Коэффициент силуэта

Коэффициент силуэта вычисляется по следующей формуле:

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)},$$

где:

- a_i — среднее расстояние от данного объекта x_i до объектов из того же кластера;
- b_i — среднее расстояние от данного объекта x_i до объектов из другого ближайшего кластера.

Индекс Калински — Харабаса

Следующий коэффициент, который мы рассмотрим, — это **индекс Калински — Харабаса**. Он показывает отношение между разбросом значений между кластерами и разбросом значений внутри кластеров и вычисляется по следующей формуле:

$$\frac{SS_B}{SS_W} \times \frac{(N-K)}{(K-1)}$$

В данной формуле:

- N — общее количество объектов;
- K — количество кластеров;
- SS_B — взвешенная межкластерная сумма квадратов расстояний;
- SS_W — внутрикластерная сумма квадратов расстояний.

Первый шаг — рассчитать взвешенную межкластерную сумму квадратов расстояний:

$$SS_B = \sum_{k=1}^{K} n_k \times \|C_k - C\|^2$$

Второй шаг — рассчитать внутрикластерную сумму квадратов.

$$\sum_{i=1}^{n_k} \|X_{ik} - C_k\|^2$$

В данной формуле:

- n_k — количество наблюдений в кластере k;
- X_{ik} — i-ое наблюдение в кластере k;
- C_k — центроид кластера k.

Такое значение мы рассчитываем для каждого кластера, а потом уже складываем их для получения значения SS_w.

Индекс Дэвиса — Болдина

Перейдём к последнему из трёх наиболее важных для нас коэффициентов — **индексу Дэвиса — Болдина**. Рассмотрим процесс его вычисления сразу по шагам, так как он реализуется достаточно сложно.

Шаг 1

Для начала вычисляем для каждого кластера следующую меру разброса значений (*в контексте этой формулы её называют компактностью*) внутри него:

$$S_k = \left\{ \frac{1}{n_k} \sum_{i=1}^{n_k} |X_{ik} - C_k|^q \right\}^{\frac{1}{q}}$$

В данной формуле:

- n_k — количество наблюдений в кластере k;
- X_{ik} — i-ое наблюдение в кластере k;
- C_k — центроид кластера k;
- q обычно принимает значение 2 (в этом случае мы рассматриваем уже привычное нам евклидово расстояние).

Шаг 2

Далее находим расстояния между центроидами кластеров (*этот показатель называют отделимостью*):

$$M_{i,j} = \|C_i - C_j\|_q$$

Шаг 3

Теперь для каждой пары кластеров вычисляем следующее отношение:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}$$

Также для каждого кластера находим максимум из полученных значений:

$$R_i \equiv \mathbf{maximum}(R_{ij})$$

Шаг 4

Усредняем значения, найденные в предыдущем пункте, — это и будет итоговое значение индекса:

$$DBI = \frac{1}{N}\sum_{i=1}^{N} R_i$$

Внутрикластерное расстояние

Для того чтобы оценить качество кластеризации, можно вычислить суммарное внутрикластерное расстояние:

$$F_0 = \sum_{k=1}^{K}\sum_{i=1}^{N}[a(x_i) = k]\rho(x_i, c_k)$$

Межкластерное расстояние

Аналогично суммарному внутрикластерному расстоянию, вводится межкластерное расстояние:

$$F_1 = \sum_{i,j=1}^{N}[a(x_i) \neq a(x_j)]\rho(x_i, x_j)$$

Отношение расстояний

Логичным образом из предыдущих двух метрик (внутрикластерного и межкластерного расстояний) мы получаем отношение расстояний:

$$\frac{F_0}{F_1} \to \min$$

ВНУТРЕННЯЯ МЕРА	ИНТЕРПРЕТАЦИЯ	ДИАПАЗОН ЗНАЧЕНИЙ	ФУНКЦИЯ В БИБЛИОТЕКЕ SKLEARN
Коэффициент силуэта	Мера того, насколько объект похож на объекты из своего собственного кластера по сравнению с объектами из других кластеров.	От -1 до 1: при качественной кластеризации значение близко к 1.	silhouette_score()
Индекс Калински — Харабаса	Показывает отношение между разбросом значений между кластерами и разбросом значений внутри кластеров. Оценка выше, когда кластеры плотные и хорошо разделены.	Любое неотрицательное значение. Чем больше значение, тем лучше.	calinski_harabasz_score()
Индекс Дэвиса — Болдина	Показывает среднюю «схожесть» между кластерами.	Не менее 0. Чем меньше значение, тем лучше.	davies_bouldin_score()
Внутрикластерное расстояние	Показывает, насколько плотно расположены объекты в кластерах.	Не менее 0. Чем меньше значение, тем лучше.	Не реализовано

Межкластерное расстояние	Показывает, насколько далеки друг от друга элементы из разных кластеров.	Не менее 0. Чем больше значение, тем лучше.	Не реализовано

Индекс Рэнда

Индекс Рэнда вычисляется по следующей формуле:

$$\mathrm{RI} = \frac{2(a+b)}{N(N-1)},$$

где:

- N — количество объектов в выборке;
- a — число пар объектов, которые имеют одинаковые метки (т. е. в фактическом разбиении находятся в одном классе) и располагаются в одном кластере;
- b — число пар объектов, которые имеют различные метки (т. е. в фактическом разбиении находятся в разных классах) и располагаются в разных кластерах.

Также используют **скорректированный индекс Рэнда (Adjusted Rand Index)**:

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}$$

Его преимущество перед обычным индексом Рэнда состоит в том, что при случайных кластеризациях его значение близко к нулю вне зависимости от количества кластеров и наблюдений.

Нормализованная взаимная информация

Следующая мера — **NMI (Normalized Mutual Information)**, или **нормализованная взаимная информация**. Она определяется следующим образом:

$$NMI(Y_{true}, Y_{pred}) = \frac{2 \times I(Y_{true}; Y_{pred})}{[H(Y_{true}) + H(Y_{pred})]}$$

Здесь:

- Y_{true} — реальные значения меток кластеров для элементов;
- Y_{pred} — предсказанные значения меток кластеров для элементов;
- $H()$ — функция, которая называется критерием информативности, или энтропией Шеннона;
- $I()$ — функция взаимной информации.

Однородность

Однородность её называют **гомогенностью**. Она показывает, насколько элементы в кластере похожи между собой, и вычисляется по следующей формуле:

$$\text{homogeneity} = 1 - \frac{H(Y_{\text{true}} \mid Y_{\text{pred}})}{H(Y_{\text{true}})}$$

Полнота

Результат кластеризации удовлетворяет требованиям полноты, если все элементы данных, принадлежащие к одному классу, оказались в одном кластере.

$$\text{completeness} = 1 - \frac{H(Y_{\text{pred}} \mid Y_{\text{true}})}{H(Y_{\text{pred}})}$$

V-мера

$$v = \frac{(1+\beta) \times \text{homogeneity} \times \text{completeness}}{(\beta \times \text{homogeneity} + \text{completeness})}$$

Напомним, что по умолчанию $\beta = 1$, но это значение можно варьировать, если хочется придать разный вес разным свойствам:

- Если однородность кластеров важнее, чем их полнота, следует указать значение $\beta < 1$. Тогда значение $\beta \times \text{homogeneity}$ в знаменателе получится небольшим и тем самым будет сильнее влиять на значение v. Чем меньше $\beta \times \text{homogeneity}$, тем выше v.
- Если однородность кластеров не особо важна, но важно, чтобы каждый кластер содержал максимальное количество похожих объектов, тогда мы регулируем значение β так, чтобы оно было больше 1.

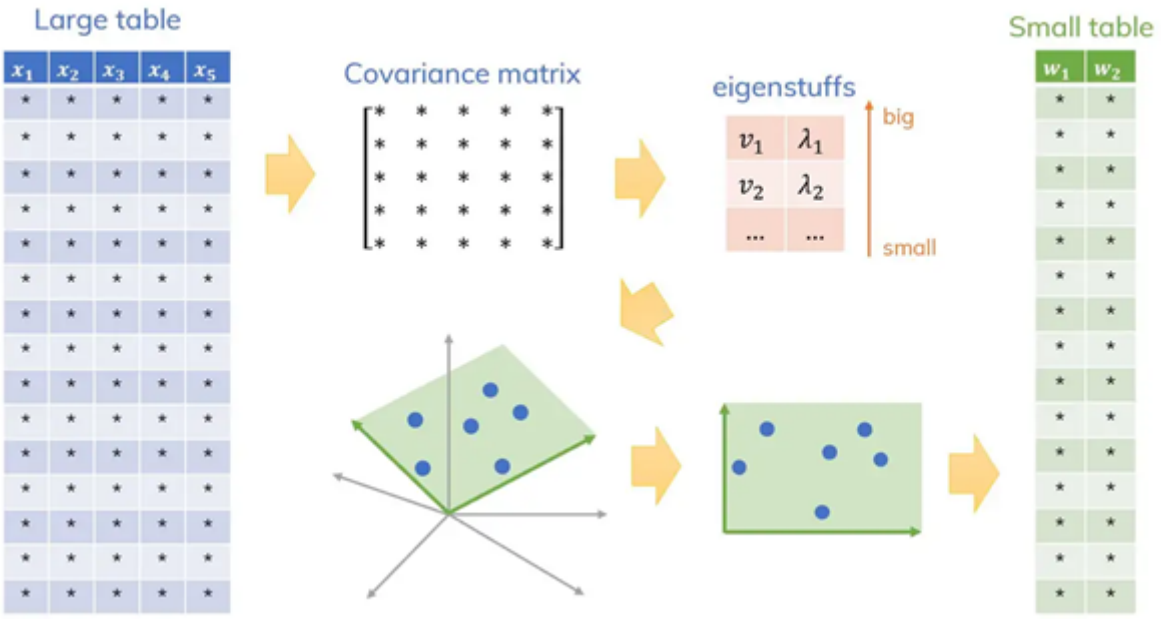
МЕТРИКА	ИНТЕРПРЕТАЦИЯ И ПРИМЕНЕНИЕ	ДИАПАЗОН ЗНАЧЕНИЙ	ФУНКЦИЯ В МОДУЛЕ METRICS БИБЛИОТЕКИ SKLEARN
Однородность (homogeneity score)	Показывает, насколько однородны получившиеся кластеры. Если в кластере оказались элементы из другого кластера, значение метрики уменьшается.	1 — идеально однородные кластеры; 0 — кластеры максимально разнородные.	`homogeneity_score`
Полнота (completeness score)	Показывает, насколько кластер заполнен объектами, которые в действительности должны принадлежать к этому кластеру.	1 — идеальное значение; 0 — объекты, которые должны образовывать один кластер, разделились на большее количество кластеров.	`completeness_score`

V-мера (V-measure)	Комбинация метрик полноты и однородности кластеров.	1 — идеально полные и однородные кластеры; 0 — полученные кластеры неоднородные, количество кластеров слишком большое.	v_measure_score
Индекс Рэнда	Показывает долю объектов датасета, которые мы правильно определили в кластер.	1 — все объекты в предсказанном кластере попали в правильные кластеры.	rand_score
Нормализованная взаимная информация	Показывает, насколько разбиение согласуется с реальными метками.	1 — максимальная согласованность, все объекты находятся в правильных кластерах; 0 — случайное разбиение.	normalized_mutual_info_score

АЛГОРИТМ РЕАЛИЗАЦИИ PCA

1. Стандартизировать данные.
2. Рассчитать ковариационную матрицу для объектов.

3. Рассчитать собственные значения и собственные векторы для ковариационной матрицы.
4. Отсортировать собственные значения и соответствующие им собственные векторы.
5. Выбрать k наибольших собственных значений и сформировать матрицу соответствующих собственных векторов.
6. Преобразовать исходные данные, умножив матрицу данных на матрицу отобранных собственных векторов.



SVD

Суть **сингулярного разложения** заключается в следующей теореме:
Любую прямоугольную матрицу A размера (n, m) можно представить в виде произведения трёх матриц:

$$A_{n\times m} = U_{n\times n} \cdot D_{n\times m} \cdot V^T_{m\times m}$$

В этой формуле:

- U — матрица размера (n, n). Все её столбцы ортогональны друг другу и имеют единичную длину. Такие матрицы называются **ортогональными**. Эта матрица содержит нормированные собственные векторы матрицы AA^T.
- D — матрица размера (n, m). На её главной диагонали стоят числа, называемые **сингулярными** числами (они являются корнями из собственных значений матриц AA^T и A^TA), а вне главной диагонали стоят нули. Если мы решаем задачу снижения размерности, то элементы этой матрицы, если их возвести в квадрат, можно интерпретировать как дисперсию, которую объясняет каждая компонента.
- V — матрица размера (m, m). Она тоже **ортогональная** и содержит нормированные собственные векторы матрицы A^TA.

t-SNE

Итак, вы уже знакомы с двумя алгоритмами для понижения размерности — PCA и SVD. Теперь давайте познакомимся с третьим — **t-SNE (стохастическое вложение соседей с t-распределением)**. Его преимущество относительно первых двух заключается в том, что он может реализовывать уменьшение размерности и разделение для данных, которые являются линейно неразделимыми.

Первый шаг в алгоритме t-SNE включает в себя измерение расстояния от одной точки до всех остальных. Изначально эти расстояния измеряются с помощью обычного евклидова расстояния, а затем сопоставляются со значениями вероятностей.

Функция плотности для нормального распределения записывается следующим образом:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Если мы опустим множитель перед экспонентой, заменим среднее арифметическое на другую точку и отмасштабируем полученное значение, то получим следующее выражение:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

Таким образом мы можем вычислить показатель, отражающий, насколько точка x_j близка к точке x_i при гауссовском распределении, сформированном с математическим ожиданием, равным x_i, и с некоторым σ, которое выбирается таким образом, чтобы у объектов в областях с большей плотностью была более маленькая дисперсия.

Теперь давайте рассмотрим оценки близости для точек, которые получились в результате снижения размерности. Найдём для них ровно тот же показатель:

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

Примечание. Считаем, что x_i переходит в y_i, x_j переходит в y_j и так далее.

Если между точками y_i и y_j будет такое же сходство, как и между изначальными точками x_i и x_j, то значения соответствующих условных вероятностей $p_{j|i}$ и $q_{j|i}$ будут эквивалентными. Чтобы оценить, насколько $q_{j|i}$ близко к $p_{j|i}$, используется **дивергенция (расстояние) Кульбака — Лейблера** (иногда его обозначают просто как KL).

Расстояние KL — это мера различий двух распределений вероятностей.

Чем ниже значение расстояния KL, тем ближе два распределения друг к другу. Расстояние KL, равное 0, подразумевает, что два рассматриваемых распределения идентичны.

После снижения размерности расстояние между двумя точками должно быть значительно больше расстояния, которое можно получить в гауссовом распределении (*потому что, например, при переходе из двухмерного пространства в одномерное, нам надо сразу учесть расстояния по двум осям на одной — для этого расстояния на одной оси должны быть достаточно большими*). Эту проблему называют **«проблемой скученности»**, и для её решения используют распределение Стьюдента.

В t-SNE используется именно распределение Стьюдента. Тогда показатель близости (вероятность) будет вычисляться следующим образом:

$$q_{i|j} = \frac{\left(1+\|y_i-y_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\|y_k-y_l\|^2\right)^{-1}}$$

Градиент функции потерь, соответственно, примет следующий вид:

$$\frac{\partial C}{\partial y_i} = 4\sum_j(p_{ij}-q_{ij})(y_i-y_j)\left(1+\|y_i-y_j\|^2\right)^{-1}$$