K-means Clustering

K-means defines an approach for fitting a dataset to some number of disjoint clusters. The quantity of clusters is predetermined by k. Given the following parameters

- Dataset $X = \{x_1, ..., x_n\}$
- The euclidean distance d
- Number of clusters k

we find k centers $\{\mu_1, ..., \mu_k\}$ that minimize the following cost function:

$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

where C_i represents the set of data points assigned to the i^{th} cluster.

Lloyd's Algorithm

An implementation of **K-means** that iteratively clusters data points into groups represented by a **centroid.**

- Lloyd's algorithm will *always* converge
- Will not always converge to the **optimal** solution

```
function Lloyd(k, X, dist_func) is
    centroids := select k random datapoints from X

    repeat until convergence do
        for each x in X do
            assign x to its closest neighbor

        centroids = compute new centers as the means of each cluster

    return clusters
```