ECM3420 - Learning from Data

Coursework 2: Clustering

Reflection

Cluster Analysis

In Question 4, I observed the clustering differences between the Standard K-Means algorithm provided by sklearn and the Incremental K-Means algorithm implemented in Question 1. The algorithms were applied to the iris dataset, after standardisation.

Looking at the results for Standard K-Means, we observe that Label 2 was clustered perfectly (all 50 data points with this target label are contained in Cluster 1). However, misclassification occurs between Cluster 2 and Cluster 3; Cluster 2 contains 39 of the 53 Label 1 data points (73.6% accuracy), and Cluster 3 contains 36 of the 47 Label 3 data points (76.6% accuracy).

With Incremental K-Means, we obtain similar results; Label 2 data points are clustered almost perfectly (98% accuracy), and similar misclassification occurs between the other two clusters. The accuracy of remaining clusters is slightly better, at 74% and 84% respectively. We expect this improvement in accuracy, but at the cost of a slightly more expensive algorithm (from the incremental updates) [1].

Now looking at the comparison figure, these results make sense. The graphs compare how data points are clustered, plotting them against two input attributes. It shows that the points in Clusters 2 and 3 have a smaller Euclidean distance between them, whilst points in Cluster 1 are significantly further away from the other clusters. This (partly) justifies why the misclassification occurs between Clusters 2 and 3 in both algorithms.

Finally, the Jaccard scores verify our results. In both versions the Incremental K-Means has a slightly higher score, since it was slightly more accurate with its clustering.

References

[1] Chakraborty S., Nagwani N.K. (2011) Analysis and Study of Incremental K-Means Clustering Algorithm. In: Mantri A., Nandi S., Kumar G., Kumar S. (eds) High Performance Architecture and Grid Computing. HPAGC 2011. Communications in Computer and Information Science, vol 169. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22577-2_46