MATH637, Spring 2019

Homework 1

Due Friday, March 8th, 9:05am

1. Read the dataset "hw1.csv".

   The dataset (hereafter denoted by D) contains 3 columns: the first two describe the components 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 labelled *scatter plot* of the dataset (similar to the one produced in the file 'Fitting with SVM' in the supplementary of Lecture 2)

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

4. 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

5. Repeat Step 4 with $coef0 = 0$.

6. Compare the performances of the classifiers produced in Step 4 and 5. What is the preferred value of $coef0$?

Instruction

- The homework is to be sent to me by email
- Send the codes in a single Python file
- Additional files (figures, comments, remarks) should be all included in another file (doc/pdf)