Connect & Discover

Postgraduate Certificate in Software Design with Artificial Intelligence

Advanced Machine Learning and Neural Networks - (AL_KSAIG_9_1)

Minor Exercise 2

Student ID: A00267948

Student Name: Daniel Foth

GIT: https://github.com/DanielsHappyWorks/aml-text-classification

Brief Description:
Using the Newsgroups dataset and the following machine learning methods

- SVM
- Naive Bayes
- Neural Network

You will need to use the techniques of stopping(removing small insignificant words eg I, the, you etc), stemming(removing the endings of words eg -ed -ing) and use of TF/IDF (Term Frequency over Item Document Frequency) to aid in the classification of the type of news report

This is a task which will require you to do some feature engineering to get decent accuracy

The submission will be the source code which will output a confusion matrix and overall accuracy of each classifier

Dataset: sklearn.datasets import fetch_20newsgroups

Contents

Data

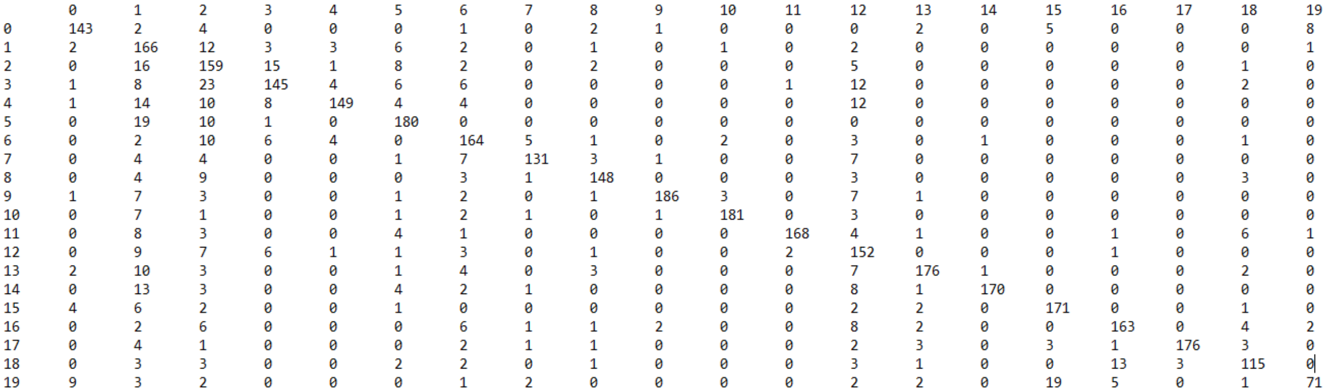
The data set contains:

Classes	20
Samples total	18846
Dimensionality	1
Features	text

SVM

Accuracy 0.8259946949602122

Confusion Matrix

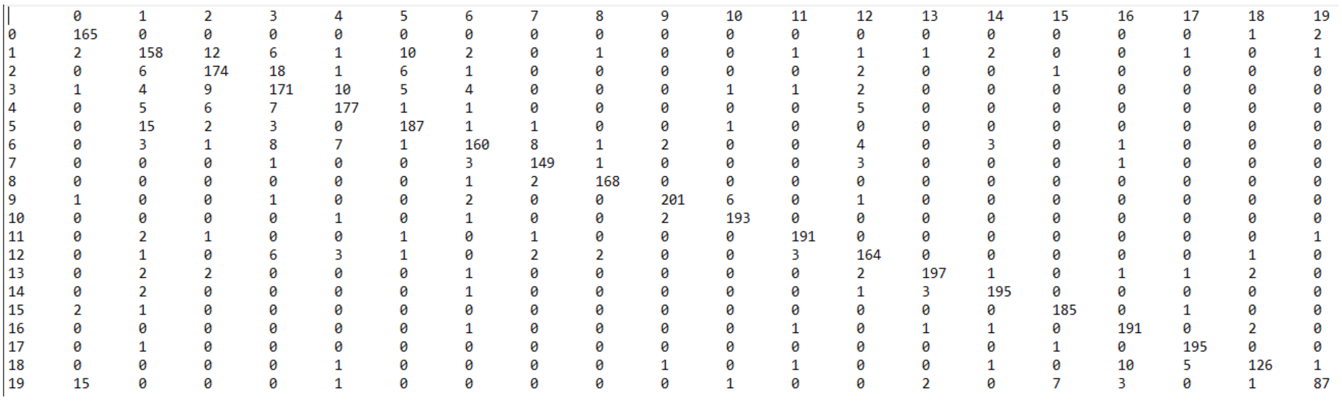


```
        0       1       2       3       4       5       6       7       8       9       10      11      12      13      14      15      16      17      18      19
0       143     2       4       0       0       0       1       0       2       1       0       0       0       2       0       5       0       0       0       8
1       2       166     12      3       3       6       2       0       1       0       1       0       2       0       0       0       0       0       0       1
2       0       16      159     15      1       8       2       0       2       0       0       0       5       0       0       0       0       0       1       0
3       1       8       23      145     4       6       6       0       0       0       0       1       12      0       0       0       0       0       2       0
4       1       14      10      8       149     4       4       0       0       0       0       0       12      0       0       0       0       0       0       0
5       0       19      10      1       0       180     0       0       0       0       0       0       0       0       0       0       0       0       0       0
6       0       2       10      6       4       0       164     5       1       0       2       0       3       0       1       0       0       0       1       0
7       0       4       4       0       0       1       7       131     3       1       0       0       7       0       0       0       0       0       0       0
8       0       4       9       0       0       0       3       1       148     0       0       0       3       0       0       0       0       0       3       0
9       1       7       3       0       0       1       2       0       1       186     3       0       7       1       0       0       0       0       0       0
10      0       7       1       0       0       1       2       1       0       1       181     0       3       0       0       0       0       0       0       0
11      0       8       3       0       0       4       1       0       0       0       0       168     4       1       0       0       1       0       6       1
12      0       9       7       6       1       1       3       0       1       0       0       2       152     0       0       0       1       0       0       0
13      2       10      3       0       0       1       4       0       3       0       0       0       7       176     1       0       0       0       2       0
14      0       13      3       0       0       4       2       1       0       0       0       0       8       1       170     0       0       0       0       0
15      4       6       2       0       0       1       0       0       0       0       0       0       2       2       0       171     0       0       1       0
16      0       2       6       0       0       0       6       1       1       2       0       0       8       2       0       0       163     0       4       2
17      0       4       1       0       0       0       2       1       1       0       0       0       2       3       0       3       1       176     3       0
18      0       3       3       0       0       2       2       0       1       0       0       0       3       1       0       0       13      3       115     0
19      9       3       2       0       0       0       1       2       0       0       0       0       2       2       0       19      5       0       1       71
```

Naive Bayes

Accuracy: 0.9108753315649868

Confusion Matrix:

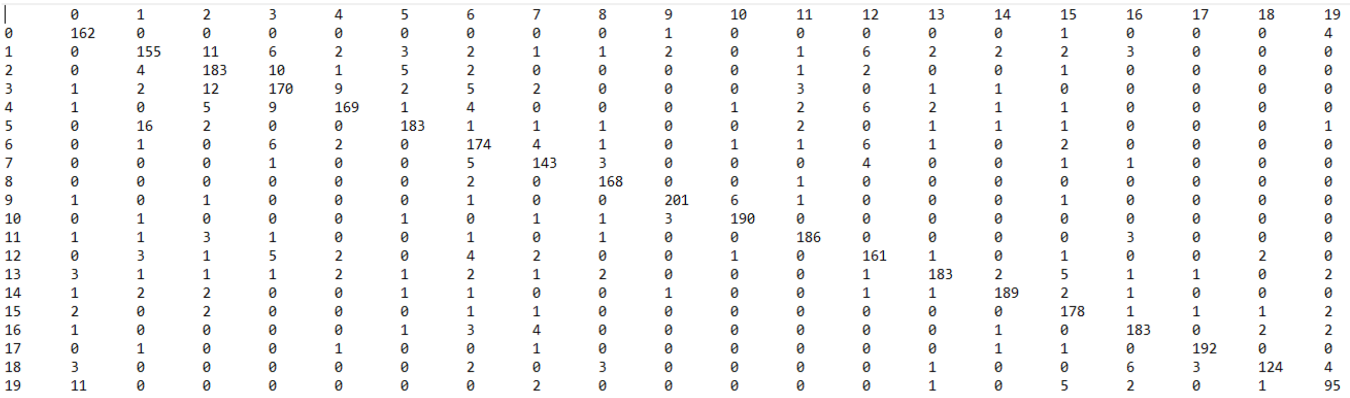


```
        0       1       2       3       4       5       6       7       8       9       10      11      12      13      14      15      16      17      18      19
0       165     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       1       2
1       2       158     12      6       1       10      2       0       1       0       0       1       1       1       2       0       0       1       0       1
2       0       6       174     18      1       6       1       0       0       0       0       0       2       0       0       1       0       0       0       0
3       1       4       9       171     10      5       4       0       0       0       1       1       2       0       0       0       0       0       0       0
4       0       5       6       7       177     1       1       0       0       0       0       0       5       0       0       0       0       0       0       0
5       0       15      2       3       0       187     1       1       0       0       1       0       0       0       0       0       0       0       0       0
6       0       3       1       8       7       1       160     8       1       2       0       0       4       0       3       0       1       0       0       0
7       0       0       0       1       0       0       3       149     1       0       0       0       3       0       0       0       1       0       0       0
8       0       0       0       0       0       0       1       2       168     0       0       0       0       0       0       0       0       0       0       0
9       1       0       0       1       0       0       2       0       0       201     6       0       1       0       0       0       0       0       0       0
10      0       0       0       0       1       0       1       0       0       2       193     0       0       0       0       0       0       0       0       0
11      0       2       1       0       0       1       0       1       0       0       0       191     0       0       0       0       0       0       0       1
12      0       1       0       6       3       1       0       2       2       0       0       3       164     0       0       0       0       0       1       0
13      0       2       2       0       0       0       1       0       0       0       0       0       2       197     1       0       1       1       2       0
14      0       2       0       0       0       0       1       0       0       0       0       0       1       3       195     0       0       0       0       0
15      2       1       0       0       0       0       0       0       0       0       0       0       0       0       0       185     0       1       0       0
16      0       0       0       0       0       0       1       0       0       0       0       1       0       1       1       0       191     0       2       0
17      0       1       0       0       0       0       0       0       0       0       0       0       0       0       0       1       0       195     0       0
18      0       0       0       0       1       0       0       0       0       1       0       1       0       0       1       0       10      5       126     1
19      15      0       0       0       1       0       0       0       0       0       1       0       0       2       0       7       3       0       1       87
```

Neural Network

Accuracy 0.8989389920424403

Confusion Matrix



```
        0       1       2       3       4       5       6       7       8       9       10      11      12      13      14      15      16      17      18      19
0       162     0       0       0       0       0       0       0       0       1       0       0       0       0       0       1       0       0       0       4
1       0       155     11      6       2       3       2       1       1       2       0       1       6       2       2       2       3       0       0       0
2       0       4       183     10      1       5       2       0       0       0       0       1       2       0       0       1       0       0       0       0
3       1       2       12      170     9       2       5       2       0       0       0       3       0       1       1       0       0       0       0       0
4       1       0       5       9       169     1       4       0       0       0       1       2       6       2       1       1       0       0       0       0
5       0       16      2       0       0       183     1       1       1       0       0       2       0       1       1       1       0       0       0       1
6       0       1       0       6       2       0       174     4       1       0       1       1       6       1       0       2       0       0       0       0
7       0       0       0       1       0       0       5       143     3       0       0       0       4       0       0       1       1       0       0       0
8       0       0       0       0       0       0       2       0       168     0       0       1       0       0       0       0       0       0       0       0
9       1       0       1       0       0       0       1       0       0       201     6       1       0       0       0       1       0       0       0       0
10      0       1       0       0       0       1       0       1       1       3       190     0       0       0       0       0       0       0       0       0
11      1       1       3       1       0       0       1       0       1       0       0       186     0       0       0       0       3       0       0       0
12      0       3       1       5       2       0       4       2       0       0       1       0       161     1       0       1       0       0       2       0
13      3       1       1       1       2       1       2       1       2       0       0       0       1       183     2       5       1       1       0       2
14      1       2       2       0       0       1       1       0       0       1       0       0       1       1       189     2       1       0       0       0
15      2       0       2       0       0       0       1       1       0       0       0       0       0       0       0       178     1       1       1       2
16      1       0       0       0       0       1       3       4       0       0       0       0       0       0       1       0       183     0       2       2
17      0       1       0       0       1       0       0       1       0       0       0       0       0       0       1       1       0       192     0       0
18      3       0       0       0       0       0       2       0       3       0       0       0       0       1       0       0       6       3       124     4
19      11      0       0       0       0       0       0       2       0       0       0       0       0       1       0       5       2       0       1       95
```

Conclusion

In conclusion processing this much text is very time consuming and makes for altering models very difficult. The best performance was achieved using Naïve Bays, but all of the models are really good as they are all over 80% accurate. Neural network came in at a close second but it’s harder to predict how it works in the background which still makes NB the preferred algorithm.