Machine Learning for Physics Research PHY391 (2024)

Homework 3

Reading. Chapter 3 of Geron (Third edition). Recommended: Chapter 3 of Acquaviva.

Problem. Use the Planet Habitability Laboratory data (phl_exoplanet_catalog.csv) that you inspected in the previous homework to build classifiers that predict planet habitability based on three features: stellar mass, orbital period, and distance.

There are three target classes in the database: 0 (non-habitable), 1 (optimistically habitable) and 2 (reasonably habitable).

You need to select the features and target classes from the database. For simplicity, treat this as a binary classification problem, combining the two habitable classes into one.

Follow a similar procedure to the one for the MNIST dataset in the Geron book to train a RandomForest and SGD classifier. Calculate the corresponding F1 scores and confusion matrices. Construct the PR curves as in Figures 3-6 and 3-8 in the Geron book and comment on which classifier is better.

PLEASE SUBMIT YOUR ANSWERS ON COURSE SITE AS EXECUTED JUPYTER NOTEBOOKS WITH EMBEDDED EXPLANATORY COMMENTS.

CLEAN UP BY DELETING IRRELEVANT CODE.