Machine Learning Lab2

Data preprocessing:

1. Firstly, please repeat lab1 to load the MNIST dataset to Matlab. As the original MNIST dataset contains 60000 digits for training, which is too large for our lab sessions. Therefore, we only use a subset of the first 1000 digits for training. Based on the lab1 sample code, you can achieve this by using the following code:

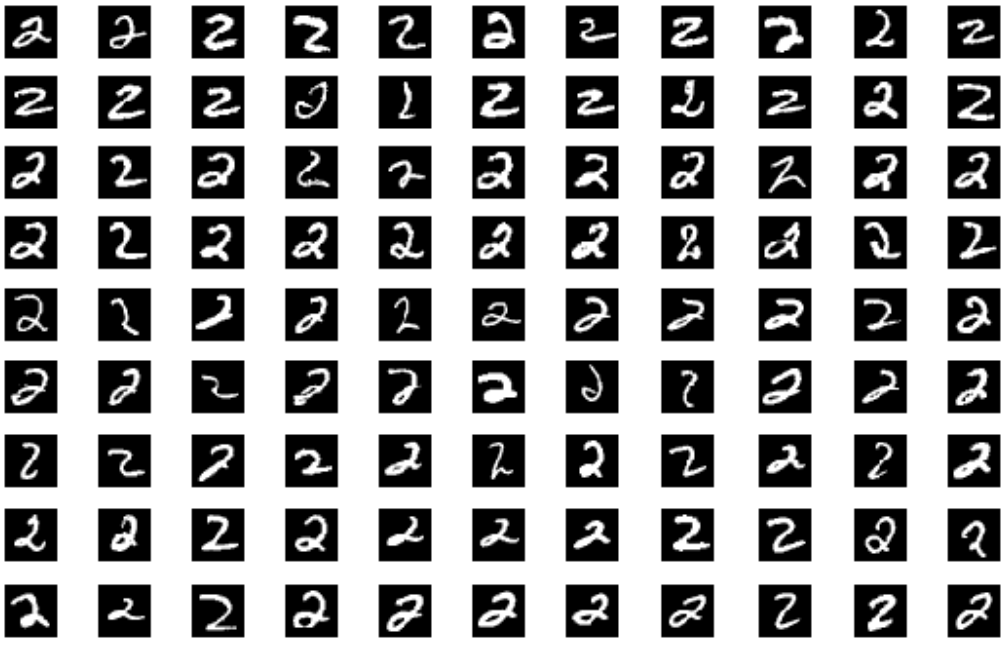
```
tr_label= tr_label(1:1000,1);
tr_feats= tr_feats(1:1000,:);
```

   To avoid repeating the above procedure in the future labs, you can directly download this small set from
   http://www.cs.nott.ac.uk/~qiu/Teaching/G53MLE/Labs/MNIST_subset.mat

2. To get a general picture of the data that we are going to process, you can count how many samples are contained in the training set for each digit. You can use the following code:

```
for i=0:9
   num_of_sample(i+1)=size(find(tr_label==i),1);
end
```

   If you look into the variable `num_of_sample`, you could see that the number of samples for each digit is roughly the same. You can also try to display all the samples for each digit contained in the training set.



3. It is sometimes necessary to normalize the data to make them more comparable. The aim of normalization is to make the Euclidean norm
   http://en.wikipedia.org/wiki/Euclidean_norm#Euclidean_norm
   of each digit equal to 1. You can judge by yourself in the future lab sessions that whether this normalization step is necessary or not.