Big Data Project Report

Praveen Obli Venkatesh, PES1UG19CS350
Raghav V Pandit, PES1UG19CS364
Rishab Kashyap BS, PES1UG19CS385
Sai Amruth Balusu, PES1UG19CS417

SSML – Spark Streaming for Machine Learning

I. Design Details

For the dataset labelled 'spam' (the goal was to classify the given news as 'spam' or 'ham'), using ML classification models from the sklearn library of python, we implemented 3 classification models –

- Gaussian Model
- Stochastic Gradient Descent Model
- Perceptron Classification Model

to determine to the best accuracy if the given news article was indeed 'spam' or 'ham'.
Initially, the dataset used was broken down into resilient distributed datasets (RDDs) and each RDD was subjected to each ML model used and was incrementally trained to provide us with the best accuracy possible for the given dataset. We also implemented minibatch k-means algorithm as a clustering method for our dataset.
We implemented the entire system on **pyspark** which is used for interactively analysing our data in a distributed environment.

II. Surface Level Implementation Details

- The dataset was broken down into multiple RDDs

- Each json RDD was converted into a pyspark data frame for analysis usage.
- Pre-processing: removed stop words and punctuations from the news articles for accurate modelling.
- Each batch of the dataset was trained with all the 3 classification models.
- Minibatch k-means algorithm was used as the clustering method for the predictions.
- Confusion matrix, precision and accuracy were used as performance metrics for the predictions.

III. Reasoning behind Design Decisions

- We used minibatch k-means algorithm as it is a hard-clustering algorithm and with binary classification, this is preferred.
- The Gaussian model, SGD classifier model and the Perceptron classification model were chosen as the ML models for the given dataset as these models were very well-suited to the type of resultant feature outputs from our dataset.

IV. Takeaway from the Project

- The usage of pyspark in order to analyse our distributed environment data is incredibly useful as it allows us to write spark applications using python but also supports data streaming and several data frame features.
- The ML models used in this project for the classification of the dataset resulted in a terrific accuracy with a future promise of the improvement possibilities that we can implement on these ML models to further increase our accuracy of predictions.