K-Means

The k-means algorithm is a technique to identify clusters of observations based on their features. Features are treated as coordinates in n-dimensional space. The goal is to identify k partitions of observations such that the with-cluster sum of squares is minimized. Otherwise stated, given k centroids that mark the center of each natural cluster, each observation in sample S can be assigned the label of the nearest centroid. To do this, the statistical objective is to:

$$argmin \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_{i,j} - \mu_j||^2$$

where the goal is to find the minimum value of the equation ($argmin$) that is defined as the sum of the distance of each point i in cluster j to its corresponding centroid of j. Distance is calculated in terms of all input features x and the j^{th} cluster centroid μ.

The technique is fairly straight forward to optimize and is one that is iterative as shown in the pseudocode below:

```
Initialize k centroids
Repeat until convergence:
  Calculate distance between each record n and centroid k
  Assign points to nearest centroid
  Update centroid coordinates as average of each feature per cluster
```

The first step involves setting k number centroids that are within the same scale as the features in the sample. For each point, calculate the distance to all centroids, then assign each point to the closest centroid. This is known as the *assignment* step. With the assignments to each of the k clusters, *update* the centroid coordinates for each cluster, the re-run the assignment step. Repeat the assignment and update steps until cluster assignments no longer change.

Under the hood

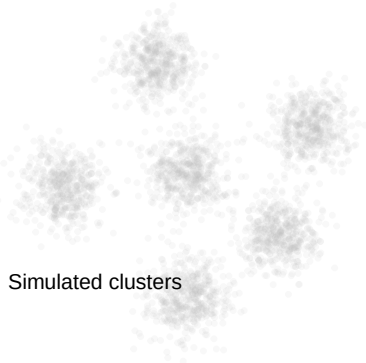
K-means are commonly used for segmenting customers to help characterize user needs, identify gene sequences that are similar, among other things. However, while it might not be apparent from the mathematics, k-means algorithms may suffer from convergence on local optima that yield unstable clusters. The algorithm will converge, but the results need to be and replicable in order for a cluster to be accurately identified. Over the last 70 years, various techniques have emerged to address stability.

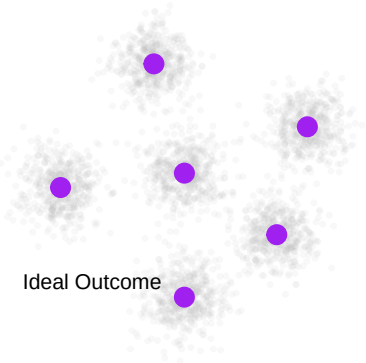
Furthermore, there tends to be lack of consensus on approaches to test the stability of identified clusters and it has only been in recent memory that stability techniques are being developed.

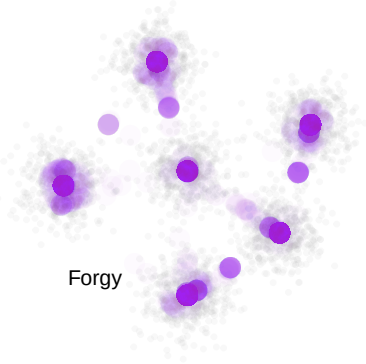
While the k-means algorithm itself is quite simple to use given the `kmeans()` method and is widely used

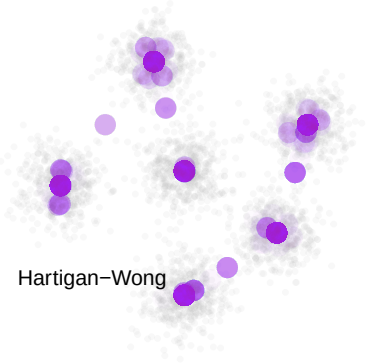
In practice

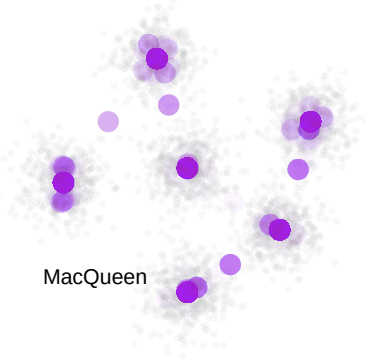
To illustate this, we will randomly generate six clusters of data in two-dimensional space. Each of these simulated clusters contain n records and a standard deviation $sd = 10$. The mean coordinates are used to place the clusters at such a distance that is sufficiently far to distinguish each cluster.


Simulated clusters


Ideal Outcome


Forgy


Hartigan−Wong


MacQueen