k-MEANS

Clustering

Let $x_1, \ldots, x_n \in \mathbb{R}^d$. Clustering is the following problem: Partition $\{x_1, \ldots, x_n\}$ into disjoint subsets called clusters such that points in the same cluster are more similar to each other than to points in other clusters.

A clustering can be represented by a cluster map, which is a function

$$C : \{1, 2, \ldots, n\} \longrightarrow \{1, 2, \ldots, k\}$$

where k is the number of clusters.

k-Means Criterion

The k-means criterion is to choose C to minimize

$$W(C) = \sum_{\ell=1}^{k} \sum_{i : C(i) = \ell} \| x_i - \bar{x}_\ell \|^2$$

where

$$\bar{x}_\ell = \frac{1}{n_\ell} \sum_{j : C(j) = \ell} x_j$$

$$n_\ell = \#\{i : C(i) = \ell\}$$

Note that k is assumed fixed and known.

$W(C)$ is sometimes called the within class scatter, because it can be shown that

$$\star \qquad W(C) = \frac{1}{2}\sum_{\ell=1}^{k}\ \sum_{i:C(i)=\ell}\left[\underbrace{\frac{1}{n_\ell}\sum_{j:C(j)=\ell}\|x_i - x_j\|^2}_{\text{Average dissimilarity to points in same cluster}}\right]$$

Most clustering algorithms can be viewed as optimizing some measure of (dis)similarity.

Establishing $\star$ is an exercise in algebra. First observe that for any ℓ

$$\|x_i - x_j\|^2 = \|(x_i - \bar{x}_\ell) - (x_j - \bar{x}_\ell)\|^2$$

$$= \langle (x_i - \bar{x}_\ell) - (x_j - \bar{x}_\ell), (x_i - \bar{x}_\ell) - (x_j - \bar{x}_\ell)\rangle$$

$$= \|x_i - \bar{x}_\ell\|^2 - 2\langle x_i - \bar{x}_\ell, x_j - \bar{x}_\ell\rangle$$

$$+ \|x_j - \bar{x}_\ell\|^2.$$

Then

$$\frac{1}{2}\sum_{\ell}\sum_{i:C(i)=\ell}\left[\frac{1}{n_\ell}\sum_{j:C(j)=\ell}\|x_i-x_j\|^2\right]$$

$$=\frac{1}{2}\sum_{\ell=1}^{k}\frac{1}{n_\ell}\Bigg[\sum_{i:C(i)=\ell}\sum_{j:C(j)=\ell}\|x_i-\bar{x}_\ell\|^2$$

$$\left.-2\sum_{i:C(i)=\ell}\sum_{j:C(j)=\ell}(x_i-\bar{x}_\ell)^T(x_j-\bar{x}_\ell)\right\}=0$$

$$+\sum_{i:C(i)=\ell}\sum_{j:C(j)=\ell}\|x_j-\bar{x}_\ell\|^2\Bigg]$$

$$=\frac{1}{2}\sum_{\ell=1}^{k}\frac{1}{n_\ell}\Bigg[n_\ell\sum_{i:C(i)=\ell}\|x_i-\bar{x}_\ell\|^2$$

$$+n_\ell\sum_{j:C(j)=\ell}\|x_j-\bar{x}_\ell\|^2\Bigg]$$

$$=W(C).\ \blacksquare$$

k-Means Algorithm

Minimizing the k-means criterion is a combinatorial optimization problem. The number of possible cluster maps C is maps C is

$$\frac{1}{k!}\sum_{\ell=1}^{k}(-1)^{k-\ell}\binom{k}{\ell}\ell^{n} \qquad \text{(Jain and Dubes, 1998)}$$

$$\begin{cases} = 34{,}105 & \text{if } n=10,\ k=4 \\ \approx 10^{10} & \text{if } n=19,\ k=4. \end{cases}$$

There is no known efficient search strategy for this space of functions. Therefore we must resort to an iterative, suboptimal algorithm.

Recall we need to solve

$$C^{*} = \arg\min_{C} \sum_{\ell=1}^{k} \sum_{i:C(i)=\ell} \|x_i-\bar{x}_\ell\|^2.$$

Now notice that

$$\bar{x}_\ell = \arg\min_{m_\ell\in\mathbb{R}^d} \sum_{i:C(i)=\ell} \|x_i-m_\ell\|^2,$$

which can be seen by writing

$$\sum_i \|x_i-m_\ell\|^2 = \sum_i \|x_i-\bar{x}_\ell+\bar{x}_\ell-m_\ell\|^2$$

$$= \underbrace{\sum_i \|x_i-\bar{x}_\ell\|^2}_{} + 2\underbrace{\sum_i (x_i-\bar{x}_\ell)^T(\bar{x}_\ell-m_\ell)}_{}$$

independent of m_ℓ $\qquad\qquad\qquad = 0$

$$+ \underbrace{\sum_{i} \| \bar{x}_\ell - m_k \|^2}_{\text{optimized by choosing } m_\ell = \bar{x}_\ell}$$

Therefore, if we define

$$W(C, \{m_\ell\}_{\ell=1}^{k}) = \sum_{\ell=1}^{k} \sum_{i : C(i) = \ell} \| x_i - m_\ell \|^2 ,$$

we have

$$C^* = \underset{C, \{m_\ell\}_{\ell=1}^{k}}{\arg\min} \; W(C, \{m_\ell\}_{\ell=1}^{k}).$$

This suggests an iterative, alternating algorithm:

- Given C, optimize $W(C, \{m_\ell\})$ w.r.t. $\{m_\ell\}$
- Given $\{m_k\}$, optimize $W(C, \{m_\ell\})$ w.r.t C

This is the k-means algorithm:

```
Initialize $m_1, \ldots, m_k \in \mathbb{R}^d$
Repeat
    For $i = 1, \ldots, n$
        $C(i) = \arg\min_{\ell} \| x_i - m_\ell \|$
End
```

```
    End
    For $\ell = 1, \dots, k$
        $m_\ell = \frac{1}{|\{i : C(i) = \ell\}|} \sum_{i : C(i) = \ell} x_i$
    End
Until clusters don't change
```

This algorithm is also known as Lloyd's algorithm or the Lloyd-Max algorithm. The same algorithm is used in the problem of quantization.

Initialization

The k-means algorithm is highly dependent on initialization. One common strategy is to initialize $m_1, \dots, m_k$ to be randomly chosen data points. It is also common to run the algorithm several times with different initializations, and take the run with smallest $W(C)$.

Unfortunately, random initialization has some problems.

- The number of iterations can be quite large in the worst case
- The final value of $W(C)$ can be quite

- The converged value of $W(C)$ can be quite far from the optimal one.

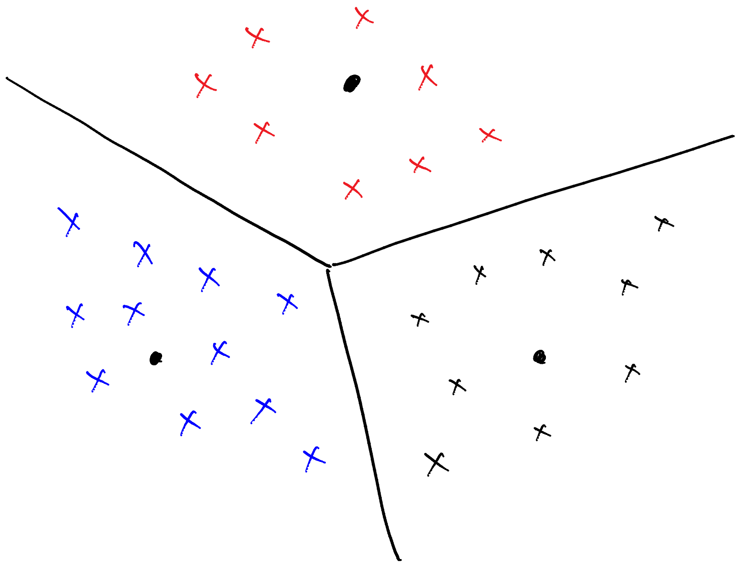
A better idea is to choose the initial $m_1, \dots, m_k$ to be far apart. A particular implementation of this idea, with guaranteed performance relative to the optimum, is called k-means++ :

1. Choose the first cluster center m_1 at random from among $x_1, \dots, x_n$
2. For each $x \in \{x_1, \dots, x_n\}$, compute $D(x)$, the distance from x to the nearest cluster center that has been selected.
3. Choose one new data point x at random as a new cluster center, with probability proportional to $D(x)^2$.
4. Repeat steps 2-3 until k centers have been selected

For more on k-means++, see the original paper by Arthur and Vassilvitskii (2007).

Cluster Geometry

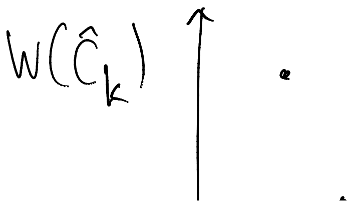
Clusters are "nearest neighbor" regions or Voronoi cells defined with respect to the cluster centers. Therefore the cluster boundaries are piecewise linear, and the clusters are convex sets.

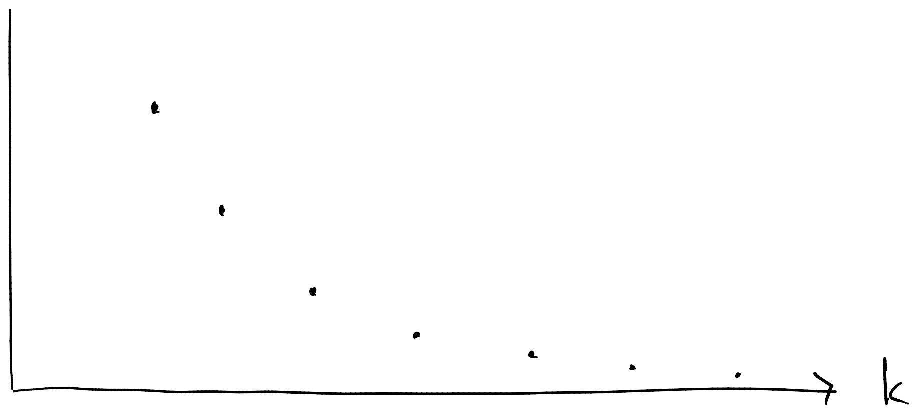


k-means will fail to identify the true clusters if at least one of them is nonconvex. k-means can be kernelized to accommodate nonconvex clusters.

Model Selection

How should k be chosen? Let $\hat{C}_k$ denote the output of k-means. One simple heuristic is to plot $W(\hat{C}_k)$ as a function of k.





The basic idea is that if k^* is the ideal cluster number, then

- If $k < k^*$, $W(\hat{C}_k) - W(\hat{C}_{k+1})$ will be relatively large
- If $k \geq k^*$, $W(\hat{C}_k) - W(\hat{C}_{k+1})$ will be relatively small

This suggests choosing k near the "knee" of the curve.

A more systematic method was developed by Kulis and Jordan in their paper "Revisiting k-means" (2007).

They suggest optimizing the following objective with respect to both C and k:

$$\sum_{\ell=1}^{k} \sum_{i:C(i)=\ell} \|x_i - \bar{x}_\ell\|^2 + \lambda k.$$

where λ is a tradeoff parameter. This criterion is derived from a nonparametric Bayesian perspective, and can be optimized (suboptimally) by a variant of the k-means algorithm.