

# 1 Кластерный анализ

## 1.1 Кластерный анализ

Цель кластерного анализа — разбить индивиды на кластеры, т.е., на группы, между которыми, в некотором смысле, расстояние больше, чем между точками внутри. Задача не формализована и, можно сказать, плохо поставлены, поэтому решается плохо.

Вообще, кластерный анализ — это ‘обучение без учителя’. Это означает, что вы не сможете формально проверить правильность результата.

Единственный вариант поставить задачу четко — это предположить какую-то статистическую модель данных и в ней находить параметры, например, по методу максимального правдоподобия (model-based clustering).

Все остальные методы — эвристические с плохо определенным (хорошо-плохо) результатом.

## 1.2 Кластерный анализ, пример model-based подхода

Предположим, что многомерная выборка — неоднородная. Но в отличие от дискриминантного анализа у нас нет признака, объясняющего эту неоднородность, и задачей является ее выявить. Тип классификации, когда есть модель, называется model-based clustering. Например, пусть наша выборка из смеси  $k$  нормальных распределений. Таким образом ее плотность имеет вид

$$p(x) = \pi_1 p(x, \mu_1, \Sigma_1) + \dots + \pi_k p(x, \mu_k, \Sigma_k), \quad (1)$$

где

$$p(x, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2} \sqrt{|\Sigma_i|}} \exp \left( -\frac{1}{2} (x - \mu_i) \Sigma_i^{-1} (x - \mu_i)^T \right) \quad (2)$$

Эта задача решается методом максимального правдоподобия. Можно выписать функцию правдоподобия (выпишите), но она имеет сложный вид и искать ее максимум по такому большому числу параметров очень непросто. Для нахождения этого максимума используется так называемый ЕМ-алгоритм (Expectation - Maximization). Мы не будем здесь его обсуждать.

## 1.3 Кластерный анализ: $k$ -means, $k$ -means++

Хотим искать кластеры  $C_1, \dots, C_k$  минимизируя следующий функционал

$$\sum_{i=1}^k \sum_{j \in C_i} \|x_j - \mu_i\|^2 \quad (3)$$

по разбиению всего пространства индивидов на  $C_j$  и по всем  $\mu_i$ . Можно делать это по следующему алгоритму:

1. Выбираем случайно  $\mu_1, \dots, \mu_k$ .
2.  $C_j$  — кластер, содержащий точки, которые лежат к  $\mu_j$  ближе, чем к остальным  $\mu_i$ .
3. Для каждого  $C_j$  пересчитываем центр  $\mu_j$  как выборочное среднее элементов из этого кластера.
4. Делаем 2 и 3 пока алгоритм не сойдется.

Проблема метода в том, что у такого функционала много локальных минимумов, и алгоритм может сойтись в значение, далекое от истинного. Метод  $k$ -means++ повторяет алгоритм, приведенный выше, но начальные значения выбираются не случайно, а следующим образом

1. Выбираем случайным образом первый центр  $\mu_1$ .
2. Считаем расстояние от всех точек до ближайшего центра  $\{\rho_i\}$ . После чего выбираем  $x_i$  как новый центр с вероятностью, пропорциональной  $\rho_i$ .
3. Пока количество центров меньше, чем  $k$ , повторяем процедуру.

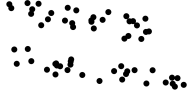



Результат функционала в  $k$ -means для данной процедуры выбора начальных центров запишем как  $J(\{C_j\}, \{\mu_j\})$ . Известно, что при некоторых условиях на форму кластеров

$$\frac{\mathbb{E}(J(\{C_j\}, \{\mu_j\}))}{J_{min}} = O(\ln k),$$

т.е. результат, в среднем, довольно близко к настоящему минимуму.

**Замечание.** *Есть результаты, что если к данным применить анализ главных компонент, то пространство, натянутое на первые  $k - 1$  главных векторов, при некоторых условиях будет близко к пространству, проходящему, через центры кластеров. Поэтому часто с помощью АГК уменьшают число признаков и потом применяют процедуру кластерного анализа.*

## 1.4 «Плохие» кластерные структуры

1.  ленточные кластеры. Внутрикластерные расстояния могут быть больше межкластерных;
2.  перекрывающиеся кластеры;
3.  кластеры, соединяющиеся перемычками и накладывающиеся на фон из редко расположенных объектов;
4.  кластеры могут отсутствовать.

Здесь мы обсуждали, что практически невозможно придумать определение кластера (не статистическое), при котором все эти кластеры будут ему удовлетворять. Вариант смеси нормальных распределений, возможно, подойдет во всех случаях.

Еще обсуждали вопрос, что для данных, где реально обособленных кластеров может и не быть (например, последняя картинка), часто кластеризацией называют сегментацию — просто нарезку на части с описанием каждого сегмента на основе значений признаков.

## 1.5 Иерархический кластерный анализ

### 1.5.1 Расстояние между точками $\rho$

Сначала нужно задать, как мы будем измерять расстояние между точками.

Самое стандартное — евклидово расстояние:  $\rho(x, y) = (\sum_i (x_i - y_i)^2)^{1/2}$ .

Расстояние городских кварталов (манхэттенское расстояние):  $\rho(x, y) = \sum_i |x_i - y_i|$ .

Расстояние Чебышёва:  $\rho(x, y) = \max_i |x_i - y_i|$ .

Процент несогласия (эта мера используется в тех случаях, когда данные являются категориальными):  $\rho(x, y) = (\#\{i : x_i \neq y_i\})/i$ .

Особый случай, если кластеризуются признаки, а не индивиды (а какая разница — такой кластерный анализ не статистическая процедура, ему все равно), то логично в качестве расстояния рассматривать корреляции. Например, 1 минус модуль корреляции или 1 минус просто корреляция, что правильнее по смыслу для задачи.

**Замечание.** Важно либо исходно стандартизовать признаки, либо измерять расстояние специальным образом. Например, использовать расстояние Махаланобиса вместо обычного евклидова, если есть предположения о форме распределения точек внутри кластера.

### 1.5.2 Примеры межкластерных расстояний

Правила слияния кластеров (linkage rule) основывается на расстояниях между кластерами.

Расстояние ближнего соседа (single linkage, кластеры в виде цепочек):

$$R^n(U, V) = \min_{u \in U, v \in V} \rho(u, v), \quad U, V \subset X;$$

расстояние дальнего соседа (complete linkage, кластеры ближе к шарикам):

$$R^l(U, V) = \max_{u \in U, v \in V} \rho(u, v);$$

групповое среднее расстояние:

$$R^g(U, V) = \frac{1}{|U||V|} \sum_{u \in U} \sum_{v \in V} \rho(u, v);$$

расстояние между центрами:

$$R^c(U, V) = \rho^2 \left( \sum_{u \in U} \frac{u}{|U|}, \sum_{v \in V} \frac{v}{|V|} \right);$$

расстояние Уорда:

$$R^w(U, V) = \frac{|U||V|}{|U| + |V|} R^c(U, V).$$

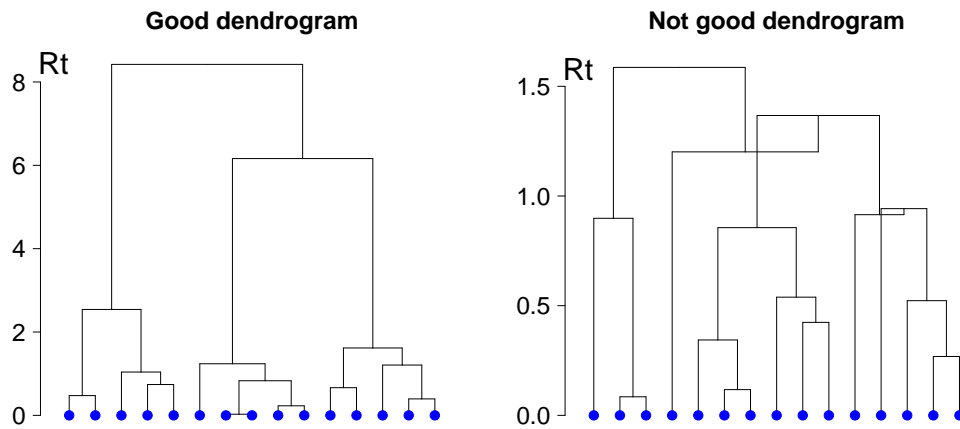
### 1.5.3 Алгоритм агломеративной иерархической кластеризации

1. Сначала все кластеры одноэлементные:  $C_1 = \{\{x_1\}, \dots, \{x_l\}\}$ ;  $R_1 = 0$ ;  
 $\forall i \neq j$  вычислить  $R(\{x_i\}, \{x_j\})$ ;
2. для всех  $t = 2, \dots, l$  ( $t$  — номер итерации)

3. найти в  $C_{t-1}$  два ближайших кластера:  
 $(U, V) = \arg \min_{U \neq V} R(U, V);$   
 $R_t = R(U, V);$
4. слить их в один кластер:  
 $W = U \cup V;$   
 $C_t = C_{t-1} \cup W \setminus \{U, V\};$
5. для всех  $S \in C_t \setminus W$
6. вычислить  $R(W, S);$

#### 1.5.4 Визуализация кластерной структуры

**Определение.** Дендрограмма — деревоподобный график, отражающий процесс последовательных слияний и структуру кластеров.



После построения дерева можно его разрезать на поддеревья по заданному расстоянию между кластерами и получить сами кластеры. Разрез делается там, где долго не было объединения кластеров (длинная ветка у дерева).

Но долго-недолго — это субъективно и зависит от выбранного расстояния. Если расстояние в квадрате, то дальние ветки искусственно удлиняются.