Applying Clustering Algorithms

Yuxuan Zhang

Westcliff University

AIT 508

Instructor: Professor Aliah Algassmi

Date: Jan 10, 2026

Applying Clustering Algorithms

The code and plots for this assignment are submitted separately in a Jupyter notebook. It can also be viewed in the assignment github repo https://github.com/yuki172/ait508-assignments/tree/main/week%201

In this assignment, I experiment with different clustering algorithms on how they organize observations in a given dataset. More specifically, I use k means clustering, agglomerative hierarchical clustering and Gaussian mixture to identify clusters in a dataset obtained from the iris dataset by extracting a subset of the features. In this report, I will discuss the experiment setup, results, behaviors specific to each algorithm and their differences.

The dataset is obtained from the iris dataset by extracting the first two features. Since the iris dataset naturally has 3 clusters, this number is used as the default value for each algorithm. However, the number of clusters is varied for each algorithm to see how each responds to the variation. For preprocessing, we perform a Z score normalization so all columns have mean 0 and variance 1. This is important for the algorithms we use because all of them are sensitive to scale.

K-means is very sensitive to outliers since the distance used is the squared Euclidean distance and outliers will distort the clusters by either shifting an existing cluster or being assigned to their own clusters. For hierarchical clustering, the sensitivity to outliers depends on the type of linkage used. For example, average linkage makes the algorithm moderately sensitive since all distances between the clusters need to be taken into account. On the other hand, single linkage is very sensitive to outliers due to possible chaining, causing dissimilar clusters to be merged. In comparison, Gaussian mixture is not as sensitive to outliers since it can - or it is allowed to - assign low probability to outliers. However, extreme outliers can still distort the distributions by affecting the mean and variance as the algorithm tries to cover such observations.

For k-means, the number of clusters is very important to the effectiveness of the algorithm. In the current example, if the number of clusters is misspecified, the resulting clusters will not correspond to - but may be subsets of - the natural clusters. This is one of the challenges for using k-means and there are methods that help make the choice of k such as the Gap statistic mentioned in Chapter 14 of Hastie et al. (2009). Hierarchical clustering does not require specifying the number of clusters when running the algorithm, but one needs to make a choice of the cut point and thus the number of clusters when making use of the result. Another parameter that’s very important to the algorithm's effectiveness is the type linkage type. As discussed in Hastie et al. (2009) and briefly alluded to in the previous section, with different linkage types, the clusters may be very different, which presents challenges in the interpretation of the clusters. Finally, Gaussian mixture also requires specifying the number of components, the correctness of which it relies on to identify the underlying distributions.

As mentioned above, all three algorithms produce similar clusters when the number of clusters is correctly specified. Among them, Gaussian mixture performs the best due to its flexibility by learning the mean and covariance of the distributions. K means clustering necessarily gives spherical clusters - due to the assignment step, which may not be suitable for this data. Hierarchical clustering with the within cluster variance linkage (Ward, 1963) produced tight clusters but is less able compared to Gaussian mixtures to learn overlaps between the natural clusters.

When fitting with the number of clusters as 3, all three algorithms found very similar clusters which approximately correspond to the natural clusters in the dataset, namely, setosa, versicolor and virginica. More precisely, the setosa cluster is identified perfectly while all methods misspecified some observations in versicolor and virginica due to the overlap in the natural clusters. This happens because none of the algorithms is designed to handle such observations.. Overall, with the known correct number of clusters, the algorithms performed well in identifying the clusters. In practice, as described above, finding the correct or reasonable number of clusters presents a challenge as an incorrect number of clusters can lead to either over or under merging which will harm interpretation. Also, although not encountered for the current case, since k-means and Gaussian mixtures are sensitive to initialization, it is a good idea to perform multiple runs, which can pose a problem for time and computing resource considerations if the dataset is large.

References

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845