Оценка качества и рекомендации по решению задачи кластеризации

План

1. Среднее внутрикластерное и межкластерное расстояние
2. Силуэт (silhouette coefficient)
3. Подбор количества кластеров по силуэту
4. Проверка наличия кластерной структуры
5. Проблема выбора хороших признаков
6. Полнота и однородность (completeness & homogeneity)
7. Оценка качества с привлечением асессоров

Среднее внутрикластерное расстояние

$$F_0 = \frac{\sum\limits_{i<j} [y_i = y_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [y_i = y_j]} \to \min .$$

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\ y_i = y} \rho^2(x_i, \mu_y) \to \min$$

Среднее межкластерное расстояние

$$F_1 = \frac{\sum\limits_{i<j}[y_i \neq y_j]\,\rho(x_i, x_j)}{\sum\limits_{i<j}[y_i \neq y_j]} \to \max$$

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max$$

Комбинируем функционалы

$$F_0 = \frac{\sum\limits_{i<j} [y_i = y_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [y_i = y_j]} \qquad F_1 = \frac{\sum\limits_{i<j} [y_i \neq y_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [y_i \neq y_j]}$$

$$F_0 / F_1 \to \min$$

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\, y_i = y} \rho^2(x_i, \mu_y) \qquad \Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu)$$

$$\Phi_0 / \Phi_1 \to \min$$

Коэффициент силуэта

- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

$$s = \frac{b - a}{max(a, b)}$$

Коэффициент силуэта

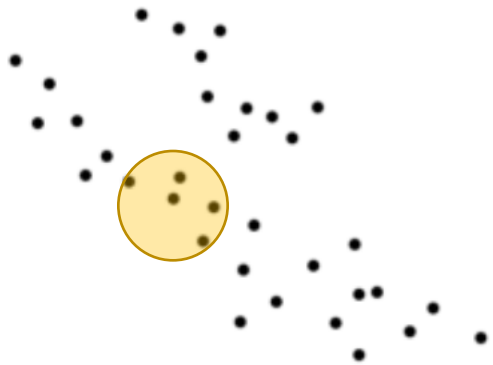
- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

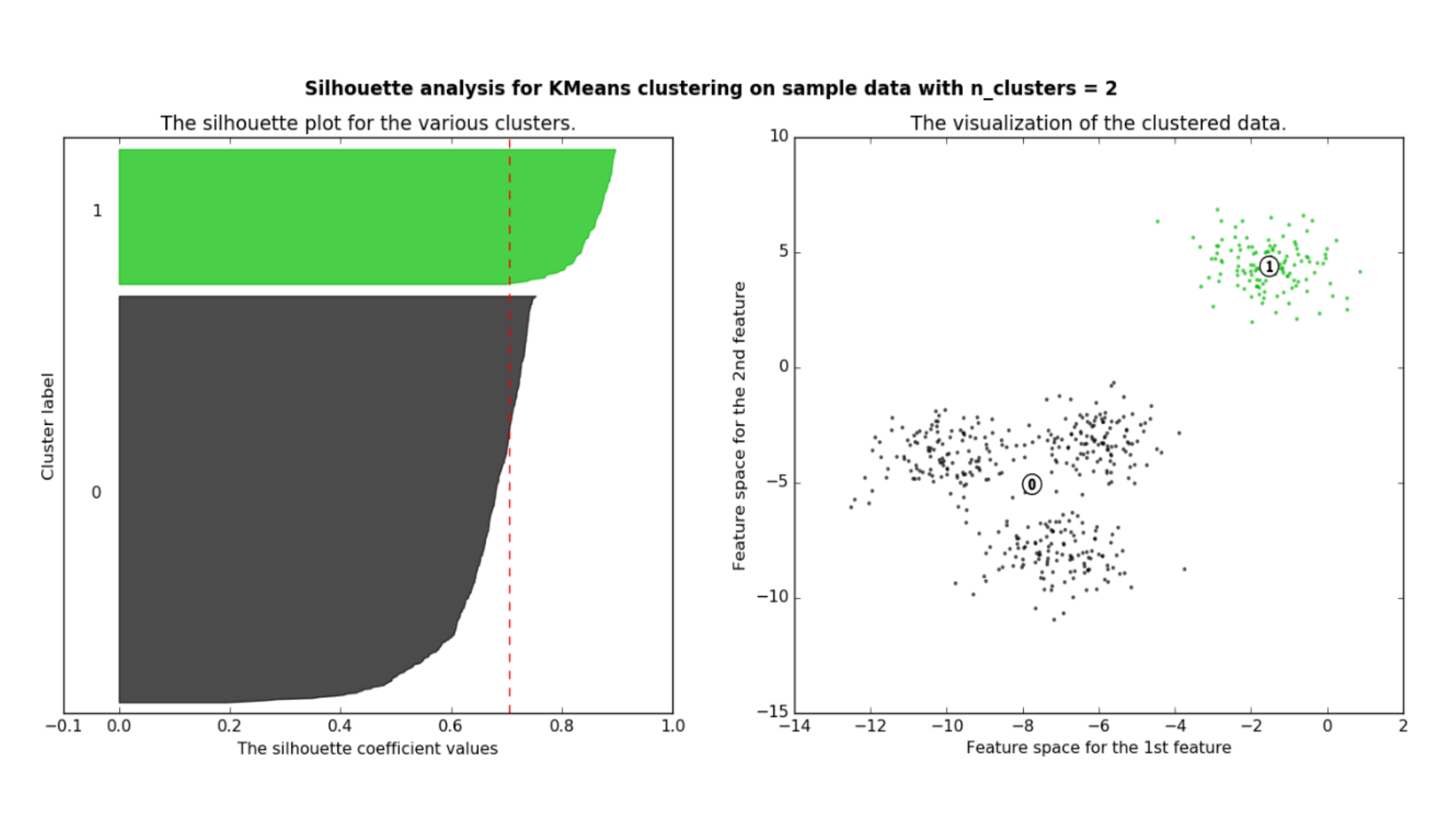
$$s = \frac{b - a}{max(a, b)}$$

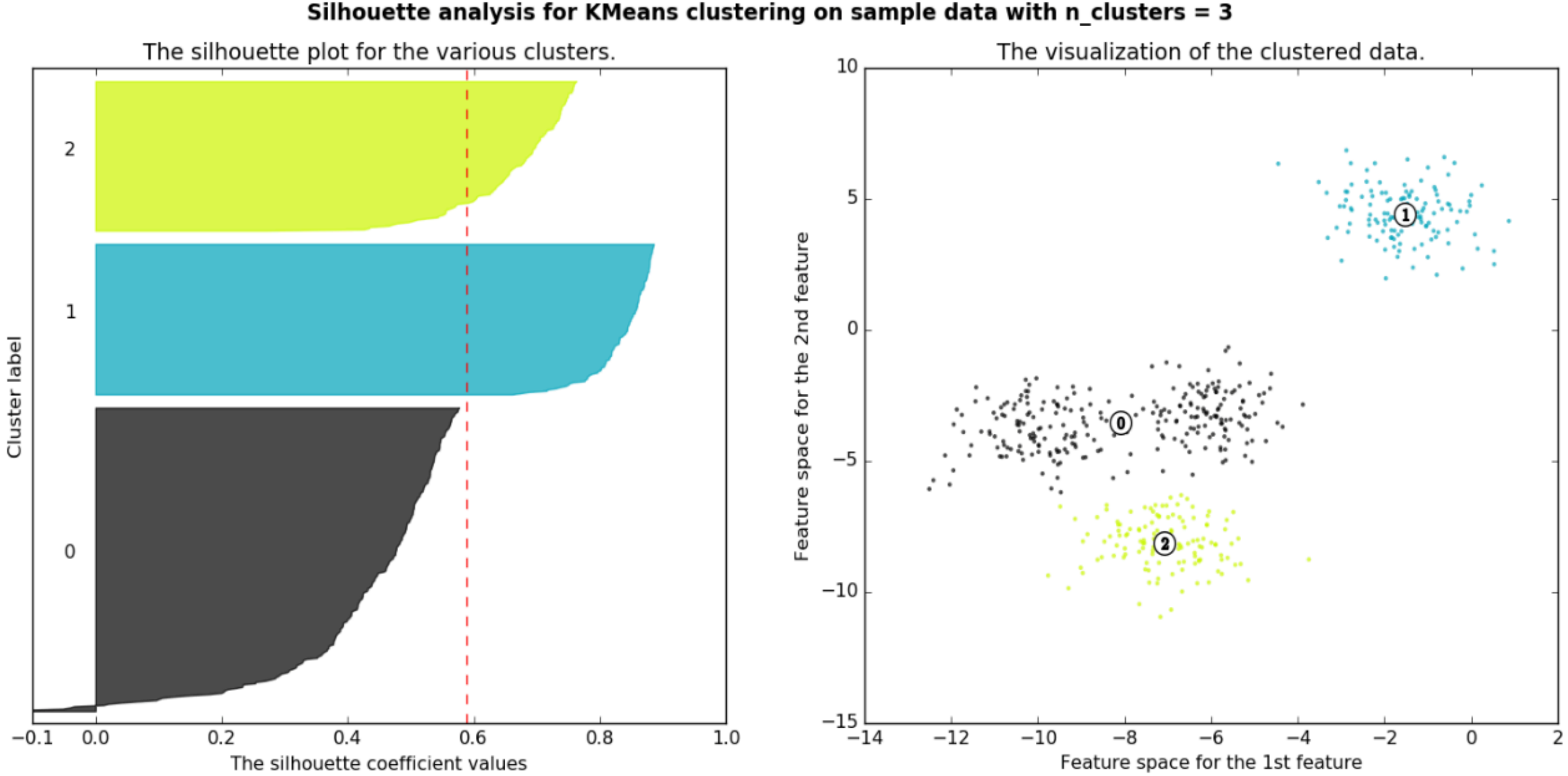
Коэффициент силуэта

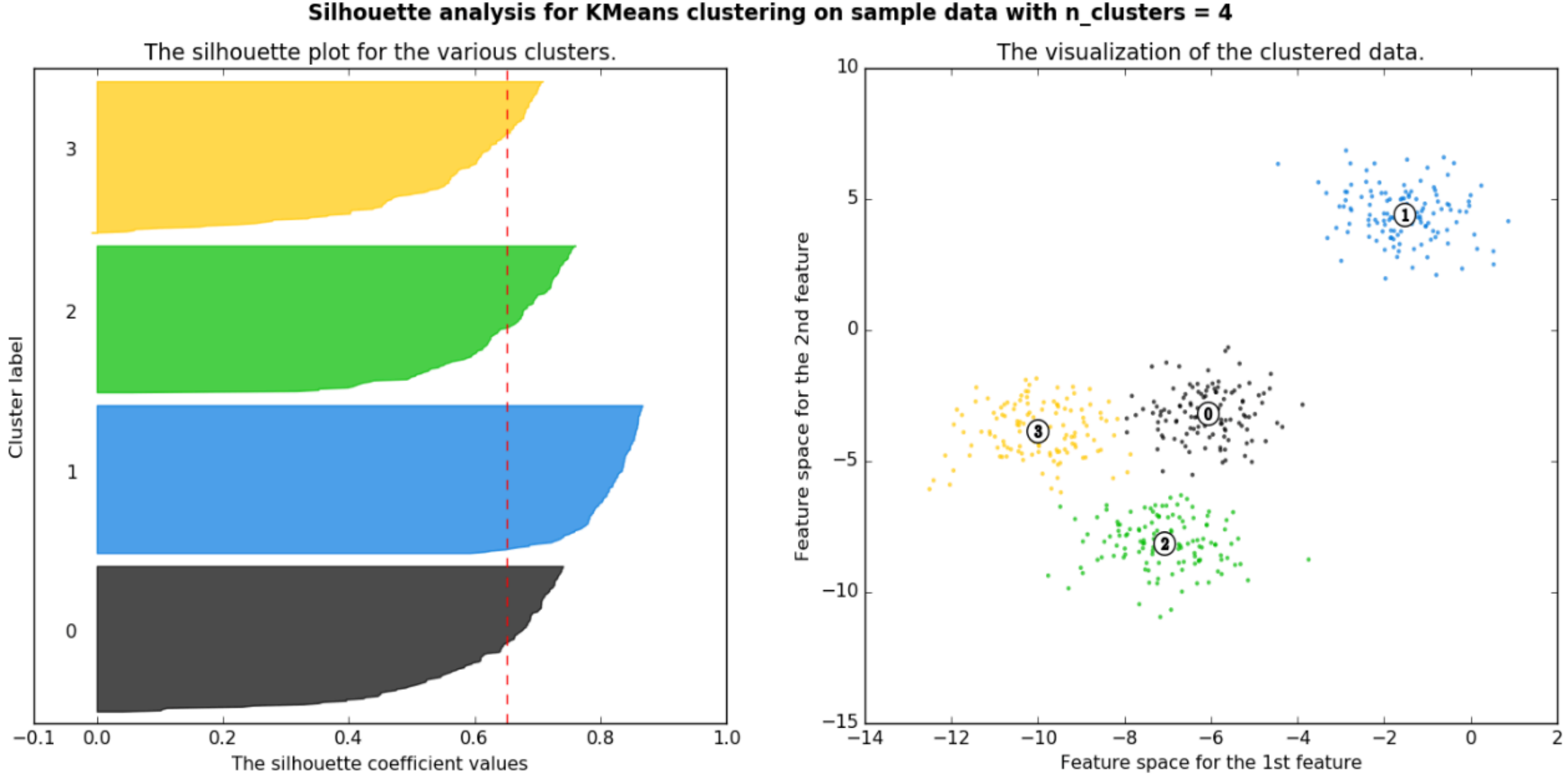
- **a**: Среднее расстояние от данного объекта до всех других объектов из того же кластера
- **b**: Среднее расстояние от данного объекта до всех объектов из *ближайшего другого кластера*

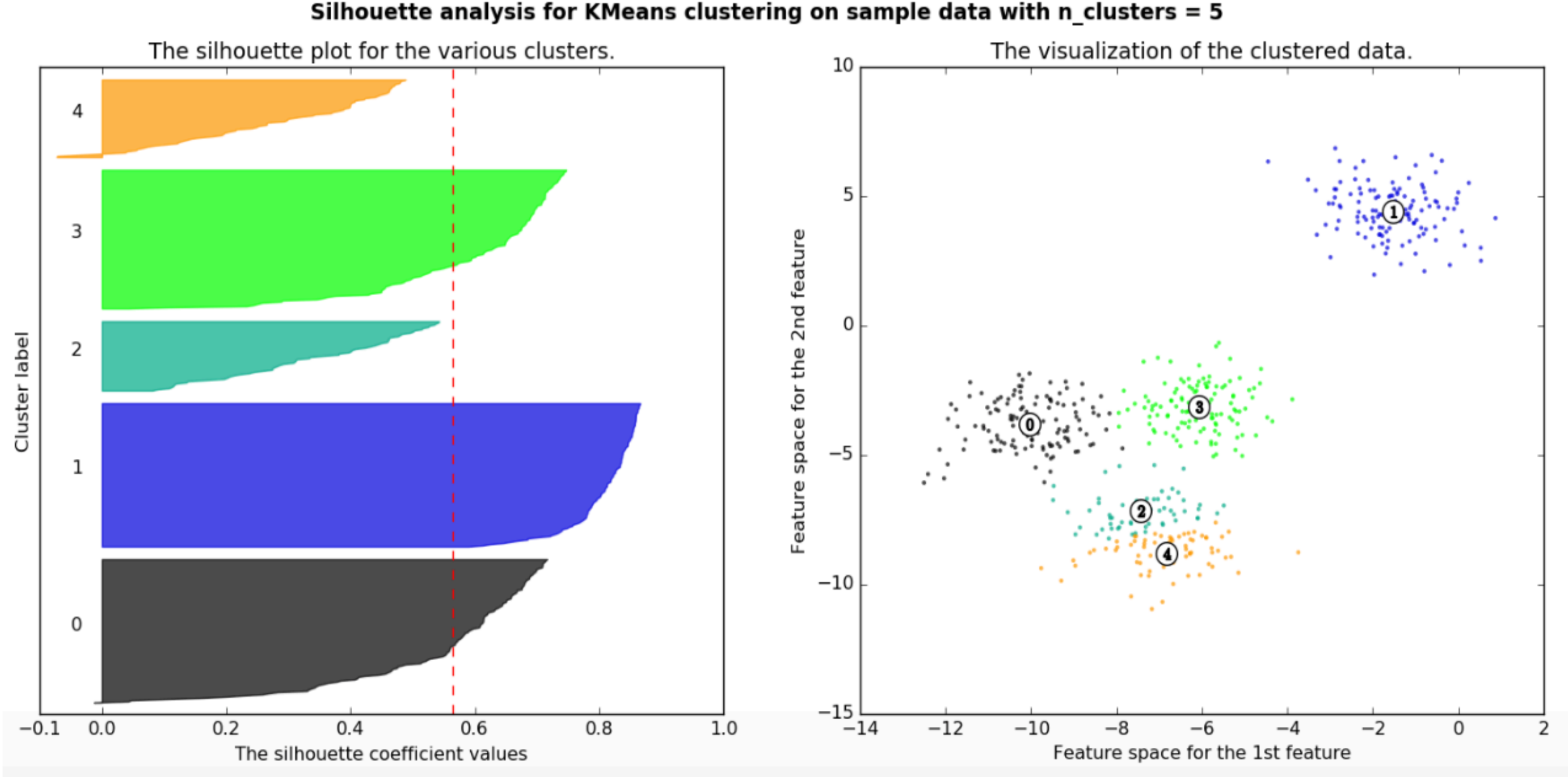
$$s = \frac{b - a}{max(a, b)}$$

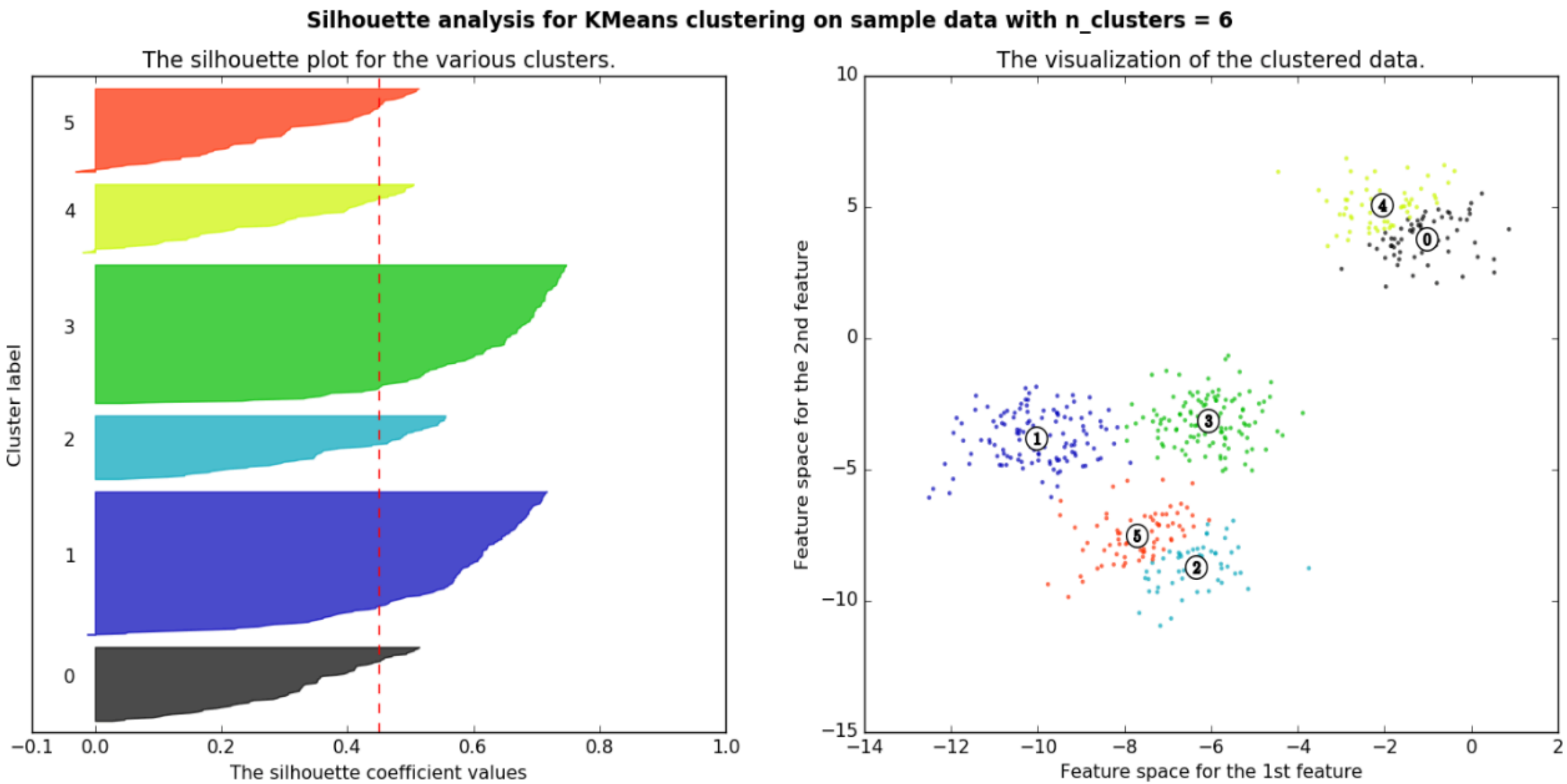



Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for the various clusters.
Cluster label
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature
1
0

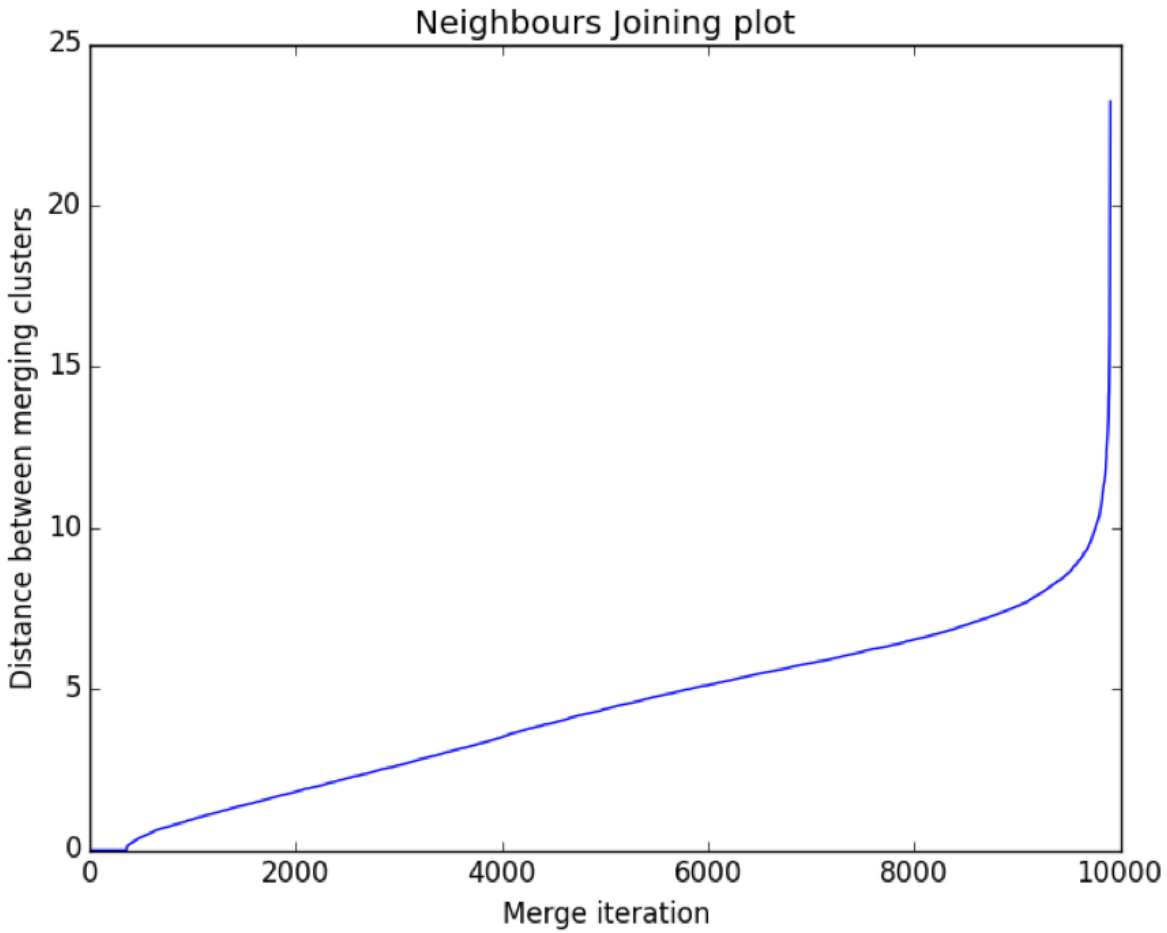

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature
0
1
2


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature
1
0
3
2


Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature
0
1
2
3
4


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6
The silhouette plot for the various clusters.
Cluster label
5
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature
4
0
1
3
5
2

Проверка наличия кластерной структуры



Проверка наличия кластерной структуры

1. Генерируем p случайных точек из равномерного распределения и p случайных из обучающей выборки
2. Вычисляем величину (статистика Хопкинса):

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

Выбор признаков

Что хотим уметь делать:

Для разных признаков понимать, насколько хорошо решена задача кластеризации

Зачем:

Тогда сможем выбирать наиболее адекватные признаки

В чем проблема:

Текущие метрики зависят от признакового пространства

Однородность, полнота, V-мера

В каких случаях значения метрик максимальны:

- **Однородность**: кластер состоит только из объектов одного класса
- **Полнота**: все объекты из класса принадлежат к одному кластеру

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = -\sum_i p_i \ln p_i$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = -\sum_i p_i \ln p_i$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = -\sum_i p_i \ln p_i$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

$$P(c) = \frac{n_c}{n}$$

Однородность, полнота, V-мера

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

$$H = -\sum_i p_i \ln p_i$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

$$P(c) = \frac{n_c}{n}$$

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} \frac{n_{c,k}}{n_{\mathrm{k}}} \cdot \log\left(\frac{n_{c,k}}{n_k}\right)$$

$$P(c|k) = \frac{n_{c,k}}{n_k}$$

Привлечение асессоров для оценки качества

Если разметки нет, можно:

1. Использовать метрики без разметки
2. Создать разметку с помощью асессоров и использовать ее
3. Предложить асессорам отвечать на вопросы вида «допустимо ли эти объекты относить в один/в разные кластеры»

Резюме

1. Среднее внутрикластерное и межкластерное расстояние
2. Силуэт (silhouette coefficient)
3. Подбор количества кластеров по силуэту
4. Проверка наличия кластерной структуры
5. Проблема выбора хороших признаков
6. Полнота и однородность (completeness & homogeneity)
7. Оценка качества с привлечением асессоров