Обучение без учителя. Кластеризация. Тематическое моделирование

Елена Гоголева, Дейвид Капаца, Анастасия Мандрикова

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Семинар по статистическому и машинному обучению

Ноябрь 2021

План доклада

1. Обучение без учителя
   - Типы задач и методы их решения
2. Кластеризация
   - Постановка задачи
   - Некорректность задачи
   - Вероятностный подход
     - EM-алгоритм для задачи разделения смеси распределений
     - k-means и его связь с EM-алгоритмом
   - Иерархическая кластеризация
     - Формула Ланса-Уильямса
     - Визуализация кластерной структуры
   - Функционалы качества кластеризации
   - Другие подходы и методы (FOREL, DBSCAN, карты Кохонена)

План доклада (продолжение)

3. Тематические модели
   - Введение
   - Вероятностная модель коллекции документов
     - Постановка задачи
     - Гипотезы
     - Предварительная обработка документов
   - PLSA
     - Стохастическое матричное разложение
     - Принцип максимума правдоподобия
     - EM-алгоритм
     - Начальное приближение
   - Критерии качества модели

Типы задач и методы их решения

В случае обучения с учителем известны как независимые переменные $X_1, ..., X_p$, так и зависимая переменная Y. В случае обучения без учителя известны лишь $X_1, ..., X_p$.

- Задачи сокращения размерности (PCA)
- Задачи визуализации данных (иерархическая кластеризация)
- Задачи кластеризации (k-means)

Постановка задачи

Дано:
X — пространство объектов;
$X^n = \{X_i\}_{i=1}^n \subset X$ — обучающая выборка;
$X_i \in \mathrm{R}^\mathrm{p}$ — объекты определяемые вектором признаков;
$\rho : X \times X \to [0, \infty)$.
Найти:
$a : X \to C$, где C — множество непересекающихся кластеров, таких что в каждый кластер попадают близкие относительно выбранной метрики ρ индивиды.

Этапы кластеризации

Общая схема кластеризации состоит из:

- выбор метрики
- разделение на кластеры
- оценка качества кластеризации
- выделение признаков которые значимы для кластеризации
- интерпретация результатов

Форма кластеров

DBSCAN

k-means

Некорректность задачи

- Задача кластеризации не формализована
- Не всегда известно число кластеров
- Результат зависит от выбранной метрики ρ
- Разнообразие критериев качества
- Выбор метода кластеризации

Вероятностный подход

Пусть, модель данных состоит из смеси k распределений:

$$p(x) = \sum_{i=1}^{k} \omega_i p_i(x)$$

Оценить по наблюдаемой выборке из $p(x)$:

- $\omega_1 \dots \omega_k$ — априорные вероятности появления обьектов из соответствующих кластеров;
- $p_1(x) \dots p_k(x)$ — плотности распределения признаков внутри кластеров.

EM - алгоритм

Предположим принадлежность $p_1(x) \dots p_k(x)$ одному семейству распределений:

$$p_i(x) = \varphi(\theta_i; x).$$

Согласно методу максимального правдоподобия:

$$\omega, \theta = \underset{\omega,\theta}{\text{argmax}} \sum_{i=1}^{n} \ln \sum_{j=1}^{k} \omega_j \varphi(\theta_j; x_i).$$

Скрытые переменные $h_{ij} = P(\theta_j|x_i)$ — это вероятность того, что индивид x_i принадлежит j смеси. Найдем по формуле Байеса:

$$h_{ij} = \frac{\omega_j \varphi(\theta_j; x_i)}{\sum\limits_{s=1}^{k} \omega_s \varphi(\theta_s; x_i)}.$$

Для любого индивида $\sum\limits_{j=1}^{k} h_{ij} = 1$.

EM - алгоритм

E - шаг:

Подставляем текущую оценку ω, θ и расчитываем скрытые переменные h_{ij}.

M - шаг:

Решение методом Лагранжа для максимизации $(\sum_{j=1}^{k} \omega_j = 1)$ даёт оценку для параметров:

$$\omega_j = \frac{1}{n} \sum_{i=1}^{n} h_{ij},$$

$$\theta_j = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} h_{ij} \ln \varphi(x_i, \theta).$$

В качестве меры близости выбрано евклидово расстояние:

$$d(x_i, x_{i'}) = \sum_{j=1}^{k} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2.$$

Минимизируем меру близости между индивидами внутри одного кластера:

$$\min_{C_1,\ldots,C_k} \left\{ \sum_{l=1}^{k} \frac{1}{|C_l|} \sum_{i,i' \in C_l} \|x_i - x_{i'}\|^2 \right\}.$$

1. Выбираем $\mu_1, \ldots, \mu_k$ – центры кластеров случайным образом.
2. Определяем принадлежность индивидов кластерам.

$$C(i) = \underset{0 \leq j \leq k}{\text{argmin}} \|x_i - \mu_j\|^2.$$

3. Для каждого кластера C_j пересчитываем центры μ_j как выборочное среднее индивидов, которые были отнесены к этому кластеру.
4. Повторяем шаги 2 и 3 пока принадлежность кластерам не перестанет изменяться.

k-means и его связь с EM - алгоритмом

Алгоритм k-средних является частным случаем для гауссовой смеси распределения с диагональными матрицами, у которых одинаковые значения на диагоналях.
В таком случае:
1) На E-шаге мы не считаем вероятности принадлежности кластерам, а приписываем каждый объект одному кластеру (вероятность принадлежности будет равна 0 или 1);
2) Форма кластеров не настраивается: они все являются сферическими.

Алгоритм агломеративной иерархической кластеризации

1. Одноэлементные кластеры:
   $C_1 = \{\{x_1\}, \ldots, \{x_n\}\};\ R_1 = 0$
   $\forall\, i \neq j$ вычислить $R(\{x_i\}, \{x_j\})$
2. для всех $t = 2, \ldots, n$ (t — номер итерации)
3. найти в C_{t-1} два ближайших кластера:
   $(U, V) = \arg\min\limits_{U \neq V} R(U, V),\ R_t = R(U, V)$
4. слить их в один кластер:
   $W = U \cup V;\ C_t = C_{t-1} \cup W \backslash \{U, V\}$
5. для всех $S \in C_t \backslash W$
6. вычислить R(W, S)

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

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

$$R(W, S) = \alpha_U U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) R(V, S)|,$$

Например,

- Расстояние Уорда:

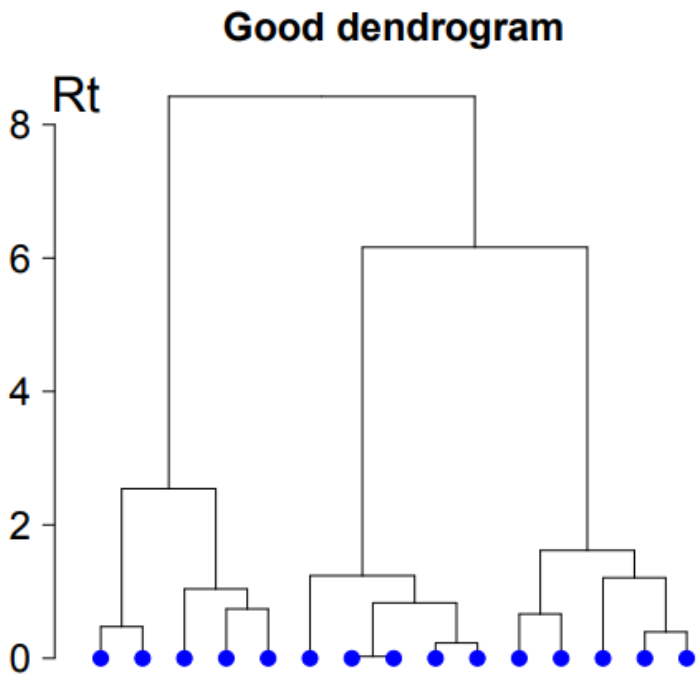
$$R(W, S) = \frac{|S||W|}{|S| + |W|} \rho^2 \left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right);$$

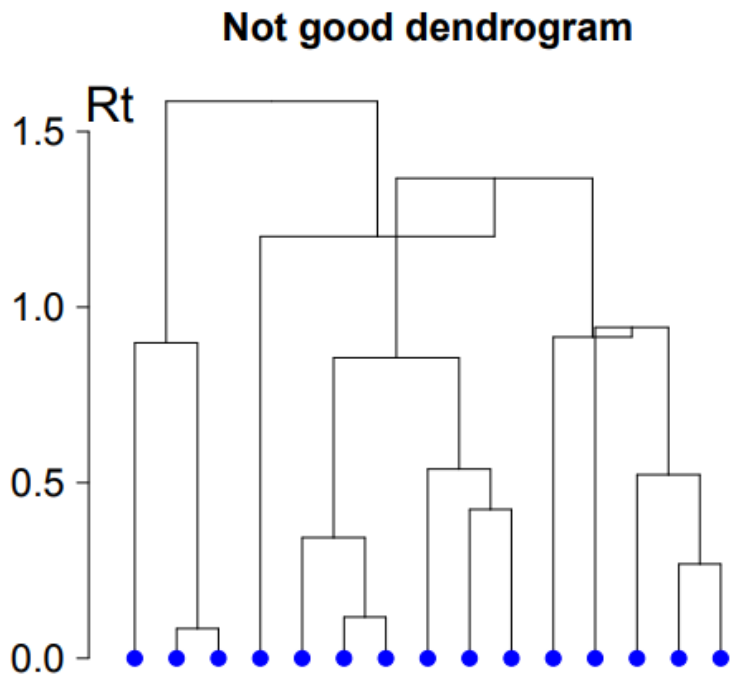
$$\alpha_U = \frac{|S| + |U|}{|S| + |W|},\ \alpha_V = \frac{|S| + |V|}{|S| + |W|},\ \beta = -\frac{|S|}{|S| + |W|},\ \gamma = 0.$$

Визуализация кластерной структуры

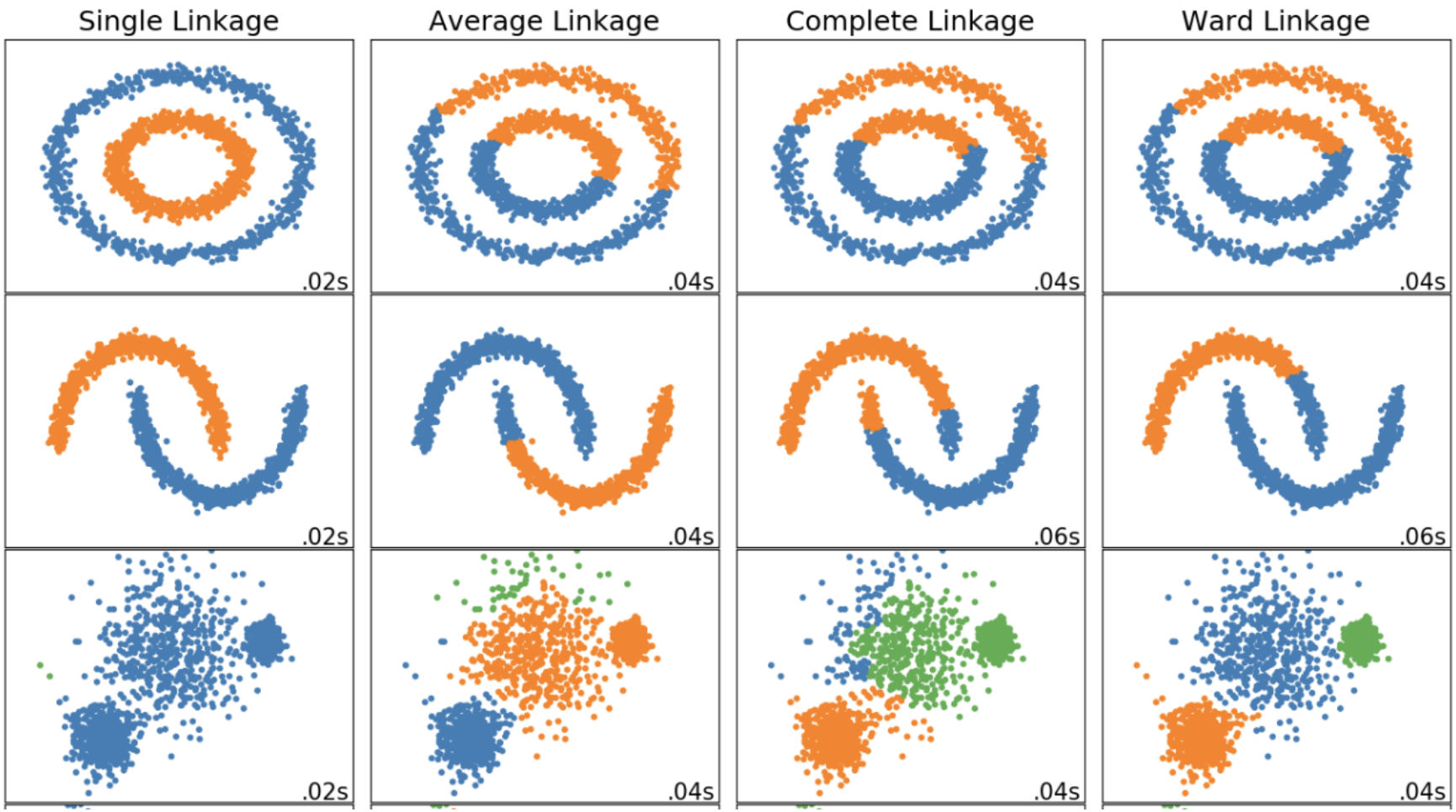
Определение

Дендрограмма — древовидный график расстояний, при которых произошло слияние кластеров на каждом шаге





Пример


Single Linkage
Average Linkage
Complete Linkage
Ward Linkage
.02s
.04s
.04s
.04s
.02s
.04s
.06s
.06s
.02s
.04s
.04s
.04s

Другие алгоритмы кластеризации (FOREL)

FOREL (ФОРмальный ЭЛемент) — алгоритм, основанный на идее объединения в один кластер объектов в областях их наибольшего сгущения.

1. Пусть $U = X_m$
2. **Пока** есть некластеризованные точки, т.е. $U \neq \varnothing$;
3. взять случайную точку $x_0 \in U$;
4. **Повторять**
5. образовать кластер с центром в x_0 и радиусом R: $K_0 = \{x_i \in U | \, \rho(x_i, x_0) \leq R\}$;
6. переместить центр x_0 в центр масс кластера: $x_0 = \frac{1}{|K_0|} \sum\limits_{x_i \in K_0} x_i$;
7. **Пока** состав кластера K_0 не стабилизируется;
8. Пометить объекты внутри сферы как кластеризованные, $U = U \backslash K_0$.

Преимущества и недостатки алгоритма FOREL

Преимущества

- Получаем двухуровневую систему кластеров;
- Кластеры могут быть произвольной формы (при добавлении модификации к построению сферы);
- Варьируя R можно управлять детальностью кластеризации.

Недостатки

- Алгоритм очень чувствителен к R и к начальному выбору точки x_0

DBSCAN

(Density-based spatial clustering of applications with noise)

Объект $x \in U$, его ϵ-окрестность $U_\epsilon(x) = \{u \in U : \rho(x, u) \leq \epsilon\}$
Каждый объект может быть одного из трёх типов:

- корневой: имеет плотную окрестность $|U_\epsilon(x)| > m$
- граничный: не корневой, но находится в окрестности корневого
- выброс: не корневой и не граничный.

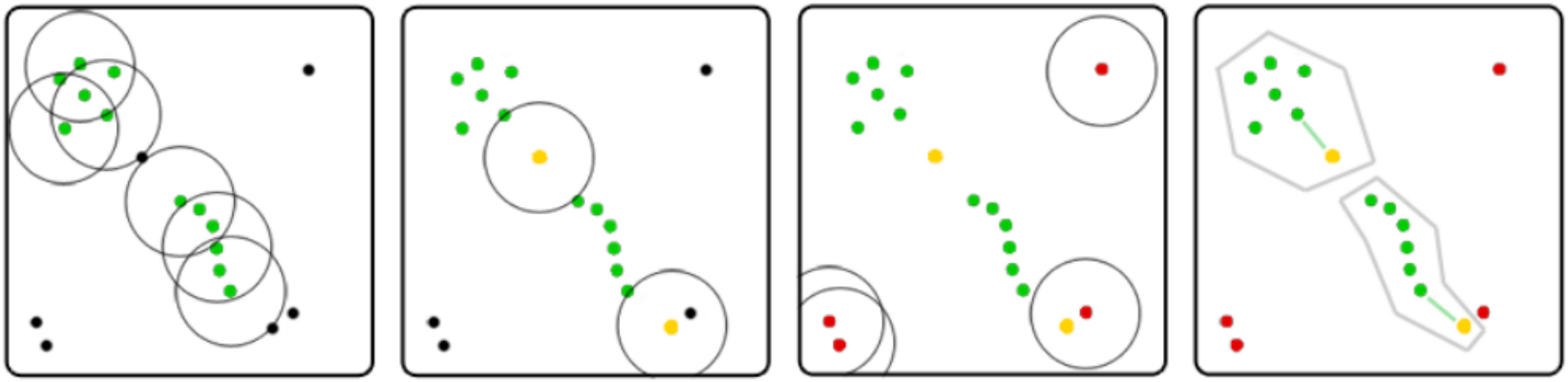


Рис.: Иллюстрация к алгоритму DBSCAN. На рисунке зелёным отмечены корневые объекты, жёлтым — граничные и красным — шумовые.

Алгоритм DBSCAN

Вход: Выборка $X^n = \{x_1, x_n\}$, параметры ϵ и m;
Выход: разбиение выборки на кластеры и шумовые выбросы;

1. $U = X_m,\ a = 0$;
2. **Пока** есть некластеризованные точки, т.е. $U \neq \varnothing$;
3. взять случайную точку $x \in U$;
4. **если** $|U_\epsilon(x)| < m$, **то**
5. пометить x как возможно шумовой;
6. **иначе**
7. создать новый кластер: $K = U_\epsilon(x),\ a = a + 1$;
8. **для всех** $x^{'} \in K$
9. **если** $|U_\epsilon(x^{'})| > m$ **то** $K = K \cup U_\epsilon(x^{'})$;
10. **иначе** пометить $x^{'}$ как граничный элемент K;
11. $a_i = a$ для всех $x^{'} \in K$;
12. $U = U \backslash K$.

Преимущества и недостатки алгоритма DBSCAN

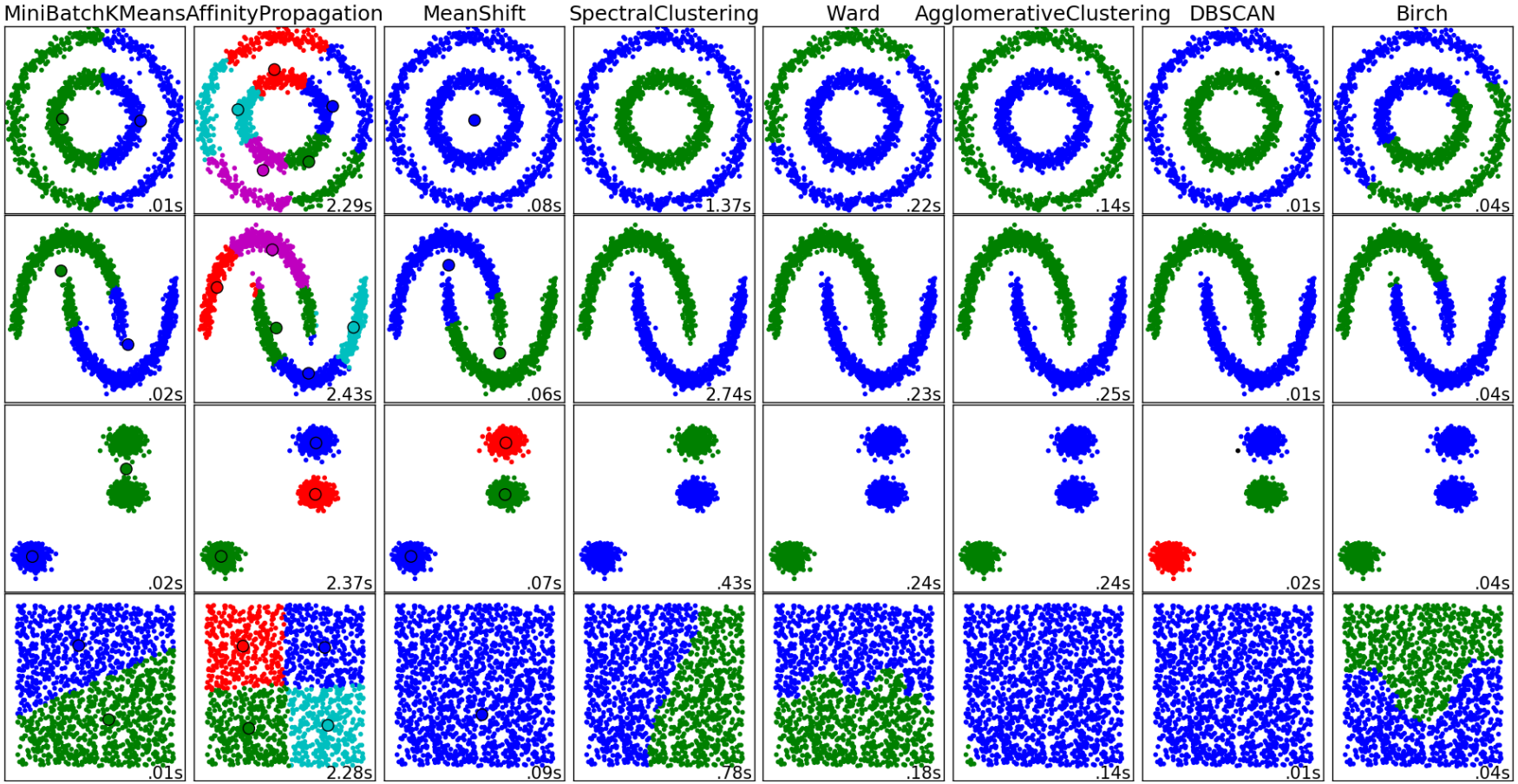
Преимущества

- Быстрая кластеризация больших данных (от $O(m \ln m)$ до $O(m^2)$ в зависимости от реализации);
- Кластеры произвольной формы;
- Явная разметка шумовых объектов;

Недостатки

- Алгоритм может неадекватно обрабатывать сильные вариации плотности данных внутри кластера, проёмы и шумовые мосты между кластерами.

Пример


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
2.29s
.08s
1.37s
.22s
.14s
.01s
.04s
.02s
2.43s
.06s
2.74s
.23s
.25s
.01s
.04s
.02s
2.37s
.07s
.43s
.24s
.24s
.02s
.04s
.01s
2.28s
.09s
.78s
.18s
.14s
.01s
.04s

Самоорганизующаяся карта Кохонена

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

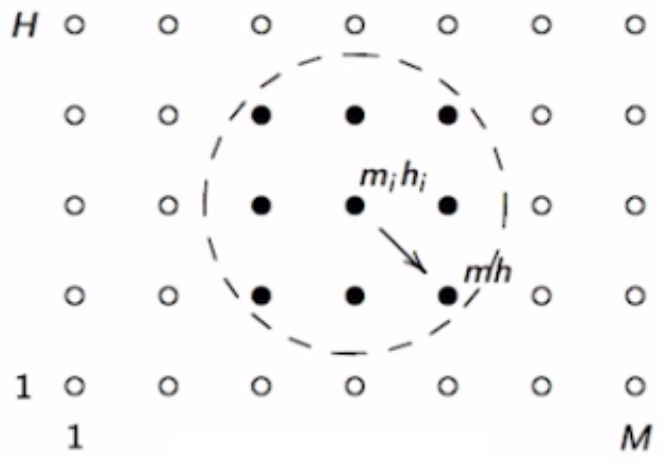
$Y = \{1, \ldots, M\} \times \{1, \ldots, H\}$ — сетка кластеров,
$r : Y \times Y$ — метрика пространства кластеров;

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

Заданы неотрицательные невозрастающие функции:

- $K(r(\cdot, \cdot), t)$ — расстояние,
- $\eta(t)$ — скорость обучения,
- $\epsilon(t)$ — окрестность, 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^n;
4. вычислить координаты ближайшего кластера:

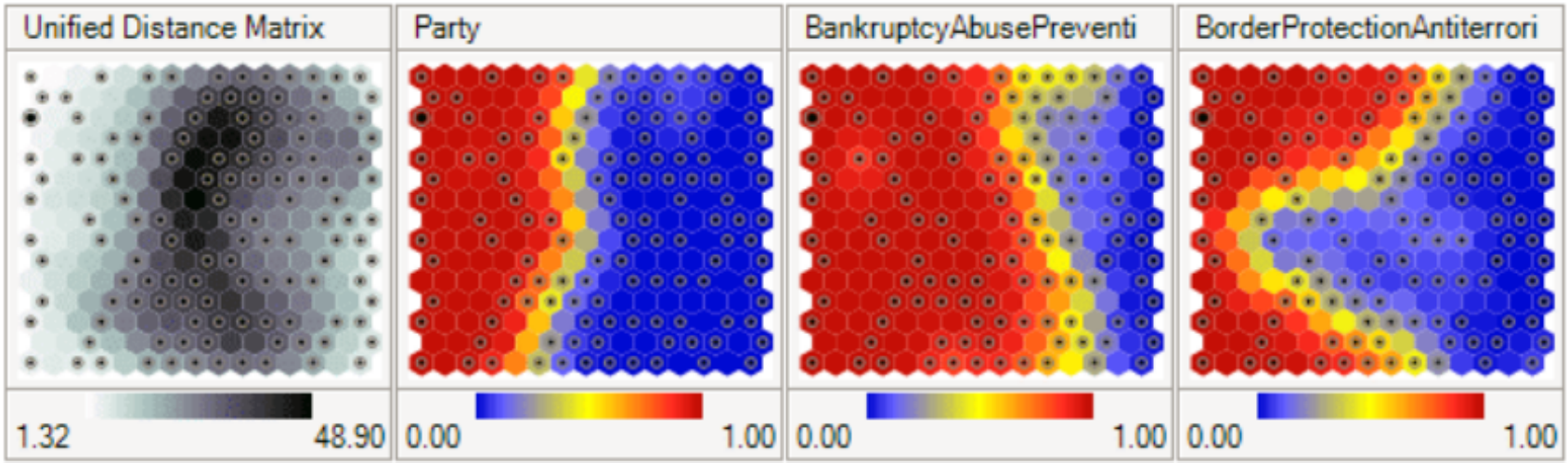
$$(m_i, h_i) = \arg\min_{(m,h)\in Y} \rho(x_i, w_{mh});$$

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}.



Преимущества и недостатки карт Кохонена

Преимущества:

- Возможность визуального анализа многомерных данных.

Недостатки:

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

Рекомендуется только для разведочного анализа данных.

Функционалы качества кластеризации

Возможна постановка задачи кластеризации: приписать номера кластеров объектам так, чтобы значение выбранного функционала качества приняло наилучшее значение.

Выделяют две группы функционалов качества внутренние и внешние:

- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i<j} \mathbf{I}_{\{y_i=y_j\}} \rho(x_i, x_j)}{\sum_{i<j} \mathbf{I}_{\{y_i=y_j\}}},$$

- Среднее межкластерное расстояние:

$$F_1 = \frac{\sum_{i<j} \mathbf{I}_{\{y_i\neq y_j\}} \rho(x_i, x_j)}{\sum_{i<j} \mathbf{I}_{\{y_i\neq y_j\}}}.$$

На практике обычно вычисляют F_0/F_1.

Функционалы качества кластеризации (Силуэт)

- Принадлежность объекта своему кластеру:

$$c(x_i) = \frac{1}{|K_i| - 1} \sum_{x_j \in K_i,\, i \neq j} \rho(x_i, x_j)$$

- Принадлежность объекта другому кластеру:

$$b(x_i) = \min_{i \neq j} \frac{1}{|K_j|} \sum_{x_z \in K_j} \rho(x_i, x_z)$$

- Силуэт объекта:

$$s(x_i) = \begin{cases} \frac{b(x_i) - c(x_i)}{\max\{c(x_i), b(x_i)\}}, & |K_i| > 1 \\ 0, & |K_i| = 1 \end{cases}$$

- Силуэт кластеризации: $S = \frac{1}{n} \sum_i s(x_i)$.

Данный функционал качества пригоден только для кластеров, которые представляют собой далеко отстоящие компактные скопления объектов.

Тематические модели

(часть 2)

Введение

Тематическое моделирование применяется к анализу текстов.

- *Тематическое моделирование* — способ построения модели коллекции текстовых документов.
- Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.
- *Вероятностная тематическая модель* описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем.

Введение

Это про выявление тематики в текстовой коллекции...

- Тема — условное распределение на множестве терминов $p(w|t)$ — вероятность термина w в теме t;
- Тематический профиль документа — условное распределение $p(t|d)$ — вероятность темы t в документе d.

Задача определения тематики коллекции документов

Дано:

- W — словарь, множество слов (терминов)
- D — множество (коллекция) текстовых документов
- n_{dw} — частота термина $w \in W$ в документе $d \in D$

Хотим найти:

- число различимых тем
- какими терминами определяется каждая тема
- к каким темам относится каждый документ

или

Найти: вероятностную тематическую модель

Гипотезы и допущения

- Гипотеза назависимости: Порядок слов в документе и порядок документов в коллекции не важны
- Гипотеза условной независимости: $p(w|d, t) = p(w|t)$
- Гипотеза разреженности: Каждый документ d и каждый термин w связан с небольшим количеством тем t.

Если не выполняется гипотеза разреженности?

- Документ относится к большому количеству тем $\rightarrow$ разобьем его на части, более однородные по тематике
- Термин относится к большому числу тем $\rightarrow$ положим, что термин является общеупотребительным и неважен для определения тематики

Вероятностная формализация постановки задачи

Базовые предположения:

каждое слово в документе связано с некоторой темой $t \in T$

$D \times W \times T$ — дискретное вероятностное пространство

D — выборка троек $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$

$d_i,\ w_i$ — наблюдаемые, t_i — скрытые (латентные)

Вероятностная модель порождения документа d:

$$p(w,d) = \sum_{t \in T} p(w,d,t) = \sum_{t \in T} p(w|t,d) \cdot p(t,d) =$$

$$= \sum_{t \in T} p(w|t,d) \cdot p(t|d) \cdot p(d) = p(d) \cdot \sum_{t \in T} p(w|t) \cdot p(t|d)$$

$$\boxed{\mathrm{p(w|d)} = \sum_{t \in T} p(w|t) \cdot p(t|d)}$$

Задача: найти $T,\ p(w|t),\ p(t|d)$.

Предварительная обработка документа

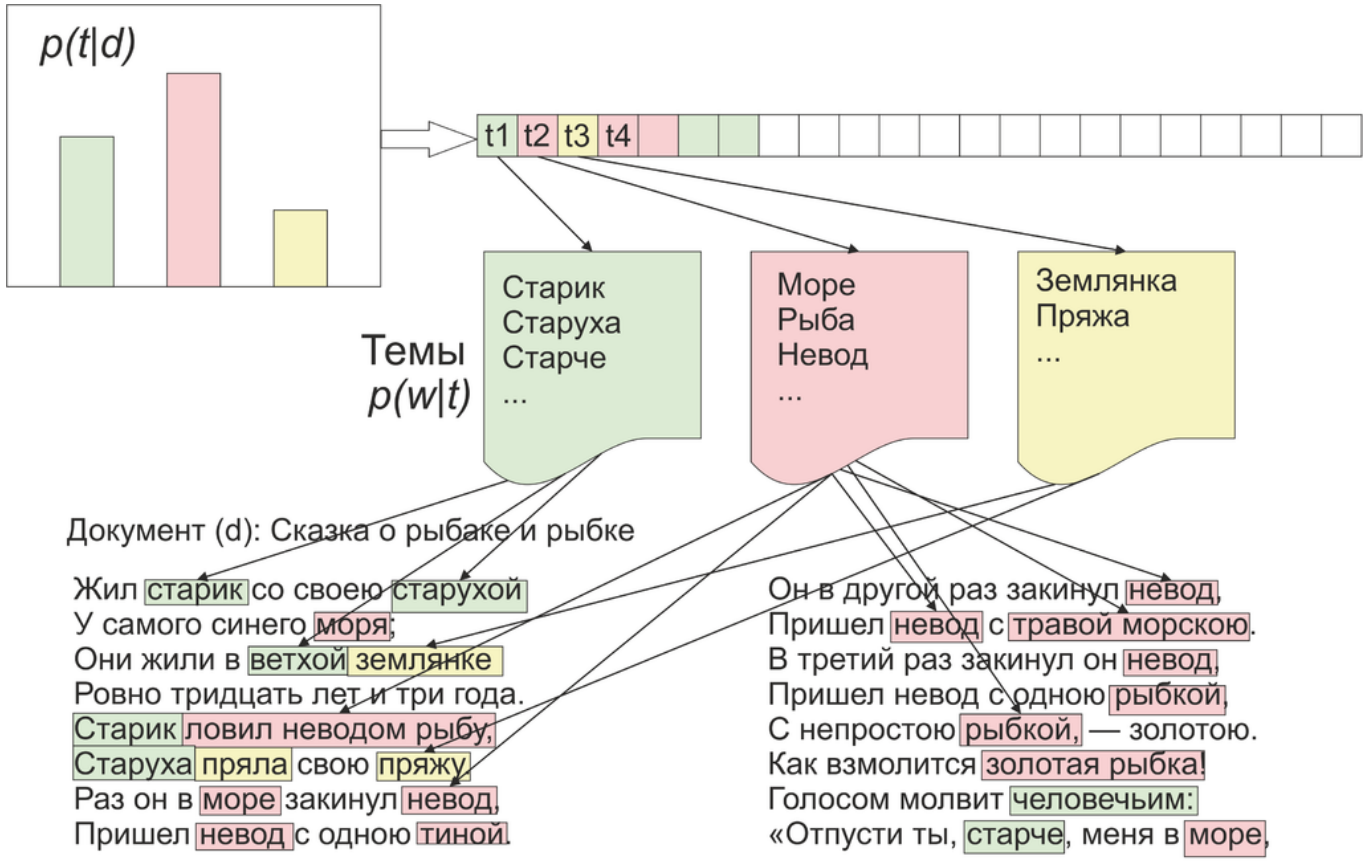
Необходимость конкретного метода обработки зависит от типа задачи, но вообще это хорошая практика.

- Лемматизация — приведение каждого слова в документе к его начальной форме.
  Трудоемкий процесс
- Стемминг — отбрасывание изменяемых частей слова.
  Большое число ошибок
- Отбрасывание стоп-слов — удаление слов, которые никак не характеризуют тему.
  Почти не влияет на длину словаря
- Отбрасывание редких слов.
  Для коллекций коротких новостных сообщений лучше не использовать
- Выделение ключевых фраз.
  Приходится привлекать экспертов

Вероятностная модель порождения документа d

Вероятностная тематическая модель:

$$p(w|d) = \sum_{t \in T} p(w|t) \cdot p(t|d)$$



Выборочный язык

Дано:

- коллекция текстовых документов D,
- n_{dw} — число вхождений термина w в документ d,
- $n_d = \sum\limits_{w \in W} n_{dw}$ — длина документа d в терминах.

Можем оценить: $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$.

Найти параметры тематической модели:

$$F = (\hat{p}(w|d))_{W \times D} \approx \sum_{t \in T} (\varphi_{wt})_{W \times T} (\theta_{td})_{T \times D}$$

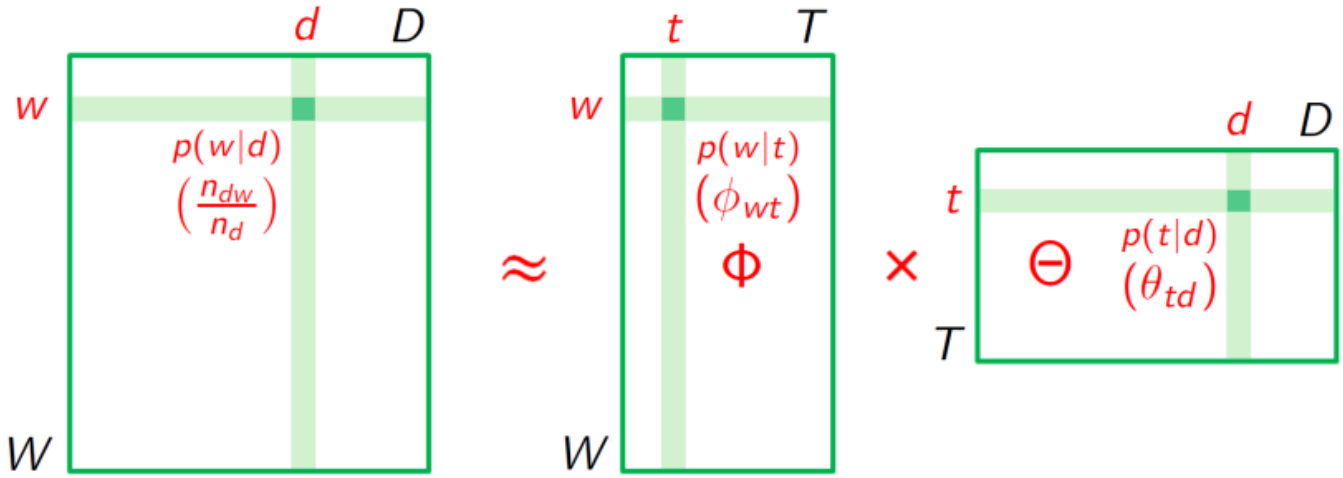
Искомые:

- $\varphi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t,
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Стохастическое матричное разложение

Задача стохастического матричного разложения:

$$F \approx \Phi\Theta \quad \Leftrightarrow \quad ||F - \Phi\Theta||^2 \to \min_{\Phi,\,\Theta}$$



Если Φ и Θ — решение, то существует матрица S ранга $|T|$:

$$F = \Phi S S^{-1}\Theta,$$

где S — матрица перестановки, а Φ, Θ тоже стохастические.

Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in W} p(d, w)^{n_{dw}}$$

где n_{dw} — число вхождений термина w в документ d,
$n = \sum_{d \in D} \sum_{w \in W} n_{dw}$ — длина коллекции в терминах.

Максимизация логарифма правдоподобия:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(w|d)p(d) \to \max_{\Phi, \Theta},$$

здесь $p(d) = n_d/n = \text{const}$, n_d — длина документа.

Приходим к задаче:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} \to \max_{\Phi, \Theta},$$

При ограничениях неотрицательности и нормировки

$$\varphi_{wt} \geq 0; \quad \sum_{w \in W} \varphi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

EM-алгоритм

Для решения задачи матричного разложения применяется **EM-алгоритм**. Он заключается в выполнении двух шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t, d, w вычисляются через φ_{wt}, θ_{td} по формуле Байеса:

$$H_{dwt} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s\in T}\varphi_{ws}\theta_{sd}}.$$

M-шаг: частотные оценки условных вероятностей вычисляются путем накопления счетчика $n_{dwt} = n_{dw}p(t|d,w)$:

$$\varphi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad \hat{n}_t = \sum_{w\in W}\hat{n}_{wt}, \quad \hat{n}_{wt} = \sum_{d\in D} n_{dw}H_{dwt};$$

$$\theta_{dt} = \frac{\hat{n}_{dt}}{\hat{n}_t}, \quad \hat{n}_d = \sum_{t\in T}\hat{n}_{dt}, \quad \hat{n}_{dt} = \sum_{w\in W} n_{dw}H_{dwt}.$$

Модификация EM-алгоритма

Можно не хранить матрицу H_{dwt} и уменьшить вычислительные затраты, получим алгоритм:

Алгоритм 2.1. PLSA-EM: рациональный EM-алгоритм для модели PLSA.

Вход: коллекция документов D, число тем $|T|$, начальные приближения Θ, Φ;
Выход: распределения Θ и Φ;

повторять
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\varphi_{wt}\theta_{td}/Z$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;
пока Θ и Φ не сойдутся;

Начальное приближение φ_t, θ_d

1. Начальное приближение можно задать нормированными случайными векторами из равномерного распределения.
2. Пройти по всей коллекции, выбрать для каждой пары (d, w) случайную тему t, вычислить частотные оценки вероятностей φ_{wt} и θ_{td} для всех d, w, t.
3. **Частичное обучение** (некоторые t известны заранее и имеются дополнительные данные о привязке некоторых d или w к t):

   Известно, что документ d относится к подмножеству $T_d \subset T$:

$$\theta_{td}^0 = \frac{1}{|T_d|}\mathbf{I}_{t \in T_d}.$$

   Известно, что подмножество терминов $W_t \subset W$ относится к теме t:

$$\varphi_{wt}^0 = \frac{1}{|W_t|}\mathbf{I}_{w \in W_t}.$$

   Известно, что некоторое множество документов $D_t \subset D$ относится к теме t:

$$\varphi_{wt}^0 = \frac{\sum_{d \in D_t} n_{dw}}{\sum_{d \in D_t} n_d}.$$

Недостатки PLSA

- Задача стохастического матричного разложения некорректно поставлена (может быть бесконечно много решений), это приводит к неустойчивости матриц Φ и Θ;
- С появлением нового d не можем вычислить p(t|d), не перестраивая модель;
- Число параметров растёт линейно по числу документов в коллекции, что может приводить к переобучению модели.

Внутренние (intrinsic) критерии качества тематической модели

Перплексия (perplexity) языковой модели $p(w|d)$:

$$\mathcal{P}(D) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right),\ n = \sum_{d\in D}\sum_{w\in d} n_{dw}.$$

Интерпретация:

- если $p(w|d) = \frac{1}{|W|}$, то $\mathcal{P} = |W|$;
- мера неопределенности в тексте.

Перплексия тестовой коллекции

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{1}{n''}\sum_{d\in D'}\sum_{w\in d''} n_{dw}\ln p(w|d)\right),\ n'' = \sum_{d\in D'}\sum_{w\in d''} n_{dw}.$$

$d = d' \cup d''$ — случайное разбиение тестового документа на две половины равной длины;

- параметры φ_{wt} оцениваются по обучающей коллекции D
- параметры θ_{td} оцениваются по первой половине d'
- перплексия вычисляется по d''

Когерентность как мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

$$\text{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{PMI}(w_i, w_j),$$

где w_i — i-й термин в порядке убывания φ_{wt}.

$\text{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information),

N_{uv} — число документов, в которых термины u, v встречаются рядом,

N_u — число документов, в которых u встретился хотя бы раз.