Machine Learning

Machine learning has been applied to many fields. You will use it to defeat criminals who try to print their own currency.

This is the Scenario:

You work for the Acme Money Analysis and Prediction Enterprises (AMAPE for short). As engineers for this company you are developing an app for the U.S. Treasury to aid in detection of counterfeit bills. You will be supplied with a data set which provides four features per bill and whether that bill was genuine or counterfeit.

You have been assigned to development AMAPE's official prediction model which will be used by anybody who accepts cash.

As with any problem you first want to study the data.

Read in the database given in the problem. It is in the CSV file data_banknote_authentication.txt provided to you. You may want to use panda read_csv which places the data in a dataframe, similar to, but not to be confused with, a dictionary. This data set contains observations based on measurements made on a number of bills. The last column is whether the bill is genuine (1) or counterfeit (0). Based on the measurements, you need to build a predictor to determine whether a bill is genuine or counterfeit.

The columns in the database are:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)

See reference on dataframes below.

For this project, use all the features. Do NOT use PCA.

Problem 1: Read the database and analyze the data.

Your analysis should include a statistical study of each variable: correlation of each variable, dependent or independent, with all the other variables. Determine which variables are most highly correlated with each other and also which are highly correlated with the variable you wish to predict.

Create a covariance matrix to show which variables are not independent of each other and which ones are best predictors of genuine money. Show the covariance matrix. Create a pair plot.

Based on this analysis you must determine what you think you will be able to do and which variables you think are most likely to play a significant role in predicting the dependent variable, in this case a genuine bill.

Your management at AMAPE want to be kept constantly updated on your progress. Write one paragraph based on these results indicating what you have learned from this analysis. We are looking for specific observations.

Name your program proj1_A.py.

Problem 2: Split your data into training and test datasets as in class. Use every method specified below. Create a report containing a table where you compare prediction percentages and based on this data choose the best method of predicting whether a bill is counterfeit. Your written analysis should be one paragraph. Your management at AMAPE expect results and to have a succinct, compelling description of your work. We are looking for specific observations.

All reports should be typed in a readable font, uploaded in pdf format, and well labeled. Anything the grader (a representative of AMAPE management) cannot find will be deemed due to omission, or poor organization, and not included in the grade.

As in most cases AMAPE management is looking to you for answers, so you will not be able to review your report with them in advance. Note: this means we will not pre-grade this report or any other assignment, turn in what you think is right. (of course you can ask questions, just not: this is what I plan to turn in, is it right?)

For each machine learning method used, find the best values for the parameters. For example, the best gamma or C or K. (Random state is NOT a parameter.) Machine learning methods to use:

Perceptron
Logistic Regression
Support Vector Machine (pick one version)
Decision Tree Learning
Random Forest
K-Nearest Neighbor

Name your program proj1_B.py.
Name your summary proj1.pdf.

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html