R CSE
PES1UG19CS467
shreyasmr38@gmail.com

Introduction

We have chosen Machine Learning with Spark MLib as our Big Data Project with Spam Ham as our data set.We have used many python libraries like pyspark , numpy , sparkdl and scikit-learn to train and test on our data set using different regression models. It simulates real world situation where large amount of data is handled to train different ML predictive models.Streams of data has been split into small batches to work with.We have experimented on our predictive models with different batch sizes to compare result from each.

Design Details and Surface Level Implementation

We streamed data from data streaming file batches of sizes 1500, 2000 and 2500 and created database for each of the batches before preprocessing the data (merging , tokenisation and vectorisation).

After this we ran incremental machine learning models(partial fit) on them to train our models(svm,log reg,mnb, k means).

We streamed the test data and made predictions using metrics like accuracy score, f1 score ,precision, recall and confusion matrix.

Reason

Logistic regression has beenused as it is a binary classifier.

Naive bayes has been used widely for spam ham classification and is appropriate for finding maximum posterior probability.

SVM has been used because it works relatively well when there is a clear margin of separation between classes and is more effective and efficient in high dimensional spaces
Kmeans is unsupervised learning method so it helps us finding pattern we might not have observed normally.Its also an eager learner and space efficient.

Takeaway from project:

Using these we can predict with high certainty whether a message is going to be spam or ham and this can save the clients a lot of time in real world