Clustering evaluation

Victor Kitov

v.v.kitov@yandex.ru

Silhuette coefficient[1]

For each object x_i define:

- s_i-mean distance to objects in the same cluster
- d_i-mean distance to objects in the next nearest cluster

Silhouette coefficient for x_i:

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

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

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

[1] 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
  - The Silhouette Coefficient is generally higher for convex clusters than other concepts of clusters
    - such as density based clusters.

Calinski-Harabaz Index[2]

- Consider K clusters. For cluster $k = 1, 2, ...K$ define
  - n_k - number of objects, c_k - centroid, C_k - indexes of objects
- Within cluster covariance matrix

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

- Between cluster covaraince matrix

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

- Calinski-Harabaz Index:

$$I = \frac{\text{tr}\, B}{\text{tr}\, W}$$

[2] 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.
  - Fast to compute
- Drawbacks
  - Index is generally higher for convex clusters than other concepts of clusters
    - such as density based clusters.

Example

Calinski-Harabaz Index=193!