Aston University
Machine Learning

Portfolio Task 2: Unsupervised Learning

Instructions:
In this assessed task, you will be applying algorithms from the second family of machine learning techniques covered in this module: unsupervised learning. The aim of this task is to test your ability to apply machine learning algorithms to well-specified tasks and to evaluate the performance of these algorithms and to use this evaluation to improve performance.

Details:
Follow the instructions below to complete the portfolio task. The task requires you to carry out some implementation in Python and to provide a short written justification of your choices, of maximum 250 words. The recommended format for submission is a Jupyter notebook, integrating your code and written justification.

Marking:
This portfolio task is worth 15% of the overall module mark.

The mark scheme for the task is as follows:

- **50-59** Solution approaches have been applied to both sub-tasks and, where requested in the task, their performance measured. The approaches taken are broadly correct but may have some flaws in application or methodology. Model evaluation and a justification of chosen approaches have been attempted but shows limited understanding.
- **60-69** Justification in sub-task 1 shows clear understanding of the properties of the chosen algorithms. Multiple solution approaches (algorithms/models/parameter sets) have been applied to the problem in sub-task 2 and have undergone evaluation. Justification for the selected approach is evidence-based and well presented.
- **70-79** The methodology used to compare solution approaches for sub-task 2 is carefully designed and leads to well-supported conclusions. Clear understanding of experimental design is demonstrated.
- **80+** As above, but with additional evidence (for both sub-tasks) of some or all of: attention to quality throughout the implementation, thorough understanding in experimental design, excellent justification.

No specific descriptors are provided for marks below the threshold of 50. Marks in the range **0-49** are allocated where the submitted work has not reached the expectation for the threshold descriptor.

Sub-task 2.1:
Download the file cluster1.csv from Blackboard. It contains 500 data points. Each data point has two features.

Using Python, apply k-means to partition the data set into three clusters. Plot a graph of the resulting clusters.

Your (fictional) colleague claims to have done the above multiple times. They claim that in their experience, k-means always converges quickly on this dataset and that the graphs that they get after each run of the algorithm (with different random seeds) show very similar clusters. They say that this shows that the three clusters found with k-means are the best possible clusters within this dataset. Do you agree with your colleague? Justify your answer based on your graph, and on your understanding of the k-means algorithm.

Sub-task 2.2:
Download the file cluster2.csv from Blackboard. It contains 1000 data points. Each data point has two features. You do not know anything about how many clusters to expect in this data. You also do not know whether the k-means algorithm or a GMM would be the best way to cluster this dataset.

Based on the principles discussed in lectures, design a methodology to choose an appropriate model and number of clusters for this dataset. Implement your methodology using Python. Explain your methodology and justify your choice of model and number of clusters. You may want to produce graphs to help support some aspects of this justification.