

COMP4702/COMP7703 - Machine Learning

Prac 5 – Clustering

Aims:

- To complement lecture material in understanding dimensionality reduction and 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

- **Q1:** In matlab, implement the k-means clustering algorithm from lectures (fig 7.3 in the text). Hand-in your code for this question.
- 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),'*');`
- **Q2:** For each of the datasets 'd' and 'e' above, run your algorithm with $k=2,3$ and 10 . You should run your algorithm several times for each value of k (since the cluster centres are initialized randomly in our implementation).
 - 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.
 - If possible, comment on any variability between results of your runs across different values of k .

Gaussian mixture models and the EM algorithm

Note that Weka contains implementations of both k-means and fitting a Gaussian mixture model using the EM algorithm (under the "Cluster" tab).

- Experiment with these using one or more of the datasets we have used in previous pracs.