# COMP4702/COMP7703 - Machine Learning

# Prac 4 – Clustering

**Aims:**

- To complement lecture material in understanding clustering techniques.
- To gain experience with simulating and implementing these techniques in software.
- To produce some assessable work for this subject.

**Procedure:**

**k-means Clustering and Gaussian mixture models**

In lectures, we have discussed k-means clustering and Gaussian mixture models. Note that Weka contains implementations of both k-means and fitting a Gaussian mixture model using the EM algorithm (under the "Cluster" tab).

Matlab also includes implementations of these in the "Statistics and Machine Learning Toolbox". Search the matlab help for some nicely documented examples. Reading through these examples is highly recommended.

→ **Q1**: Apply the k-means clustering algorithm (as implemented in Matlab or something else if you prefer) to the heightweight dataset (first and second columns). Plot the resulting cluster centres together with the data. Colour the data according to class label (third column; hint: sort the data!).

**Mean Shift Clustering**

→ **Q2:** In matlab, implement the mean shift clustering algorithm as discussed in lectures and papers. Hand-in your code for this question. To do this, use a "flat" kernel function (you will need to specify the value for the radius parameter, $\lambda$). You can choose to implement as either a "blurring" or "non-blurring" process.

→ To test your algorithm, create some 2-D datasets using matlab's Gaussian random number generator randn:

```
a = randn(200,2);
b = a + 4;
c = a;
c(:,1) = 3*c(:,1);
c = c - 4;
d = [a; b];
e = [a; b; c];
plot(a(:,1),a(:,2),'+');
hold on
plot(b(:,1),b(:,2),'o');
plot(c(:,1),c(:,2),'*');
```

→ **Q3:** For each of the datasets ‘d’ and ‘e’ above, run your algorithm with three different, suitable values of λ.

- Plot the cluster centres over the plot of the data (produced similar to that shown above), for one of the “typical” run results from your algorithm.
- Comment on any variability between results of your runs across different values of λ.