## Clustering evaluation

Victor Kitov

v v kitov@yandex ru

# Silhuette coefficient<sup>1</sup>

For each object  $x_i$  define:

• *s<sub>i</sub>*-mean distance to objects in the same cluster

• *d<sub>i</sub>*-mean distance to objects in the next nearest cluster Silhouette coefficient for *x<sub>i</sub>*:

$$Silhouette_i = rac{d_i - s_i}{\max\{d_i, s_i\}}$$

Silhouette coefficient for  $x_1, ... x_N$ :

$$\textit{Silhouette} = rac{1}{N}\sum_{i=1}^{N}rac{d_i-s_i}{\max\{d_i,s_i\}}$$

<sup>1</sup>Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65.

## Discussion

- Advantages
  - The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering.
  - Scores around zero indicate overlapping clusters.
  - The score is higher when clusters are dense and well separated.
- Disadvantages
  - complexity  $O(N^2D)$ 
    - use feature space indexing or random subsampling
  - The Silhouette Coefficient is generally higher for convex clusters than other concepts of clusters
    - such as density based clusters.

#### Calinski-Harabaz Index<sup>2</sup>

- Consider K clusters. For cluster k = 1, 2, ... K define
  - nk number of objects, ck centroid, Ck indexes of objects
- Within cluster covariance matrix

$$W = rac{1}{N-K} \sum_{k=1}^{K} \sum_{x \in C_k} (x - c_k) (x - c_k)^T$$

Between cluster covaraince matrix

$$B = rac{1}{K-1}\sum_{k=1}^{K} n_k \left(c_k - c
ight) \left(c_k - c
ight)^{T}$$

• Calinski-Harabaz Index:

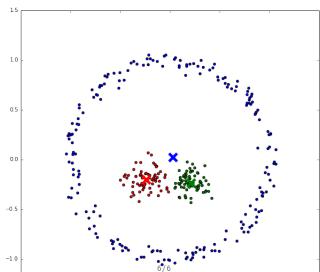
$$I = \frac{\operatorname{tr} B}{\operatorname{tr} W}$$

<sup>2</sup>Caliński, T., & Harabasz, J. (1974). "A dendrite method for cluster analysis". Communications in Statistics-theory and Methods 3: 1-27.

#### Discussion

- Advantages
  - The score is higher when clusters are dense and well separated.
  - Complexity O(ND)
- Drawbacks
  - Index favours convex clusters

## Example



Calinski-Harabaz Index will be small here.