BIG DATA ANALYTICS

Please check
https://github.com/TenseGor1lla/SVM-MapReduce.git
for the code.

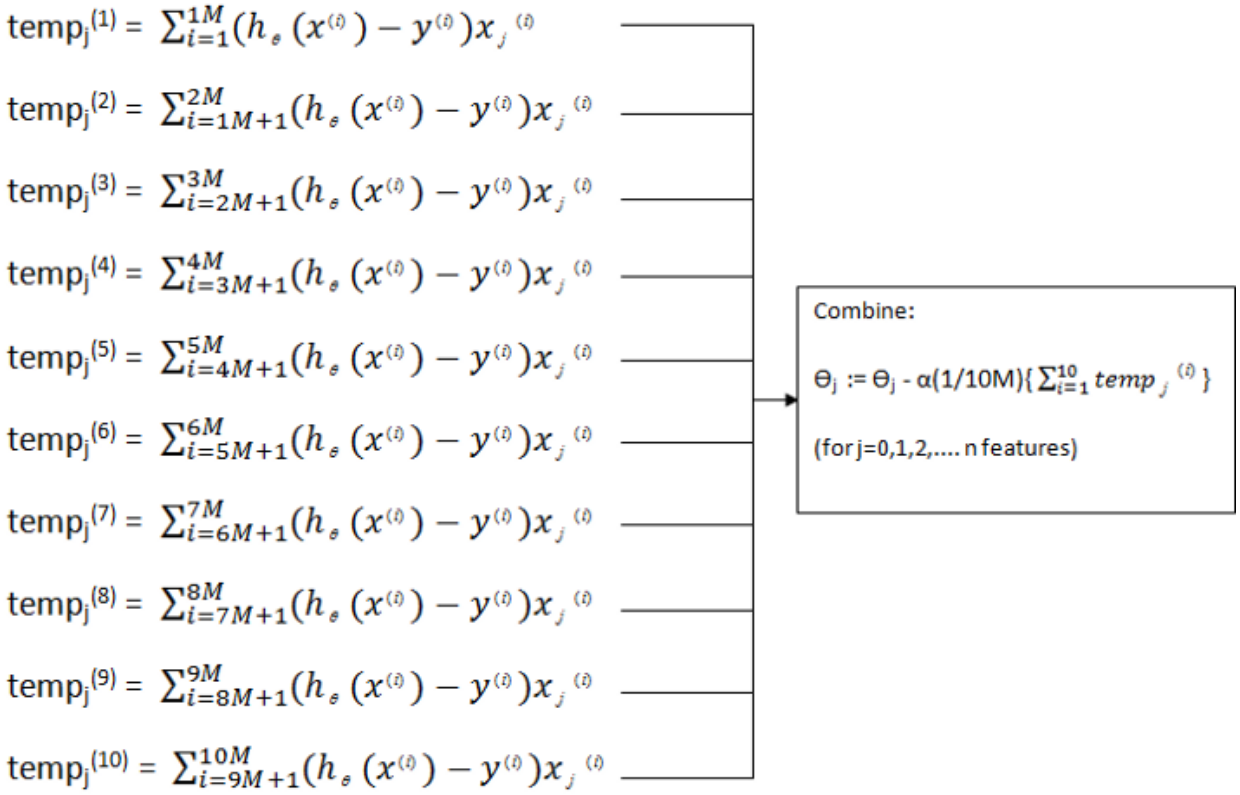
INNOVATIVE ASSIGNMENT

SVM USING MAPPER AND REDUCER

Some Machine Learning problems are just too big to run on one machine, sometimes maybe you just have such a large amount of data, (for instance you have 100 million training examples) that you would not want to run all that through a single computer, irrespective of what algorithm you are using. To combat this problem, a different approach to large scale Machine Learning known as the **"Map-Reduce"** approach, was developed by Jeffrey Dean and Sanjay Ghemawat. With the idea of Map-Reduce we would be able to scale learning algorithms to large machine learning problems, much larger than is possible with Batch Gradient Descent or Stochastic Gradient Descent.

In Map-Reduce we split the training set into a convenient number of subsets. Each of these subsets serve as an input for n different machines. Each of these machines will now run the Batch Gradient

Descent learning rule for their respective subset of data.



ALGORITHM

The **main algorithm** can basically be broken down into 4 steps:

1. Basic setup and initialization of the weights and biases
2. Map the class labels from {0, 1} to {-1, 1}
3. Perform gradient descent for *n* iterations, which involves the computation of the

gradients and updating the weights and biases accordingly.

4. Make the final prediction

Objective function

$$\mathcal{J}(w,b) = \lambda \frac{1}{2}||w||^2 + \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(w \cdot x_i + b))$$

Mapper: Computing the gradients

```
Mapper<Object, Text, Text, TextArrayWritable>
```

Reducer: Aggregating the gradients calculated by Combiner

FILES

weights.txt
bias.txt
SVM.java
Predictions.java

Dataset

- **Randomly generated 2D points with label as 0 or 1**

weights.txt:

- **by default set to zero**
- **weights gets updated as we train the SVM**

SVM.java:

- **To train the model**

Predictions.java:

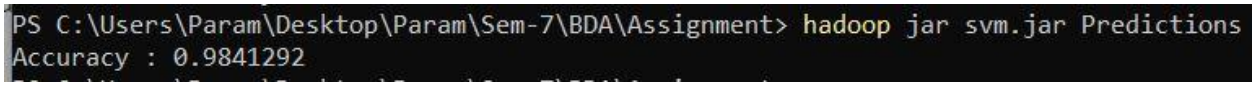
- **To predict the accuracy of the testing data**

```
PS C:\Users\Param\Desktop\Param\Sem-7\BDA\Assignment> hadoop jar svm.jar Predictions
Accuracy : 0.4996712
```

Accuracy of testing data before training

```
PS C:\Users\Param\Desktop\Param\Sem-7\BDA\Assignment> hadoop jar svm.jar SVM
2022-11-23 10:59:08,505 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2022-11-23 10:59:09,188 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2022-11-23 10:59:09,202 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
Param/.staging/job_1669176864178_0006
2022-11-23 10:59:09,320 INFO input.FileInputFormat: Total input files to process : 1
2022-11-23 10:59:09,386 INFO mapreduce.JobSubmitter: number of splits:7
2022-11-23 10:59:09,492 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669176864178_0006
2022-11-23 10:59:09,494 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-23 10:59:09,642 INFO conf.Configuration: resource-types.xml not found
2022-11-23 10:59:09,643 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-23 10:59:09,694 INFO impl.YarnClientImpl: Submitted application application_1669176864178_0006
2022-11-23 10:59:09,727 INFO mapreduce.Job: The url to track the job: http://DESKTOP-H59A7L1:8088/proxy/application_1669
176864178_0006/
2022-11-23 10:59:09,728 INFO mapreduce.Job: Running job: job_1669176864178_0006
2022-11-23 10:59:17,866 INFO mapreduce.Job: Job job_1669176864178_0006 running in uber mode : false
2022-11-23 10:59:17,867 INFO mapreduce.Job:  map 0% reduce 0%
2022-11-23 10:59:39,501 INFO mapreduce.Job:  map 17% reduce 0%
2022-11-23 10:59:40,523 INFO mapreduce.Job:  map 25% reduce 0%
2022-11-23 10:59:46,765 INFO mapreduce.Job:  map 43% reduce 0%
2022-11-23 10:59:52,851 INFO mapreduce.Job:  map 53% reduce 0%
2022-11-23 10:59:55,904 INFO mapreduce.Job:  map 58% reduce 0%
2022-11-23 10:59:56,916 INFO mapreduce.Job:  map 63% reduce 0%
2022-11-23 10:59:57,934 INFO mapreduce.Job:  map 86% reduce 0%
```

Training model for 3 epochs



Accuracy of testing data after training

By

- 19BCE254 (Neel Shah)
- 19BCE255 (Param Shah)
- 19BCE301 (Harshwardhan Yadav)