Machine Learning Homework 5-2
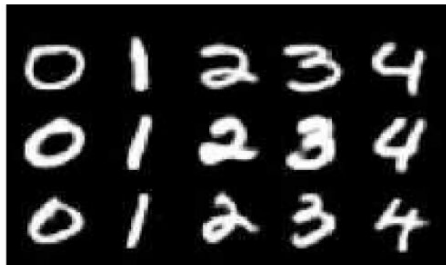
SVM

I. SVM on MNIST dataset

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

- Training data
    - o **X_train.csv** is a 5000x784 matrix. Every row corresponds to a 28x28 gray-scale image.
    - o **Y_train.csv** is a 5000x1 matrix, which records the class of the training samples.
- Testing data
    - o **X_test.csv** is a 2500x784 matrix. Every row corresponds to a 28x28 gray-scale image.
    - o **Y_test.csv** is a 2500x1 matrix, which records the class of the test samples.
- What you are going to do
    - o Part1: Use different kernel functions (linear, polynomial, and RBF kernels) and have comparison between their performance.
    - o Part2: 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 the 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. (There are lots of sources on the internet, just google for it.)
    - o Part3: 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. Report

- Submit a report in pdf format. The report should be written in English.
- Report format:
    - a. code with detailed explanations (20%)
        - Paste the screenshot of your functions with comments and explain your code. For example, show the formula of different kernel functions and the process you search for the kernel parameters, etc.
        - Note that if you don't explain your code, you cannot get any points in section b and c either.
            - Part1 (5%)
            - Part2 (10%)
            - Part3 (5%)
    - b. experiments settings and results (20%)
        - Show everything we asked you to show
            - Part1 (5%)
            - Part2 (10%)
            - Part3 (5%)
    - c. observations and discussion (10%)
        - Anything you want to discuss, such as trying different user-defined kernel functions and comparing the performance.

III. Turn in

1. Report (.pdf)
2. Source code

You should zip source code and report in one file and name it like ML_HW5-2_yourstudentID_name.zip, e.g. ML_HW5-2_0856XXX_王小明.zip.

P.S. If the zip file name has format error or the report is not in pdf format, there will be a penalty (-10). Please submit your homework before the deadline, late submission is not allowed.

- Packages allowed in this assignment:

You are only allowed to use LIBSVM library, numpy, scipy.spatial.distance, and package for visualizing results. Official introductions can be found online.
Important: scikit-learn is not allowed.