Data Mining and Machine Learning

Clustering III

Dr. Edgar Acuna
Department of Mathematics

University of Puerto Rico - Mayaguez
academic.uprm.edu/eacuna

Cluster Validation

Internal Indexes. Do not require knowing a previous assignment of classes. Mainly these indexes are Statistics based on sums of squares between clusters and within clusters. The number of K clusters is the one that maximizes or minimizes one of these indices (Milligan, G.W. & Cooper,M.C., 1985). Among the main ones are the Dunn index, the Davies-Bouldin Index, the Calanski-Harabaz index and the average silhouette width.

Determining the number of components of a mixture of distributions is the same as determine the number of clusters. So, one can use the AIC criteria (Akaike Information Criterion) and BIC (Bayesian Information Criterion) to find out the number of clusters.

Davies-Bouldin's Index (1979)

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where}$$

$$R_i = \max_{j=1\ldots nc, i\neq j}\left(R_{ij}\right), \; i = 1\ldots n_c$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d\left(v_i, v_j\right), \; s_i = \frac{1}{\|c_i\|}\sum_{x\in c_i} d\left(x, v_i\right)$$

Where,

- d(x,y) is the Euclidean distance between x and y.
- c_i is the cluster i.
- v_i is the centroid of cluster c_i
- $\|c_i\|$ refers to the norm of c_i

The best number of clusters nc would be the one with the lowest Davies-Bouldin' index.

Calanski-Harabasz index (1974)

It is defined by:
CH(k) = [B(k)/(k − 1)]/W(k)/(n − k)

where k denotes the number of clusters, and B(k) and W(k) denote the between and within cluster sums of squares of the partition, respectively.

$$B[k] = \sum_{i=1}^{k} n_i d^2(c_i, c)$$

$$W[k] = \sum_i \sum_{x \in Ci} d^2(x, c_i)$$

where ci is the centroid of the cluster Ci , c is the centroid of the whole data, and ni is the number of points in the cluster Ci.
An optimal number of clusters is then defined as a value of k that maximizes CH(k).

Silhouette plots

Silhouette plots, (Rousseeuw 1987) could be used for:

- Determining the number of clusters.
- Evaluate how good the observations were assigned into the clusters.

The **silhouette width** if the i-th observation is defined by:

$sil_i = (b_i - a_i)/ \max(a_i, b_i)$

Where, a_i is the average distance between i and all the other observations that belong to same cluster as i and b_i denotes the smallest average distance between i and the observations that belong to other clusters

.

The silhouette value goes from -1 to +1.

Silhouette plots (cont.)

The observations with a big silhouette width are well grouped while the ones with a small silhouette width tend to be placed in the middle of 2 clusters.

For a given number of clusters, K, the average silhouette width of the conglomerate configuration will be simply the average of sil_i over all observations. That is to say,

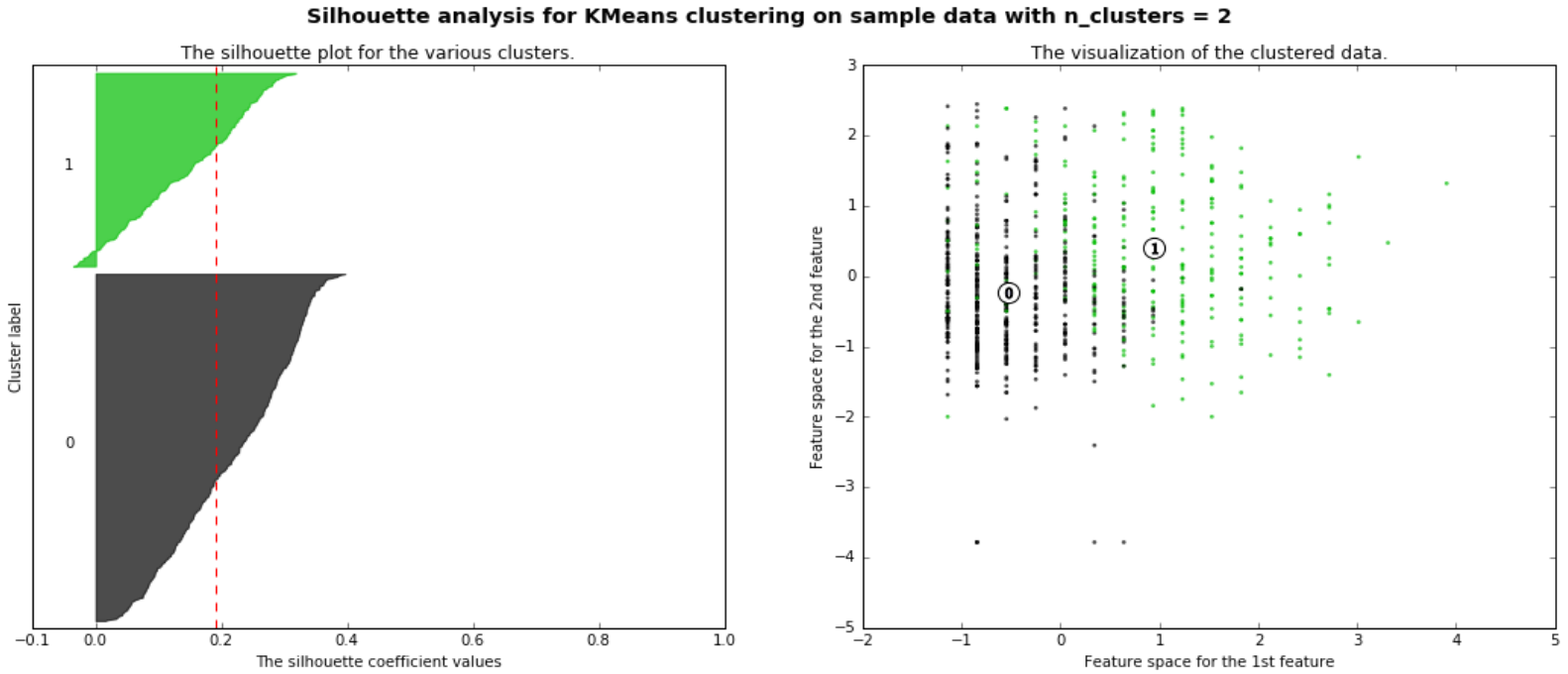
$$\bar{s} = \frac{\sum_i sil_i}{n}$$

Kaufman and Rousseeuw (1990) suggested estimating the optimum number of cluster K for which the average silhouette width is as large as possible (close to 1).

Computing Internal measures in Python

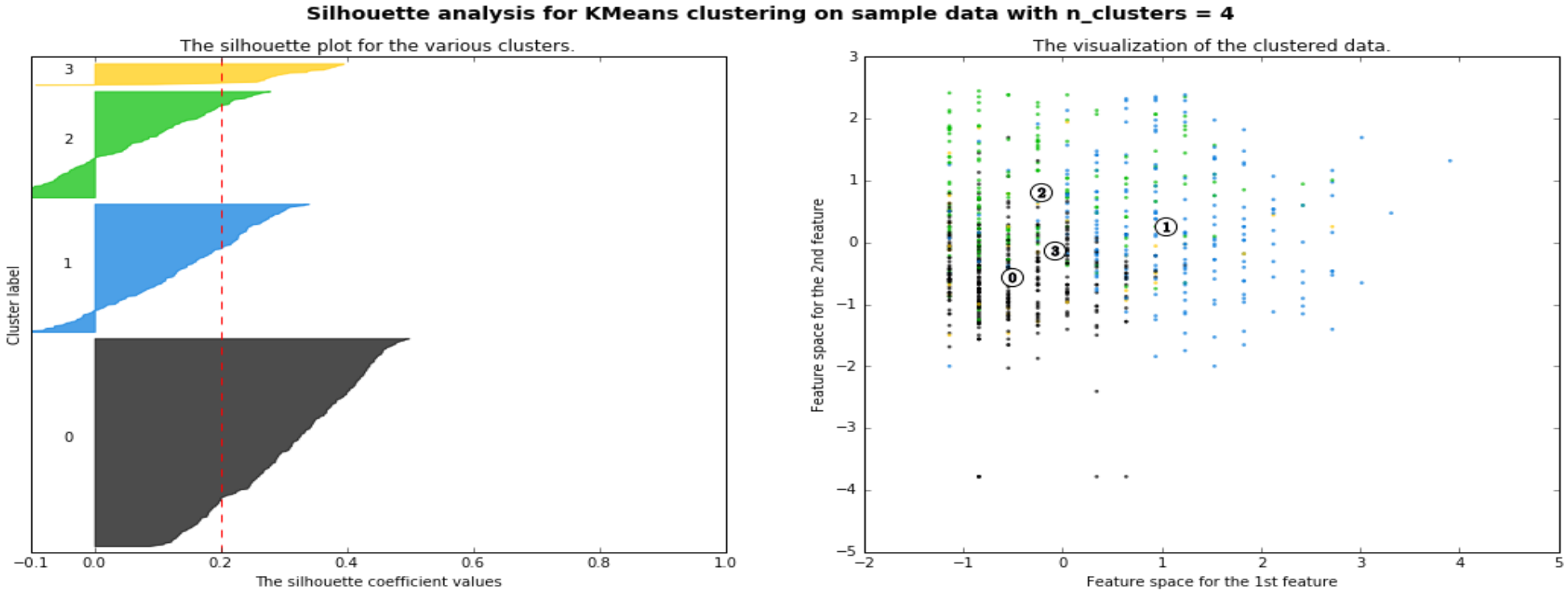
The module scikit-learn includes a submodule metrics that contains functions to compute Davies-Bouldin, Calinski-Harabasz, and Silhouette indexes.

Silhouette plots for clustering Diabetes (k=2)



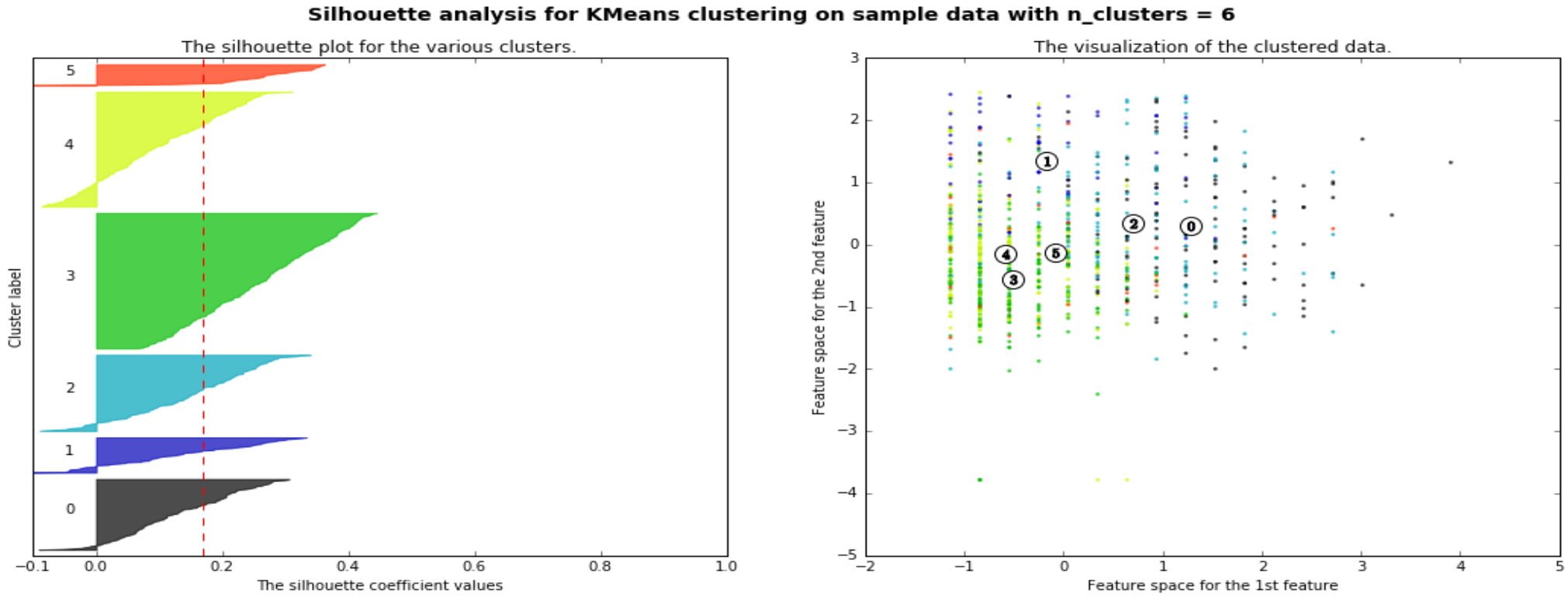
'For n_clusters =', 2, 'The average silhouette_score is :', 0.19214441257334328

Silhouette plots for clustering Diabetes (k=4)



For n_clusters =', 4, 'The average silhouette_score is :', 0.20174890536020607

Silhouette plots for clustering Diabetes (k=6)



For n_clusters =', 6, 'The average silhouette_score is :', 0.1691344540040223 6

External Indexes

Suppose we have 2 partitions of n objetos $\mathbf{x}_1$, ..., $\mathbf{x}_n$: the Rth partitions of U = {u_1, ..., u_R} and the partition of V in C-classes: V={v_1, ...,v_C}, usually one of them is known in advance. The external indexes measure the agreement between the partitions and can be expressed in terms of a contingency table with entries nij that represents the number of objects that are in both clusters u_i y v_j, i = 1,...,R, j = 1,...,C . Let

$n_{i.} = \sum_{j=1}^{c} n_{ij}$ y $n_{j.} = \sum_{i=1}^{R} n_{ij}$

which denote the sums of rows and columns of the contingency table. Let

$$Z = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}^2$$

Given the U and V partitions

a: number of pairs of objects that are in the same cluster in both U and V.

c: number of pairs of objects that are in the same cluster in V but not in U.

b: number of pairs of objects that are in the same cluster in U but not in V.

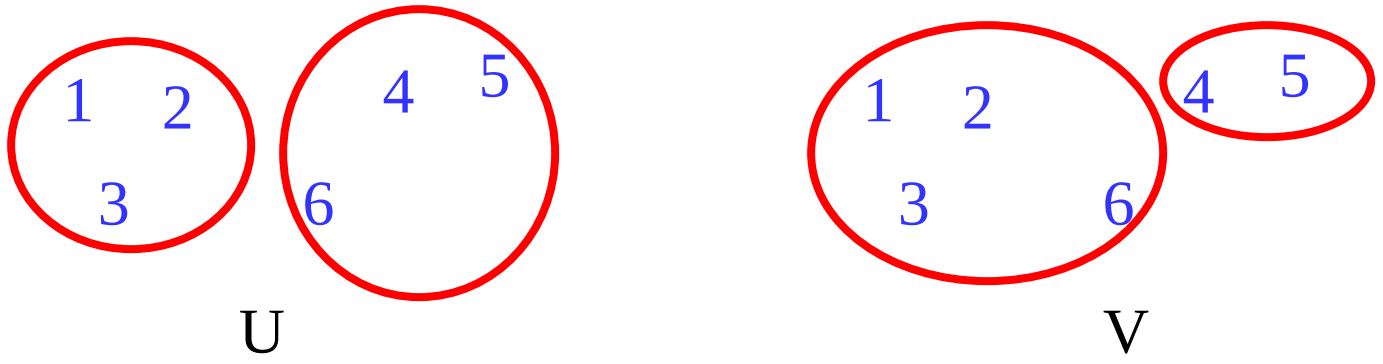
d: number of pairs of objects that are in different clusters in both U and V.

$m_1=a+b = \sum_{i=1}^{R}\binom{n_{i.}}{2}$ number of pairs of objects in the same cluster in U.

$m_2=a+c = \sum_{j=1}^{C}\binom{n_{.j}}{2}$ number of pairs of objects in the same cluster in V.

Example

a= 4, b =2, c=3, d=6, m_1=6, m_2=7, Z=65



Note that M=a+b+c+d= $\binom{n}{2}$

- Rand(1971)

$$Rand = 1 + \frac{(z - (1/2)(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2))}{\binom{n}{2}} = \frac{a+d}{\binom{n}{2}}$$

- Jaccard

$$Jac = \frac{(z-n)}{\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2 - Z - n} = \frac{d}{b+c+d}$$

- Fowlkes and Mallows

$$FM = \frac{(1/2)(z-n)}{[\sum_{i=1}^{R} \binom{n_{i.}}{2} \sum_{j=1}^{C} \binom{n_{.j}}{2}]^{1/2}} = \frac{d}{\sqrt{m_1 m_2}}$$

A Rand, Jaccard and FM value close to 1 indicates a good grouping.

Hubert's and Arabie(1985) proposed an adjusted rand Index Γ.

Γ= (rand index-expected index)/(max index- expected index)

External measures in Python

The module scikit-learn includes a submodule metrics that contains functions to compute Adjusted Rank Index and Fowlkes-Mallows.

Other Measures

- Purity
- Entropy
- Medida F
- FOM=Figure of Merit (Yeung and Ruzzo,2001)
- The Gap Statistics (Tibshirani,2000).
- **Clest (Dudoit & Fridlyand, 2002)**