Lecture Pattern Analysis

Part 09: Model Selection for K-Means

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
May 13, 2021

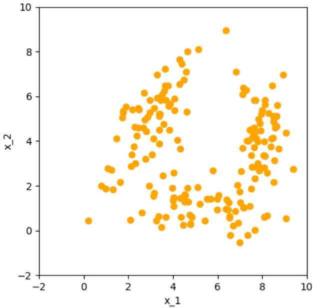
Introduction

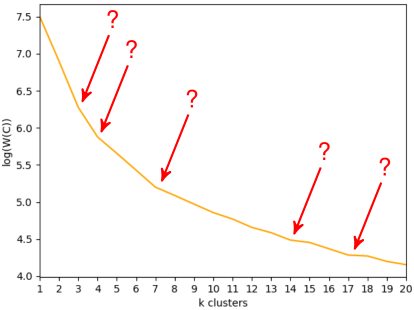
- Clustering is unsupervised, and does not provide an objective function for model selection
- So, specifically for k-means: what k shall we choose?
- Even if the application demands, e.g., the "3 most important clusters", $k = 3$ could be a poor choice if the intrinsic number of clusters is larger
- In this lecture, we investigate the **Gap-Statistics** as a statistical way to determine k[1]
- The idea is to
  - examine the k-means optimization criterion, the **Within-Cluster Distance** $W(C)$, for different k,
  - and to select the smallest k for which $W(C)$ is substantially better than the $W(C)$ of $k + 1$ clusters

[1] The gap statistics is covered in the book by Hastie/Tibshirani/Friedman Sec. 14.3.11

Examining the Within-Cluster-Distance $W(C)$

- Investigate the progression of $W(C)$ for different k
- For increasing k, $W(C)$ has to decrease (exceptions are bad local minima):





- Hence, the optimum k can not be found by searching for the minimal $W(C)$
- An alternative is the "elbow method", to search for a elbow on the curve
- However, which elbow is significant? At $k = \{3, 4, 7, 14, 17\}$?

Gap Statistics

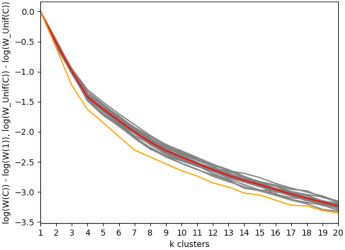
- Tibshirani *et al.* propose to relate $W(C)$ of our samples to the $W(C)$ of an artificially created reference
- This reference are clusterings of uniform sample distributions
- More specifically:
  1. Draw B sets of uniformly distributed samples (Tibshirani uses $B = 20$)
  2. On those distributions, calculate for different k the mean of the log of $W(C)$, denote the result $\log(W_{\text{unif}}(C))$
  3. For k clusters, calculate the gap $G(k)$ as the difference between the reference $\log(W_{\text{unif}}(C))$ and our log-within cluster distances $\log(W(C))$
  4. Select the optimum k as

$$k^* = \underset{k}{\operatorname{argmin}} \{k | G(k) \geq G(k+1) - s'_{k+1}\} \tag{1}$$

where $s'_{k+1} = s_k \cdot \sqrt{1 + 1/B}$ is an unbiased estimate of the standard deviation s_k of $\log(W_{\text{unif}}(C))$

Example: Within-Cluster Distances on the Uniform Distribution

- Offset-corrected $\log(W(C))$ (orange) and $\log(W_{\text{unif}}(C))$ (red), and the $B = 20$ individual reference curves (gray):



- Gaps and standard deviations for curve differences. $k^* = 3$ is selected: