# COMP 4983: Lab Exercise #11

Instructions:

In this lab, you will

- compute principal components on paper for a trivial dataset
- visualize the $k$-Nearest Neighbors ($k$-NN) classifier decision boundaries of principal component analysis (PCA) dimensionality reduced images
- determine the minimum number of principal components using the elbow in the scree plot

## Part 1: Principal Component Analysis (on paper)

In this part of the lab, you will compute principal components for a trivial dataset.

Consider a trivial matrix $X$ of dimension $5 \times 4$ consisting of the following five (5) samples:

| Sample | Input Vector |
|---|---|
| $x_1$ | (0, 4, 2, 2) |
| $x_2$ | (2, 4, 2, 0) |
| $x_3$ | (2, 3, 3, 0) |
| $x_4$ | (1, 4, 2, 1) |
| $x_5$ | (-1, 4, 2, 0) |

The eigen decomposition of $X^T X$, assuming $X$ has been centered, returned the following $V$ and $D$ matrices:

$$V = \begin{bmatrix} -0.22 & 0 & -0.9 & 0.37 \\ -0.67 & -0.71 & 0.2 & 0.09 \\ 0.67 & -0.71 & -0.2 & -0.09 \\ 0.22 & 0 & 0.32 & 0.92 \end{bmatrix} \text{ and } D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 7.85 & 0 \\ 0 & 0 & 0 & 2.75 \end{bmatrix}$$

a) Center matrix $X$ to have zero mean.
b) Apply principal component analysis to reduce the dimension from 4 features to 2 features.
c) Project $x_1$ and $x_3$ onto columns of $W_2$.

## Part 2: Dimensionality Reduction with PCA

In this part of the lab, you will

- apply principal component analysis to images of handwritten digits, 1-5, from the provided MNIST dataset to reduce the dimension of the images to 2 features
- apply the $k$-NN classifier to classify the PCA dimensionality reduced images and visualize the $k$-NN classifier decision boundaries of the PCA dimensionality reduced images
- determine the minimum number of principal components using the elbow in the scree plot

In the provided MNIST dataset, each sample is an 8 pixel by 8 pixel grayscale image of a handwritten digit, 1-5, reshaped into a row vector of 64 elements, where each pixel value is an integer ranging between 0 (white) and 16 (black).

Steps:

1) Download the handwritten digit dataset, *data_lab11.csv*, from BCIT Learning Hub (Content | Laboratory Material | Lab 11) and save it in your working directory. The dataset, *data_lab11.csv*, contains 601 rows (including a header row) and 65 columns. Each row of 65 columns contains the 64 pixels of a handwritten digit image, followed by the digit, 1-5, that this image corresponds to.
2) Download the script, *PCA_lab11.py*, from BCIT Learning Hub (Content | Laboratory Material | Lab 11) and save it as *PCA_lab11.py* in your working directory. This script contains the function **plot_decision_regions()** which plots the decision boundaries between different classes for a given classifier.
3) Add to your script, *PCA_lab11.py*, to read from *data_lab11.csv*.
4) Split the dataset into training and test sets, with the first 75% of the dataset for training and the remaining 25% for testing.
5) Apply PCA on the training set to reduce the dimension of the training set from 64 features to 2 features using **PCA.fit_transform()** from **sklearn. decomposition**.
6) Apply PCA on the test set to reduce the dimension of the test set from 64 features to 2 features using **PCA.transform()** from **sklearn. decomposition**. Note that in this step, **PCA.transform()** is used instead of **PCA.fit_transform()**.
7) Apply the $k$-NN classifier on the dimensionality reduced training set with the number of neighbors, $k = 5$.
8) Evaluate and output the error rate (percentage of misclassifications) of the $k$-NN classifier on the PCA dimensionality reduced test set.
9) Plot the PCA dimensionality reduced training set and the decision boundaries of the $k$-NN classifier using the **plot_decision_regions()** function. Include in your plot, a terse descriptive title, x-axis label, y-axis label and a legend.
10) Plot the scree plot of all the principal components by initializing **PCA()** with **n_components=None**.
11) Determine the elbow in the scree plot.
12) Repeat Steps (5) to (8), but this time, reduce the dimension of the training set and test set to the minimum number of principal components (features) as determined using the elbow in Step (11).