智能软件开发
方向基础

第八章 聚类(clustering)
--PART1. 聚类的引入与算法评价

张朝晖

2023.3~6

序号	内容
1	概述
2	机器学习的基本概念
3	模型的选择与性能评价
4	数据的获取、探索与准备
5	近邻模型------分类、回归
6	决策树模型------分类、回归
7	集成学习------分类、回归
8	(朴素)贝叶斯模型------分类
9	聚类
10	特征降维及低维可视化(PCA, t-SNE)
11	总复习

主要内容

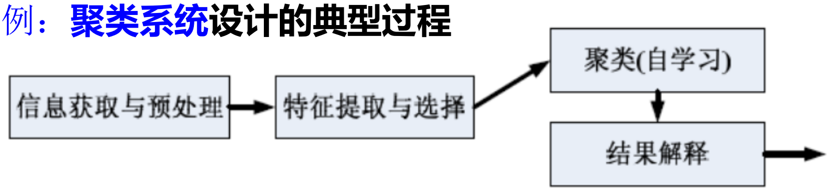
1. 聚类的引入

2. 聚类算法的评价

问题的引入

非监督式学习(密度函数估计、聚类、降维...)

例：**聚类系统设计的典型过程**



适用：

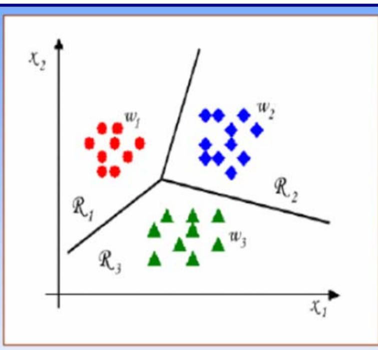
(1)大型数据挖掘：大量未标记数据训练分类器；
人工标记分组结果

(2)揭示观测数据的内在结构特性

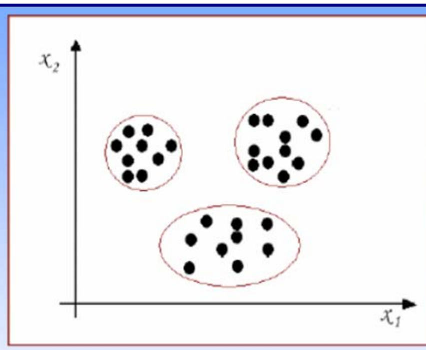
(3)是分类或其它学习任务的前驱阶段
提取数据的基本特征，进一步用于分类...

河北师范大学软件学院
Software College of Hebei Normal University

分类(Classification)与聚类(Clustering)



Given labeled training patterns, construct decision boundaries or partition the feature space



Given some patterns, discover the underlying structure (categories) in the data

什么是聚类(clustering)

To find a structure in a collection of unlabeled data.

- ***A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".***

- ***A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.***

- ***High intra-class(cluster) similarity(簇内高相似)***
  Low inter-class(cluster) similarity(簇间低相似)

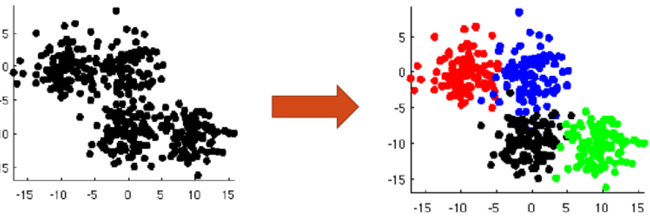
聚类问题的描述

输入： 无标签数据集 $\boldsymbol{D}=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\}, \boldsymbol{x}_i=[x_{i1} \quad \cdots \quad x_{id}]^T \in R^d$

要生成的簇的数目 k

输出： k个互不相交的簇$\{\boldsymbol{C}_l \mid l=1,\dots,k\}$

$$\begin{cases} \boldsymbol{C}_i \bigcap_{i\neq j} \boldsymbol{C}_j = \Phi \\ \bigcup_{l=1}^{k} \boldsymbol{C}_l = \boldsymbol{D} \end{cases}$$



λ_j --样本$\boldsymbol{x}_j \in \boldsymbol{D}$的簇标记，$j\in\{1,2,\dots,k\}$

数据集$\boldsymbol{D}=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\}$的标签集合$\boldsymbol{\lambda}=\{\lambda_1,\dots,\lambda_m\}$

聚类有很多典型应用，如：

- 相似功能的基因分组
- 相似政见的个体划分
- 相似主题的文档划分

…

河北师范大学软件学院
Software College of Hebei Normal University

聚类任务的遵循步骤及有关问题

- 特征选择及样本描述
  选择什么样的特征？是否需要规范化预处理？
- 近邻测度
  如何度量样本之间的“相似”或“相异”
- 聚类准则
  依赖于专家对“可判别”的解释，聚类准则应以蕴涵于数据集内的类的类型为基础。
- 聚类算法
  选择特定的算法，用于揭示数据集的聚类结构
- 聚类性能评价、结果的解释

近邻测度与聚类准则

A. 向量(点)之间的测度

(1) Dissimilarity Metric

-- dist

$d: \boldsymbol{X} \times \boldsymbol{X} \to \Re$

对于$\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{X}$, $dist(\bullet, \bullet)$须满足：

$$\begin{cases} \text{非负性：} \forall \boldsymbol{x}, \boldsymbol{y},\ dist(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \\ \text{同一性：} dist(\boldsymbol{x}, \boldsymbol{y}) = 0 \text{当且仅当} \boldsymbol{x} = \boldsymbol{y} \\ \text{对称性：} dist(\boldsymbol{x}, \boldsymbol{y}) = dist(\boldsymbol{y}, \boldsymbol{x}) \\ \text{直递性：} dist(\boldsymbol{x}, \boldsymbol{y}) \leq dist(\boldsymbol{x}, \boldsymbol{z}) + dist(\boldsymbol{y}, \boldsymbol{z}) \end{cases}$$

(2) Similarity Metric -- s

$s: \boldsymbol{X} \times \boldsymbol{X} \to \Re$

对于$\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{X}$, $s(\bullet)$须满足：

$$\begin{cases} \exists s_0 \quad -\infty < s(\boldsymbol{x}, \boldsymbol{y}) \leq s_0 < +\infty \qquad \forall \boldsymbol{x}, \boldsymbol{y} \\ \qquad\qquad s(\boldsymbol{x}, \boldsymbol{x}) = s_0 \\ s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x}) \\ s(\boldsymbol{x}, \boldsymbol{z}) s(\boldsymbol{y}, \boldsymbol{z}) \leq [s(\boldsymbol{x}, \boldsymbol{z}) + s(\boldsymbol{y}, \boldsymbol{z})] s(\boldsymbol{x}, \boldsymbol{y}) \end{cases}$$

例：向量之间的几种典型距离度量

有序属性之间的距离

对于$\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X} \subset \boldsymbol{R}^n \qquad \boldsymbol{x} = [x_1 \ \cdots \ x_n]^T \quad \boldsymbol{y} = [y_1 \ \cdots \ y_n]^T$

A. 闵可夫斯基距离 (*Minkowski distance*)

$$dist_{mk}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

加权闵可夫斯基距离 (*Weighted Minkowski distance*)

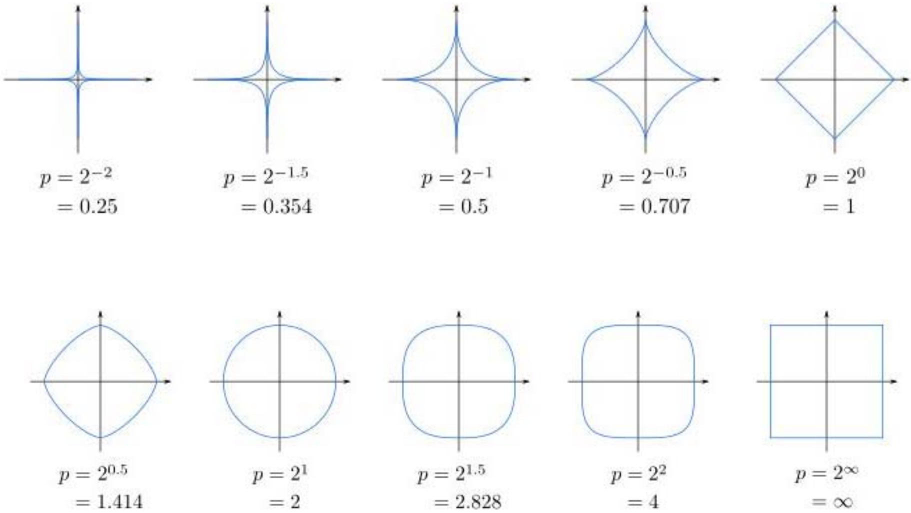
$$dist_{wmk}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} w_u |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

$$w_u \geq 0, \qquad \sum_{u=1}^{n} w_u = 1$$

B. 曼哈顿距离 (*Manhattan distance*) $\quad dist_{man}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_1 = \sum_{i=1}^{n} |x_i - y_i|$

例：不同p值下，平面坐标系内到原点的距离为1的点的轨迹

$$\{\boldsymbol{x} \mid \|\boldsymbol{x}\|_p = 1\}$$



对于$\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X} \subset \boldsymbol{R}^n \quad \boldsymbol{x} = [x_1 \ \cdots \ x_n]^T \ \ \boldsymbol{y} = [y_1 \ \cdots \ y_n]^T$

B. 曼哈顿距离(*Manhattan distance*)

$$dist_{man}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_1 = \sum_{i=1}^{n} |x_i - y_i|$$

C. 欧式距离(*Eculidean distance*)

$$dist_{ed}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{\frac{1}{2}}$$

D. 切氏距离(*Chebyshev distance*)

$$dist_{che}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_\infty = \max_{1 \le i \le n} |x_i - y_i|$$

E. 马氏距离(*Mahalanobis distance*)

$$d_{mah}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{y})}$$

F. Camberra距离(*Lance* 距离,*Williams* 距离)

$$dist_{cam}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i + y_i|} \qquad x_i, y_i \ge 0 且 x_i + y_i \ne 0$$

例：向量之间的几种典型距离度量

无序属性之间的距离

$m_{u,a}$ ——样本集内关于特征u取离散值a的样本数

$m_{u,a,i}$ ——样本集的第i簇中，特征u上取离散值为a的样本数

k ——样本集划分的聚类簇数目

[1]若描述样本的特征为d个离散特征，对于任意两个样本$\boldsymbol{x}_i, \boldsymbol{x}_j$

基于特征**VDM距离**，可以得到样本$\boldsymbol{x}_i, \boldsymbol{x}_j$之间距离：

$$\boldsymbol{MinkovDM}_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)=\left(\sum_{u=1}^{\mathrm{d}} \boldsymbol{VDM}_p\left(x_{iu}, x_{ju}\right)\right)^{\frac{1}{p}}$$

其中：特征u的两个离散值a,b之间的**VDM距离**$(Value\ \ Difference\ \ Metric)$

$$\boldsymbol{VDM}_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$$

[2]基于混合属性(d_c个有序属性以及$d-d_{\mathrm{c}}$个无序属性) 的样本之间距离

$$\boldsymbol{MinkovDM}_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)=\left(\sum_{u=1}^{d_c}\left|x_{iu}-x_{ju}\right|^p+\sum_{u=d_c+1}^{d} \boldsymbol{VDM}_p\left(x_{iu}, x_{ju}\right)\right)^{\frac{1}{p}}$$

例：向量之间的相似性度量

对于$\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X} \subset \boldsymbol{R}^n$ $\quad \boldsymbol{x} = [x_1 \ \cdots \ x_n]^T \quad \boldsymbol{y} = [y_1 \ \cdots \ y_n]^T$

A. 内积 $\quad s_{inner}(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$

B. 余弦相似度 $\quad s_{\cos ine}(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$（**注意区分余弦距离**）

C. Pearson相关系数 $\quad r_{Pearson}(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{(\boldsymbol{x} - \bar{x}\boldsymbol{1})^T (\boldsymbol{y} - \bar{y}\boldsymbol{1})}{\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \|\boldsymbol{y} - \bar{\boldsymbol{y}}\|}$

其中 $\quad \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \quad \bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

D. Tanimoto测度 $\quad s_T(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^T \boldsymbol{y}} = \dfrac{\boldsymbol{x}^T \boldsymbol{y}}{\|x - y\|^2 + \boldsymbol{x}^T \boldsymbol{y}}$

B. 样本与集合之间的测度

方式1.集合中所有样本对近邻测度$\wp(\boldsymbol{x}, \boldsymbol{C})$均有贡献。

设观测样本$\boldsymbol{x}, \boldsymbol{y}$之间近邻测度为$\wp(\boldsymbol{x}, \boldsymbol{y})$

则样本$\boldsymbol{x}$与聚类（或簇）C之间的近邻函数$\wp(\boldsymbol{x}, \boldsymbol{C})$，可以是：

最大近邻函数 $\quad \wp_{\max}^{ps}(\boldsymbol{x}, \boldsymbol{C}) = \max_{\boldsymbol{y} \in \boldsymbol{C}} \wp(\boldsymbol{x}, \boldsymbol{y})$

最小近邻函数 $\quad \wp_{\min}^{ps}(\boldsymbol{x}, \boldsymbol{C}) = \min_{\boldsymbol{y} \in \boldsymbol{C}} \wp(\boldsymbol{x}, \boldsymbol{y})$

平均近邻函数 $\quad \wp_{\text{avg}}^{ps}(\boldsymbol{x}, \boldsymbol{C}) = \dfrac{1}{n_C} \sum_{\boldsymbol{y} \in \boldsymbol{C}} \wp(\boldsymbol{x}, \boldsymbol{y}) \qquad n_C$为集合$C$的势

方式2.近邻性以样本x与集合C的表示之间的近邻性度量。

--集合的表示通常有点、超平面、超球面等.

C. 两集合之间的近邻函数

设两观测样本$\boldsymbol{x}, \boldsymbol{y}$之间近邻测度为$\mathscr{D}(\boldsymbol{x}, \boldsymbol{y})$

对于给定的两个向量集合$\boldsymbol{D}_i, \boldsymbol{D}_j$,近邻函数常见：

最大近邻函数 $\mathscr{D}_{\max}^{ss}\left(\boldsymbol{D}_i, \boldsymbol{D}_j\right)=\max\limits_{\boldsymbol{x}\in\boldsymbol{D}_i, \boldsymbol{y}\in\boldsymbol{D}_j} \mathscr{D}(\boldsymbol{x}, \boldsymbol{y})$

最小近邻函数 $\mathscr{D}_{\min}^{ss}\left(\boldsymbol{D}_i, \boldsymbol{D}_j\right)=\min\limits_{\boldsymbol{x}\in\boldsymbol{D}_i, \boldsymbol{y}\in\boldsymbol{D}_j} \mathscr{D}(\boldsymbol{x}, \boldsymbol{y})$

平均近邻函数 $\mathscr{D}_{avg}^{ss}\left(\boldsymbol{D}_i, \boldsymbol{D}_j\right)=\dfrac{1}{n_{D_i} n_{D_j}} \sum\limits_{\boldsymbol{x}\in\boldsymbol{D}_i} \sum\limits_{\boldsymbol{y}\in\boldsymbol{D}_j} \mathscr{D}(\boldsymbol{x}, \boldsymbol{y})$

其中$n_{D_i} n_{D_j}$为集合$\boldsymbol{D}_i, \boldsymbol{D}_j$的势

均值近邻函数 $\mathscr{D}_{mean}^{ss}\left(\boldsymbol{D}_i, \boldsymbol{D}_j\right)=\mathscr{D}\left(\boldsymbol{m}_{D_i}, \boldsymbol{m}_{D_j}\right)$

$\boldsymbol{m}_{D_i}, \boldsymbol{m}_{D_j}$是关于集合$\boldsymbol{D}_i, \boldsymbol{D}_j$的点描述，如均值点、中值等。

D. 聚类准则

- 类内距离准则
- 类间距离准则
- 基于类内、类间距离的准则函数
- 基于模式与类核的距离的准则函数

主要内容

1. 聚类的引入

2. 聚类算法的评价

评价的意义

(1)避免所发现的数据结构源自噪声干扰

(2)不同聚类算法的比较

(3)两个聚类集合(**two sets of clusters**)的比较

(4)两个聚类的比较

--> **聚类趋向：验证给定数据集是否具有聚类结构；发现数据中真实的结构**

评价的几个角度

- 明确给定数据集合中“聚类的趋势”
  如：区分给定数据集内是否存在非随机性“结构”

- “外部评价”--将聚类分析的结果与给定的结果(带有类别标签的专门数据)比较

- “内部评价”--评估聚类分析的结果是否与数据结构相符，而无需参考外部信息- 只借助数据本身

- 比较不同聚类算法的分析结果，以确定哪种聚类算法更好

- 确定正确的“聚类数目”

评价的几种类型(*types of validation measures*)

(1)外部评价(*external validation*)

需要关于研究对象相关领域的先验知识

如：一个预定义的划分

不足 强化了研究者的主观猜测；
会忽略某些与之前认识不符的现象；
导致错过发现新规律新模式的机会

(2)内部评价(*internal validation*)

基于数据本身内在的信息，量化分析

河北师范大学软件学院
Software College of Hebei Normal University

外部评价的一些常见指标

Clustering metrics

See the Clustering performance evaluation section of the user guide for further details.

The `sklearn.metrics.cluster` submodule contains evaluation metrics for cluster analysis results. There are two forms of evaluation:

- supervised, which uses a ground truth class values for each sample.
- unsupervised, which does not and measures the 'quality' of the model itself.

`metrics.adjusted_mutual_info_score`(...[, ...])	Adjusted Mutual Information between two clusterings.
`metrics.adjusted_rand_score`(labels_true, ...)	Rand index adjusted for chance.
`metrics.calinski_harabasz_score`(X, labels)	Compute the Calinski and Harabasz score.
`metrics.davies_bouldin_score`(X, labels)	Compute the Davies-Bouldin score.
`metrics.completeness_score`(labels_true, ...)	Completeness metric of a cluster labeling given a ground truth.
`metrics.cluster.contingency_matrix`(...[, ...])	Build a contingency matrix describing the relationship between labels.
`metrics.cluster.pair_confusion_matrix`(...)	Pair confusion matrix arising from two clusterings.
`metrics.fowlkes_mallows_score`(labels_true, ...)	Measure the similarity of two clusterings of a set of points.
`metrics.homogeneity_completeness_v_measure`(...)	Compute the homogeneity and completeness and V-Measure scores at once.
`metrics.homogeneity_score`(labels_true, ...)	Homogeneity metric of a cluster labeling given a ground truth.
`metrics.mutual_info_score`(labels_true, ...)	Mutual Information between two clusterings.
`metrics.normalized_mutual_info_score`(...[, ...])	Normalized Mutual Information between two clusterings.
`metrics.rand_score`(labels_true, labels_pred)	Rand index.
`metrics.silhouette_score`(X, labels, *[, ...])	Compute the mean Silhouette Coefficient of all samples.
`metrics.silhouette_samples`(X, labels, *[, ...])	Compute the Silhouette Coefficient for each sample.
`metrics.v_measure_score`(labels_true, ...[, beta])	V-measure cluster labeling given a ground truth.

给定数据集$\boldsymbol{D}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}, \boldsymbol{x}_i=[x_{i1} \quad \cdots \quad x_{id}]^T \in \boldsymbol{R}^d$

若 $\begin{cases} \text{由聚类算法给出的簇划分结果} & \boldsymbol{C}=\{\boldsymbol{C}_1,\ldots,\boldsymbol{C}_r\} \\ \text{参考模型给出的簇划分结果} & \boldsymbol{C}^*=\{\boldsymbol{C}_1^*,\ldots,\boldsymbol{C}_s^*\} \end{cases}$

数据集$\boldsymbol{D}$内各样本相应簇标记值集合 $\begin{cases} \boldsymbol{\lambda}=\{\lambda_1,\ldots,\lambda_N\} \\ \boldsymbol{\lambda}^*=\{\lambda_1^*,\ldots,\lambda_N^*\} \end{cases}$

数据集$\boldsymbol{D}$内样本两两配对，定义：

$$a=|\boldsymbol{SS}| \qquad \boldsymbol{SS}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i=\lambda_j, \lambda_i^*=\lambda_j^*, i<j\}$$

$$b=|\boldsymbol{SD}| \qquad \boldsymbol{SD}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i=\lambda_j, \lambda_i^*\neq\lambda_j^*, i<j\}$$

$$c=|\boldsymbol{DS}| \qquad \boldsymbol{DS}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i\neq\lambda_j, \lambda_i^*=\lambda_j^*, i<j\}$$

$$d=|\boldsymbol{DD}| \qquad \boldsymbol{DD}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i\neq\lambda_j, \lambda_i^*\neq\lambda_j^*, i<j\}$$

并且 $\begin{cases} \boldsymbol{A}=\boldsymbol{SS}\cup\boldsymbol{SD}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i=\lambda_j, i<j\} \\ \boldsymbol{B}=\boldsymbol{SS}\cup\boldsymbol{DS}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i^*=\lambda_j^*, i<j\} \end{cases}$ $\quad |\boldsymbol{A}\cap\boldsymbol{B}|=a, |\boldsymbol{A}\cup\boldsymbol{B}|=a+b+c$

$$d=|\boldsymbol{DD}|=C_N^2+\sum_{i=1}^{r}\sum_{j=1}^{s}C_{n_{ij}}^2-\sum_{i=1}^{r}C_{a_i}^2-\sum_{j=1}^{s}C_{b_j}^2 =\frac{1}{2}\left[N^2+\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}^2-\sum_{i=1}^{r}a_i^2-\sum_{j=1}^{s}b_j^2\right]$$

$$a=|\boldsymbol{SS}|=\sum_{i=1}^{r}\sum_{j=1}^{s}C_{n_{ij}}^2 =\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{1}{2}n_{ij}(n_{ij}-1)$$

$$b=|\boldsymbol{SD}|=\sum_{j=1}^{s}C_{b_j}^2-\sum_{i=1}^{r}\sum_{j=1}^{s}C_{n_{ij}}^2 =\frac{1}{2}\left[\sum_{j=1}^{s}b_j^2-\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}^2\right]$$

$$c=|\boldsymbol{DS}|=\sum_{i=1}^{r}C_{a_i}^2-\sum_{i=1}^{r}\sum_{j=1}^{s}C_{n_{ij}}^2 =\frac{1}{2}\left[\sum_{i=1}^{r}a_i^2-\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}^2\right]$$

$$\boxed{\begin{aligned} &a+b+c+d=C_N^2 \\ &a+d=\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}^2+C_N^2-\frac{1}{2}\left[\sum_{i=1}^{r}a_i^2+\sum_{j=1}^{s}b_j^2\right] \\ &b+c=\frac{1}{2}\left[\sum_{i=1}^{r}a_i^2+\sum_{j=1}^{s}b_j^2\right]-\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}^2 \end{aligned}}$$

$$\begin{cases} \boldsymbol{A}=\boldsymbol{SS}\cup\boldsymbol{SD}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i=\lambda_j, i<j\} \\ \boldsymbol{B}=\boldsymbol{SS}\cup\boldsymbol{DS}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j) \mid \lambda_i^*=\lambda_j^*, i<j\} \\ |\boldsymbol{A}\cap\boldsymbol{B}|=a \\ |\boldsymbol{A}\cup\boldsymbol{B}|=a+b+c \end{cases}$$

河北师范大学软件学院
Software College of Hebei Normal University

基于上述定义，给出用于聚类性能度量的常见**外部指标**

[1] ***Jaccard*** 系数（***Jaccard Coefficient***，简称 ***JC***，雅卡尔系数）

$$JC=\frac{|A\cap B|}{|A\cup B|}=\frac{a}{a+b+c}\in[0,1]$$

Jaccard 距离（***Jaccard Distance***，简称 ***JD***）　$JD=1-JC$

[2] ***FM*** 指数（***Fowlkes and Mallows Index***，简称 ***FMI***）

$$FMI=\sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}}\in[0,1]$$

*sklearn.metrics.**fowlkes_mallows_score**(labels_true, labels_pred, *, sparse=False)*

[3] ***Rand*** 指数（***Rand Index***，简称 ***RI***）

$$RI=\frac{a+d}{a+b+c+d}=\frac{2(a+d)}{N(N-1)}\in[0,1]$$

[4] 调整后的 ***Rand*** 指数（***Ajusted Rand Index***，简称 ***ARI***）

$X \backslash Y$	簇标签 Y_1	Y_2	$\ldots$	Y_s	Sums
真 X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
实 X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
标 $\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
签 X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	

为确保"在聚类结果随机产生的情况下，指标应该接近零"，引入调整兰德系数(Adjusted rand index)，以获取更高的区分度

列联表(contingency table)

n_{ij}----真实类别标签为i、簇标签为j的样本数目

总的样本数 $N=\sum_i\sum_j n_{ij}$

a_i———真实标签为i的样本数目，$a_i=\sum_j n_{ij}$

b_j———簇标签为j的样本数目，　$b_j=\sum_i n_{ij}$

$$ARI = \frac{RI - E[RI]}{max[RI] - E[RI]} = \frac{\sum_i \sum_j C_{n_{ij}}^2 - \frac{\sum_i C_{a_i}^2 \sum_j C_{b_j}^2}{C_N^2}}{\frac{1}{2}\left[\sum_i C_{a_i}^2 + \sum_j C_{b_j}^2\right] - \frac{\sum_i C_{a_i}^2 \sum_j C_{b_j}^2}{C_N^2}}$$

$$RI = \frac{\sum_i \sum_j C_{n_{ij}}^2}{C_N^2}$$

注意： $C_0^2 = C_1^2 = 1$

$$E[RI] = \left(\frac{\sum_i C_{a_i}^2}{C_N^2}\right)\left(\frac{\sum_j C_{b_j}^2}{C_N^2}\right)$$

$$E\left[\sum_i \sum_j C_{n_{ij}}^2\right] = \frac{\sum_i C_{a_i}^2 \sum_j C_{b_j}^2}{C_N^2}$$

$$max[RI] = \frac{1}{2}\left[\frac{\sum_i C_{a_i}^2 + \sum_j C_{b_j}^2}{C_N^2}\right]$$

		簇标签				
	$X \backslash Y$	Y_1	Y_2	$\dots$	Y_s	Sums
真	X_1	n_{11}	n_{12}	$\dots$	n_{1s}	a_1
实	X_2	n_{21}	n_{22}	$\dots$	n_{2s}	a_2
标	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
签	X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	a_r
	Sums	b_1	b_2	$\dots$	b_s	

sklearn.metrics.adjusted_rand_score(*labels_true*,*labels_pred*

[5]完整性(*Completeness*)

- 该指标最大值为1.
- 越接近1，越好。

$$Completeness = 1 - \frac{H(\lambda|C)}{H(\lambda)} = 1 - \frac{-\sum_i \frac{a_i}{N} \sum_j \frac{n_{ij}}{a_i} log\left(\frac{n_{ij}}{a_i}\right)}{-\sum_j \frac{b_j}{N} log\left(\frac{b_j}{N}\right)}$$

簇划分熵

以类别为条件的簇划分熵

		簇标签				
	$X \backslash Y$	Y_1	Y_2	$\dots$	Y_s	Sums
真	X_1	n_{11}	n_{12}	$\dots$	n_{1s}	a_1
实	X_2	n_{21}	n_{22}	$\dots$	n_{2s}	a_2
标	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
签	X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	a_r
	Sums	b_1	b_2	$\dots$	b_s	

对于一种聚类结果，如果来自同类的样本都聚为同一簇，就说这种聚类结果满足完整性。

*sklearn.metrics.**completeness_score**(labels_true, labels_pred)*

[6]同质性(***homogeneity***)

➢ 该指标最大值为1.
➢ 越接近1，越好。

$$Homogeneity = 1 - \frac{H(C|\lambda)}{H(C)} = 1 - \frac{-\sum_j \frac{b_j}{N}\sum_i \frac{n_{ij}}{b_j} log\left(\frac{n_{ij}}{b_j}\right)}{-\sum_i \frac{a_i}{N} log\left(\frac{a_i}{N}\right)}$$

类划分熵

以簇划分为条件的类划分熵

X \ Y	簇标签 Y_1	Y_2	$\ldots$	Y_s	Sums
真 X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
实 X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
标 $\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
签 X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	

对于一种聚类结果，如果聚为同一簇的样本都来自相同类，就说这种聚类结果满足**同质性**。

sklearn.metrics. homogeneity _score(labels_true, labels_pred)

[7]***V - Measure***

$$V = \frac{1}{\frac{\beta}{1+\beta} \times \frac{1}{completeness} + \frac{1}{1+\beta} \times \frac{1}{\mathrm{homogeneity}}} = \frac{(1+\beta) \times \mathrm{homogeneity} \times completeness}{\beta \times \mathrm{homogeneity} + completeness}$$

X \ Y	簇标签 Y_1	Y_2	$\ldots$	Y_s	Sums
真 X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
实 X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
标 $\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
签 X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	

➢ 是关于完整性、同质性的调和平均
➢ 该指标最大值为1.
➢ 越接近1，越好。

*sklearn.metrics.v_measure_score(labels_true, labels_pred, *, beta=1.0)*

[8]标准互信息($Normalized\ Mutual\ Infomation, NMI$)

用于衡量基于聚类的数据划分结果与类别分布的吻合程度。
由表可得两种信息熵:

➢ 类别分布的信息熵

$$H(C) = -\sum_{i=1}^{r} P_i log P_i$$
$$= -\sum_{i=1}^{r} \frac{a_i}{N} log\left(\frac{a_i}{N}\right)$$

➢ 簇分布的信息熵

$$H(J) = -\sum_{j=1}^{s} q_j log q_j$$
$$= -\sum_{j=1}^{s} \frac{b_j}{N} log\left(\frac{b_j}{N}\right)$$

X \ Y		簇标签 Y_1	Y_2	$\ldots$	Y_s	Sums
真	X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
实	X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
标	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
签	X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums		b_1	b_2	$\ldots$	b_s	

河北师范大学软件学院 Software College of Hebei Normal University

类别分布与簇标签分布之间的互信息:

$$\mathrm{MI}(C,J) = -\sum_{i=1}^{r}\sum_{j=1}^{s} P(i,j) log\frac{P(i,j)}{p_i q_j}$$
$$= -\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}}{N} log\frac{\frac{n_{ij}}{N}}{\frac{a_i}{N}\frac{b_j}{N}} = -\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}}{N} log\frac{n_{ij}N}{b_j a_i}$$

➢ NMI取值区间[0,1]
➢ 该指标最大值为1.
➢ 越接近1，越好。

标准互信息:

$$\mathrm{NMI}(C,J) = \frac{\mathrm{MI}(C,J)}{\sqrt{H(C)H(J)}}$$

X \ Y		簇标签 Y_1	Y_2	$\ldots$	Y_s	Sums
真	X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
实	X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
标	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
签	X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums		b_1	b_2	$\ldots$	b_s	

河北师范大学软件学院 Software College of Hebei Normal University

[9]调整后的互信息($\boldsymbol{Adjusted\ Mutual\ Infomation}, AMI$)

AMI取值位于区间[-1,1]，取值越大越好。

若 $\mathrm{NMI}(C,J)=\dfrac{\mathrm{MI}(C,J)}{\frac{H(C)+H(J)}{2}}$， 则 $\mathrm{AMI}=\dfrac{MI-E[MI]}{\frac{H(C)+H(J)}{2}-E[MI]}$

若 $\mathrm{NMI}(C,J)=\dfrac{\mathrm{MI}(C,J)}{\sqrt{H(C)H(J)}}$， 则 $\mathrm{AMI}=\dfrac{MI-E[MI]}{\sqrt{H(C)H(J)}-E[MI]}$

若 $\mathrm{NMI}(C,J)=\dfrac{\mathrm{MI}(C,J)}{\max\{H(C),H(J)\}}$， 则 $\mathrm{AMI}=\dfrac{MI-E[MI]}{\max\{H(C),H(J)\}-E[MI]}$

		簇标签				
	$X \diagdown Y$	Y_1	Y_2	$\dots$	Y_s	Sums
真实标签	X_1	n_{11}	n_{12}	$\dots$	n_{1s}	a_1
	X_2	n_{21}	n_{22}	$\dots$	n_{2s}	a_2
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	a_r
	Sums	b_1	b_2	$\dots$	b_s	

$$E[MI(C,J)]=\sum_{i=1}^{r}\sum_{j=1}^{s}\left\{\sum_{n_{ij}=max\{0,a_i+b_j-N\}}^{min\{a_i,b_j\}}\frac{n_{ij}}{N}log\left(\frac{n_{ij}N}{b_ja_i}\right)P\left(\boldsymbol{Table}|a_i,b_j,n_{ij}\right)\right\}$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{s}\left\{\sum_{n_{ij}=max\{0,a_i+b_j-N\}}^{min\{a_i,b_j\}}\frac{n_{ij}}{N}log\left(\frac{n_{ij}N}{b_ja_i}\right)\frac{a_i!b_j!(N-a_i)!(N-b_j)!}{N!n_{ij}!(a_i-n_{ij})!(b_j-n_{ij})!(N-a_i-b_j+n_{ij})!}\right\}$$

$$P\left(\boldsymbol{Table}|a_i,b_j,n_{ij}\right)$$

$$=\frac{C_N^{n_{ij}}C_{N-n_{ij}}^{b_j-n_{ij}}C_{N-n_{ij}}^{a_i-n_{ij}}}{C_N^{a_i}C_N^{b_j}}$$

$$=\frac{a_i!\,b_j!\,(N-a_i)!\,(N-b_j)!}{N!\,n_{ij}!\,(a_i-n_{ij})!\,(b_j-n_{ij})!\,(N-a_i-b_j+n_{ij})!}$$

		簇标签				
	$X \diagdown Y$	Y_1	Y_2	$\dots$	Y_s	Sums
真实标签	X_1	n_{11}	n_{12}	$\dots$	n_{1s}	a_1
	X_2	n_{21}	n_{22}	$\dots$	n_{2s}	a_2
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	a_r
	Sums	b_1	b_2	$\dots$	b_s	

$$MI(C,J)\le min\{H(C),H(J)\}\le\sqrt{H(C)H(J)}\le\frac{H(C)+H(J)}{2}\le max\{H(C),H(J)\}\le H(C,J)$$

河北师范大学软件学院
Software College of Hebei Normal University

内部评价的一些常见评价指标

Clustering metrics

See the Clustering performance evaluation section of the user guide for further details.

The `sklearn.metrics.cluster` submodule contains evaluation metrics for cluster analysis results. There are two forms of evaluation:

- supervised, which uses a ground truth class values for each sample.
- unsupervised, which does not and measures the 'quality' of the model itself.

`metrics.adjusted_mutual_info_score`(…[, …])	Adjusted Mutual Information between two clusterings.
`metrics.adjusted_rand_score`(labels_true, …)	Rand index adjusted for chance.
`metrics.calinski_harabasz_score`(X, labels)	Compute the Calinski and Harabasz score.
`metrics.davies_bouldin_score`(X, labels)	Compute the Davies-Bouldin score.
`metrics.completeness_score`(labels_true, …)	Completeness metric of a cluster labeling given a ground truth.
`metrics.cluster.contingency_matrix`(…[, …])	Build a contingency matrix describing the relationship between labels.
`metrics.cluster.pair_confusion_matrix`(…)	Pair confusion matrix arising from two clusterings.
`metrics.fowlkes_mallows_score`(labels_true, …)	Measure the similarity of two clusterings of a set of points.
`metrics.homogeneity_completeness_v_measure`(…)	Compute the homogeneity and completeness and V-Measure scores at once.
`metrics.homogeneity_score`(labels_true, …)	Homogeneity metric of a cluster labeling given a ground truth.
`metrics.mutual_info_score`(labels_true, …)	Mutual Information between two clusterings.
`metrics.normalized_mutual_info_score`(…[, …])	Normalized Mutual Information between two clusterings.
`metrics.rand_score`(labels_true, labels_pred)	Rand index.
`metrics.silhouette_score`(X, labels, *[, …])	Compute the mean Silhouette Coefficient of all samples.
`metrics.silhouette_samples`(X, labels, *[, …])	Compute the Silhouette Coefficient for each sample.
`metrics.v_measure_score`(labels_true, …[, beta])	V-measure cluster labeling given a ground truth.

给定数据集$D=\{x_1,\ldots,x_m\}, x_i=[x_{i1} \ \cdots \ x_{id}]^T \in R^d$

若由聚类给出的簇划分结果 $\mathcal{C}=\{C_1,\ldots,C_k\}$

并且 $\begin{cases} dist(\cdot,\cdot) \text{——两样本点之间距离} \\ \mu=\frac{1}{|C|}\sum_{x\in C} x \text{——任意簇} C\in\mathcal{C} \text{的中心点.} \end{cases}$

$\forall C\in\mathcal{C},$
簇C内样本间的平均距离

$$avg(C)=\frac{2}{|C|(|C|-1)}\sum_{1\le i<j\le|C|} dist(x_i,x_j)$$

簇C内样本间的最远距离

$$diam(C)=\max_{1\le i<j\le|C|} dist(x_i,x_j)$$

簇C_i,C_j样本间最近距离 $d_{\min}(C_i,C_j)=\min_{x_i\in C_i,x_j\in C_j} dist(x_i,x_j)$

簇C_i,C_j中心点之间距离 $d_{\text{cen}}(C_i,C_j)=dist(\mu_i,\mu_j)$

基于上述定义，给出用于聚类性能度量的常见**内部指标**

sklearn.metrics.davies_bouldin_score(X, labels)

[1] DBI ($Davies-Bouldin\ \ Index$, 戴维森-堡丁指数)

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg(C_i)+avg(C_j)}{d_{\text{cen}}(C_i,C_j)}\right)$$

DBI值越小越好.

对于每个给定类别i，找到与其它类之间的最大比值

$$avg(C)=\frac{2}{|C|(|C|-1)}\sum_{1\le i<j\le|C|} dist(x_i,x_j)$$

$$d_{\text{cen}}(C_i,C_j)=dist(\mu_i,\mu_j)$$

$[2]$方差比$\begin{pmatrix} \textbf{Calinski and Harabasz score}, \\ \textbf{Variance Ratio Criterion} \end{pmatrix}$

$$\textbf{Variance Ratio Criterion} = \frac{\textbf{the within} - \textbf{cluster dispersion}}{\textbf{the between} - \textbf{cluster dispersion}}$$

$\textbf{Variance Ratio Criterion} \geq 0$，该值越小越好。

sklearn.metrics.calinski_harabasz_score(X, labels)

$[3]$ **Dunn**指数(**Dunn Index**, 简称**DI**)

$$\boldsymbol{DI} = \min_{1 \leq \boldsymbol{i} \leq \boldsymbol{k}} \left(\frac{\min_{j \neq i} \boldsymbol{d}_{\min}\left(\boldsymbol{C}_i, \boldsymbol{C}_j\right)}{\max_{1 \leq \boldsymbol{l} \leq \boldsymbol{k}} \boldsymbol{diam}\left(\boldsymbol{C}_l\right)} \right) \qquad \boldsymbol{DI} \in [0, \infty)$$

DI值越大越好.

minimal intercluster distance

maximal intracluster distance

河北师范大学软件学院
Software College of Hebei Normal University

[4](平均)Silhouette宽度(average Silhouette width)

A. Silhouette值--样本x_i的Silhouette宽度--**基于样本**

sklearn.metrics.silhouette_samples(X, labels, *, metric='euclidean', **kwds)

$$S_i=\frac{b_i-a_i}{\max\{b_i,a_i\}}\qquad S_i\in[-1,1]$$

a_i--样本x_i与同类中其它样本的平均距离

b_i--样本x_i与其它与之最近"簇"的所有样本的平均距离

S_i值越接近1，表明样本x_i所在"簇"具有很好聚集性

S_i值越接近-1，表明样本x_i错分至其目前所在"簇"

该样本只是位于某两"簇"之间的某个地方

S_i=0，表明样本x_i也可以分至与其目前所在"簇"最近的那"簇"中

[4](平均)Silhouette宽度(average Silhouette width)

B.聚类C_k的平均Silhouette宽度 --**针对每簇**

$$S(C_k)=\frac{1}{N_k}\sum_{x_i\in C_k}S_i$$

$$\begin{cases}S(C_k)\in[-1,1]\\ C_k\in C\end{cases}$$

[4] (**平均**) *Silhouette*宽度 (*average* *Silhouette width*)

***C.Silhouette*宽度(剪影宽度，轮廓系数)**

--**整个数据集**所有样本平均***Silhouette***值

sklearn.metrics.silhouette_score()

$$Silhouette\text{宽度} = \frac{1}{|C|}\sum_{C_k \in C} S(C_k) = \frac{1}{N}\sum_{i=1}^{N} S_i$$

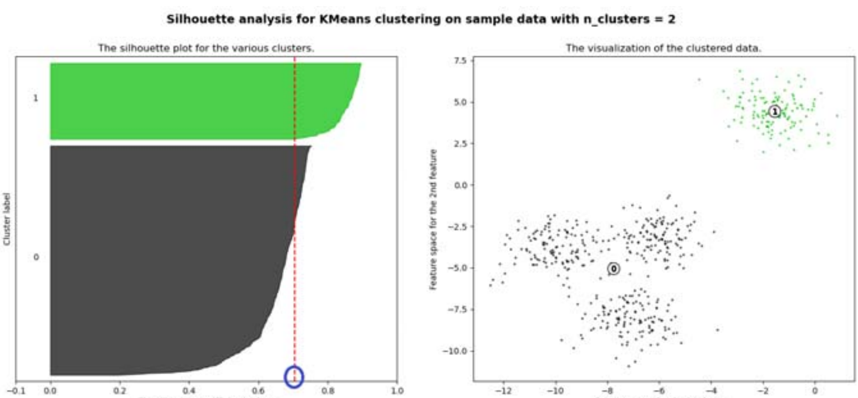
"***Silhouette*宽度**" 可用来：

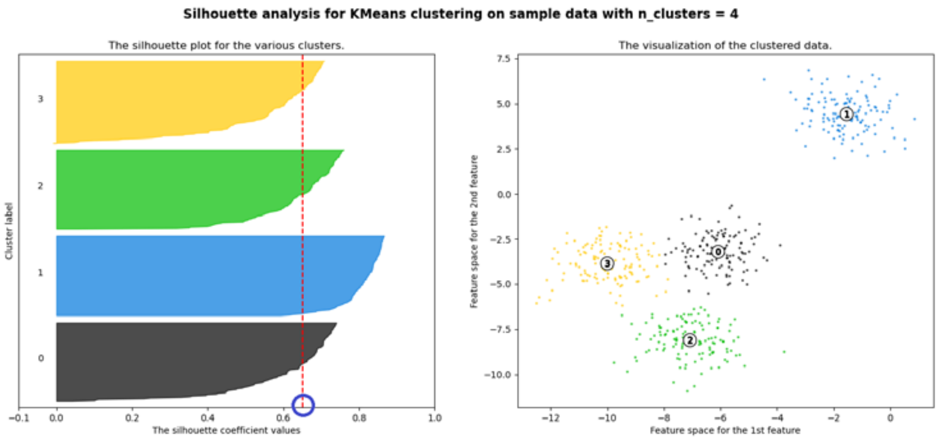
A--评价聚类的有效性，越接近于1越好

B--确定聚类数目的多少

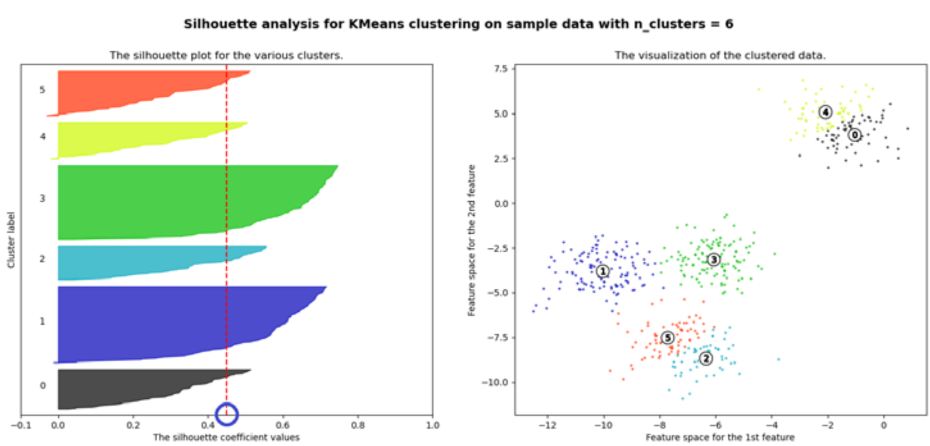
例：利用平均剪影宽度辅助选择聚类数目

n_clusters = 2 The average silhouette_score =0.7049787496083262
n_clusters = 4 The average silhouette_score =0.6505186632729437
n_clusters = 6 The average silhouette_score =0.4504666294372765





河北师范大学软件学院
Software College of Hebei Normal University



河北师范大学软件学院
Software College of Hebei Normal University