Aston University
Machine Learning

Unsupervised Learning (1)

Learning Outcomes:
In the current unit, we studied two clustering algorithms: the k-means algorithm and the EM algorithm for optimizing the parameters of a mixture model. In this lab, we will deepen our understanding of the algorithms by implementing them and applying them to some test data sets.

Instructions:
Download the datasets pop1.csv – pop4.csv. Each contains a population of 2D variables, generated from a mixture of Gaussians. We will try to separate each into two clusters using the k-means and the EM algorithm.

Create a scatter plot of each dataset. What do you think the clusters should be? Do you foresee any issues when applying the clustering algorithms?

Task 1:
In this task, we will implement the k-means algorithm. Recall that the pseudocode for k-means is as follows:

```
BEGIN
    INITIALISE k cluster centres
    REPEAT UNTIL converged DO
        FOR EACH data point DO
            ASSIGN data point to closest cluster centre
        OD
        FOR EACH cluster centre DO
            REPLACE cluster centre with mean of assigned data points
        OD
    OD
END
```

A few things to keep in mind when implementing the algorithm:

- k (the number of cluster centres) should be a parameter of your algorithm.
- A cluster centre is a point in feature space. Your cluster centres, therefore, should have the same dimensionality as your features. In lectures, we discussed a good method for choosing initial cluster centres.
- We can test for convergence either by testing whether the cluster centres remain the same in two iterations or by testing whether the allocation of data points to clusters remains the same. The latter is easier in practice as the former requires the comparison of double valued vectors.
- When we refer to “closest” in k-means, we mean closest in Euclidean distance in feature space.

Once you have implemented the k-means algorithm, apply it to the test data sets with k=2. Plot your results, giving the points allocated to each cluster a different colour. Are the results what you expected? How can you explain them?

Task 2 (Challenging):
In this task, we will implement the EM algorithm to fit Gaussian Mixture models to the datasets.

We are going to try to use what we have learned in lectures to interpret the information in the course text: *Murphy, Machine Learning : A Probabilistic Perspective*. You should be able to access the e-book using Aston LIS's Smart Search (https://www2.aston.ac.uk/library/smartsearch).

You can find a description of the EM algorithm for fitting mixture models in the course text starting on p348 and for GMMs starting on p350. The key things to remember are that:

- The algorithm proceeds iteratively, like k-means. We start off by choosing some initial model parameters (see the lecture notes for the model parameters for each distribution) and then iteratively perform the **E step** and the **M step** until convergence.
- The formal aim of the **E step** is to calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$, but we don't actually need to calculate it directly. Instead, we just need to calculate the responsibilities ($r_{i,k}$) for use in the **M step**.
- The **M step** uses the responsibilities to update model parameters. For a GMM, the maximum likelihood estimations of these parameters can be found in equations 11.28, 11.31 and 11.32.

Again, once you have implemented the EM algorithm, apply it to the test data sets to fit 2 Gaussians to the data. For your visualisation, assign each point to the cluster with highest responsibility. Are the results what you expected? How can you explain them?