Неконтролируемое обучение (обучение без учителя)

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Неконтролируемое обучение

- Неконтролируемое обучение (обучение без учителя, unsupervised learning) – это метод машинного обучения, при котором модель обучается на немаркированных/неразмеченных данных.
- Используется для задач, в которых известны описания множества объектов (обучающей выборки), и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами.

Задача обучения без учителя

Необходимо найти взаимосвязи, зависимости, закономерности, существующие между объектами.

Имеются:

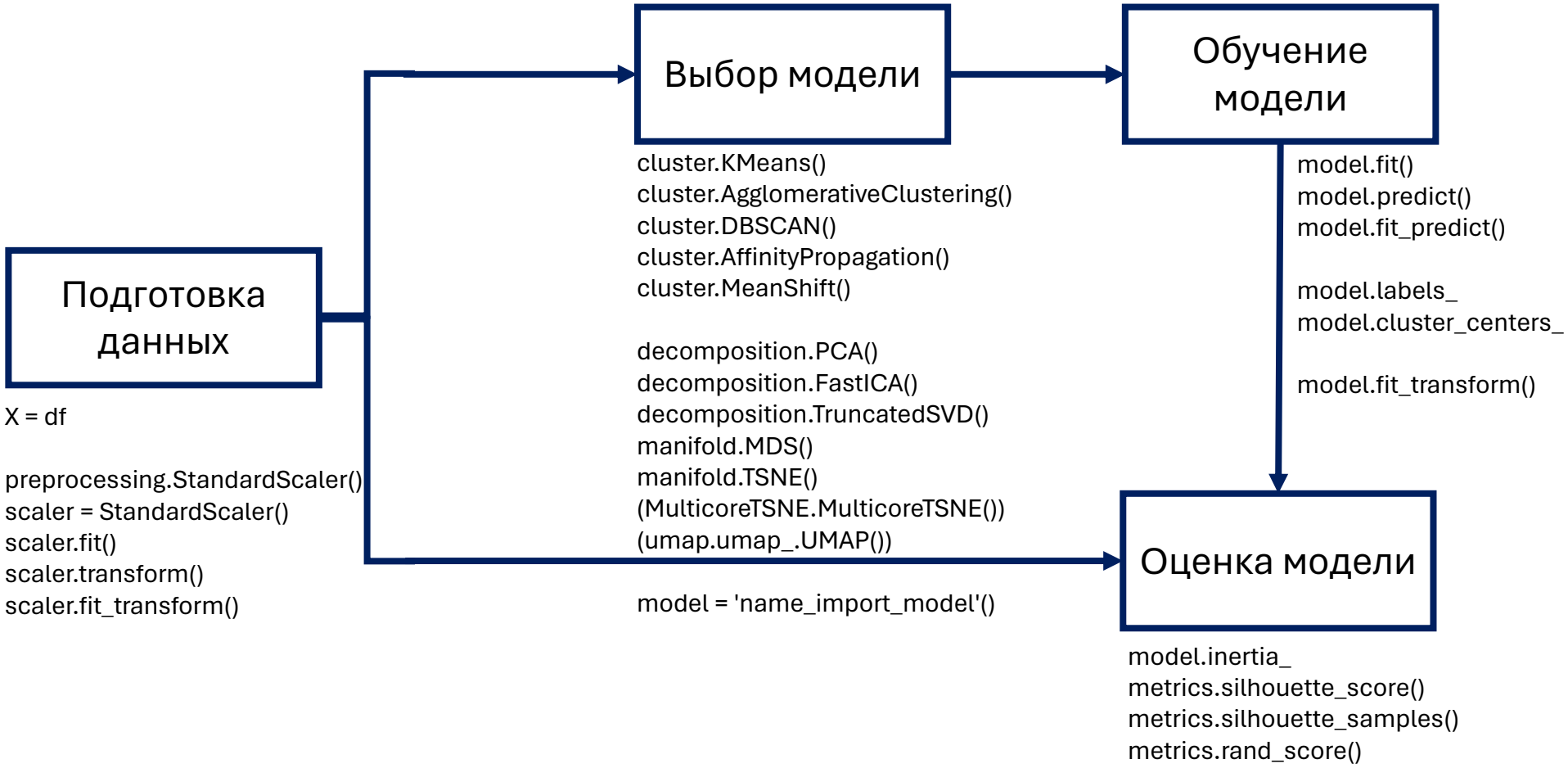
- Множество объектов (ситуаций) X со своими признаковыми описаниями.
- Между объектами x_i существуют зависимости.

На основе имеющихся прецедентов необходимо построить алгоритм a, способный найти взаимосвязи, зависимости, закономерности, существующие между объектами x_i.

Алгоритмы/модели неконтролируемого обучения

- кластеризация (clustering) заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.
- поиск ассоциативных правил (association rules learning). Исходные данные представляются в виде признаковых описаний. Требуется найти такие наборы признаков, и такие значения этих признаков, которые особенно часто (неслучайно часто) встречаются в признаковых описаниях объектов.
- снижение размерности (dimensionality reduction) заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки. В классе линейных преобразований наиболее известным примером является метод главных компонент.
- Также используется в задачах фильтрации выбросов, построения доверительной области, заполнения пропущенных значений, визуализации.

Процесс неконтролируемого обучения (sklearn)



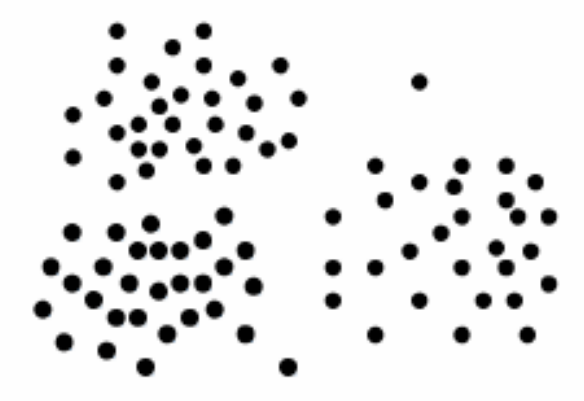
Кластеризация

Нужно сгруппировать все похожие объекты на группы, подобрав такой алгоритм, который для каждого объекта проставит метку так, чтобы объекты с одной меткой были в определенном смысле похожи друг на друга, а объекты с разными метками отличались.

Для проведения кластеризации существуют разные методы. Универсального метода кластеризации нет!

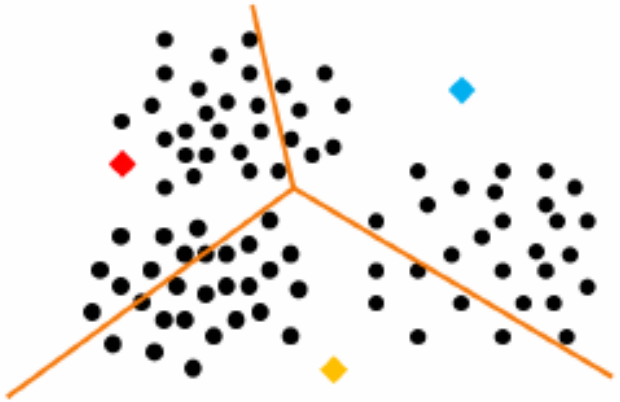
k-means (метод k-средних)

Шаг 1. Выбрать количество кластеров k, которое нам кажется оптимальным для наших данных

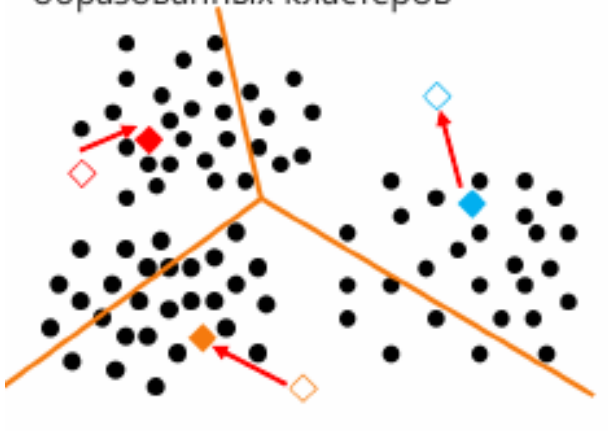


Шаг 2. Определить среди наших объектов те, которые будут являться центрами наши k кластеров (можно случайно)

Шаг 3. Для каждого объекта посчитать, к какому центру он ближе



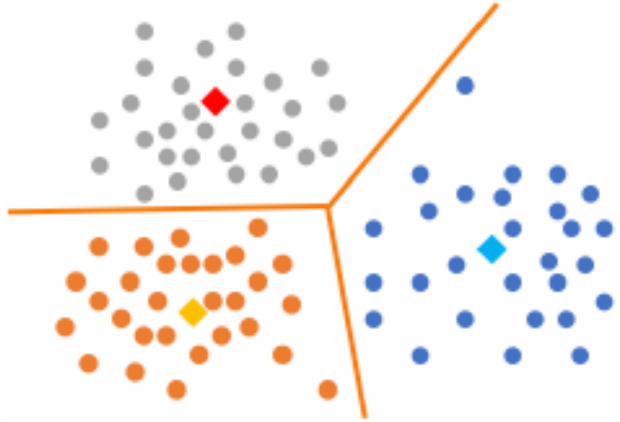
Шаг 4. Пересчитать центры образованных кластеров



Шаг 5. Повторяем шаги 3 и 4

либо фиксированное число раз, либо до тех пор, пока смещение центра относительно предыдущего положения превышает какое-то заранее заданного небольшого значения

Шаг 6. Итоговый результат кластеризации



Метод локтя – выбор числа кластеров

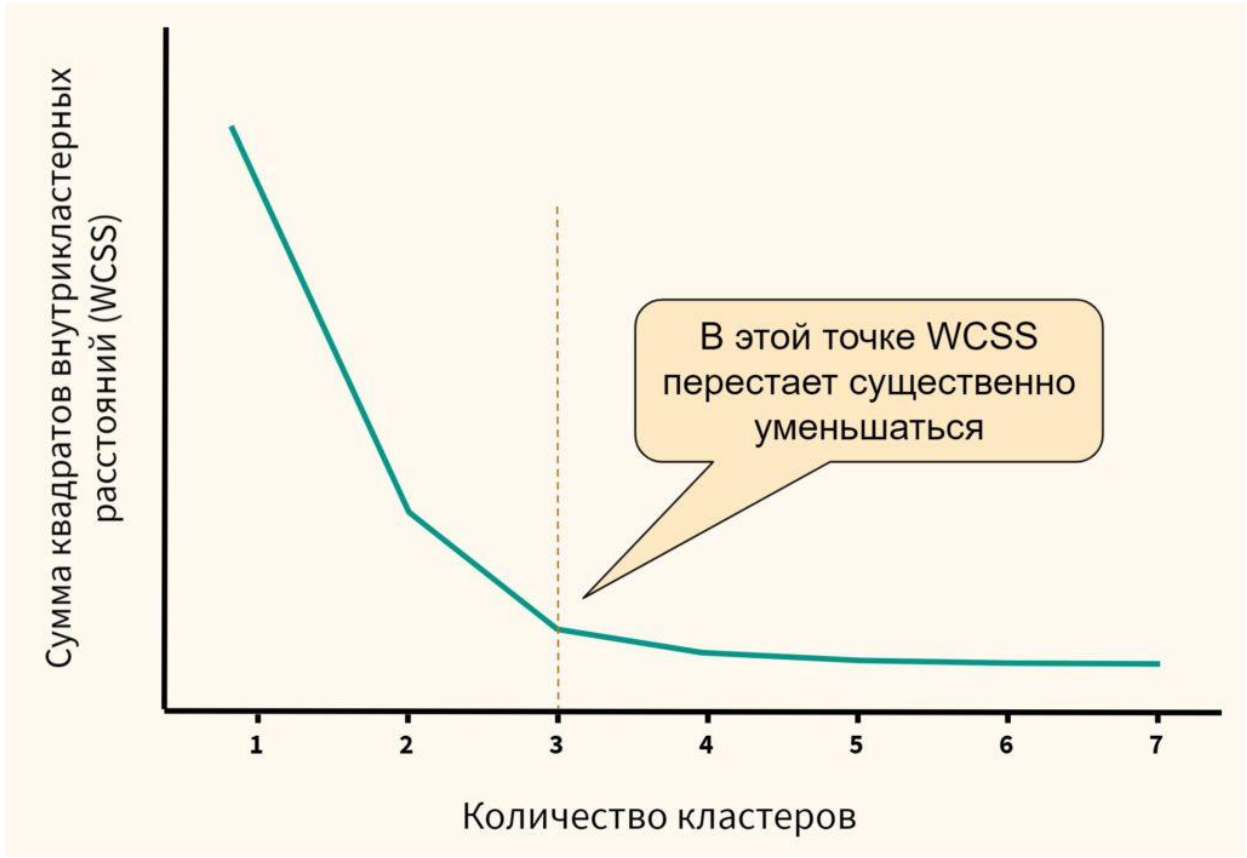
Цель метода локтя — минимизировать сумму квадратов внутрикластерных расстояний до центра кластера (within-cluster sum of squares, WCSS, функция потерь) J:

Кол-во кластеров
Кол-во наблюдений
i-ое наблюдение
центроид j-ого кластера

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \min(||x_i^{(j)} - c_j||)^2$$

Функция расстояния

Функция потерь
(еще говорят целевая функция, objective function)



Коэффициент силуэта - метрика качества кластеризации

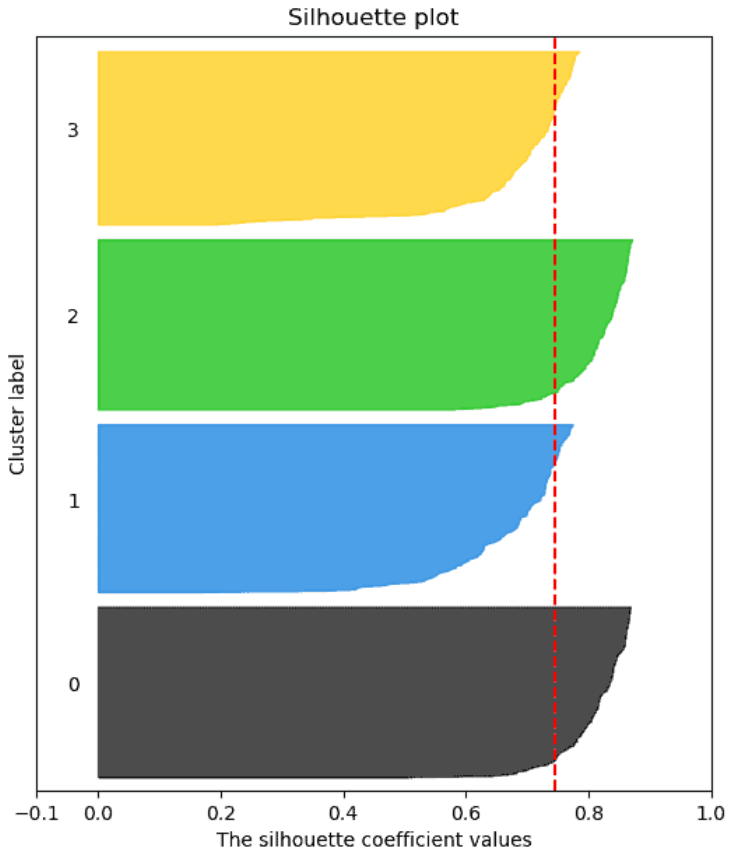
Силуэтом выборки называется средняя величина силуэта объектов данной выборки.
Силуэт показывает насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров.

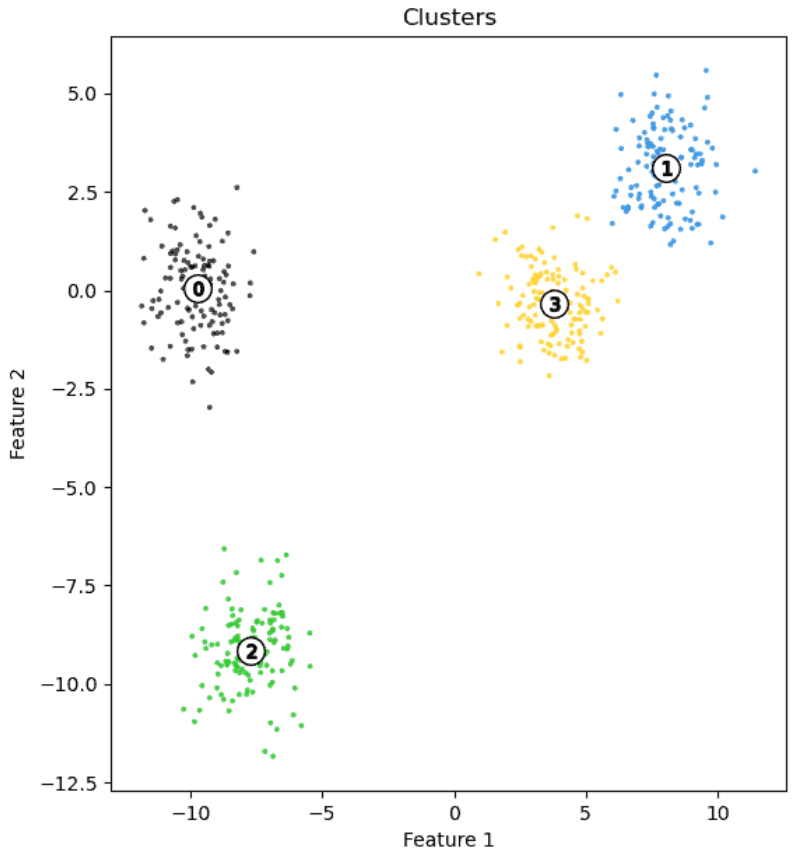
$$s = \frac{b - a}{\max(a, b)}$$

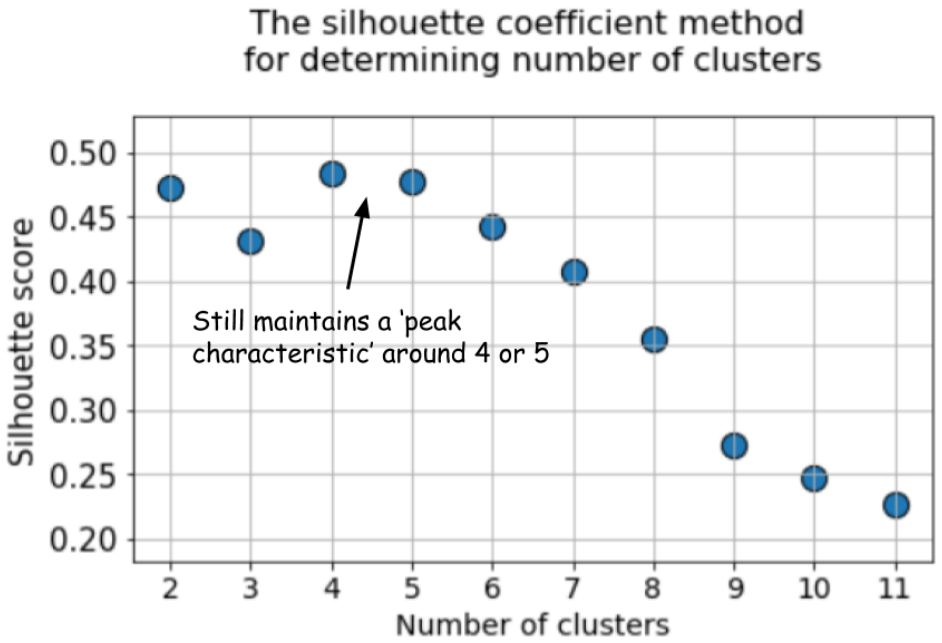
a — среднее расстояние от данного объекта до объектов из того же кластера;
b — среднее расстояние от данного объекта до объектов из ближайшего кластера.
Лежит в диапазоне [−1,1]:

- Значения, близкие к -1, соответствуют плохим (разрозненным) кластерам;
- Значения, близкие к 0, говорят о том, что кластеры пересекаются и накладываются друг на друга;
- Значения, близкие к 1, соответствуют "плотным" четко выделенным кластерам.

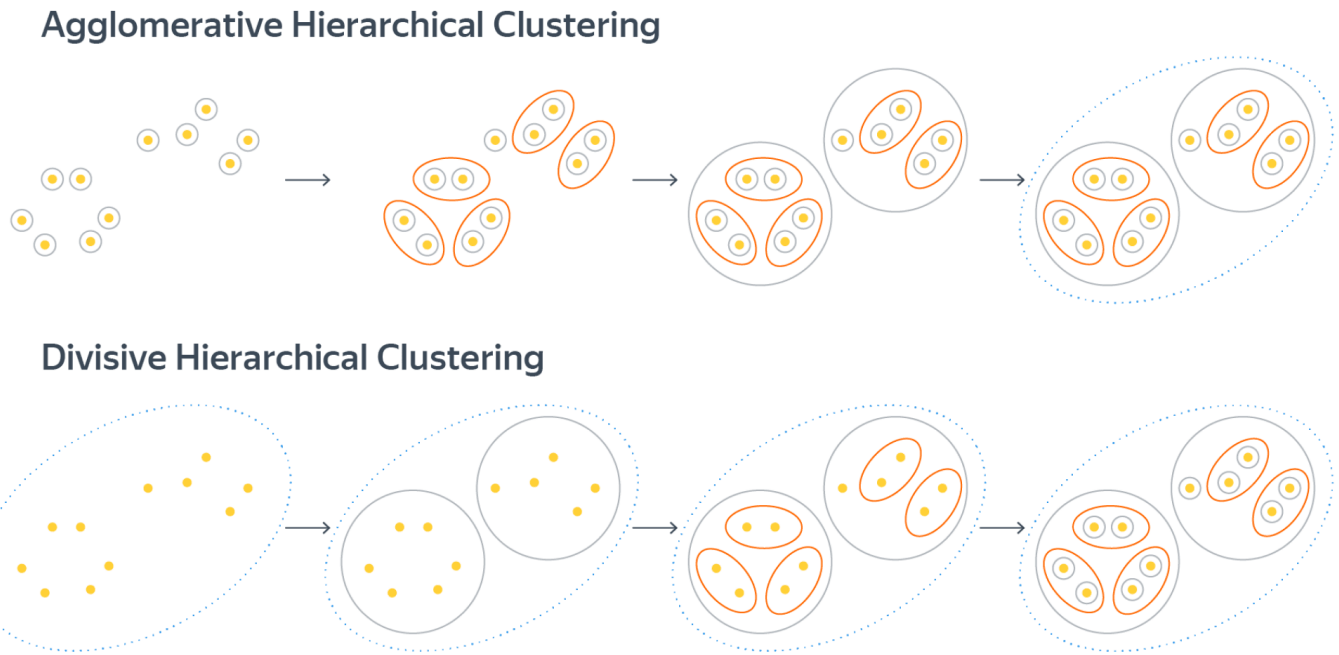
Коэффициент силуэта - выбор числа кластеров

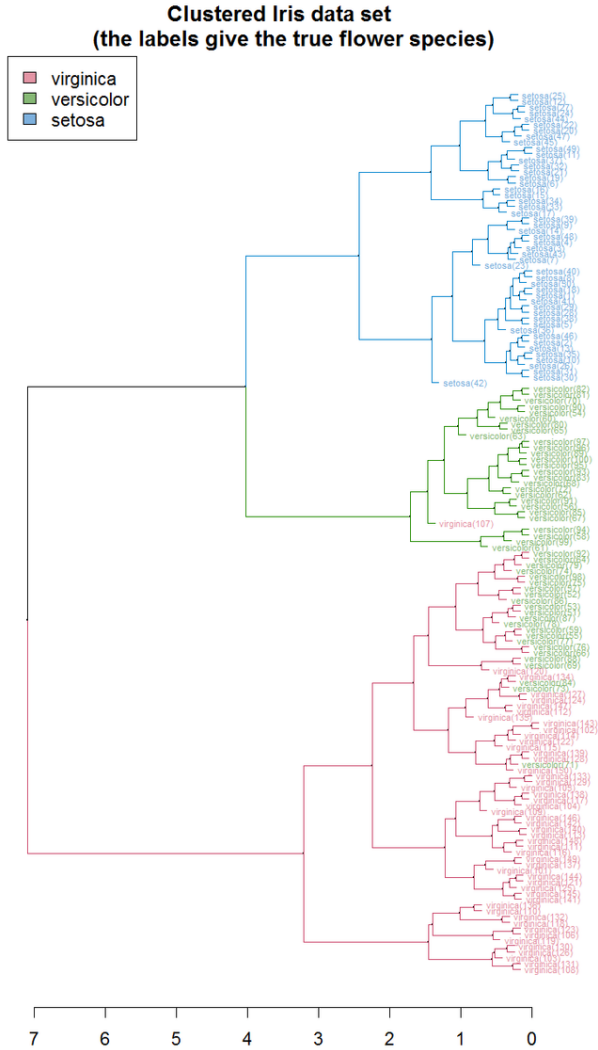






Иерархическая агломеративная кластеризация





Дендрограмма

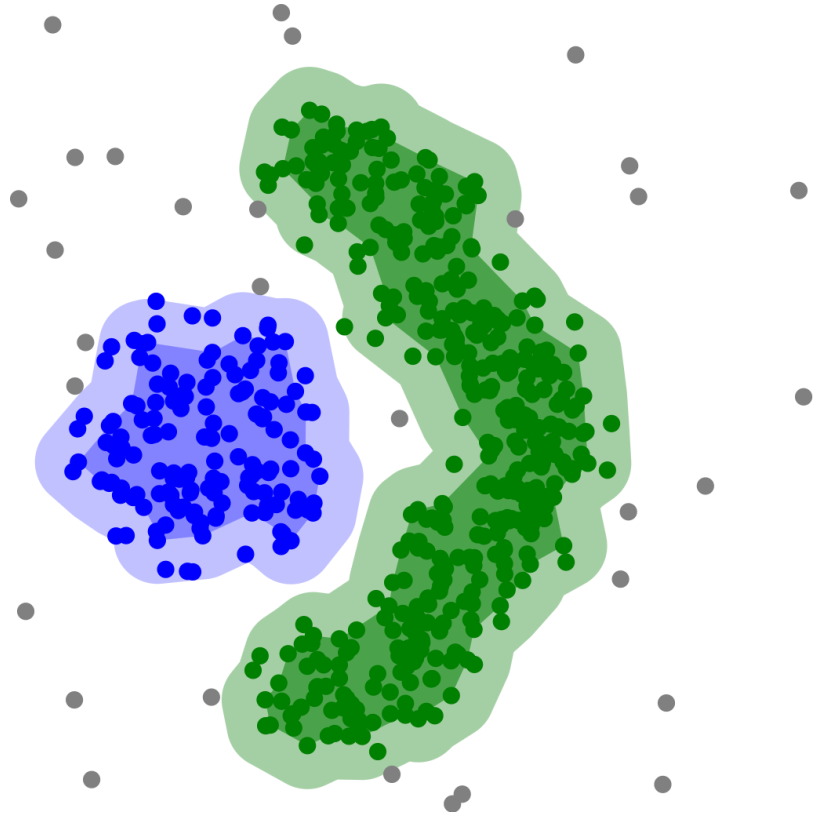
Расстояние между двумя кластерами A и B может оцениваться как среднее расстояние между элементами этих кластеров d(a, b):

$$\frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

DBSCAN

Это алгоритм кластеризации, основанной на плотности — если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями), и помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко).

- Не требует задания числа кластеров.
- Устойчив к выбросам.
- Может найти нелинейно разделимые кластеры.



Снижение размерности

Уменьшение размерности данных — это подход упрощения сложных наборов данных для облегчения их обработки. Методы уменьшения размерности позволяют представить многомерные данные в пространстве меньшей размерности, т.е. меньшим количеством измерений (столбцов) при сохранении наиболее важной информации.

Линейные методы

Позволяют уменьшить размерность данных при сохранении наиболее важной информации. Они фиксируют исходные закономерности и линейные взаимосвязи в данных, предлагая способ их представления в пространстве меньшего размера. Основные линейные методы:

Principal Component Analysis (PCA) – анализ главных компонент;

Independent Component Analysis (ICA) – анализ независимых компонент;

Truncated Singular Value Decomposition (TruncatedSVD) - усеченная декомпозиция сингулярных значений.

Нелинейные методы

Пытаются зафиксировать более сложные нелинейные взаимосвязи в данных и представить их в виде пространств меньшей размерности. Основные нелинейные методы:

Multidimensional Scaling (MDS) - многомерное масштабирование;

t-Distributed Stochastic Neighbor Embedding (T-SNE) - стохастическое вложение соседей с t-распределением

Uniform Manifold Approximation and Projection (UMAP)

Использование Scikit-learn (sklearn) для кластеризации и снижения размерности

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import seaborn as sns
import matplotlib.pyplot as plt

X = df
scaler = StandardScaler()
X = scaler.fit_transform(X)
X = pd.DataFrame(data = X, columns = df.columns)

kmeans = KMeans(n_clusters = 3)
cluster = kmeans.fit_predict(X)

pca2D = PCA(n_components=2)
pca_2D = pca2D.fit_transform(X)
pca2D_df = pd.DataFrame(data = pca_2D, columns = ['x', 'y'])
pca2D_df['cluster'] = cluster

sns.scatterplot(x='x', y='y', hue='cluster', data=pca2D_df)
plt.title("cluster+PCA")
plt.show()
```