EE769 Introduction to Machine Learning (Jan 2022 edition)

Electrical Engineering, Indian Institute of Technology Bombay

Programming Assignment – 2 : 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) You may use libraries such as scikit-learn, and need not code anything from scratch.
d) 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'.

Objective: Learn various steps and due diligence needed to train successful classification models.

Background: Some experiments were conducted on mice to see if a treatment of Down's syndrome works or not. Mice were divided into control and diseased (genotype), treated or untreated and whether it shows a particular behavior or not (treatment_behavior). Readings for 77 proteins were recorded for the mice, but some of the readings were discarded if they seemed unreliable (out of range). Your job is to develop a pre-processing pipeline and a classifier, and also find out which subset of proteins is important in predicting which class. Specifically:

1. Read the data directly from https://www.ee.iitb.ac.in/~asethi/Dump/MouseTrain.csv (do not upload)
2. Perform exploratory data analysis to find out:
    a. Which variables are usable, and which are not?
    b. Are there significant correlations among variables?
    c. Are the classes balanced?
3. Develop 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)
4. Select metrics that you will use, such as accuracy, F1 score, balanced accuracy, AUC etc. Remember, you have two separate classification tasks – one is binary, the other has four classes. You may have to do some reading about multi-class classification metrics.
5. Using five-fold cross-validation find the reasonable (I cannot say "best" because you have two separate classifications to perform) hyper-parameter settings for the following model types:
    a. Elastic net logistic regression (L1 and L2 weight)
    b. Neural network with single ReLU hidden layer and Softmax output (number of neurons, weight decay)
    c. SVM (a few kernels, their hyper-parameters such as width, and slack penalty)
    d. Random forest (number of trees, max tree depth, max number of variables per node)
6. Check feature importance for each model to see if the same proteins are important for each model. Read up on how to find feature importance.
7. See if removing some features systematically will improve your models (e.g. using recursive feature elimination https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html).
8. Finally, test a few promising models on the test data: https://www.ee.iitb.ac.in/~asethi/Dump/MouseTest.csv
9. Write your observations and thoughts
10. Write your references as well as other classmates outside of your team with whom you discussed