MATH637, Spring 2023

Homework 4

Due Monday, April 24th, 11:59pm

Notes:

- This homework is a group assignment (one submission per group)
- There will be no Colab template for this assignment. The groups are supposed to create the Colab notebook themselves and submit their Colab notebook (which contains the corresponding codes, figures, and conclusions).

(2%) Read the dataset "hw4.csv", accessible on the webpage.

The dataset (hereafter denoted by D) contains 3 columns: the first two describe the components of a two-dimensional vector $X \in \mathbb{R}^2$, and the third one is the binary (0-1) label y associated with X.

(2%) Produce a labeled *scatter plot* of the dataset

(2%) Use the function sklearn.model_selection.KFold to shuffle and split the dataset into 10 smaller datasets: $D_1, D_2, \ldots, D_{10}$

(3%) For each of the dataset D_i, we will use D_i as the *test set* to test the accuracy of the algorithm, while the rest of the dataset is used as the *training set* to construction the classifier.

Specifically, for each D_i:

- use the function sklearn.svm.SVC to construct a binary classifier (using the Support Vector Machine algorithm) with parameters
  - kernel='poly'
  - degree =2
  - C=1
  - $coef0 = 1$

  to fit the training set $D \setminus D_i$.
- Compute the accuracy of the classifier in predicting the label of examples in the test set D_i

(2%) Repeat Step 4 with $coef0 = 0$.

(1%) Compare the performances of the classifiers produced in Steps 4 and 5. What is the preferred value of $coef0$?