IN-STK5000 Assignment 2

Aurora Poggi, Fábio Rodrigues Pereira, Nick Walker
Assignment Group 9

September 2021

1 Data set information

We chose the ionosphere dataset, and we compared random forest and logistic regression classifiers. The dataset contains 224 ”good” datapoints and 126 ”bad” data points.

A system in Goose Bay, Labrador, collected this radar data. A phased array of 16 high-frequency antennas with a total transmitted power of 6.4 kilowatts makes up this system. Free electrons in the ionosphere were the intended targets. Radar returns ”good” for indications of ionosphere structure. Otherwise, it returns ”bad” for that do not contains signals through the ionosphere.

The time of a pulse and the pulse number were used as inputs in an autocorrelation function that was applied to the received signals. For the Goose Bay system, there were 17 pulse numbers. The complex values given by the function resulting from the complex electromagnetic signal are described by two attributes per pulse number in this database.

2 Pipeline, Models, and Method

We used the sklearn Pipeline object to construct our data processing pipeline, using three different normalization methods from scikit-learn (”std”, ”maxabs”, and ”minmax”). We compared two models: Random Forest and Logistic Regression classifiers. The Random Forest models were trained and evaluated with and without bootstrapping, with two different criteria (”gini” and ”entropy”) and with 100, 300, and 500 iterations. The Logistic Regression models were trained and evaluated with L1 and L2 loss and again with 100, 300, and 500 iterations. The pipeline and model training code can be found on Github.[1]

[1] https://github.com/fabiorodp/IN_STK5000_Adaptive_methods_for_data_based_decision_making/tree/main/assignment2

3 Artificial Dataset Results

We used the generator class to generate 500 points of artificial data with 34 dimensions and good and bad labels distributed as the original ionosphere data. The model with the highest accuracy was the random forest classifier with 85% accuracy on average. This number is calculated as an average with 10-fold cross-validation. As this data is generated randomly from a normal distribution, we expected that the performance of the models in practice would differ significantly, most likely achieving better performance because the data is not random (and the actual data points were likely not generated from a normal distribution).

4 Ionosphere Dataset Results

We compared the artificial data results with the ionosphere dataset using the same algorithms as before. In this case, we test two cases, in the first we use oversampling on the smaller class (the ”bad” data points) so that the classes are balanced. In the second case, we use the data without balancing the classes. On the real dataset the most accurate algorithm was again the random forest classifier in both cases. With the ionosphere data the model achieved 93.7 accuracy on average through cross-validation when the classes were unbalanced, with 10 different folds as before. When the data was balanced, the model achieved on average 97.9% accuracy. As expected, both of these results are substantially better than the performance of the algorithms on the artificial dataset.