E1 213 Pattern Recognition and Neural Networks

Home Work Assignment: 3

Due Date: **One week before final examination date**

In this assignment there are 5 problems. Three of these involve real data and the remaining two involves 2D synthetic data. All are classification problems. As in the previous assignments, you are supposed to learn and test classifiers.

In this assignment you will explore non-linear classifiers. Specifically, on each problem you should explore two different nonlinear classifiers such as a neural network and an SVM. You should explore different architectures for the neural network and different kernel functions for the SVM.

In data sets where training and test data are given separately, you use training data to learn classifier and test data for testing it. If the data is not separated like that, you can explore different options such as splitting data into training and test or using cross validation.

You can implement the learning algorithms on any platform you want (C, C++, MATLAB, Python etc.). You are welcome to use codes that are freely available from any source. You are not required to submit any codes/implementation. A document containing some general useful information on how to implement ML algorithms is also provided. I thank Mr. Deep Patel and Mr. Santhosh Babu G for preparing this document.

Like in earlier assignments, you need to submit a report summarizing your exploration of the data sets. For each problem, briefly describe what all are implemented and then present all the results obtained. Discuss all points from your results that you consider surprising or interesting. The final submission should be in the form of a short PDF file.

The grading depends on whether or not you have done all explorations that are asked for, how you presented the results, your discussion of results and whether you have done some exploration on your own.

The problems are described below.

1. This is a 2-class problem with 2-dimensional feature space. All relevant

files are in the subdirectory called 2class-Synthetic. The class conditional densities are uniform over some region in $\Re^2$ and the relevant regions are shown in an image file. There are three data files. The data may have what is called label noise. That is class labels given in the data may be incorrect. The three data files correspond to 0%, 20% and 40% label noise. Note that the data may not be separable when there is label noise. (Explore SVM with polynomial and Gaussian kernel on this in addition to any other classifier)

2. This is a 5-class classification problem with two dimensional feature space. Three data sets are given. (Image files of the data are also provided for you to visualize the data). Here also there is label noise. In one data set, there is no noise and the classes are separable. In the other two, the data is not separable because there is label noise. All data sets are in the subdirectory, "Board".

3. The data set for this problem is the MNIST data that is widely used for testing neural network algorithms. The data consists of images of hand-written numerals. It is a 10-class problem. There are two data sets here:

   - MNIST: 50000 training + 10000 test samples
   - MNIST-rot: 12000 training + 10000 test samples

   Each sample in MNIST is a hand-written digit image of dimension 28x28. MNIST-rot is a rotated version of the MNIST which also has the same dimension as that of MNIST. You are required to learn a classifier using MNIST data. You get extra credit by exploring the MNIST-rot data also for learning a better classifier.

   All data files are in subdirectory "MNIST". All data files are "mat" files.

4. This is a 2-class classification problem. The data consists of (preprocessed) fMRI recordings of 34 patients with Alzeimer's desease and 47 normal subjects in rest state. The problem is to classify a fMRI recording as healthy or not. The data is given as one mat file per subject. The recoding gives information about activity in different brain regions as a function of time. One uses some discretization of the brain regions which is referred to as parcellation. Thus, each pattern would be an

"image" of dimension $T \times N$ where T is the number of time ticks and N is the number of brain regions. The activity itself is represented by a real number. Each mat file in the data has the following:

- sub_id : Subject Id
- tc_rest_aal : 140×116 (time $\times$ Brain Region)
- tc_rest_power: 140×264 (time $\times$ Brain Region)

tc_rest_aal and tc_rest_power are two different parcellation of the same recordings.

The data is in the subdirectory "Neuro_dataset". There is one mat file per subject. There are two subdirectories of "Neuro_dataset" that contain data from the two classes.

Explore feedforward neural network and SVM for this biary classification problem. At the minimum, learn the classifier for one parcellation. You would get extra credit for exploring both parcellations and for exploring combining the two parcellations.

5. The last problem is 5-class classification problem. The data set is a widely used benchmark data for epileptic seizure detection. The pattern here is essentially an EEG recording. From the EEG recording, a 178-dimensional feature vector is extracted. There are five classes here. (class-1: recording with eyes open, class-2: recording with eyes closed, class-3: recording from healthy brain area; class-4: recording includes brain area with tumour, class-5: recording of an epileptic seizure). For us, the semantics of the classes are irrelevant. You are to explore a neural network and an SVM to learn this 5-class classifier.

   The data is in the subdirectory "EEG". It consists of 11500 training samples where each is a 178 dimensional feature vector. The data is given as a csv file. Please ignore the first column in the csv file. the next 178 columns give the feature vector. The last column gives the class label.