EE769 Introduction to Machine Learning (Jan 2021 edition)

Electrical Engineering, Indian Institute of Technology Bombay

Programming Assignment – 1 : ML for Smart Monkeys

Instructions:

a) Only submit ipython notebooks. The notebook should be a complete code plus report with copious comments, references and URLs, outputs, critical observations, and your reasoning to choose next steps.
b) Use good coding practices such as avoiding hard-coding, using self-explanatory variable names, using functions (if applicable). This will also be graded.
c) Cite your sources if you use code from the internet. Also clarify what you have modified. Ensure that the code has a permissive license or it can be assumed that academic purposes fall under 'fair use'.

Problem statements:

1. Regression and out-of-distribution prediction:
    a. Download the wine quality datasets from https://archive.ics.uci.edu/ml/datasets/Wine+Quality
    b. Explore, visualize, and pre-process the data as appropriate.
    c. Train and validate (either using bootstrapping or cross-validation), and test two separate models, varying at least one hyperparameter for at least three of the following types of models:
        i. L1 (LASSO) regularized linear regression (**mandatory**; comment on feature elimination)
        ii. L2 (ridge) regularized linear regression or elastic net (both L1 and L2)
        iii. Random forest
        iv. Support vector regression
        v. Neural network with single hidden layer (output layer should have linear activation)
    d. Search the net about how to determine the importance of each variable, and find the importance in the final models tried. Comment on whether the same variables are important for different models.
    e. Test the model for red with data from white and vice versa, and comment on whether the model for red wines is applicable to white wines and versa or not.
2. Classification:
    a. Download the data to predict Down syndrome in mice from https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression#
    b. The prediction problem is to either predict the genotype (binary) or the class using the gene expression variables from DYRK1A_N to CaNA_N.
    c. Explore, visualize, and pre-process the data as appropriate, including developing a strategy to deal with missing variables. You can choose to impute the variable. The recommended way is to use multivariate feature imputation (https://scikit-learn.org/stable/modules/impute.html)
    d. Train and validate (either using bootstrapping or cross-validation), and test two separate models, varying at least one hyperparameter for at least three of the following types of models:
        i. L1 (LASSO) regularized logistic regression (**mandatory**, including comments on feature elimination)
        ii. L2 (ridge) regularized logistic regression or elastic net (both L1 and L2)
        iii. Random forest
        iv. Support vector classification
        v. Neural network with single hidden layer (output layer should be have softmax activation)
    e. See if removing some features systematically will improve your models (e.g. using L1 regularization or recursive feature elimination https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html).