Machine Learning for Physics Research PHY391 (2024)

Homework 2

Reading. Go over Chapter 2 in Geron (Third edition) to get an idead of what is involved in training a ML model, even if you do not understand all the details yet.

Problem 1. Using the California housing data example, study how the performance of a certain ML method (e.g. RandomForestRegressor) depends on feature rescaling. For example, what happens if we do not rescale the data at all? Or if we use a rescaling function different than a log?

Problem 2. Geron, Third edition, Chapter 2, Excercise 1: Try to build a regressor for the California housing datase using a support vector machine regressor (sklearn.svm.SVR) with various hyperparameters, such as kernel=”linear” (with various values for the C hyperparameter) or kernel=”rbf” (with various values for the C and gamma hyperparameters). Note that support vector machines don’t scale well to large datasets, so you should probably train your model on just the first 5,000 instances of the training set and use only 3-fold cross-validation, or else it will take hours. Don’t worry about what the hyperparameters mean for now; we’ll discuss them in Chapter 5. How does the best SVR predictor perform?

Problem 3. (Modified from Acquaviva, Chapters 2 and 3). Inspect the data obtained from the Planet Habitability Laboratory website (phl_exoplanet_catalog.csv) after first loading them into a jupyter notebook. We will later try to predict planet habitability based on three features: stellar mass, orbital period, and distance. This will be a classification problem with 0 (non-habitable), 1 (optimistically habitable) and 2 (reasonably habitable) as the target classes. You don’t need to train a model for this problem yet.

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

CLEAN UP BY DELETING IRRELEVANT CODE.