Обучение без учителя. Разделение смеси распределений. Кластеризация.

Сальников Дмитрий Игоревич, гр. 622

Санкт-Петербургский государственный университет
Прикладная математика и информатика

Санкт-Петербург
2017г.

Обучение без учителя

Особенности:

- известны только описания множества объектов X и обучающей выборки $X^l = \{x_i\}_{i=1}^l \subset X$;
- отсутствует отклик (зависимая переменная);
- требуется обнаружить внутренние взаимосвязи, зависимости, закономерности между объектами.

Обучение без учителя

Типы задач:

- Задачи кластеризации;
- Задачи поиска правил ассоциации;
- Задачи сокращения размерности;
- Задачи визуализации данных.

Анализ главных компонент

- позволяет сократить размерность входных данных;
- является средством визуализации данных.

Дано:
$X = \mathbb{R}^n$, $X^l = [X_1, \ldots, X_n]$, $\overline{X_i} = 0$, $i = \overline{1:n}$.

Задача:

- аппроксимировать данные подпространством меньшей размерности $s < n$:

$$\sum_{i=1}^{n} dist^2(x_i, \alpha_s) \longrightarrow \min_{\substack{\alpha_s \subset \mathbb{R}^n, \\ dim(\alpha_s)=s}}.$$

Анализ главных компонент

- позволяет сократить размерность входных данных;
- является средством визуализации данных.

Дано:
$X = \mathbb{R}^n$, $X^l = [X_1, \ldots, X_n]$, $\overline{X_i} = 0$, $i = \overline{1:n}$.

Задача:

- аппроксимировать данные подпространством меньшей размерности $s < n$:

$$\sum_{i=1}^{n} dist^2(x_i, \alpha_s) \longrightarrow \min_{\substack{\alpha_s \subset \mathbb{R}^n, \\ dim(\alpha_s)=s}}.$$

Анализ главных компонент

$S = \frac{1}{l-1} X^T X$ – выборочная ковариационная матрица.

SVD(S):

$$X^T_{l\times n} = U_{n\times k} \Lambda^{\frac{1}{2}}_{k\times k} V^T_{l\times k},\ k = \sum_{j=1}^{n} I_{\{\lambda_j \neq 0\}} \leq \min(l-1, n),$$

$U = [U_1, \ldots, U_k]$, $U_i \perp U_j$, $||U_j|| = 1$ – матрица главных направлений (перехода к новым признакам),
$\Lambda = diag(\lambda_1, \ldots, \lambda_k)$, $\lambda_1 \geq \cdots \geq \lambda_k > 0$ – диагональная матрица ненулевых собственных чисел матрицы S,
$V = [V_1, \ldots, V_k]$, $V_i \perp Vj$, $||V_j|| = 1$ – факторные векторы,

$Z = [Z_1, \ldots, Z_k] = XU$, $Z_j = \sqrt{\lambda_j} V_j$ – матрица главных компонент (новых признаков).

Анализ главных компонент

$S = \frac{1}{l-1} X^T X$ – выборочная ковариационная матрица.

SVD(S):

$X^T_{l\times n} = U_{n\times k} \Lambda^{\frac{1}{2}}_{k\times k} V^T_{l\times k}$, $k = \sum_{j=1}^{n} I_{\{\lambda_j \neq 0\}} \leq \min(l-1, n)$,

$U = [U_1, \ldots, U_k]$, $U_i \perp U_j$, $||U_j|| = 1$ – матрица главных направлений (перехода к новым признакам),

$\Lambda = diag(\lambda_1, \ldots, \lambda_k)$, $\lambda_1 \geq \cdots \geq \lambda_k > 0$ – диагональная матрица ненулевых собственных чисел матрицы S,

$V = [V_1, \ldots, V_k]$, $V_i \perp Vj$, $||V_j|| = 1$ – факторные векторы,

$Z = [Z_1, \ldots, Z_k] = XU$, $Z_j = \sqrt{\lambda_j} V_j$ – матрица главных компонент (новых признаков).

Анализ главных компонент

Подпространство, натянутое на $[U_1, \ldots, U_s]$, $s \leq k$ решает поставленную задачу.

Компоненты, соответствующие большим λ_j, объясняют больший процент дисперсии:

$$Var(Z_j) = \lambda_j,\ j = 1, \ldots, k;\ \sum_{j=1}^{n} Var(X_j) = \sum_{j=1}^{k} \lambda_j.$$

$\frac{\lambda_j}{\sum_{s=1}^{k} \lambda_s}$ – доля объясненной дисперсии компоненты Z_j.

Методы выбора числа главных компонент

- фиксируем минимальную долю объясненной дисперсии S_{min}, тогда $k' = \arg\min_s \left\{ \frac{\sum_{j=1}^{s} \lambda_j}{\sum_{j=1}^{k} \lambda_j} \geq S_{min} \right\}$;
- правила Кайзера: $k' = \sum_{j=1}^{k} I_{\left\{\lambda_j > \frac{1}{k}\sum_{s=1}^{k}\lambda_s\right\}}$ (собственные числа, большие среднего значения);
- правило сломанной трости: $k' = \arg\max_s \left\{ \frac{\lambda_1}{\sum_{j=1}^{k}\lambda_j} > \frac{1}{k}\sum_{j=1}^{k}\frac{1}{j}, \ldots, \frac{\lambda_s}{\sum_{j=1}^{k}\lambda_j} > \frac{1}{k}\sum_{j=s}^{k}\frac{1}{j} \right\}$ ($\frac{1}{k}\sum_{j=s}^{k}\frac{1}{j}$ – МО упорядоченных по убыванию длин кусков трости длины 1, случайно поломанной на k частей);
- выбор количества компонент на усмотрение эксперта.

Стандартизация данных

Перед применением АГК рекомендуется привести данные к одной шкале; простой путь – стандартизация признаков.

Пример

Пусть ковариационная матрица S имеет вид

$$S = \begin{pmatrix} a^2 & a\rho \\ a\rho & 1 \end{pmatrix},$$

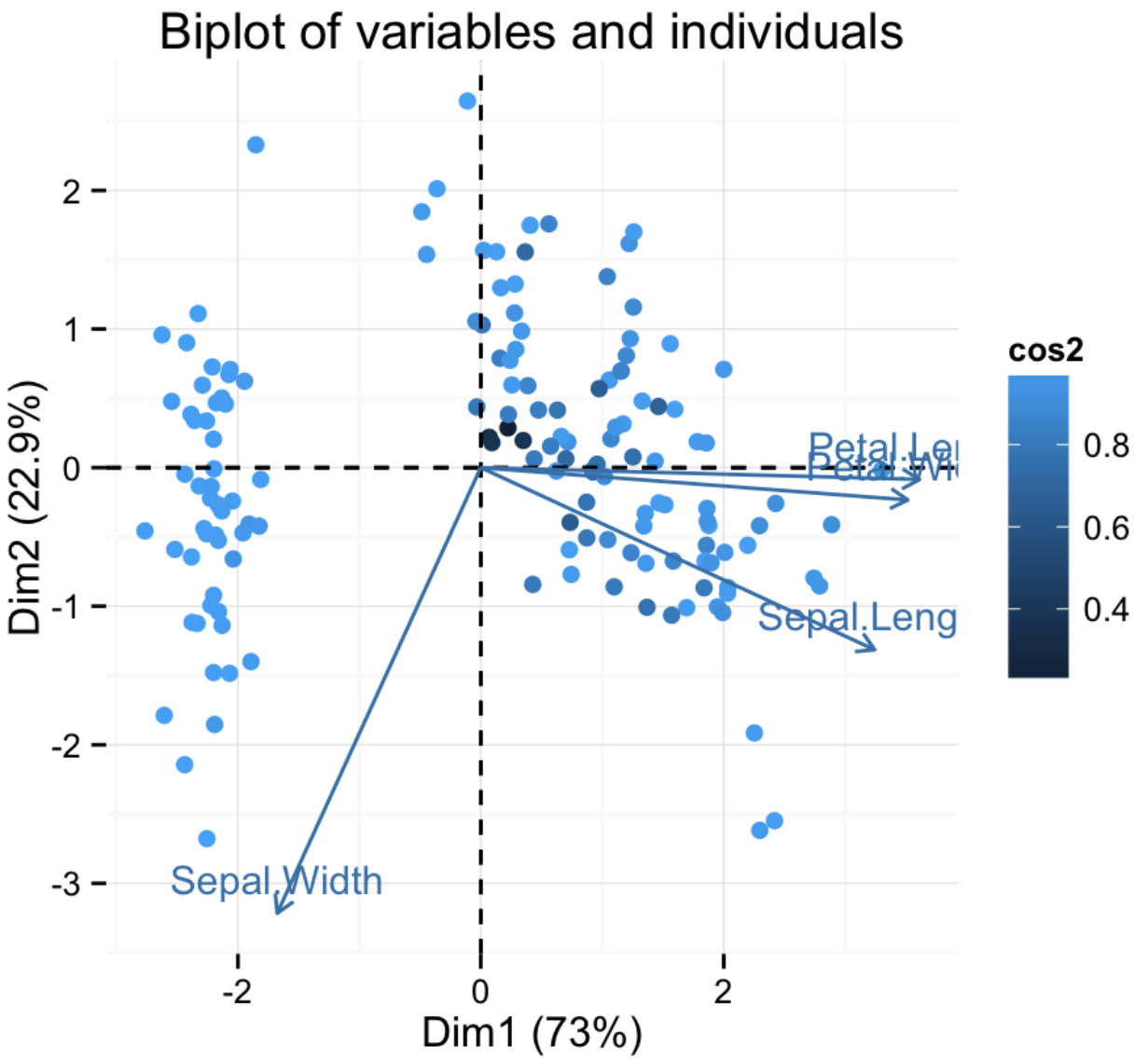
тогда при $a \to \infty$

$$\frac{u_{11}}{u_{12}} \to \frac{a}{\rho},$$

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} \to 1.$$

Вклад первого признака X_1 в U_1 будет большим, а компонента Z_2 будет иметь малую долю объясненной дисперсии.

Пример визуализации



Пояснения к графику

В плоскости первых двух компонент $span(U_1, U_2)$ (объясняющих больше 95% дисперсии) изображены объекты и исходные орты с координатами (U_{i1}, U_{i2}), $i = \overline{(1:n)}$.

Шкала справа показывает, насколько хорошо объекты описываются плоскостью $span(U_1, U_2)$. Согласно формуле

$$\cos^2(\angle(x_i, span(U_1, U_2))) = \left(\frac{x_i U_1}{||x_i||\,||U_1||}\right)^2 + \left(\frac{x_i U_2}{||x_i||\,||U_2||}\right)^2 = \\ = (z_{i1}^2 + z_{i2}^2)/||x_i||^2$$

объекты, для которых $\cos^2$ мал, плохо описываются плоскостью $span(U_1, U_2)$. Такие объекты окрашены в темный цвет.

Выводы

- Метод не требует выполнения каких-либо гипотез и может быть применен к любым данным;
- требует предварительного центрирования и масштабирования;
- позволяет значительно сократить размерность пространства признаков;
- Проекции на первые два или три главных направления позволяют наглядно визуализировать данные.

Разделение смеси распределений

Гипотеза (о вероятностной природе данных)

X^l – *независимы, одинаково распределены с плотностью*

$$p(x) = \sum_{j=1}^{k} w_j p_j(x;\theta_j),\ \sum_{j=1}^{k} w_j = 1,\ w_j \geq 0,\ j = \overline{1:k}.$$

Задачи:

1. зная k оценить $\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$;
2. оценить k.

Аналитическое решение задачи 1:

$$L(\Theta) = \ln \prod_{i=1}^{l} p(x_i) = \sum_{i=1}^{l} \ln \sum_{j=1}^{k} w_j p_j(x_i;\theta_j) \to \max_{\Theta}.$$

Разделение смеси распределений

Гипотеза (о вероятностной природе данных)

X^l – *независимы, одинаково распределены с плотностью*

$$p(x) = \sum_{j=1}^{k} w_j p_j(x;\theta_j),\ \sum_{j=1}^{k} w_j = 1,\ w_j \geq 0,\ j = \overline{1:k}.$$

Задачи:

1. зная k оценить $\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$;
2. оценить k.

Аналитическое решение задачи 1:

$$L(\Theta) = \ln \prod_{i=1}^{l} p(x_i) = \sum_{i=1}^{l} \ln \sum_{j=1}^{k} w_j p_j(x_i;\theta_j) \to \max_{\Theta}.$$

Разделение смеси распределений

Гипотеза (о вероятностной природе данных)

X^l – *независимы, одинаково распределены с плотностью*

$$p(x) = \sum_{j=1}^{k} w_j p_j(x;\theta_j),\ \sum_{j=1}^{k} w_j = 1,\ w_j \geq 0,\ j = \overline{1:k}.$$

Задачи:

1. зная k оценить $\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$;
2. оценить k.

Аналитическое решение задачи 1:

$$L(\Theta) = \ln \prod_{i=1}^{l} p(x_i) = \sum_{i=1}^{l} \ln \sum_{j=1}^{k} w_j p_j(x_i;\theta_j) \to \max_{\Theta}.$$

Теорема (необходимые условия экстремума)

Θ – точка локального экстремума $L(\Theta)$, если она удовлетворяет следующей системе уравнений:

$$g_{ij} = [P(j|x_i)] = \frac{w_j p_j(x_i;\theta_j)}{\sum\limits_{s=1}^{k} w_s p_s(x_i;\theta_s)}, \qquad \begin{matrix} i = \overline{1:l}, \\ j = \overline{1:k}; \end{matrix} \qquad \textit{(E–шаг)}$$

$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{l} g_{ij} \ln p_j(x_i;\theta), \qquad j = \overline{1:k}. \qquad \textit{(M–шаг)}$$

$$w_j = \frac{1}{l} \sum_{i=1}^{l} g_{ij},$$

EM–алгоритм

1. задать k, δ, начальные $\Theta = (w_j, \theta_j)_{j=1}^{k}$ и $\{g_{ij}\}_{i,j=1}^{l,k}$;
2. повторять
3. E–шаг (expectation):
   для всех $i = 1, \ldots, l$, $j = 1, \ldots, k$

$$g_{ij}^0 = g_{ij}; \qquad g_{ij} = \frac{w_j p_j(x_i; \theta_j)}{\sum_{s=1}^{k} w_s p_s(x_i; \theta_s)};$$

4. M–шаг (maximization):
   для всех $j = 1, \ldots, k$

$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{l} g_{ij} \ln p_j(x_i; \theta); \qquad w_j = \frac{1}{l} \sum_{i=1}^{l} g_{ij};$$

5. пока $\max_{i,j} |g_{ij} - g_{ij}^0| > \delta$;

EM–алгоритм

Достоинства:

- сводит сложную многоэкстремальную задачу к максимизации правдоподобия по компонентам смеси;
- довольно быстрая скорость сходимости.

Проблемы:

- чувствителен к начальному приближению;
- проблема выбора числа компонент смеси.

Добавление и удаление компонент смеси

- Если много объектов x_i имеют низкие правдоподобия $p(x_i)$, то создаем $k+1$-ю компоненту и по этим объектам стоим ее начальное приближение;
- (Иерархический EM) Для каждой компоненты $j = 1 \dots k$ смотрим на соответствующие ей объекты x_i $(\arg\max\limits_{j=1\dots k} g_{ij})$. Если много таких x_i имеют низкие правдоподобия $p_j(x_i)$, то расщепляем компоненту на две;
- если у j-ой компоненты низкий w_j – удаляем ее;
- регуляризация: $L(\Theta) - \tau \sum\limits_{j=1}^{k} \ln w_j \to \max$, тогда

θ_j – те же, а $w_j = \left(\frac{1}{l} \sum\limits_{i=1}^{l} g_{ij} - \tau \right)_+$.

Добавление и удаление компонент смеси

- Если много объектов x_i имеют низкие правдоподобия $p(x_i)$, то создаем $k+1$-ю компоненту и по этим объектам стоим ее начальное приближение;
- (Иерархический EM) Для каждой компоненты $j = 1 \dots k$ смотрим на соответствующие ей объекты x_i $(\arg\max\limits_{j=1\dots k} g_{ij})$.

  Если много таких x_i имеют низкие правдоподобия $p_j(x_i)$, то расщепляем компоненту на две;
- если у j-ой компоненты низкий w_j – удаляем ее;
- регуляризация: $L(\Theta) - \tau \sum\limits_{j=1}^{k} \ln w_j \to \max$, тогда

  θ_j – те же, а $w_j = \left(\frac{1}{l} \sum\limits_{i=1}^{l} g_{ij} - \tau \right)_+$.

Добавление и удаление компонент смеси

- Если много объектов x_i имеют низкие правдоподобия $p(x_i)$, то создаем $k+1$-ю компоненту и по этим объектам стоим ее начальное приближение;
- (Иерархический EM) Для каждой компоненты $j = 1 \dots k$ смотрим на соответствующие ей объекты x_i $(\arg\max\limits_{j=1\dots k} g_{ij})$.

  Если много таких x_i имеют низкие правдоподобия $p_j(x_i)$, то расщепляем компоненту на две;
- если у j-ой компоненты низкий w_j – удаляем ее;
- регуляризация: $L(\Theta) - \tau \sum\limits_{j=1}^{k} \ln w_j \to \max$, тогда

  θ_j – те же, а $w_j = \left(\frac{1}{l} \sum\limits_{i=1}^{l} g_{ij} - \tau \right)_+$.

Добавление и удаление компонент смеси

- Если много объектов x_i имеют низкие правдоподобия $p(x_i)$, то создаем $k+1$-ю компоненту и по этим объектам стоим ее начальное приближение;
- (Иерархический EM) Для каждой компоненты $j = 1 \dots k$ смотрим на соответствующие ей объекты x_i $(\arg\max\limits_{j=1\dots k} g_{ij})$.

  Если много таких x_i имеют низкие правдоподобия $p_j(x_i)$, то расщепляем компоненту на две;
- если у j-ой компоненты низкий w_j – удаляем ее;
- регуляризация: $L(\Theta) - \tau \sum\limits_{j=1}^{k} \ln w_j \to \max$, тогда

  θ_j – те же, а $w_j = \left(\frac{1}{l} \sum\limits_{i=1}^{l} g_{ij} - \tau \right)_+$.

Модификации EM–алгоритма

Обобщенный EM: на M-шаге при вычислении $\theta_j = \arg\max_\theta \sum_{i=1}^{l} g_{ij} \ln p_j(x_i;\theta)$ достаточно лишь сместиться в направлении максимума, не добиваясь высокой точности.
Преимущество: уменьшается время работы при сопоставимом качестве решения.

Стохастический EM: на M-шаге предварительно $\forall x_i$ промоделировать принадлежность к одному из классов X_j согласно распределению с весами g_{ij}, $j = 1 \ldots k$.
Затем максимизировать невзвешенные правдоподобия $\theta_j = \arg\max_\theta \sum_{x_i \in X_j} \ln p_j(x_i;\theta)$.

Преимущество: меньше зависит от начального приближения, способен ближе подбираться к глобальному максимуму, довольно быстро работает.

Модификации EM–алгоритма

Обобщенный EM: на M-шаге при вычислении $\theta_j = \arg\max_\theta \sum_{i=1}^{l} g_{ij} \ln p_j(x_i;\theta)$ достаточно лишь сместиться в направлении максимума, не добиваясь высокой точности.
Преимущество: уменьшается время работы при сопоставимом качестве решения.

Стохастический EM: на M-шаге предварительно $\forall x_i$ промоделировать принадлежность к одному из классов X_j согласно распределению с весами g_{ij}, $j = 1 \ldots k$.
Затем максимизировать невзвешенные правдоподобия $\theta_j = \arg\max_\theta \sum_{x_i \in X_j} \ln p_j(x_i;\theta)$.
Преимущество: меньше зависит от начального приближения, способен ближе подбираться к глобальному максимуму, довольно быстро работает.

Смесь нормальных распределений

Гипотеза (о пространстве объектов и форме кластеров)

$X = \mathbb{R}^n$, $x = (x_1, \ldots, x_n)$, $\theta_j = (\mu_{j1}, \ldots, \mu_{jn}, \sigma_{j1}, \ldots, \sigma_{jn})$,

$$p_j(x; \theta_j) = (2\pi)^{-\frac{n}{2}} (\sigma_{j1} \cdots \sigma_{jn})^{-1} \exp\left(-\frac{1}{2} \sum_{s=1}^{n} \sigma_{js}^{-2} |x_s - \mu_{js}|^2\right)$$

$\mu_j = (\mu_{j1}, \ldots, \mu_{jn})$ – *центр кластера* j,
$\Sigma_j = \operatorname{diag}(\sigma_{j1}^2, \ldots, \sigma_{jn}^2)$ – *диагональная матрица ковариаций.*

Тогда на M-шаге:

$$\mu_{js} = \frac{1}{l w_j} \sum_{i=1}^{l} g_{ij} x_{is}, \; j = \overline{1:k}, \; s = \overline{1:n},$$

$$\sigma_{js}^2 = \frac{1}{l w_j} \sum_{i=1}^{l} g_{ij} (x_{is} - \mu_{js})^2, \; j = \overline{1:k}, \; s = \overline{1:n}.$$

Смесь нормальных распределений

Гипотеза (о пространстве объектов и форме кластеров)

$X = \mathbb{R}^n$, $x = (x_1, \ldots, x_n)$, $\theta_j = (\mu_{j1}, \ldots, \mu_{jn}, \sigma_{j1}, \ldots, \sigma_{jn})$,

$$p_j(x; \theta_j) = (2\pi)^{-\frac{n}{2}} (\sigma_{j1} \cdots \sigma_{jn})^{-1} \exp\left(-\frac{1}{2} \sum_{s=1}^{n} \sigma_{js}^{-2} |x_s - \mu_{js}|^2\right)$$

$\mu_j = (\mu_{j1}, \ldots, \mu_{jn})$ – *центр кластера* j,
$\Sigma_j = \operatorname{diag}(\sigma_{j1}^2, \ldots, \sigma_{jn}^2)$ – *диагональная матрица ковариаций.*

Тогда на M-шаге:

$$\mu_{js} = \frac{1}{l w_j} \sum_{i=1}^{l} g_{ij} x_{is}, \; j = \overline{1:k}, \; s = \overline{1:n},$$

$$\sigma_{js}^2 = \frac{1}{l w_j} \sum_{i=1}^{l} g_{ij} (x_{is} - \mu_{js})^2, \; j = \overline{1:k}, \; s = \overline{1:n}.$$

Кластеризация

Дано:

- X – пространство объектов;
- $X^l = \{x_i\}_{i=1}^l$ – обучающая выборка;
- $\rho : X \times X \to [0, \inf)$ – функция расстояния между объектами;

Найти:

- Y – множество кластеров;
- $a : X \to Y$ – алгоритм кластеризации, такой что:
  - каждый кластер состоит из близких объектов;
  - объекты разных кластеров существенно различны.

Некорректность задачи кластеризации

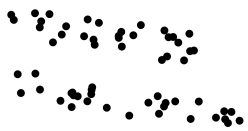
Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров $|Y|$, как правило, неизвестно заранее;
- результат кластеризации существенно зависит от метрики ρ, задаваемой субъективно.

Цели кластеризации

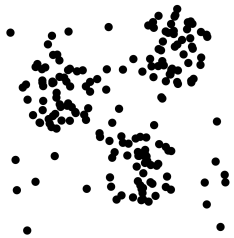
- Упростить дальнейшую обработку данных, разбить множество X^l на группы схожих объектов, чтобы работать с каждой группой в отдельности (задачи классификации, регрессии, прогнозирования);
- сократить объем хранимых данных, оставив по одному представителю от каждого кластера (задачи сжатия данных);
- выделить нетипичные объекты, которые не подходят ни к одному из кластеров (задачи одноклассовой классификации);
- построить иерархию множества объектов (задачи таксономии).

«Плохие» кластерные структуры

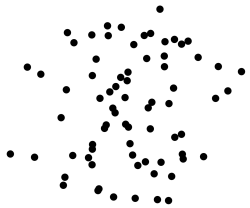


ленточные кластеры. Внутрикластерные расстояния могут быть больше межкластерных;

перекрывающиеся кластеры;



кластеры, соединяющиеся перемычками и накладывающиеся на фон из редко расположенных объектов;



кластеры могу отсутствовать.

Нечеткий алгоритм c-средних

Обозначения:

$\mu = \{\mu_j\}_{j=1}^{k}$ – центры кластеров,

$U = \{U_{ij}\}_{l \times k}$ – матрица принадлежности объектов x_i кластеру j,

$\sum\limits_{j=1}^{k} U_{ij} = 1,\ i = \overline{1:l},$

$m > 1$ – параметр «размытости»; Обычно полагают $m = 2$.

Решаемая оптимизационная задача:

$$\sum_{i=1}^{l}\sum_{j=1}^{k} U_{ij}^{m}\rho(x_i, \mu_j) \to \min_{U,\,\mu_j}.$$

При $m \to 1$ вырождается в алгоритм k-средних, при этом

$$\forall\, i = 1, \ldots, l\ \exists\, j_i \in \{1, \ldots, k\}\colon U_{ij_i} = 1,\ U_{ij} = 0,\ j \neq j_i.$$

Нечеткий алгоритм c-средних

Обозначения:

$\mu = \{\mu_j\}_{j=1}^{k}$ – центры кластеров,

$U = \{U_{ij}\}_{l \times k}$ – матрица принадлежности объектов x_i кластеру j,

$\sum\limits_{j=1}^{k} U_{ij} = 1,\ i = \overline{1:l},$

$m > 1$ – параметр «размытости»; Обычно полагают $m = 2$.

Решаемая оптимизационная задача:

$$\sum_{i=1}^{l} \sum_{j=1}^{k} U_{ij}^{m} \rho(x_i, \mu_j) \to \min_{U,\, \mu_j}.$$

При $m \to 1$ вырождается в алгоритм k-средних, при этом $\forall\, i = 1, \ldots, l\ \exists\, j_i \in \{1, \ldots, k\}\colon U_{ij_i} = 1,\ U_{ij} = 0,\ j \neq j_i.$

Нечеткий алгоритм c-средних

Обозначения:

$\mu = \{\mu_j\}_{j=1}^{k}$ – центры кластеров,

$U = \{U_{ij}\}_{l \times k}$ – матрица принадлежности объектов x_i кластеру j,

$\sum_{j=1}^{k} U_{ij} = 1,\ i = \overline{1:l},$

$m > 1$ – параметр «размытости»; Обычно полагают $m = 2$.

Решаемая оптимизационная задача:

$$\sum_{i=1}^{l} \sum_{j=1}^{k} U_{ij}^{m} \rho(x_i, \mu_j) \to \min_{U,\, \mu_j}.$$

При $m \to 1$ вырождается в алгоритм k-средних, при этом

$$\forall\ i = 1, \ldots, l\ \exists\ j_i \in \{1, \ldots, k\}\colon U_{ij_i} = 1,\ U_{ij} = 0,\ j \neq j_i.$$

Нечеткий алгоритм c-средних

1. Задать k, m, ϵ и начальные μ_j, $j = 1, \ldots, k$;
2. повторять
3. вычислить принадлежность x_i к каждому кластеру:

$$P_{ij} = \rho(x_i, \mu_j);$$
$$U_{ij} = \frac{1}{\sum_{s=1}^{k} \left(\frac{P_{ij}}{P_{is}}\right)^{\frac{1}{m-1}}}; \qquad i = 1, \ldots, l,\ j = 1, \ldots, k;$$

4. пересчитать центры кластеров:

$$\mu_j = \frac{\sum_{i=1}^{l} U_{ij}^m x_i}{\sum_{i=1}^{l} U_{ij}^m},\ j = 1, \ldots, k;$$

5. пока $||U - U^0|| > \epsilon$, либо $\max\limits_{j=1,\ldots,k} \rho(\mu_j, \mu_j^0) > \epsilon$;

Нечеткий алгоритм c-средних

Проблемы:

- чувствительность к начальным приближениям;
- чувствительность к выбросам;
- выбор числа кластеров k;
- в общем случае обладает плохо исследованной сходимостью.

Решения:

- несколько случайных кластеризаций;
- постепенное наращивание k (аналогично EM-алгоритму);
- центры кластеров должны значительно различаться;
- алгоритм k-средних сходится за конечное число шагов;
- нечеткие алгоритмы решают проблему выбросов и нивелируют чувствительность к начальным данным.

Нечеткий алгоритм c-средних

Проблемы:

- чувствительность к начальным приближениям;
- чувствительность к выбросам;
- выбор числа кластеров k;
- в общем случае обладает плохо исследованной сходимостью.

Решения:

- несколько случайных кластеризаций;
- постепенное наращивание k (аналогично EM-алгоритму);
- центры кластеров должны значительно различаться;
- алгоритм k-средних сходится за конечное число шагов;
- нечеткие алгоритмы решают проблему выбросов и нивелируют чувствительность к начальным данным.

Агломеративная иерархическая кластеризация

Алгоритм Ланса–Уильямса:

1. Сначала все кластеры одноэлементные:
   $C_1 = \{\{x_1\}, \ldots, \{x_l\}\}$; $R_1 = 0$;
   $\forall i \neq j$ вычислить $R(\{x_i\}, \{x_j\})$;
2. для всех $t = 2, \ldots, l$ (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 $R(W, S)$ по формуле Ланса–Уильямса;

Формула Ланса–Уильямса

Позволяет обобщить большинство способов определить расстояние между кластерами

$$R(W, S),\ W = U \cup V,\ U, V, S \subset X,$$

зная расстояния

$$R(U, S),\ R(V, S),\ R(U, V).$$

Формула Ланса–Уильямса:

$$R(U \cup V, S) = \alpha_U \cdot R(U, S) + \alpha_V \cdot R(V, S) + \\ + \beta \cdot R(U, V) + \gamma \cdot |R(U, S) - R(V, S)|, \quad \alpha_U, \alpha_V, \beta, \gamma \in \mathbb{R}.$$

Формула Ланса–Уильямса

Позволяет обобщить большинство способов определить расстояние между кластерами

$$R(W, S),\ W = U \cup V,\ U, V, S \subset X,$$

зная расстояния

$$R(U, S),\ R(V, S),\ R(U, V).$$

Формула Ланса–Уильямса:

$$R(U \cup V, S) = \alpha_U \cdot R(U, S) + \alpha_V \cdot R(V, S) + \\ + \beta \cdot R(U, V) + \gamma \cdot |R(U, S) - R(V, S)|, \quad \alpha_U, \alpha_V, \beta, \gamma \in \mathbb{R}.$$

Примеры межкластерных расстояний

Расстояние ближнего соседа :

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

расстояние дальнего соседа :

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

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 :

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

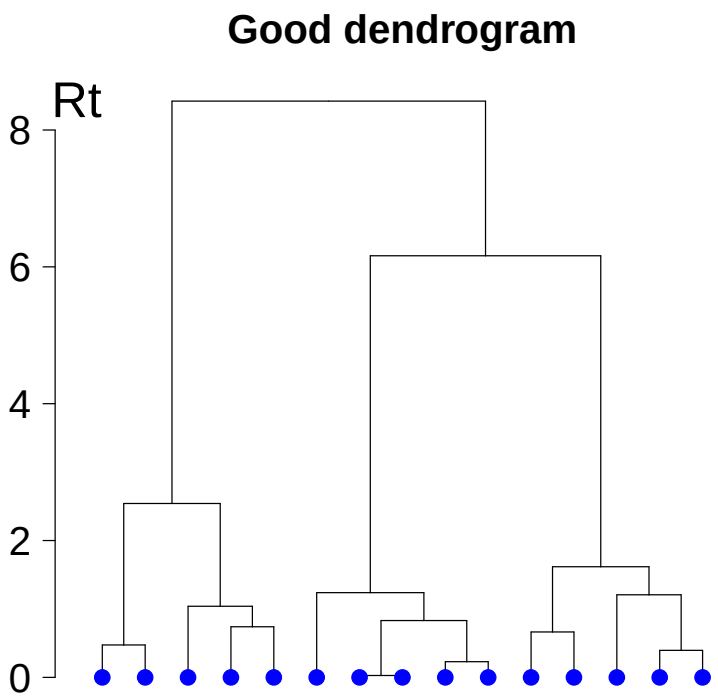
расстояние Уорда :

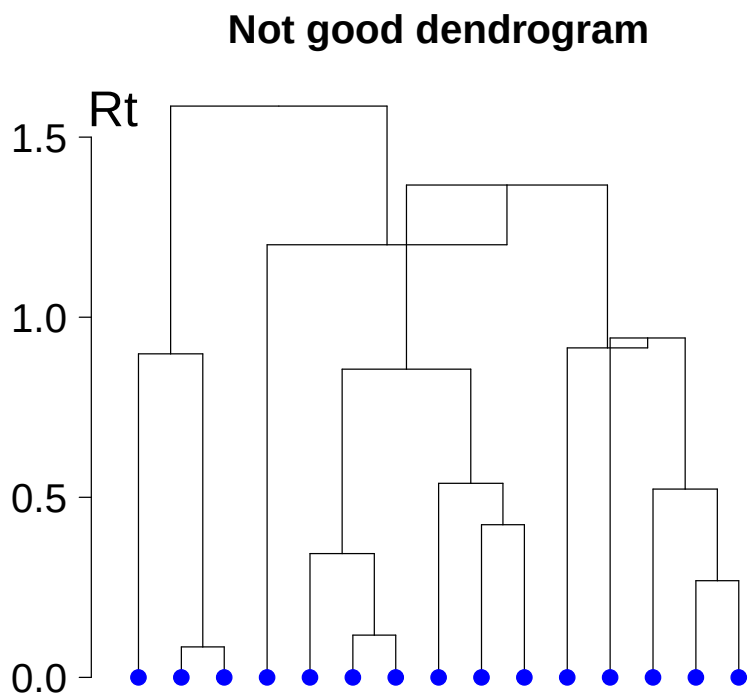
$$R^w(U,V) = \frac{|U||V|}{|U| + |V|} R^c(U,V).$$

Визуализация кластерной структуры

Определение

Дендрограмма – деревоподобный график, отражающий процесс последовательных слияний и структуру кластеров.





Свойство монотонности

Определение

Функция расстояния между кластерами R монотонна, если при каждом слиянии расстояние между объединяемыми кластерами не убывает: $R_2 \leq R_3 \leq \cdots \leq R_l$.

Теорема (Миллиган)

Функция расстояния между кластерами R монотонна, если

$$\alpha_U \geq 0,\ \alpha_V \geq 0,\ \alpha_U + \alpha_V + \beta \geq 1,\ \min(\alpha_U, \alpha_V) + \gamma \geq 0.$$

Если R монотонна, то дендрограмма не имеет самопересечений.

R^c не монотонно; R^n, R^l, R^g, R^w – монотонны.

Свойства сжатия и растяжения

Определение

Функция межкластерного расстояния R сжимающая, если $\forall t\ R_t \leq \rho(\mu_U, \mu_V)$ и растягивающая – если $\forall t\ R_t \geq \rho(\mu_U, \mu_V)$, где μ_U, μ_V – центры соответствующих кластеров.
Иначе R сохраняет метрику пространства.

Свойство растяжения способствует «разрежеванию» верхних уровней дендрограммы, что упрощает выбор числа кластеров.

R^n – сжимающее;
R^l, R^w – растягивающие;
R^g, R^c – сохраняют метрику пространства.

Быстрый (редуктивный) алгоритм Ланса–Уильямса

1. $C_1 = \{\{x_1\}, \ldots, \{x_l\}\}$; $R_1 = 0$;
   $\forall i \neq j$ вычислить $R(\{x_i\}, \{x_j\})$;
2. выбрать начальное значение параметра δ;
   построить $P(\delta) = \{(U, V) | U, V \in C_1, R(U, V) \leq \delta\}$;
3. для всех $t = 2, \ldots, l$
4. если $P(\delta) = \varnothing$, то увеличить δ так, чтобы $P(\delta) \neq \varnothing$;
5. $(U, V) = \underset{(U,V) \in P(\delta)}{\arg\min} R(U, V)$;
   $R_t = R(U, V)$;
6. $W = U \cup V$;
   $C_t = C_{t-1} \cup W \setminus \{U, V\}$; $P(\delta) = P(\delta) \setminus (U, V)$;
7. для всех $S \in C_t \setminus W$
8. вычислить $R(W, S)$ по формуле Ланса–Уильямса;
9. если $R(W, S) \leq \delta$, то $P(\delta) = p(\delta) \cup (W, S)$;

Применимость быстрого алгоритма

Теорема

Если функция расстояния между кластерами R является редуктивной, то быстрый алгоритм приводит к той же кластеризации, что и исходный алгоритм.

Теорема (Диде и Моро, 1984)

Функция расстояния между кластерами R редуктивная, если:

$$\alpha_U \geq 0,\ \alpha_V \geq 0,\ \alpha_U + \alpha_V + \min(\beta, 0) \geq 1,\ \min(\alpha_U, \alpha_V) + \gamma \geq 0.$$

Определение свойства редуктивности опустим за ненадобностью.

R^c – не редуктивная; R^n, R^l, R^g, R^w – редуктивные.

Стратегия выбора параметра δ:

1. Если $|C_t| \le n_1$, то $P(\delta) = \{(U, V) | U, V \in C_t\}$;
2. иначе выбрать n_2 случайных расстояний $R(U, V)$; назначить δ минимальным из них;

- n_1, n_2 влияют только на скорость, но не на результат кластеризации; можно положить $n_1 = n_2 = 20$.

Общие рекомендации по иерархической кластеризации:

- лучше пользоваться R^w – расстоянием Уорда;
- лучше пользоваться быстрым алгоритмом;
- определять число кластеров по максимуму $|R_{t+1} - R_t|$, тогда итоговое число кластеров $= |C_t|$.

Стратегия выбора параметра δ:

1. Если $|C_t| \leq n_1$, то $P(\delta) = \{(U, V) | U, V \in C_t\}$;
2. иначе выбрать n_2 случайных расстояний $R(U, V)$; назначить δ минимальным из них;

- n_1, n_2 влияют только на скорость, но не на результат кластеризации; можно положить $n_1 = n_2 = 20$.

Общие рекомендации по иерархической кластеризации:

- лучше пользоваться R^w – расстоянием Уорда;
- лучше пользоваться быстрым алгоритмом;
- определять число кластеров по максимуму $|R_{t+1} - R_t|$, тогда итоговое число кластеров $= |C_t|$.

Самоорганизующаяся карта Кохонена

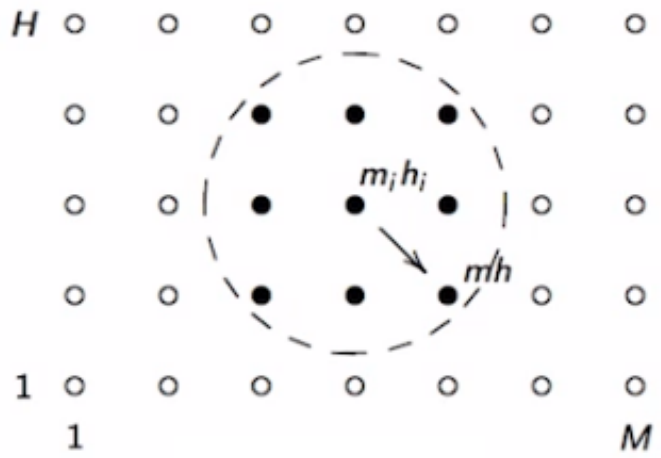
X, $\rho : X \times X$ – метрика пространства объектов;

$Y = \{1, \ldots, M\} \times \{1, \ldots, H\}$ – сетка кластеров,
$r : Y \times Y$ – метрика пространства кластеров;

Каждому узлу (m, h) приписан нейрон Кохонена $w_{mh} \in X$;

Заданы неотрицательные невозрастающие функции
$K(r(\cdot, \cdot), t)$ (расстояния), $\eta(t)$ (скорости обучения),
$\epsilon(t)$ (окрестности), где t – номер итерации;

$\upsilon_{\epsilon(t)}(m_i, h_i)$ – $\epsilon(t)$-окрестность (m_i, h_i) в метрике r:



Самоорганизующаяся карта Кохонена

1. задать начальные w_{mh}, $m = \overline{1:M}$, $h = \overline{1:H}$;
2. повторять
3. выбрать случайным образом x_i из X^l;
4. вычислить координаты ближайшего кластера:

$$(m_i, h_i) = \arg\min_{(m,h)\in Y} \rho(x_i, w_{mh});$$

5. для всех $(m, h) \in \upsilon_\epsilon(m_i, h_i)$
6. сделать шаг стохастического градиентного спуска:

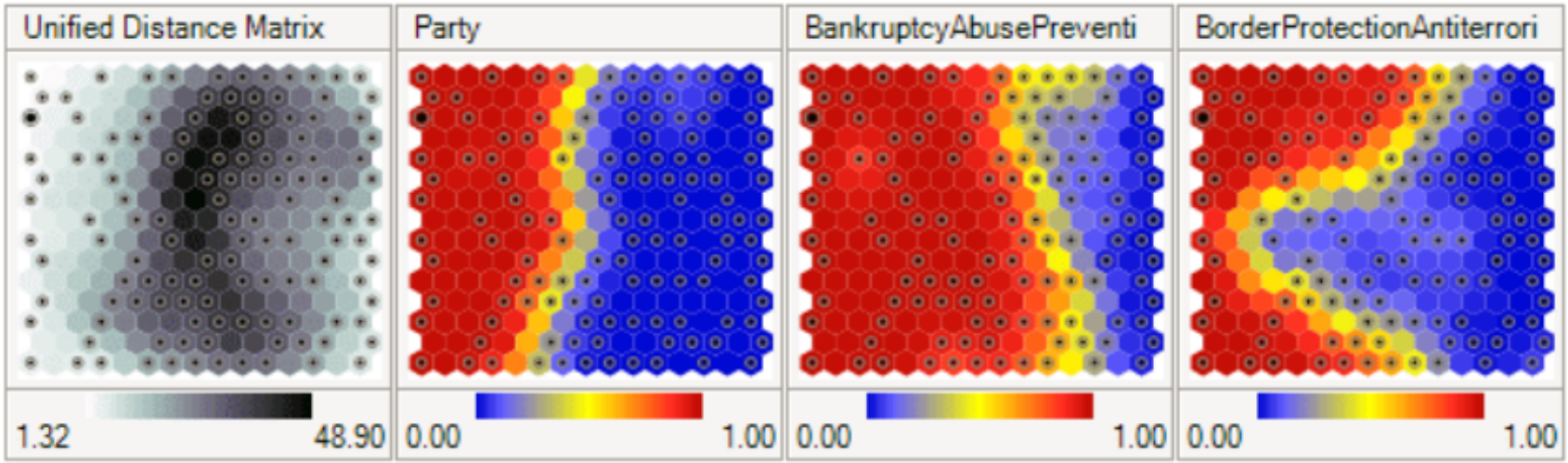
$$w_{mh} = w_{mh} + \eta(t)(x_i - w_{mh})K\left(r((m_i, h_i), (m, h))), t\right);$$

7. пока кластеризация не стабилизируется;

Интерпретация карт Кохонена

Два типа графиков – цветных карт $M \times H$:

- Цвет узла (m, h) – локальная плотность в точке (m, h) – среднее расстояние до k ближайших точек выборки;
- По одной карте на каждый признак:
  цвет узла (m, h) – значение j-й компоненты вектора w_{mh}.



Рекомендации по выбору параметров

Инициализация w_{mh}:

- случайными значениями;
- случайно выбранными x_i;
- Линейная инициализация: наложить сетку Y на плоскость первых двух главных компонент, тогда w_{mh} присвоить соответствующие вектора и исходного пространства X.

Задание функций:

- $\epsilon(t) = \sigma(t)$, где $\sigma(t)$ – Гауссова функция.
- $K(r(\cdot,\cdot),t) = e^{-\frac{r(\cdot,\cdot)}{2\sigma(t)}}$ или $\begin{cases} const, & r(\cdot,\cdot) < \sigma(t); \\ 0, & \text{иначе;} \end{cases}$
- $\eta(t) = At + B$ или $\frac{A}{t+B}$, A, $B = const$;

Рекомендации по выбору параметров

Инициализация w_{mh}:

- случайными значениями;
- случайно выбранными x_i;
- Линейная инициализация: наложить сетку Y на плоскость первых двух главных компонент, тогда w_{mh} присвоить соответствующие вектора и исходного пространства X.

Задание функций:

- $\epsilon(t) = \sigma(t)$, где $\sigma(t)$ – Гауссова функция.
- $K(r(\cdot,\cdot),t) = e^{-\frac{r(\cdot,\cdot)}{2\sigma(t)}}$ или $\begin{cases} const, & r(\cdot,\cdot) < \sigma(t); \\ 0, & \text{иначе}; \end{cases}$
- $\eta(t) = At + B$ или $\frac{A}{t+B}$, A, $B = const$;

Достоинства и недостатки карт Кохонена

Достоиства:

- Возможность визуального анализа многомерных данных.

Недостатки:

- **Субъективность.** Карта зависит не только от кластерной структуры данных, но и...
  - от свойств функций K, η, ϵ;
  - от начальных значений w_{mh};
  - от случайного выбора x_i в ходе итераций.
- **Искажения.** Близкие объекты исходного пространства могут переходить в далекие точки на карте, и наоборот.

Рекомедуется только для разведочного анализа данных.