SSML---Spark-Streaming-for-Machine-Learning

Team No.: BD2_015_019_097_309

Design Details:

1. Bernoulli Naïve Bayes
   Implements the Naïve Bayes training and classification algorithms for data that is distributed according to multi-variate Bernoulli Distribution.
2. Multinomial Naïve Bayes
   Implements the Naïve Bayes algorithm for multinomially distributed data and is one of the two classic Naïve Bayes variant used in text classification.
3. Stochastic Gradient Descent Classifier
   SGD is an optimization algorithm often used in ML applications to find the model parameters that correspond to the best fit between the predicted and actual output.

Surface Level Implementation Detail:

1. **Streaming**
   We started off by streaming the train csv to localhost, from where we read the stream data using a socket created by Spark context. The stream data is received in the form of RDDs (Resilient Distributed Datasets) on which we perform pre-processing batch wise.
2. **Pre-processing**
   The steps involved in pre-processing are:

1. Replacing URL pattern with URL
2. Replacing usernames with user
3. Replacing characters except digits and alphabets
4. Replacing 3 or more consecutive letters by 2 letters
5. Removing extra spaces
6. Replacing quotes with space
7. Converting to lower case
8. Removing stop words
9. Tokenizing text (Splitting a sentence into words)
10. Stemming text (Reduces similar text to a common form)

3. **Model building and clustering**

   We built 3 models: Binomial Naïve Bayes, Multinomial Naïve Bayes, Stochastic Gradient Descent Classifier for classification.

   For the purpose of clustering, we make use of mini batch k-means.

4. **Incremental learning**

   We performed incremental learning on the 3 models by training them on each batch as it arrives. We made use of sklearn’s partial_fit() for this.

5. **Storing and loading model**

   We stored the models in .sav files after training them by using pickle.dump. For predicting on the test dataset, we load the models using pickle.load.

6. **Testing the model**

   We streamed the test data and preprocessed the tweets, then we used the three models to predict the sentiments of these tweets.

Reason Behind Design Decision:

We went with Bernoulli Naïve Bayes as it is extremely fast when compared with other models. It can handle irrelevant features well. It produces results which are self explanatory and also makes real-time predictions.

Multinomial Naïve Bayes is easy to implement as we only need to calculate probability. You can use this algorithm on both continuous and discrete data.

SGD Classifier is efficient and easy to implement. It also converges faster for larger datasets as it causes updates to the parameters more frequently.

We used mini-batch K-means as it is the best clustering algorithm that can be used for incremental learning.

Takeaway from the Project:

Through this project, we were able to learn, study and understand multiple models and also see the difference between their operation and how to choose suitable models given a problem statement. We learnt how to work with streaming data using Spark and train models incrementally.