AC50001 Introduction to Data Mining and Machine Learning   
**AC50001.2 Assignment: Classification and Clustering***Author: Vladislavs Ignatjevs (120015095)*

During last week I was working on Classification and Clustering assignment in attempt to create a system capable of dimension reduction using LDA, PCA and clustering using K-Means, and classification using SVM with a linear kernel, SVM with RBF kernel and Neural Network classifier. All these created components were tested/trained on limited MNIST handwritten digit database (only 3 numbers, 100 samples each).

**Q1. Principal Component Analysis (PCA)**

Applying PCA on data was the first classification problem I attempted to implement. PCA is a good way to compress data in regards to pattern recognition improvement. PCA transforms data to new, linearly uncorrelated coordinate system. It has its advantages and disadvantages. For example, if speaking about advantages, PCA is fast and easy to implement. However, if we overwrite data after applying PCA, it will not be possible to get the exact original data back, since PCA clears up the data that it considers unnecessary. As the result, we might lose significant information during the classification process. Apart from that, there are situations where we cannot apply PCA, since it works on linear spaces. For example, if the data is manifold structured, PCA will not perform an analysis in respect to expected results. The other drawback of PCA I found out comes from inability of Linear PCA to provide the information about class labels. LDA, on the other hand, solves this problem.

At the beginning I spent lots of time because I understood the task in a wrong way. Rather than applying PCA to all number data, I attempted to apply PCA on every number separately. This gave interesting results when plotted results. That was what I believe an attempt to reduce dimensions from 100 to 2 and create an interesting interpretation of most common features for numbers one, five and eight. Due to this being unsuitable for the task. I am not attaching the result of these application here, since it is not applicable for the task given, but I backed up the unused code into folder called “Unused”.

PCA can be summarized in the following steps:

1. Subtracting the mean for individual dimensions
2. Calculating covariance matrix
3. Calculating Eigen values and Eigen vectors
4. Transformation into new data

For the task allocated, I started with importing the data set and concatenating all the data sets into one larger matrix and transposing it:

%combine all data

all\_data = [digit\_one digit\_five digit\_eight]';

Then, I calculated the mean for this matrix and subtracted it from individual dimensions:

%substract the mean for individual dimenstions

all\_data = all\_data - repmat(meanAll, size(all\_data,1),1);

After doing that, it was possible to calculate covariance matrix (to see how much data vary from the mean in respect to each other )

**PCA with K-Means**

**Q2. Linear Discriminant Analysis (LDA)**

Linear Discriminant Analysis on multi-class classification problem.

Linear Discriminant Analysis on 2-class classification problem.

Can pick up class labels

LDA is not guaranteed to perform better than PCA

**Q3. 2 class problem classification**

**Support Vector Machine (SVM)**

**SVM with linear kernel**

**SVM with RBF (radial basis function) kernel**

**Neural Network classifier**