PA 2018 - 05

Mean shift: Representative of the "Mode-seeking" algorithms = Related Application⇒clustering of Data

K-means:

- Obtain(guess) an initial distribution of cluster centers
- for each data point, identify the closest cluster center
- each cluster center is replaced by the coordinaterweise average of all data point that are closest to it
- repeat until convergence

Greedy algorithm / greedy search

it is a strictly local opination
Biggest advantage: speed
Biggest disadvantage: no brain

Two very common (dis-)similarity criterion for clustering's is the "within" or "intra" cluster distance W(C) ($\Rightarrow Eq.14.28$) and the "between" or "inter"-cluster distance B(C) ($\Rightarrow Eq.14.28$)
$\Rightarrow$ the k-Means algorithm minimizes the within-cluster distances(greedily)($\Rightarrow Eq.14.31 - 14.33$)

How can we determine a reasonable Parameter K?

- low-dimensional dataset: Look at the data
- High-dimensional data set: Need a catamitical criterion
- K-means minimize $W(C)$,
- $W(C)$ decreases for increasing k, for $k = N : W(C) = 0$ "every point is in its own cluster"
- One trick to determine K is Tibshiranis "gap statistics" stop at the k where $G(K) \leq G(k+1) - s_{k+1}^{,}$ where $G(k) = log(w(C_k)) - log(W(C_1)))$ $G(K)$ negative number
- $S_{k+1}^{,}$ is the standard der. of the outcomes for $k+1$ clusters
- A second trick is to create a reference verve from the uniform distribution consider $W(C_k^{uniform})$
- $\Rightarrow$ compute the ratio of $W(C_K))$ over $W(C_K^{uniform})$ and pick the minimum (or an early minimum)

Example for agglomentive clustering

by "Efficient Graph-based Image Segmentation"