CSC4007 – Advanced Machine Learning
Vien Ngo, EEECS, QUB

Lab 8: K-Means Clustering with Cross-Validation and Applications

In this lab, we will use k-fold cross-validation to select the hyper-parameter k for k-means clustering (note that the two hyperparameters k in these two algorithms are different). We will also investigate some applications of clustering for: image compression or color quantization. An example code (using scikit-learn library) is also provided to see how k-means clustering can be used to discover non-linear cluster boundaries if using with other advanced machine learning techniques.

The implementation is with Jupyter Notebook. We will use only Numpy and Matplotlib libraries (some demonstrating example uses scikit-learn). This lab will run on week 8. If you could not finish them during this lab, you can finish at home. Through this lab, you will practice to write code to further understand and re-implement the k-means clustering algorithm learned in Lecture 06.

STEP 1: Download these files: *clustering2.ipynb* (main code) and *qub.jpg* (image file).

STEP 2: Open the notebook file: "*clustering2.ipynb*" and learn how to write code in Python (using Numpy only) to implement the k-means clustering algorithm with cross-validation. (The code also contains an example of using scikit-learn to apply k-means for discovering non-linear cluster boundaries).

STEP 3: Learn example applications of k-means in "*clustering2.ipynb*" for image image compression/color quantization (there are both code by scikit-learn and code developed from scratch).