Lecture Pattern Analysis

Part 06: K-Means

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
May 7, 2021

K-Means at a Glance

- K-means is arguably the most well-known clustering algorithm[1]
- **Hard-clustering** method, i.e., each sample gets a discrete cluster label assigned
- Idea: minimize Euclidean **Within-Cluster Distance** $W(C)$:

$$W(C) = \frac{1}{2} \cdot \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 \tag{1}$$

$$= \sum_{k=1}^{K} N_k \cdot \sum_{C(i)=k} \left\| \mathbf{x}_i - \boldsymbol{\mu}_k \right\|^2 \ , \tag{2}$$

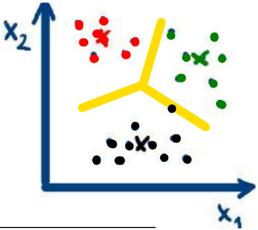
where K is the total number of clusters, $C(i)$ the cluster ID for sample $\mathbf{x}_i$, N_k the number of points in cluster k, and $\boldsymbol{\mu}_k$ the mean of all points in cluster k.

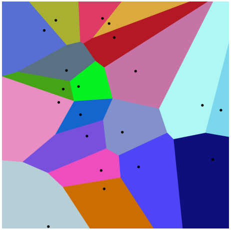
[1] Literature references are, e.g., Hastie/Tibshirani/Friedman Sec. 14.3.6 or Bishop Sec. 9.1

K-Means Algorithm

1. Initialization: set K cluster centers in sample space (e.g., randomly selected)
2. Assign each sample to the nearest cluster center w.r.t. Euclidean distance
3. Calculate the mean of each cluster from its assigned samples
4. goto 2) until convergence

- Remarks:
  - K-means is locally optimal $\rightarrow$ different initializations $\stackrel{?}{=}$ different results
  - The clusters partition the space, the partitioning is called Voronoi tesselation[2]





[2] Picture on the right is from wikipedia (CC BY-SA 4.0): https://upload.wikimedia.org/wikipedia/commons/5/54/Euclidean_Voronoi_diagram.svg

Example Run for $k = 3$, Random Starting Positions