

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

❶ Обучение без учителя

- Типы задач и методы их решения

❷ 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)

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

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

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

Дано:

X — пространство объектов;

$X^n = \{X_i\}_{i=1}^n \subset X$ — обучающая выборка;

$X_i \in \mathbb{R}^p$ — объекты определяемые вектором признаков;

$\rho : X \times X \rightarrow [0, \infty)$.

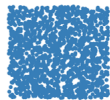
Найти:

$a : X \rightarrow C$, где C — множество непересекающихся кластеров, таких что в каждый кластер попадают близкие относительно выбранной метрики ρ индивиды.

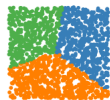
Общая схема кластеризации состоит из:

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

DBSCAN



k-means



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

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 объектов из соответствующих кластеров;
- $p_1(x) \dots p_k(x)$ — плотности распределения признаков внутри кластеров.

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

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

Согласно методу максимального правдоподобия:

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

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

$$h_{ij} = \frac{\omega_j p_j(\theta_j; x_i)}{\sum_{s=1}^k \omega_s p_s(\theta_s; x_i)}.$$

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

Е - шаг:

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

М - 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:

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

1. Выбираем μ_1, \dots, μ_k – центры кластеров случайным образом.
2. Определяем принадлежность индивидов кластерам.

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

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.

Алгоритм k-средних является частным случаем для гауссовой смеси распределения с одинаковыми диагональными матрицами и дисперсиями близкими к нулю.

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.

- 1 Одноэлементные кластеры:
 $C_1 = \{\{x_1\}, \dots, \{x_n\}\}; R_1 = 0$
 $\forall i \neq j$ вычислить $R(\{x_i\}, \{x_j\})$
- 2 для всех $t = 2, \dots, n$ (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 $R(W, S)$

Формула Ланса–Уильямса

Позволяет обобщить большинство способов определить расстояние между кластерами

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

зная расстояния

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

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

Например,

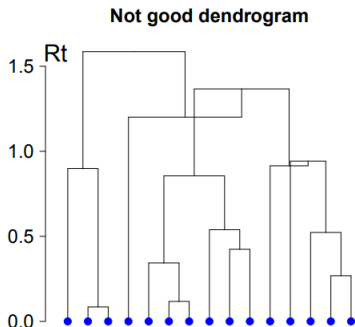
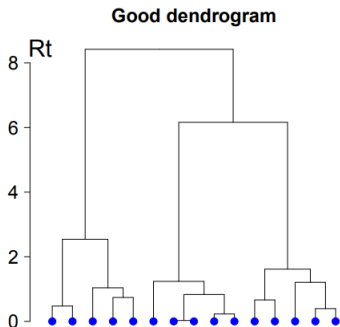
- Расстояние Уорда:

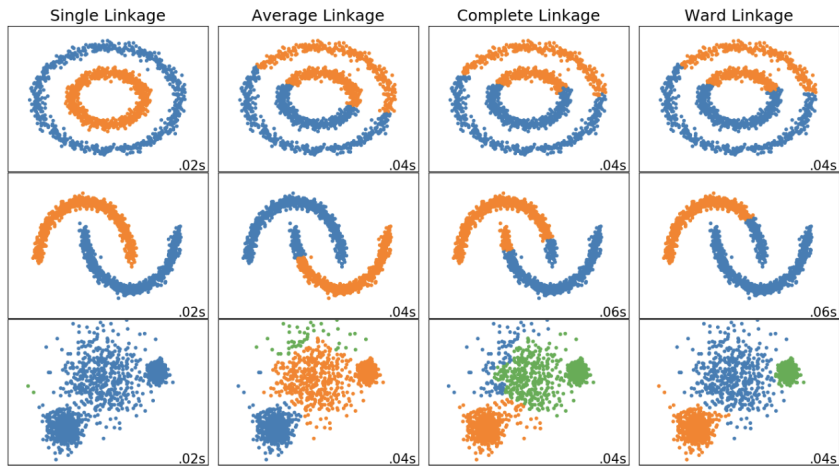
$$R(W, S) = \frac{|S||W|}{|S| + |W|} \rho^2 \left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right);$$

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

Определение

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





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. $U \neq \emptyset$;
 - 3 взять случайную точку $x_0 \in U$;
 - 4 **Повторять**
 - 5 образовать кластер с центром в x_0 и радиусом R :
 $K_0 = \{x_i \in U \mid \rho(x_i, x_0) \leq R\}$;
 - 6 переместить центр x_0 в центр масс кластера:
 $x_0 = \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i$;
 - 7 **Пока** состав кластера K_0 не стабилизируется;
 - 8 Пометить объекты внутри сферы как кластеризованные,
 $U = U \setminus K_0$.

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

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.

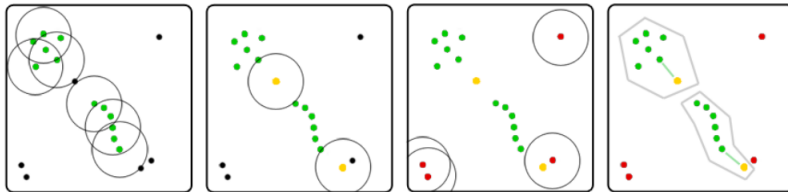


Рис.: Иллюстрация к алгоритму DBSCAN. На рисунке зелёным отмечены корневые объекты, жёлтым — граничные и красным — шумовые.

Вход: Выборка $X^n = \{x_1, \dots, x_n\}$, параметры ϵ и m ;

Выход: разбиение выборки на кластеры и шумовые выбросы;

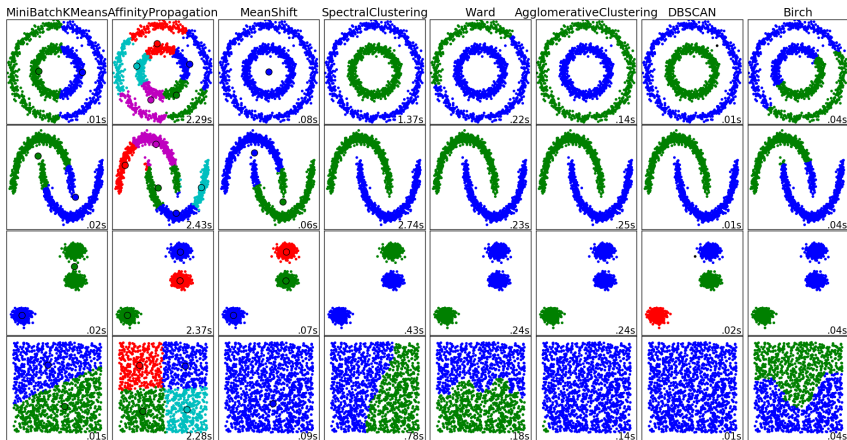
- 1 $U = X_m, a = 0$;
- 2 Пока есть некластеризованные точки, т.е. $U \neq \emptyset$;
- 3 взять случайную точку $x \in U$;
- 4 если $|U_\epsilon(x)| < m$, то
- 5 позначить x как возможно шумовой;
- 6 иначе
- 7 создать новый кластер: $K = U_\epsilon(x), a = a + 1$;
- 8 для всех $x' \in K$
- 9 если $|U_\epsilon(x')| > m$ 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 $U = U \setminus K$.

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.



Самоорганизующаяся карта Кохонена

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

$Y = \{1, \dots, M\} \times \{1, \dots, H\}$ — сетка кластеров,

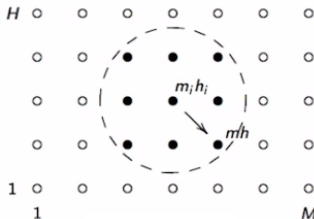
$r: Y \times Y$ — метрика пространства кластеров;

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;

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



- 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:

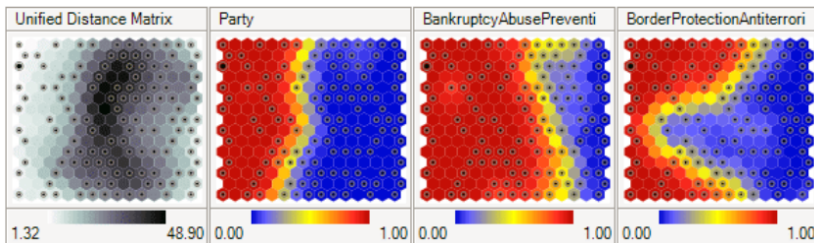
$$(m_i, h_i) = \arg \min_{(m, h) \in Y} \rho(x_i, w_{mh});$$

- 5 **для всех** $(m, h) \in v_\epsilon(m_i, h_i)$
- 6 сделать шаг стохастического градиентного спуска:
 $w_{mh} = w_{mh} + \eta(t)(x_i - w_{mh})K(r[(m_i, h_i), (m, h)], t)$;
- 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:

- 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.

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 .

- 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)

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.

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 .

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

или

Найти: вероятностную тематическую модель

- Гипотеза независимости: 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

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:

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

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

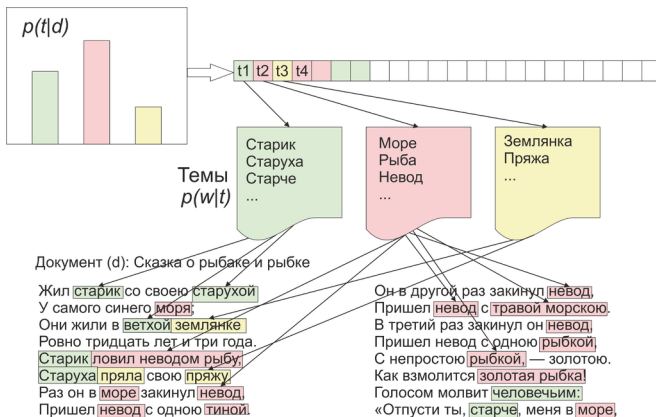
Задача: найти $T, p(w|t), p(t|d)$.

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 тематическая модель:

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



Дано:

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

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

Найти параметры тематической модели:

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

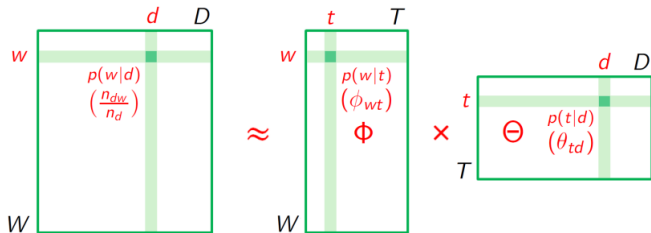
Искомые:

- $\varphi_{wt} = p(w|t)$ — 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:

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



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 ,

$n = \sum_{d \in D} \sum_{w \in d} n_{dw}$ — длина коллекции в терминах.

Максимизация логарифма правдоподобия:

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

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

Приходим к задаче:

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

При ограничениях неотрицательности и нормировки

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

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

Е-шаг: условные вероятности тем $p(t|d, w)$ для всех t, d, w вычисляются через $\varphi_{wt}, \theta_{td}$ по формуле Байеса:

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

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

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

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

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 ;

1 **повторять**

2 обнулить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t$ для всех $d \in D, w \in W, t \in T$;

3 **для всех** $d \in D, w \in d$

4 $Z := \sum_{t \in T} \varphi_{wt} \theta_{td}$;

5 **для всех** $t \in T$ таких, что $\varphi_{wt} \theta_{td} > 0$

6 увеличить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t$ на $\delta = n_{dw} \varphi_{wt} \theta_{td} / Z$;

7 $\varphi_{wt} := \hat{n}_{wt} / \hat{n}_t$ для всех $w \in W, t \in T$;

8 $\theta_{td} := \hat{n}_{dt} / n_d$ для всех $d \in D, t \in T$;

9 **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 множество документов $D_t \subset D$ относится к теме t :

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

- 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 $p(w|d)$:

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

Интерпретация:

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

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

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

$d = d' \cup d''$ — 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 (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.