MACHINE LEARNING: SCIKIT-LEARN

INTRODUCTION

In this assignment, you will develop a machine learning algorithm to get the best possible predictive model for breast cancer, given some diagnostic information. We will use the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which includes 569 rows (patients).

O.L. Mangasarian, W.N. Street and W.H. Wolberg.
Breast cancer diagnosis and prognosis via linear programming.
Operations Research, 43(4), pages 570-577, July-August 1995.

QUESTIONS

Use the dataset along with scikit-learn and machine learning approaches to design a learning algorithm capable of classifying breast cancer cases. Report the accuracy of your best algorithm.

1) Read the information about the dataset in the wdbc.names file.
2) Import the dataset using pandas. Summarize the dataset.
3) Establish a 20% 'test' sample by randomly splitting your data into 80/20.
4) Create numpy matrices from the pandas dataframe.
5) Calculate the null information rate for this dataset using a dummy classifier ('most frequent')
6) Try a variety of approaches to classification for this problem.
   a. At a minimum, implement these methods (picking whatever variables you want to pick – all or a subset – read the .names file to understand the variables!):
      i. logistic regression with no regularization
      ii. L1 penalty logistic regression (LASSO), standardize the inputs first
      iii. L2 penalty logistic regression (Ridge), standardize the inputs first
      iv. Elastic net penalty logistic regression, standardize the inputs first
      v. Random Forest
      vi. Gradient tree boosting: Classification
   b. For each of the above approaches, clearly identify the hyperparameters you need to optimize. Use k-folds cross-validation to optimize the hyperparameters.