EEE 6608: Machine Learning and Pattern Recognition

Final assignment

Due date: April 09, 2022 (the portal will close at 10am**)**

Final assignment instructions

1. Load the digits dataset: `from sklearn.datasets import load_digits`
2. Implement principal component analysis (PCA) using `svd()` or `eig()`. Note that you need to choose the reduced dimension such that at least 90% of the energy is captured in the new coordinate system. See class lecture on how to do it.
3. Implement k-means clustering algorithm from scratch.
4. Compare clustering performance **with and without** PCA:
    a. Variation of clustering performance with number of samples
    b. Variation of clustering performance with number of clusters K
    c. Variation of clustering performance with different initializations of the cluster centers (show at least 5)
5. For the performance index, you must use the loss function of k-means objective and the Calinski-Harabasz Index. You can add more if you feel like so.
6. You will have to submit a report on this assignment that will contain your **code, result plots and a brief discussion**. You can add other sections in the report if you deem necessary.

```
import numpy as np
import h5py
def load_dataset():
    train_dataset = h5py.File('datasets/train_happy.h5', "r")
    test_dataset = h5py.File('datasets/test_happy.h5', "r")

    train_set_x_orig = np.array(train_dataset["train_set_x"][:])
    train_set_y_orig = np.array(train_dataset["train_set_y"][:])

    test_set_x_orig = np.array(test_dataset["test_set_x"][:])
    test_set_y_orig = np.array(test_dataset["test_set_y"][:])

    classes = np.array(test_dataset["list_classes"][:])

    train_set_y_orig = train_set_y_orig.reshape((1, train_set_y_orig.shape[0]))
    test_set_y_orig = test_set_y_orig.reshape((1, test_set_y_orig.shape[0]))

    return train_set_x_orig, train_set_y_orig, test_set_x_orig, test_set_y_orig, classes
```