Assignment 3

Saleha Iqbal Hussain

2011CS10252

# Question 1

## Part1

Average accuracy: - 99.42

## Part2

Random Accuracy – 12.5%

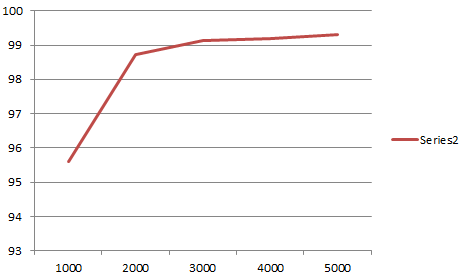
Increase – 7.95 times

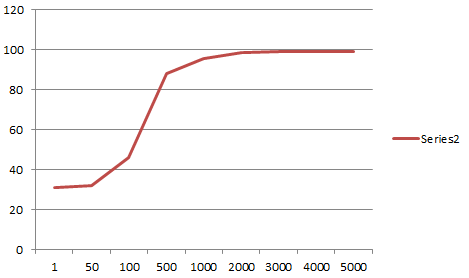
## Part 3

No, each class has its own probability of words. Words in each class are determined from a set of examples. And thus this special case of examples just adds to the frequency if words. If the words used with a high frequency in one class over other, the classification would and should go in favor of that class. If both have nearly same frequency of words, then it doesn’t matter.

A problem will arise if these examples use words that have not been used before in any examples.

## Part 4





Very steep learning curve. It learns very quickly. It achieves 90% accuracy in about 500 examples. Rest was marginal improvement.

## Part5

The confusion matrix is attached.

Order

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 'talk.politics.mideast' | 'rec.motorcycles' | 'rec.sport.baseball' | 'rec.sport.hockey' | 'talk.politics.misc' | 'talk.religion.misc' | 'talk.politics.guns' | 'rec.autos' |
| 'talk.politics.mideast' | 932 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 'rec.motorcycles' | 1 | 995 | 1 | 0 | 0 | 0 | 0 | 2 |
| 'rec.sport.baseball' | 2 | 0 | 991 | 1 | 1 | 0 | 0 | 1 |
| 'rec.sport.hockey' | 0 | 0 | 2 | 995 | 0 | 0 | 0 | 0 |
| 'talk.politics.misc' | 0 | 1 | 0 | 0 | 763 | 1 | 1 | 0 |
| 'talk.religion.misc' | 2 | 0 | 0 | 0 | 2 | 621 | 2 | 0 |
| 'talk.politics.guns' | 2 | 0 | 0 | 0 | 8 | 6 | 906 | 1 |
| 'rec.autos’ | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 985 |

Highest diagonal entry – motorcycles and hockey

Highest non-diagonal entry – guns confused for politics.misc

# Question 2

## Part1

Code is attached

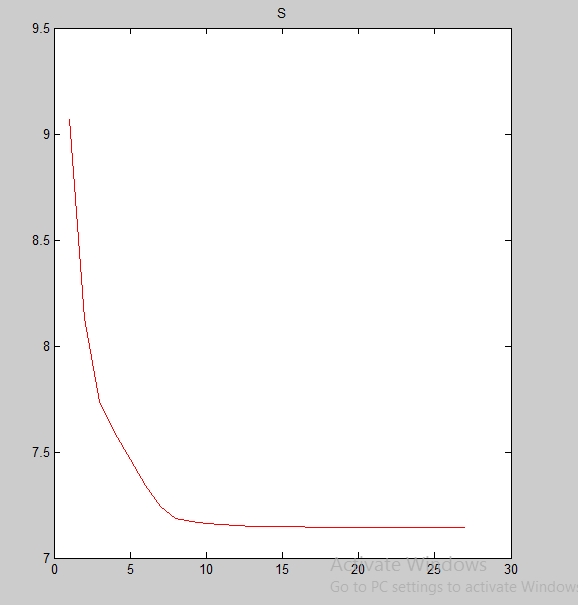
## Part2

Algorithm is implemented, code is attached.

Convergence – May converge in as less as 12 iterations, may need more than 30, depends upon initial cluster selected. But often converges in less than 30.

## Part3

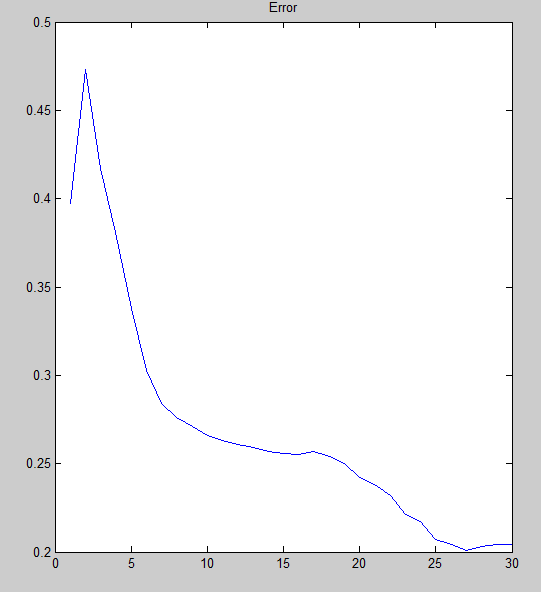
Decreases monotonically. Value near 7 on convergence.



## Part4

Decreases as a rule, may not be monotonic. On convergence error is nearly 0.2 ie 20 percent.

Some are classified more easily than others. For example – 7 and 1 are often mistaken for each other.



# Question 3

## Part1

Prob: -6.1833e+03

Matrices are generated as part of the code

## Part2

E – Splitting data values based on obtained probabilities for missing parameters.

M – Find new probability tables from the data sample obtained

Stopping criteria – error difference on test set becomes 0;

## Part3

Negative slope - Log likelihood decreases with more missing data, as is expected. Also, the diff of slope is decreasing.

# Question4