K-means clustering

k-means clustering is a simple and popular algorithm for the *unsupervised clustering problem*, the task of grouping a set of observations so that a group is "similar" within itself and "dissimilar" to other groups. k-means partitions n observations into k clusters, with each observation belonging to the nearest mean of the cluster. This problem is NP-hard in general, but there are heuristics which guarantee convergence to a local optimum.

The standard heuristic (known as *Lloyd's algorithm*) is the following:

Algorithm 1 Lloyd's algorithm for k-means clustering

```
generate an initial set of k means
while not converged do
    assign all data points to nearest Euclidean-distance mean
    calculate new means to as the centroids of the observations in the cluster
end while
```

There is a choice of initialization method. The *Forgy method* randomly picks k observations as initial means, while the *Random Partition method* randomly picks a cluster for each observation.

The Lloyd's algorithm is a heuristic, so it does not guarantee a global optimum. Furthermore, there exists sets of points in which it converges in exponential time. However, it has been shown to have a smoothed polynomial running time, and in practice converges quickly.