Clustering Evaluation

There are a large number of clustering performance evaluation tools available. If the labels of the clusters are known then there are metrics like the adjusted Rand index and the Normalized mutual information score. When the labels for the clusters are not known the silhouette score and Davis-Bouldin may be used. To see other methods, look to scikit-learn's cluster evaluation page. We use silhouette scores in these materials because the numerical range is intuitive.

- -1 - indicates incorrect clustering
- 0 - highly overlapping clusters
- 1 - dense well-separated clusters

Historically the concept of inertia or the within cluster sum-of-squares has been used for model selection. Generally, it is discussed in the the context of the *k*-means algorithm, but the metric has a number of drawbacks as far as model selection. Inertia makes the assumption that clusters are convex and isotropic. More importantly though it is not a normalized metric. We recommend any of the methods provided by scikit-learn metrics submodule. Another method that is often discussed in the same context as inertia is the elbow method as a heuristic for determining the number of clusters. These same type of plots can be drawn with other metrics, but they are subjective and can be unreliable.

Important: When thinking about the appropriate number of clusters it is possible that there is more than one answer depending on the perspective. If we clustered visible light on planet earth at night in euclidean space then there would be valid cluster assignments at the continent level, at the country level and at the city level. Keeping this in mind can help ensure you model pipelines are flexible. It can help to think about clusters labels as a probabilistic assignment rather than a hard label.

When comparing across algorithms it is important that the comparisons be between results that contain roughly the same number of clusters. While not definitive, there have been some hints with the running wholesale example that an appropriate number of clusters is around five. For several of the algorithms discussed we can set this value directly. Another alternative is trying to visualize the data using the dimension reduction techniques. The labels designated with colors can provide insight into the cluster assignments. This technique is generally appropriate for the EDA part of the workflow.