Advanced Programming - Mini Project 2

Michael Erik Vesterli - miev@itu.dk
Emil Christian Lynegaard - ecly@itu.dk

May 14, 2018

1 Introduction

The goal of this project is to utilize a combination of the glove word-vector dataset (Pennington et al., 2014), and a data set consisting of amazon product reviews (He and McAuley, 2016), to train a classifier to predict the accompanying star-rating of a amazon product review based on the review comment. Specifically, we will use the Perceptron algorithm(Wikipedia, 2018), to train a classifier on a training data set, which will then be evaluated on a corresponding test data set.

2 Testing

For testing we have used a set of 10,261 reviews related to Musical Instruments from He and McAuley (2016). Due to time restriction, we only managed to run the classifier using a limited set of parameters. Our implementation and results use a neural network layout consisting of 2 hidden layers with 4 and 5 nodes respectively. Furthermore we have only tested it using 100 iterations. Despite the small size of the neural network, our tests still yield very promising results as discussed in section 3.

3 Results

We have tested the perceptron classifier using a basic 0.9/0.1 training/testing data distribution and using 10-fold cross validation on a similar distribution. As shown in table 1, the two varying testing methods yield very similar results, which is somewhat expected, as the 10-fold cross validation uses the best model

Run	**cpu time**	**clock time**	**f1-basic**	**f1-10fold-cv**
1	3811.58s	652.41s	0.8241	0.8212
2	3741.26s	587.82s	0.8381	0.8115
3	3444.61s	523.59s	0.8115	0.8262

Table 1: Result of running the program 3 times on a 4-physical, 8-logical core machine.

out of the 10 trained ones. Interestingly, this means that the best model returning by the cross validator, is capable of yielding similar results to a model trained on 9 times the data it was trained on.

By inspecting the average performance of all the models trained by Spark's Cross Validator, we see that the average performance of all the models, was within 1% of the best model, indicating that 10% of the total data is generally sufficient to achieve the results reported in table 1.

From the results presented in table 1, it is clear that using Spark's machine learning libraries offer a very high degree of parallelization, as we see CPU times around 5-6 times as large as clock times.

4 Conclusion

Based on the results found in this report, the Perceptron algorithm in combination has shown to be a powerful combination, in managing to efficiently train a neural network to yield very promising results on somewhat limited amounts of data.

References

R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *WWW, 2016*, 2016.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Wikipedia. Perceptron — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Perceptron&oldid=838780414`, 2018. [Online; accessed 14-May-2018].