Fabian Hansch Mauritzson

Text Analytics

April 25, 2021

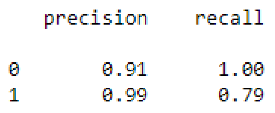
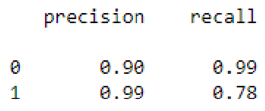
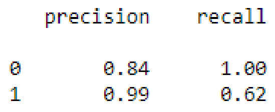
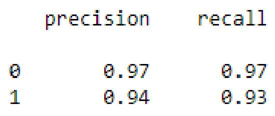
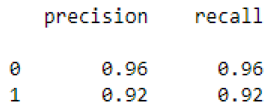
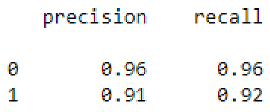
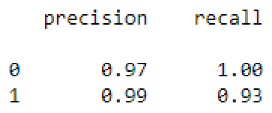
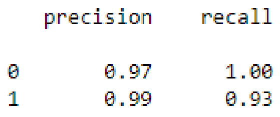
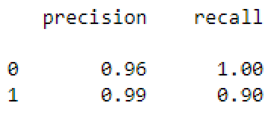
Assignment 3

Question 1:

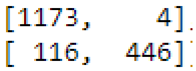
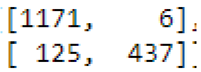
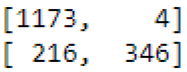
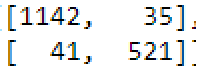
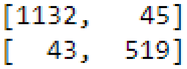
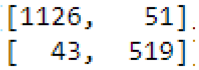
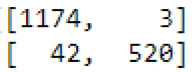
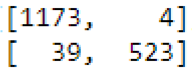
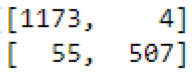
Naive Bayes was the worst at predicting negatives. This can be seen in that for all feature representations, Naive Bayes had many more false negatives than any other classifier. Decision Tree was the worst at predicting positives. In all feature representations, Decision Trees had a lot more false positives than any other classifier. Random Forest was on the same level as Naive Bayes in positives, and on the same level as Decision Tree in terms of negatives.

All the feature representations had in common that when used with Naive Bayes had a significantly lower recall in terms of detecting spams. The same drastic difference could not be found with the other classifiers, but there was a drop in recall between spam and non-spam with all classifiers. Binary representation was the one with the most consistent precision and recall. There were combinations where the precision was higher for detecting spam than non-spam.

Precision and Recall:

	Binary	Frequency	TF-IDF
Naive Bayes	 precision recall 0 0.91 1.00 1 0.99 0.79	 precision recall 0 0.90 0.99 1 0.99 0.78	 precision recall 0 0.84 1.00 1 0.99 0.62
Decision Tree	 precision recall 0 0.97 0.97 1 0.94 0.93	 precision recall 0 0.96 0.96 1 0.92 0.92	 precision recall 0 0.96 0.96 1 0.91 0.92
Random Forest	 precision recall 0 0.97 1.00 1 0.99 0.93	 precision recall 0 0.97 1.00 1 0.99 0.93	 precision recall 0 0.96 1.00 1 0.99 0.90

Confusion Matrix:

	Binary	Frequency	TF-IDF
Naive Bayes	 [1173, 4] [116, 446]	 [1171, 6] [125, 437]	 [1173, 4] [216, 346]
Decision Tree	 [1142, 35] [41, 521]	 [1132, 45] [43, 519]	 [1126, 51] [43, 519]
Random Forest	 [1174, 3] [42, 520]	 [1173, 4] [39, 523]	 [1173, 4] [55, 507]

Question 2:

	Binary	Frequency	TF-IDF
Naive Bayes	4*100 + 116*5 = 980	6*100 + 125*5 = 1225	4*100 + 216*5 = 1480
Decision Tree	35*100 + 41*5 = 3705	45*100 + 43*5 = 4715	51*100 + 43*5 = 5315
Random Forest	3*100 + 42*5 = 510	4*100 + 39*5 = 595	4*100 + 55*5 = 675

This table shows that the combination of Binary representation and Random Forest has the lowest total cost. The interesting part of this analysis is that the classifiers have a larger impact on the cost than feature representations. This can be seen in that 1-3 of the lowest costs are all using Random Forest, while the highest costs combination are all using Naive Bayes. The difference between Random Forest + Binary and Random Forest + Frequency must also be discussed. Random Forest + Binary outperformed with only one classification in a testing dataset of over 1700 data points. This means that depending on how the split between training and testing turned out might have impacted the final result since they were so close in comparison.