DS222 - Assignment 2

Nashez Zubair [*] [1]

1. Model

1.1. Preprocessing

The Word Hashing technique was used to convert the textual data to numbers suitable for predictions. The stopwords were also removed and stemming was also performed on the data.

1.2. Classification Function

The actual program is an implementation of an L2 regularized softmax function for multiclass classification. It has a weight matrix (parameters) of size $|vocabulary| * |classcount|$ and uses Stochastic Gradient Descent for updates.

2. Implementation

2.1. Framework selection

TensorFlow was used for implementation of both the local and distributed versions. This choice was made over frameworks like Hadoop because it would unnecessarily go to the disk for writing in between iterations and Spark because for that I would have to go through Glint (**?**) which was not installed on Turing, would not go well for asynchronous implementation and also has much less support from the community over TensorFlow due to it being a relatively newer parameter server implementation.

2.2. Programming Language

Python was used as compared to my first preference of Java as I painfully discovered for the last assignment that Java is not as suitable for such ML tasks as compared to Python.

2.3. Local Mode

Even the local implementation is in the matrix format as we needed to parallelize this for the other settings. There are 3 settings for this implementation:

- Constant learning rate at 0.01
- Increasing learning rate as: $rate = rate + (0.005)/epochs$
- Decreasing learning rate as: $rate = rate - (0.005)/epochs$

The accuracy for these are shown below in 1. The accuracy value for large full dataset are reasonable but the small dataset overfits terribly.

Table 1. Observations

DataSet	Constant	Decrement	Increment
Full	59.61	54.99	63.26
Small	97.5	76.8	98.8

3. Help Sought

- Consulted Machine Learning by KL Murphy (**?**) for equations for multi class Logistic Regression classification
- Aakash clarified some doubts
- Shivansh helped out with TensorFlow related problems

[*]Equal contribution [1]M.Tech, CDS. Correspondence to: <nashezzubair@iisc.ac.in>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).