# Homework 4

**Principal Component Analysis (PCA)** is an unsupervised machine learning algorithm that learns a subspace to project a set of multi-dimensional points on with minimum information loss. Given a set of vectors:

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

PCA first calculates the mean-centered vectors

$$X_\mu = \begin{bmatrix} x_1^T - x_\mu \\ x_2^T - x_\mu \\ \vdots \\ x_n^T - x_\mu \end{bmatrix}$$

where $x_\mu = (x_1^T + x_2^T + \cdots + x_n^T)/n$.

After that, PCA calculates the covariance matrix $cov(X_\mu) = X_\mu^T X_\mu$ which is a $n \times n$ matrix.

Next, it calculates the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors of the covariance matrix are vectors that point to the directions with the maximum spread of points. If we project the data on a subspace constructed by a number of eigenvectors (as bases), we will end up with projections of the original points with minimum information loss.

To do so, we first extract the eigenvectors $\mathbf{v}$ and eigenvalues $\lambda$. Next, if we want to project the data on a subspace of dimension $M$, we select the eigenvectors corresponding to the $M$ largest eigenvalues and construct a matrix $T = [v_1 \ v_2 \ \dots \ v_M]$.

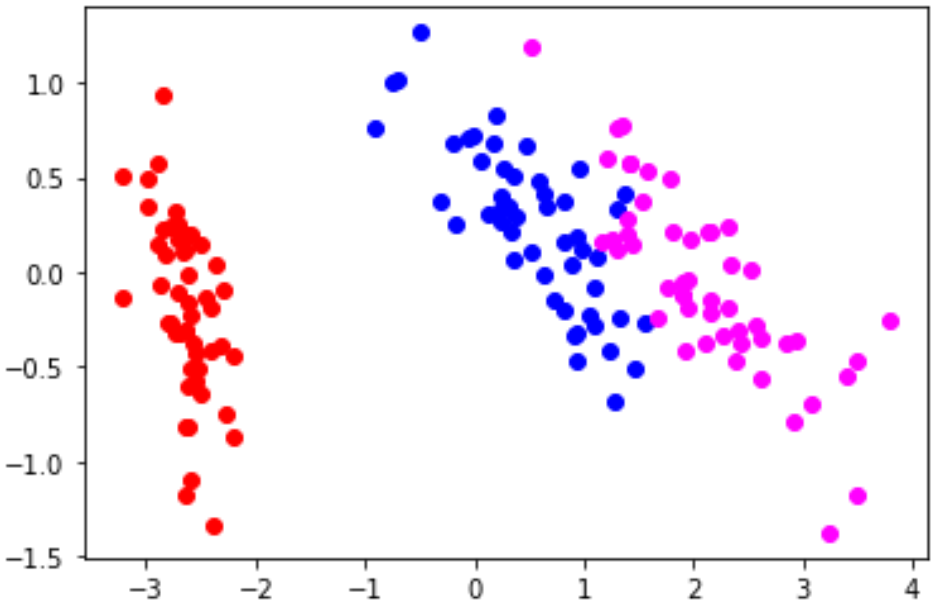
To get the projected points $X_M$, we simply multiply the transpose of matrix $T$ ( $T^T$) by the original mean-centered points matrix $X_\mu$; $X_M = T^T X_\mu$.

Finally, to plot these points on a 2-D plane, you have to choose $M = 2$, or on a 3-D space, you have to choose $M = 3$. Then you can scatter plot the points and see the results.

## Question 1 - PCA (Hard)

The Iris  data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a $150 \times 4$ `numpy.ndarray` The rows being the samples and the columns being:  Sepal Length, Sepal Width, Petal Length and Petal Width.

In this question, you will load the Iris dataset from the `sklearn.datasets` module using the `load_iris()` function. This function returns an object with the data $X$ stored in `load_iris().data` and the labels stored in `load_iris().target`. Your job is to use the PCA technique explained above to project the 4-D data of this dataset onto a 2D plane and scatter plot the points with each sample have a different color based on its label (since there are 3 labels, you will use 3 colors). An example of the output is shown below:




**Side Note**: you need to install the library scikit-learn in order to load the dataset. Also, please write the PCA code yourself. Do not use already written implementations of the algorithm in scikit-learn or any other library. Any answer that doesn't follow the shown steps above will not be considered correct.