CS4248 Assignment 2

Yin Yumeng
A0105408N

1 Introduction

In this assignment, I use a Java Program to perform ontext- sensitive spelling correction.
I implemented a logistic regression classifier for context-sensitive spelling correction, I specially use stochastic gradient ascent learning algorithm.

2 Implementing details

2.1 Actual features used for training and testing

- Surrounding words: All word types contained in the training file with stop words removed.
- Collocations: The collocations words used for the program are all two word phrases before and after the target word.

2.2 Parameters for logistic regression classifier:

- learningRate: 0.1
- maxIterations: 20000
- threshold:0.002

2.3 Details when implementing logistic regression with gradient ascent:

- For each iterations ran in program, program will go through all training samples. Here stochastic gradient ascent is used that only one sample is used to update weights array. To train each sample of data, function to update weight value at a specific place is listed as following.

```
for (int i=0; i<x.size(); j++) {
        weights[i] = weights[i] + learningRate * (label - predicted) *
x.get(i);
}
```

Hera formula used is listed as following:

$$w_{i(k+1)} \leftarrow w_{i(k)} + \alpha \cdot x_i^j \cdot (y^j - \frac{1}{1 + e^{-w_k \cdot x^j}})$$

And the $||\nabla l(w)||$ is computed everytime after all trainings are complete for an iterations.
The iterations will stop only when either the maxIterations is met or when

$||\nabla l(w)||$ is less than a predefined threshold.

- Predicted label is computed by using formula.

$$P(y = 1|x) = \frac{1}{1 + e^{-w \cdot x}}$$

$$w \cdot x = w_0 x_0 + w_1 x_1 + \cdots + w_{n-1} x_{n-1}$$

3 Evaluation of training and testing

3.1 Evaluation of how learning rate affect the learning model $||\nabla l(w)||$

Learning curve plot using the training data.
Learning rate here is the parameter that need tuning.
Here the threshold is not used

- maxIterations: 20000

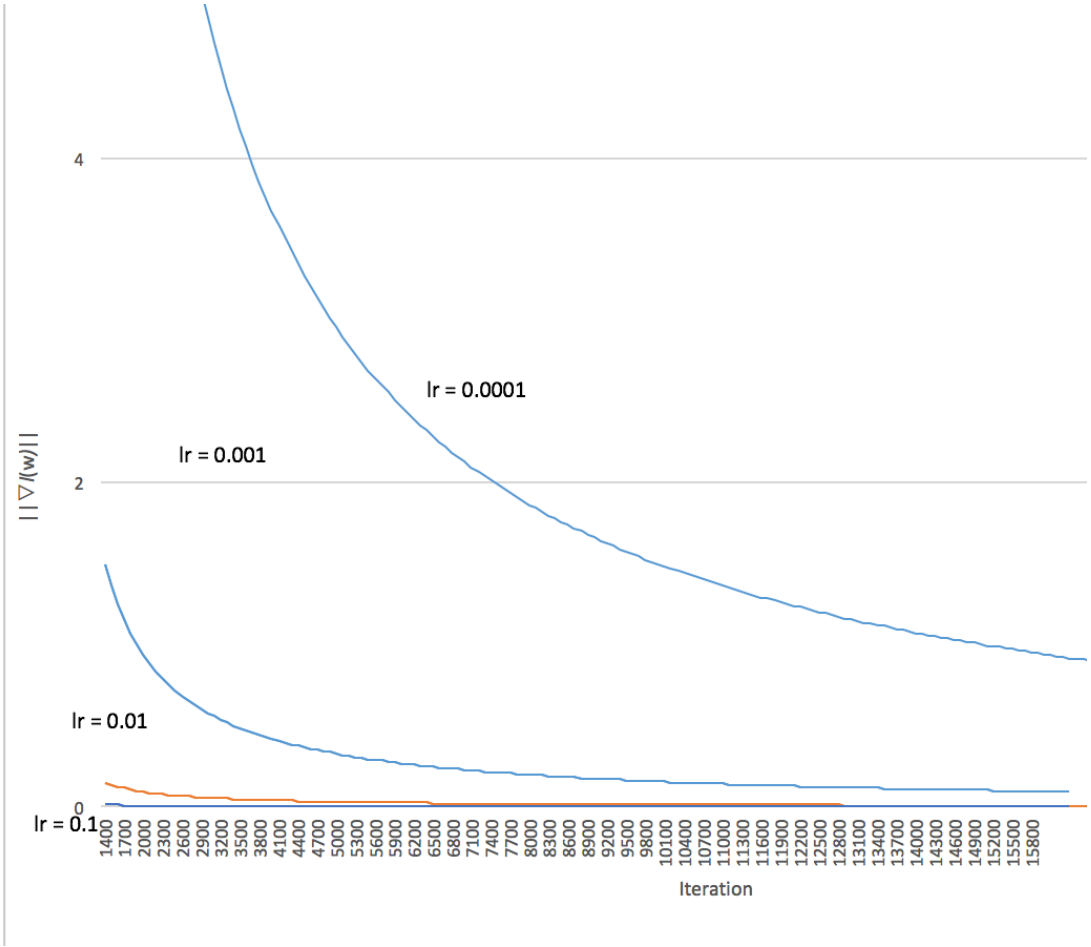

chart1

The line chart that plot the changing of $||\nabla l(w)||$ corresponding to the iteration numbers (start plot from the 1400 iteration and end at 15800 iteration) is shown above.

Learning rate	Training time(min)	Iterations run	||▽l(w)||(final)
0.0001	24	20000	0.757
0.001	24.3	20000	0.076
0.01	24.65	20000	0.0076
0.1	24	20000	0.00076

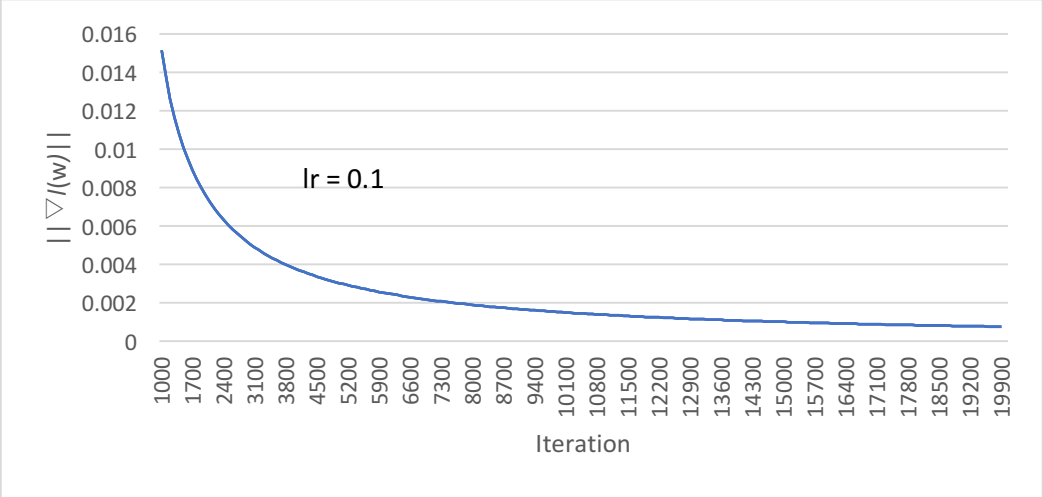

chart2

From the chart we can see that with the iteration increasing, the $||\nabla l(w)||$ is act in downward trend.
With the learning rate increase by some extent from 0.0001 to 0.001 to 0.1, the larger the learning rate the faster the training process. However, too large learning rate will cause the training process to be less stable some time. So that learning rate can not be too large or too small.

For following test I continue use the learning rate as 0.1 and from the chart2 the threshold will be set as value 0.002 in the following evaluation experiments.

3.2 Evaluation of how the chosen Collocations words affect the accuracy

- learningRate: 0.1
- maxIterations:4000
- doesn't use threshold for the experiment

Collocations	Training time(min)	Testing time (s)	Iterations run	Testing accuracy
C-2,-1 and C1,2	4.5	0.694	4000	77%
C-3,-1 and C1,3	4.4	0.802	4000	76%
C-2,-1 and C1,2 C-3,-2 and C2,3	6.1	0.700	4000	79%
All two word phrases before or after the target	17.5	1.193	4000	85%

From the previous experiment, the best case should be use all of two word phrases in the sentence although the time used is longer than using only one two word phrases before and after the target. After the experiment, I also set the threshold to 0.004 since the featureVector used is about twice size as before.

3.3 Final training and testing result on three different confusion sets

- learningRate: 0.1
- maxIterations: 20000
- threshold:0.004

Training and Testing	Interations(Train)	Training time(min)	Testing time(s)	Test accuracy
{ adapt, adopt }	3320	8.31	1.15	85%
{ bought, brought }	3347	12	1.147	85%
{ peace, piece }	3325	12	1.124	96%

4 Conclusion

It seems that within a extract extent(0.0001 to 0.01), the larger the learning rate the faster the model is trained. Then after one value, the larger the learning rate, the less accuracy the model is trained.

In conclusion, experiments with parameters tuned should be done to get the most proper parameters that used to train the model so that for the extract kind of training data it will reach the better result within a less time period.