CS-345/M45 Lab Class 2 **Release date: 15/10/2019**
Due date: 30/10/2019

This lab is about utilizing unsupervised learning to cluster data. We will be implementing the k-means and GMM clustering algorithms on some example data by adding our own code to a skeleton jupyter notebook.

☐ Task 2.1 – Fisher Iris Dataset

The first task is to get used to the provided dataset and explore the observed features. The data is contained in the zip file `Iris.zip` on BlackBoard, download and extract the zip file to your local directory.

- Load the data and visualise the data with a scatter plot.

☐ Task 2.2 – k-Means

The second task is to cluster our data with k-means. The purpose of unsupervised clustering is to attempt to identify groupings within the data in a data-driven approach. We will use all of the feature dimensions within the Iris dataset in order to cluster the observations we have.

- For this task we will implement the k-means algorithm utilising the `lab2.ipynb` notebook, clustering our data into k clusters, before visualising the results.

☐ Task 2.3 – Gaussian Mixture Models

The third task is to cluster our data with GMMs. For this task we will implement the GMM algorithm utilising the skeleton notebook, clustering our data with the posterior probabilities of g Gaussian distributions.

- Use the skeleton code from `lab2.ipynb` to implement the GMM algorithm and visualise the results.
- You will need the following equations from the slides on GMM clustering:

$$P(j)^{new} = \frac{1}{N}\sum_{n}^{N} p^{old}(j|x^n) \tag{1}$$

$$\mu_j^{new} = \frac{\sum_n^N p^{old}(j|x^n)x^n}{\sum_n^N p^{old}(j|x^n)} \tag{2}$$

$$\sum_j^{new} = \frac{\sum_n^N p^{old}(j|x^n)(x^n - \mu_j^{new})(x^n - \mu_j^{new})^T}{\sum_n^N p^{old}(j|x^n)} \tag{3}$$

☐ Challenge Task 2.4

Look into the effect of varying the hyper-parameters. Try increasing the number of k or g. Try a new dataset; Kaggle.com has a plethora of public datasets for machine learning applications. See if you can find an interesting one that is suitable for clustering and try it out.