This presentation will include 3 parts. I first played around with the classifiers with some artificial datasets. Then I applied some of them on my own dataset and some text datasets we’ve seen before. And then I explored BERT a little bit.

The 10 artificial datasets are generated from 5 different random functions, ranging from a complete mess, to clusters with clear boundaries. And I run classification task on each of the 10 datasets with each of the 8 classifiers.

This is the performance of the classifiers on the first dataset, that mess we’ve seen before. Each two rows is the result from one classifier. And the MLP here is multilayer perceptron. No surprise, the error rates are very close to 0.5.

Then let’s see a dataset looks like this. We can see that here random forest, neural network and gradient boosting all did a quite decent job, but the error rate of naïve bayes, SVC and logistic are very bad. This shows us that these classifiers seem to be not able to draw a flexible boundary.

And this is the performance of the classifiers on the nineth and the tenth dataset. And everybody seems to be doing a great job here! And I also found that when I first train the classifiers on the nineth dataset, then train them again on the tenth dataset, the performance will be even better.

Now let’s come to the part with my own data. The text I used are tweets that I scrapped from the official accounts of 10 different media. This is a chart I found from the internet, I randomly picked 5 media from the left column, and another 5 from the right column.

I scrapped one thousand tweets from each of these media’s account. And I only kept the relevant variables, though actually only the Text, the Media and the Left columns are used.

So first, after some basic preprocessing and vectorization. I visualized the results from PCA with two components. So red is for right and blue is for left. PCA seems to be not doing a very good job here.

Then I run some logistic regressions. As the number of included PCA component increases, the test accuracy increased from 60% to 73%. And the logistic regression with L1 regularization gives a test accuracy of 84%. Which is impressive!

However, the decision tree and random forest gave a complete different pircture. From the confusion matrices, we can see that what they were doing was basically label every tweets as “left”. Through the random forest did a slightly better job.

And the KNN did just the opposite as the decision tree and the random forest. So these three classifiers basically don’t work at all. They are even outperformed by logistic regression with just 10 PCA components.

And then we come to neural network. First, we can notice that unsurprisingly, it did a pretty excellent job. The accuracy on the test set is 82.4%. However, we also see that the neural network didn’t do a better job than logistic regression with regularization.

I also compared the two with other evaluation measures like precision, recall and F-measure. It seems that neural network is more balanced on the precision score, but in all the other measures, logistic regression is better.

So the conclusion here is: A classifier as simple as logistic regression can do a very good job on certain task. But also notice that the neural network classifier here is just a very simple one, without any tuning.

Then I moved on to other datasets that we have seen before. This is the newsgroups data. Support Vector Machine, logistic regression, and neural network are the best ones.

This is the small senate release press data. We want to distinguish between releases from Obama and Clinton. And the winner here is random forest. SVC, logistic, decision tree and gradient boosting all did a very good job.

Now we get to the large senate press data, with five senators. Our stars here are decision tree, random forest and gradient boosting. And it seems that different classifiers are good at identifying different senator.

Finally, spams versus non-spams. This task seems to be a difficult one. And support vector machine and neural network got the lowest error rates. But we also notice that most of the classifiers tend to be more precise in identifying non-spams, quite conservative. Only logistic regression reached a higher precision rates in spams.

And then I played around with BERT, on the most beginner friendly pipeline, sentiment analysis. I applied it on my media tweets dataset. And then I calculated the average proportion of tweets with positive sentiment by skewness group, and by media. It seems that both left and right are unhappy most of the time, though left media are slightly more optimistic.

…… And fourth, since the classifiers are applied on vectorized words, it makes it difficult to see what is happening inside the black box. So, I would be happy to know if there are some ways to open it.