Dino Fejzulovic

February 22nd 2022

Machine Learning Assignment 3A

**The name and link to the data set:** <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>

**Average accuracy: 0.7142%**

**Average probability for correct Predictions: 0.9209**

**Average probability for Incorrect Predictions: 0.9335**

**Performance of Naïve Bayes to your best reported performance of kNN and Decision Trees**

| GaussNB | kNN | DecisionTree |
| --- | --- | --- |
| 0.9209 | 2 | 0.7172 |

I believe that the Naive Bayes algorithm did much better than both the kNN and decision tree as the average accuracy scores have beaten the former 2 algorithms. The kNN was recorded as a 2 but under normal conditions it would have probably performed at the 0.7% range. The GaussNB is a linear classifier unlike the Knn and tends to perform faster and more accurately at small datasets , with scale and complexity of course KNN tends to drop in performance while Naive Bayes algorithms flourish with Big Data. Overall, the GaussNBB performed well.

**Discuss the average probability scores and give an example of how you might use the probability score to provide more useful information than simply a classification label.**

1. Average probability for correct Predictions: 0.9209
2. Average probability for Incorrect Predictions: 0.9335

Unfortunately my average probability for a correct prediction was lower than my average for an incorrect prediction. This may be because of the lack of size in the data and the trouble classifying such a complex set of number that may or may not have any correlation to each other ( it is medical data after all, a real time statistic and not a 2 months planned analysis). With a bigger dataset with a more refined objective we could find a lower accuracy and performance but much more correct predictions.

Since, The Gaussian Naive Bayes are best used in cases when all our features are continuous .. I would try to figure out if it was possible to do a binary classification with this dataset like how we did in the Spam/Ham classifier in class, because the classifier label does not show us that much.