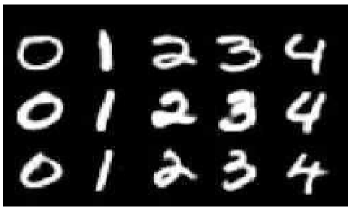
Machine Learning Homework 5

Support Vector Machine

Due day 23:59PM 20th May

I. SVM on MNIST dataset (60% in total):

Use SVM models to tackle classification on images of hand-written digits (digit class only ranges from 0 to 4, as figure shown below).



Training data and testing data are both provided:

- ✧ Training:

  X_train.csv is a 5000x784 matrix. Every row corresponds to a 28x28 gray-scale image.

  Y_train.csv is a 5000x1 matrix, which records the class of the training samples.

- ✧ Testing:

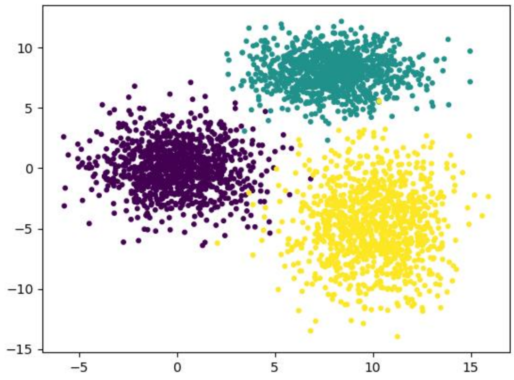
  X_test.csv is a 2500x784 matrix. Every row corresponds to a 28x28 gray-scale image.

  T_test.csv is a 2500x1 matrix, which records the class of the test samples.

- ◆ What you are going to do:
  1. (10%) Use different kernel functions (linear, polynomial, and RBF kernels) and have comparison between their performance.
  2. (20%) Please use C-SVC (you can choose by setting parameters in the function input, C-SVC is soft-margin SVM). Since there are some parameters you need to tune for, please do the **grid search** for finding parameters of best performing model. For instance, in C-SVC you have a parameter C, and if you use RBF kernel you have another parameter γ , you can search for a set of (C,γ) which gives you best performance in cross-validation. (lots of sources on internet, just google for it)
  3. (30%) Use linear kernel+RBF kernel together (therefore a new kernel function) and compare its performance with respect to others. You would need to find out how to use a user-defined kernel in libsvm.

II. Find out support vectors (20% in total):

Use SVM models to classify a 2D dataset into 3 clusters and show the support vectors. The scatter map of the data points is shown as following:



Training data is provided:

Plot_X is a 3000x2 matrix. Every row corresponds to the coordinates of a data point.

Plot_Y is a 3000x1 matrix, which records the class of the data points.

◆ What you are going to do:

1. (20%) Train SVM model with different kernel functions (linear, polynomial, RBF and linear+RBF kernels) and visualize the result.

   Detail of the visualization:

   - Use different colors to show different clusters.
   - All the data samples are shown by “dots”, the “support vectors” that you obtained from your model should be shown with different symbols, e.g. square, triangle, cross.
   - 4 figures in total (linear, polynomial, RBF, linear+RBF)

III. Report (20% in total):

Submit a report in **pdf** format for showing your **code with detailed explanations**, giving **detailed discussion** on experiments as well as your observations. The report should be written in **English**.

You should zip source code and report in one file and name it like ML_HW5_yourstudentID_name.zip, e.g. ML_HW5_0756005_鄭家期.zip.

IV. Packages allowed in this assignment:

You are only allowed to use LIBSVM library, numpy and scipy.spatial.distance.

Official introductions can be found online.

Important: scikit-learn is not allowed.