FYS-STK3155/4155 Applied Data Analysis and Machine Learning - Project 3

Lotsberg, Bernhard Nornes
Nguyen, Anh-Nguyet Lise

https://github.com/liseanh/FYS-STK4155-project3

November - December 2019

Abstract

Whole page: 6.24123in Column: 3.01682in

1 Introduction

In popular culture the neural network is probably the most well known form of machine learning. In recent years many other statistical learning methods have proven themselves as well however. In this project we compare the performance of neural networks and gradient boosting in the case of binary classification. In addition to these, we also use the much simpler k-nearest neighbours method as a baseline for classification performance.

2 Data

The data set we will analyse in this project is the MAGIC Gamma Telescope data set retrieved from the UCI Machine Learning Repository, which was generated by a Monte Carlo (MC) program described by D. Heck et. al. [2] to simulate high energy gamma particle registration in a Cherenkov gamma telescope. The set consists of ten explanatory variables and a binary response variable `class` which specifies whether the measured photons resulted from a gamma particle (`class = g`) or a hadron (`class = h`). The entire data set consists of 19020 instances with no missing values, with outcome distribution as shown in Figure 1. The explanatory and response variables are defined as the following by the UCI Machine Learning Repository [1]:

1. `fLength`: continuous # major axis of ellipse [mm]
2. `fWidth`: continuous # minor axis of ellipse [mm]
3. `fSize`: continuous # 10-log of sum of content of all pixels [in #phot]
4. `fConc`: continuous # ratio of sum of two highest pixels over `fSize` [ratio]
5. `fConc1`: continuous # ratio of highest pixel over `fSize` [ratio]
6. `fAsym`: continuous # distance from highest pixel to center, projected onto major axis [mm]
7. `fM3Long`: continuous # 3rd root of third moment along major axis [mm]
8. `fM3Trans`: continuous # 3rd root of third moment along minor axis [mm]
9. `fAlpha`: continuous # angle of major axis with vector to origin ¡¡¡¡¡¡¡ HEAD [deg]
10. `fDist`: continuous # distance from origin

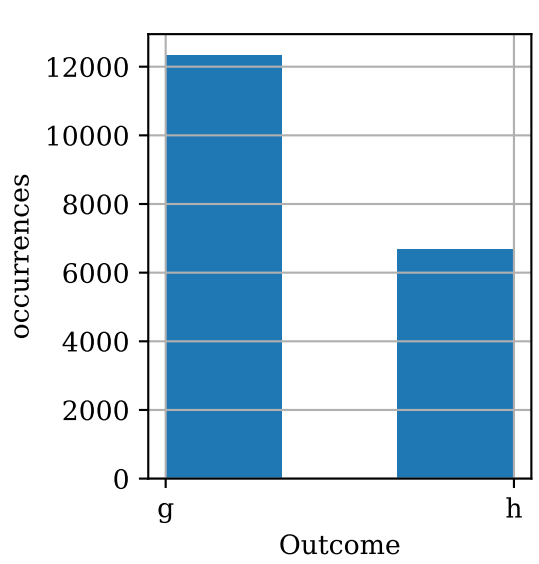

Figure 1: Frequencies of the outcomes g and h in the data set. The numbers of instances for the categories were 12332 and 6688 for g and h respectively.

to center of ellipse [mm] ======= [deg]

11. `fDist`: continuous # distance from origin to center of ellipse [mm] ¿¿¿¿¿¿¿ b5b8dbc733c8eba89a02bc2fad511a931945ffa8
12. `class`: g, h # gamma (signal), hadron (background)

3 Methods

3.1 *k*-Nearest Neighbour (kNN)

3.2 Multilayer Perceptron (MLP)

3.3 Gradient Boosting

3.4 Model evaluation

confusion matrix and f1 score

4 Results

5 Discussion

6 Conclusion

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[2] Dieter Heck, G Schatz, J Knapp, T Thouw, and JN Capdevielle. CORSIKA: a Monte Carlo code to simulate extensive air showers. Technical report, 1998.

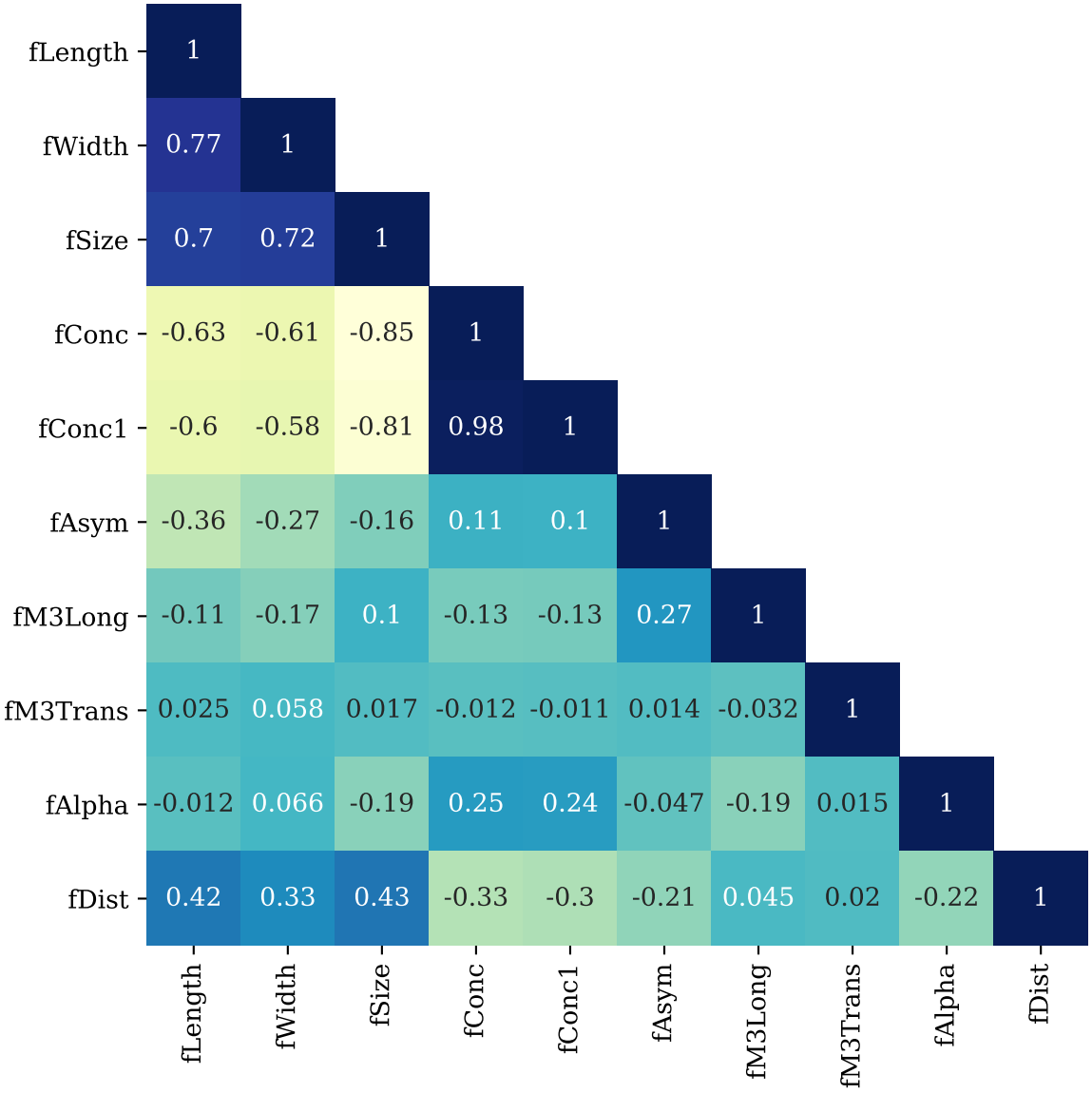

Figure 2: Correlation matrix of the features in the train set. Upper triangle excluded for readability.

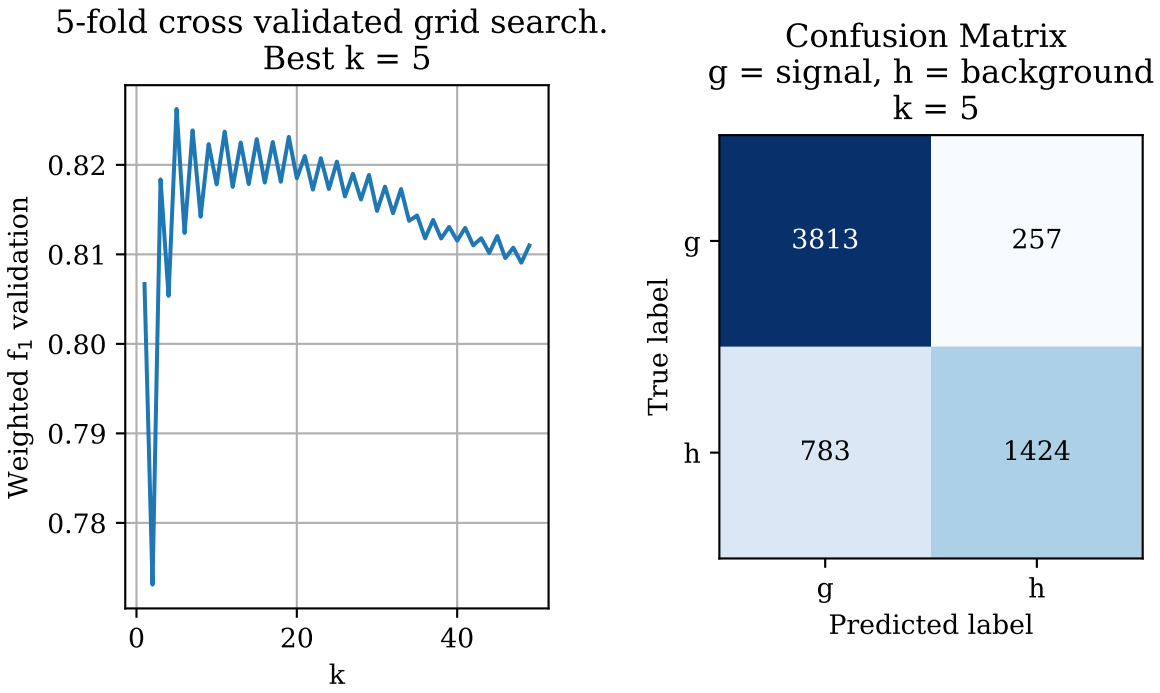

Figure 3: Results from tuned kNN using cross validation. The confusion matrix was found using the test set.

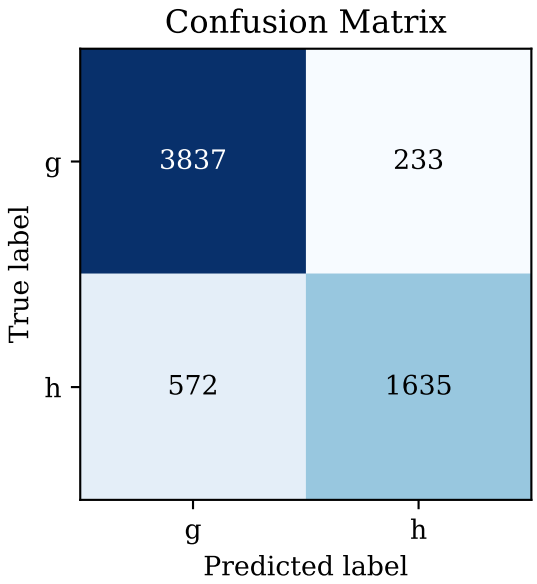

Figure 4: Confusion matrix of the neural network model applied to the test set.

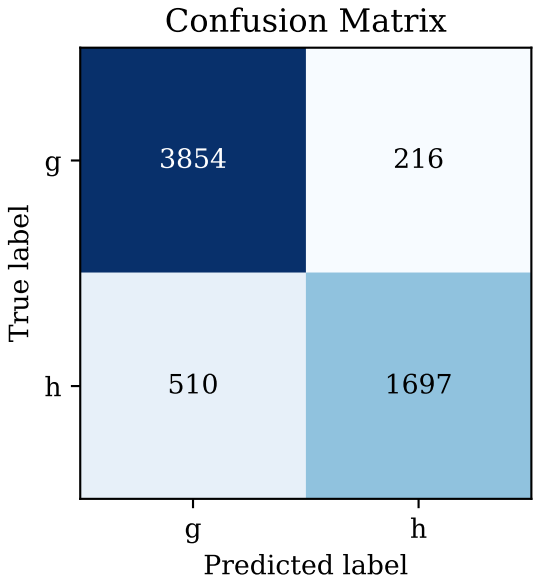

Figure 5: Confusion matrix of the gradient boosted model applied to the test set.