Modern Machine Learning
Computer Assignment #7

1. *PCA for dimensionality reduction:* Implement the PCA algorithm for data compression.

   Background: The PCA algorithm can be summarized in the following steps that are performed repeatedly until convergence:

   - **Input:** Training set $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M] \in \mathbb{R}^{N \times M}$.
   - *Mean normalization:* Perform $\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{M}\sum_{j=1}^{M}\mathbf{x}_j$
   - *Eigendecomposition:* Perform eigendecomposition to the correlation matrix $\mathbf{C} = \frac{1}{M}\mathbf{X}\mathbf{X}^T = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$
   - Stack the k eigenvectors associated with the k largest eigenvalues (sorted in descending order) and stack them into a matrix $\mathbf{U}_k \in \mathbb{R}^{N \times K}$
   - Find the k-dimensional representation of $\mathbf{X}$ as $\mathbf{Y}_k = \mathbf{U}_k^T\mathbf{X}$ (where $\mathbf{X}$ is assumed to be mean normalized)

   The data used in this exercise is a portion of the MINST database. It contains handwritten digits 0-9. The size of the images is 28 × 28 pixels, these pixels are vectorized producing features of size 1 × 784. Your program will employ PCA to reduce the dimension of the feature vectors from 1 × 1024 to 1 × 100. Examples of the MNIST images can be seen in Figure 1.

   Python files: The script PCA_example.py uses the Python function *princomp*$(\cdot)$ to perform dimensionality reduction. In this assignment, you could use 'scipy' package to load MNIST dataset. You are allowed to use the functions of the 'numpy' package, .e.g., 'linalg.eig()', to derive the eigenvalues and eigenvectors.
   Submission guidelines: Your submission should include:

Figure 1: Example images from MINST database

- A unique **zip folder**, which should include a modified version of PCA_example.py, which includes your own version of $princomp(\cdot)$:
  - $evectors, representation = princomp(MNIST, numpc)$ has two inputs and two outputs.
  - Inputs: 'MNIST' is the input dataset with dimension $M \times N$, where M is the number of samples, N is the number of features. M = 1000 and N = 784 in this assignment. 'numpc' is the number of principal components.
  - Outputs: 'evectors' consists of 'numpc' eigenvectors corresponding to the 'numpc' largest eigenvalues, the dimension of which is $M \times numpc$. 'representation' is the representation of MNIST, the dimension of which is $numpc \times N$.

  Please rename the modified file PCA_example.py replacing the word 'example' in the provided script with your last name. For example PCA_smith.py. **This should be the main function**.
- A pdf file with a figure of the error $\frac{1}{M}\|\mathbf{X}_{norm} - \mathbf{X}_{rec}\|_F^2$ as a function of the number of principal components k, where k starts from 10 and ends at 320 with step being 10.

MATLAB files: The script PCA_example.m uses the matlab function $pca(\cdot)$ to perform dimensionality reduction.
Submission guidelines: Your submission should include:

- A unique **zip folder**, which should include a modified version of PCA_example.m and the following functions that implement your own version of PCA:
  - $\mathbf{X}_{norm}$ = normalize_features($\mathbf{X}$), where $\mathbf{X}_{norm}$ is the mean-normalized version of matrix $\mathbf{X}$
  - $[\mathbf{U}, \mathbf{\Lambda}]$ = myeig($\mathbf{X}_{norm}$) is a function that returns the matrix $\mathbf{U}$ containing the eigenvectors associated with the eigenvalues of the diagonal matrix $\mathbf{\Lambda}$. Recall that eigendecomposition is performed over the correlation matrix of $\mathbf{X}_{norm}$ (equivalent to the covariance matrix of $\mathbf{X}$)
  - $[\mathbf{Y}_k]$ = project_data($\mathbf{U}, \mathbf{X}_{norm}, k$) is a function that returns the projection of the data onto the k eigenvectors associated with the k largest eigenvalues, which is defined as $\mathbf{Y}_k$.
  - $[\mathbf{X}_{rec}]$ = recover_data($\mathbf{U}, \mathbf{Y}_k, k$) is a function that returns the reconstructed data $\mathbf{X}_{rec}$ using the projection $\mathbf{Y}_k$

  Please rename the modified file PCA_example.m replacing the word 'example' in the provided script with your last name. For example PCA_smith.m. **This should be the main function**.

- A pdf file with a figure of the error $\frac{1}{M}\|\mathbf{X}_{norm} - \mathbf{X}_{rec}\|_F^2$ as a function of the number of principal components k, where k starts from 10 and ends at 320 with step being 10.