Приложения EM-алгоритма

Практикум на ЭВМ, весна 2018

Попов Артём Сергеевич

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

1 марта 2018 г.

EM-алгоритм в общем виде

X — наблюдаемые переменные; T — скрытые; Θ — параметры.

Задача максимизации *неполного* правдоподобия:

$$\ln p(X|\Theta) = \ln \int p(X, T|\Theta) dT \to \max_{\Theta}$$

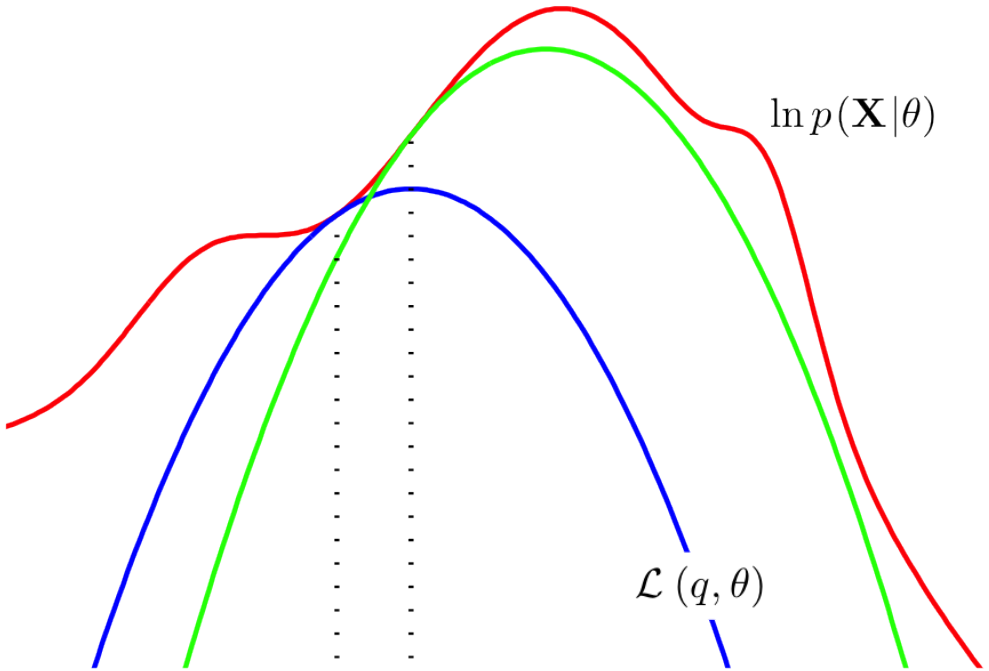
Преобразуем:

$$\ln p(X|\Theta) = \int \ln p(X|\Theta) q(T) dT = \int q(T) \ln \frac{p(X, T|\Theta)}{p(T|X, \Theta)} \frac{q(T)}{q(T)} dT$$

$$= \underbrace{\int q(T) \ln \frac{p(X, T|\Theta)}{q(T)} dT}_{L\{q,\Theta\}} + \underbrace{\int q(T) \ln \frac{q(T)}{p(T|X, \Theta)} dT}_{KL(q(T)||p(T|X,\Theta))} \geq L\{q, \Theta\}$$

- **E-step:** $KL(q(T)||p(T|X, \Theta)) \to \min\limits_{q(T)} \Leftrightarrow q(T) = p(T|X, \Theta)$
- **M-step:** $L\{q, \Theta\} \to \max\limits_{\Theta} \Leftrightarrow \mathbb{E}_{q(T)} \ln p(X, T|\Theta) \to \max\limits_{\Theta}$

Геометрическая интерпретация оптимизационного процесса



- **E-step:** $KL(q(T)||p(T|X,\Theta)) \to \min\limits_{q(T)} \Leftrightarrow q(T) = p(T|X,\Theta)$
- **M-step:** $L\{q,\Theta\} \to \max\limits_{\Theta} \Leftrightarrow \mathbb{E}_{q(T)} \ln p(X,T|\Theta) \to \max\limits_{\Theta}$

Модель смеси распределений

Наблюдаемые переменные:
$X = x_1, x_2, \dots x_N$ – выборка из смеси распределений

Скрытые переменные:
$T = t_1, t_2, \dots t_N$ – номера компонент смеси

Параметры:
$\Theta = \theta_1, \dots \theta_k, w_1, \dots w_k$

Модель: K компонент, каждая имеет свое распределение:

$$p(x_i|t_i = j, \Theta) = p_j(x_i|\theta_j)$$

Компоненты выбираются с весами:

$$p(t_i = j|\Theta) = w_j;$$

Если бы видели скрытые переменные...

Максимизировали бы полное правдоподобие:

$$\begin{cases} \ln p(X, T|\Theta) = \sum_{i=1}^{N} \ln p(x_i, t_i|\Theta) = \\ \qquad = \sum_{i=1}^{N} \sum_{j=1}^{k} [t_i = j] \ln w_j p_j(x_i|\theta_j) \to \max\limits_{\Theta} \\ \sum_{j=1}^{k} w_j = 1, \quad w_j \geq 0 \end{cases}$$

Оценки:

$$\sum_{i=1}^{N} [t_i = j] \ln p_j(x_i|\theta_j) \to \max\limits_{\theta_j}; \qquad w_j = \frac{1}{N} \sum_{i=1}^{N} [t_i = j]$$

... Но мы их не видим.

ЕМ-алгоритм для максимизации неполного правдоподобия:

$$\ln p(X|\Theta) = \sum_{i=1}^{N} \ln p(x_i|\Theta) = \sum_{i=1}^{N} \ln \sum_{j=1}^{k} p(x_i, t_j|\Theta) =$$

$$= \sum_{i=1}^{N} \ln \sum_{j=1}^{k} p(x_i|t_j, \Theta) p(t_j|\Theta) = \sum_{i=1}^{N} \ln \sum_{j=1}^{k} w_j p_j(x_i|\theta_j)$$

$$\begin{cases} \sum_{i=1}^{N} \ln \sum_{j=1}^{k} w_j p_j(x_i|\theta_j) \to \max\limits_{\Theta} \\ \sum_{j=1}^{k} w_j = 1, \quad w_j \geq 0 \end{cases}$$

EM-алгоритм для разделения смеси

- **E-шаг:** оцениваем апостериорные распределения на скрытые переменные по формуле Байеса:

$$p(t_i = j|x_i, \Theta) = \frac{p(t_i = j|\Theta)p(x_i|t_i = j, \Theta)}{p(x_i|\Theta)} = \frac{w_j p_j(x_i|\theta_j)}{\sum_{s=1}^{k} w_s p_s(x_i|\theta_s)}$$

- **M-шаг:** максимизируем м.о. полного правдоподобия:

$$\begin{cases} \mathbb{E}_{q(T)} \ln p(X, T|\Theta) = \mathbb{E}_{q(T)} \sum_{i=1}^{N} \ln p(x_i, t_i|\Theta) = \\ \qquad = \sum_{i=1}^{N} \sum_{j=1}^{k} p(t_i = j|x_i, \Theta) \ln w_j p_j(x_i|\theta_j) \to \max\limits_{\Theta} \\ \sum_{j=1}^{k} w_j = 1, \qquad w_j \geq 0 \end{cases}$$

ЕМ-алгоритм как способ решения системы уравнений

Теорема (необходимые условия экстремума)

Точка $\Theta = (w_j, \theta_j)_{j=1}^k$ локального экстремума $p(X|\Theta)$ удовлетворяет системе уравнений относительно Θ и $G = (g_{ij})$, $g_{ij} = p(t_i = j|x_i, \Theta)$:

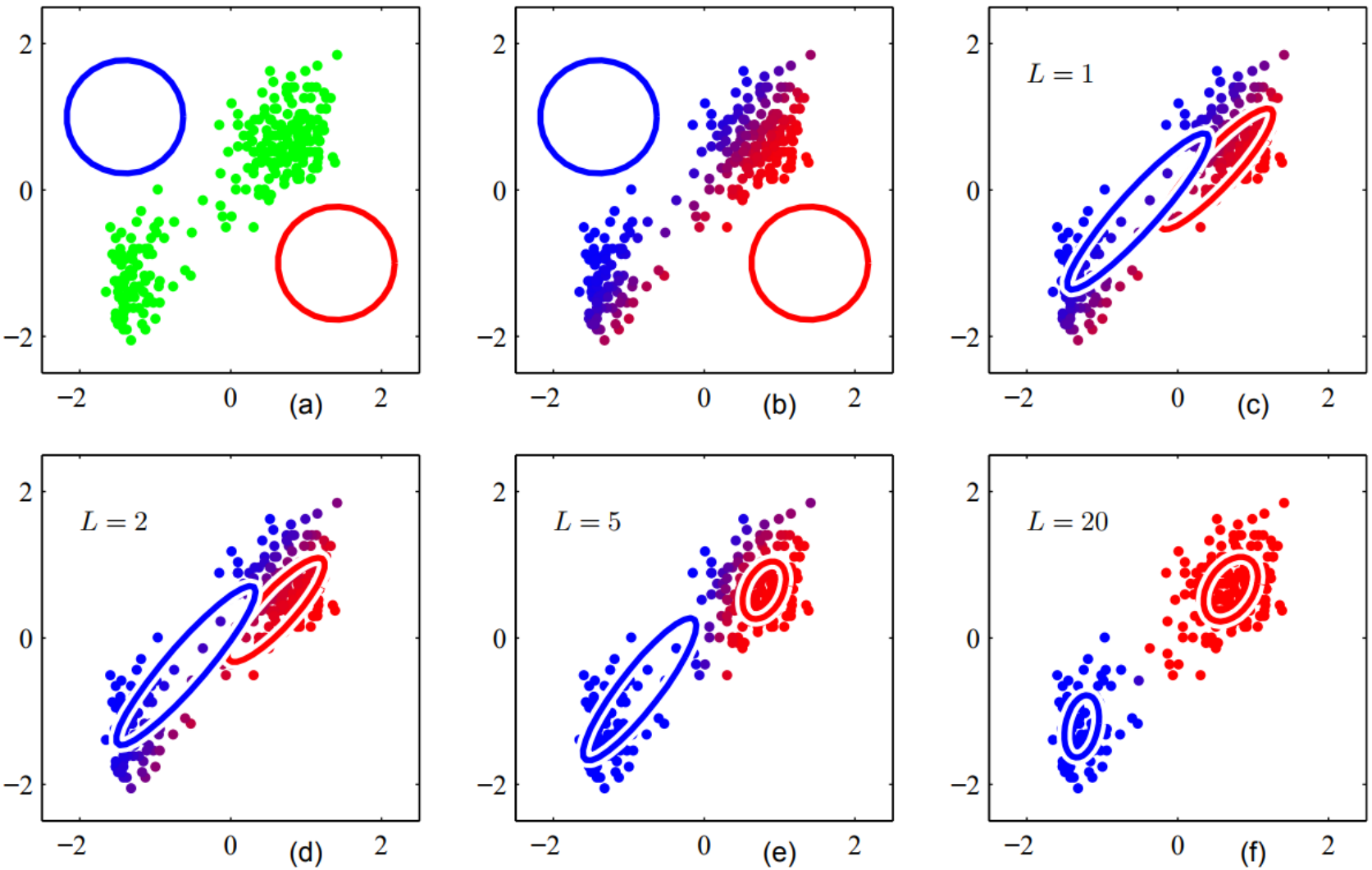
$$g_{ij} = \frac{w_j p_j(x_i|\theta_j)}{\sum_{s=1}^k w_s p_s(x_i|\theta_s)}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k; \text{ (E)}$$

$$\theta_j = \arg\max_\theta \sum_{i=1}^m g_{ij} \ln p_j(x_i|\theta), \quad j = 1, \ldots, k; \text{ (M)}$$

$$w_j = \frac{1}{m} \sum_{i=1}^m g_{ij}, \quad j = 1, \ldots, k. \text{ (M)}$$

ЕМ-алгоритм — это метод простых итераций для её решения

EM-алгоритм по шагам



Постановка задачи кластеризации

Дано:

$X = \{x_1, \dots x_N\}$ — обучающая выборка

Найти:

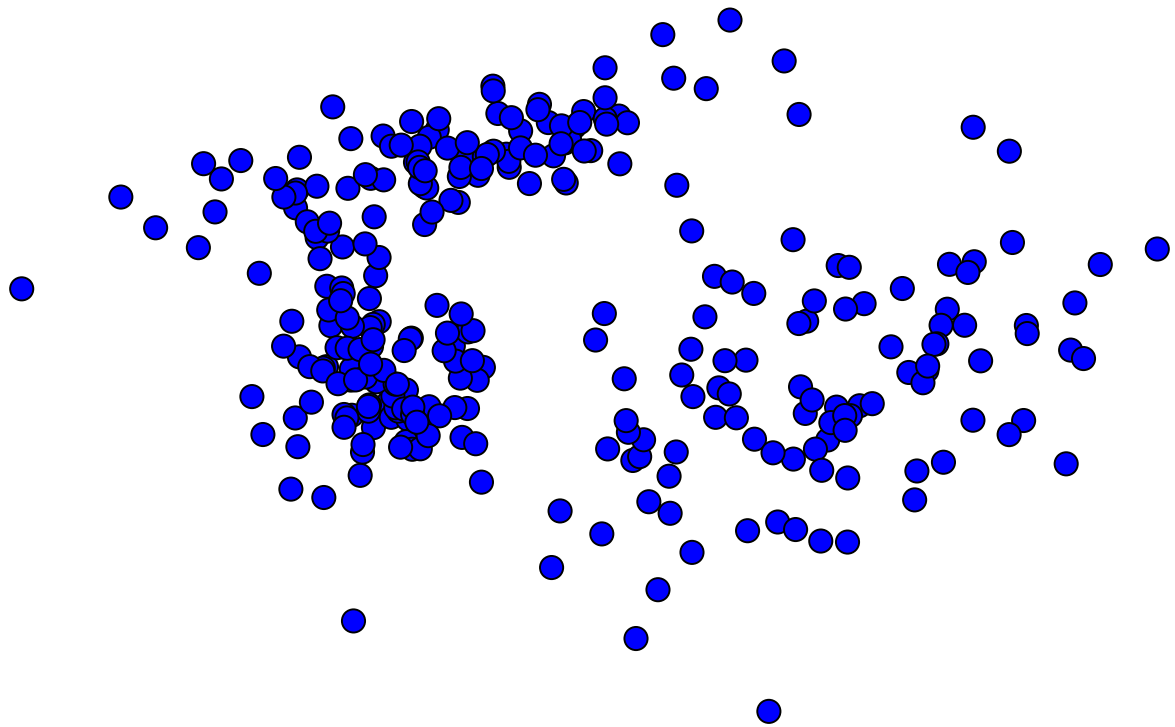
Y — множество кластеров

$a : X \to Y$ — сопоставление объектов кластерам

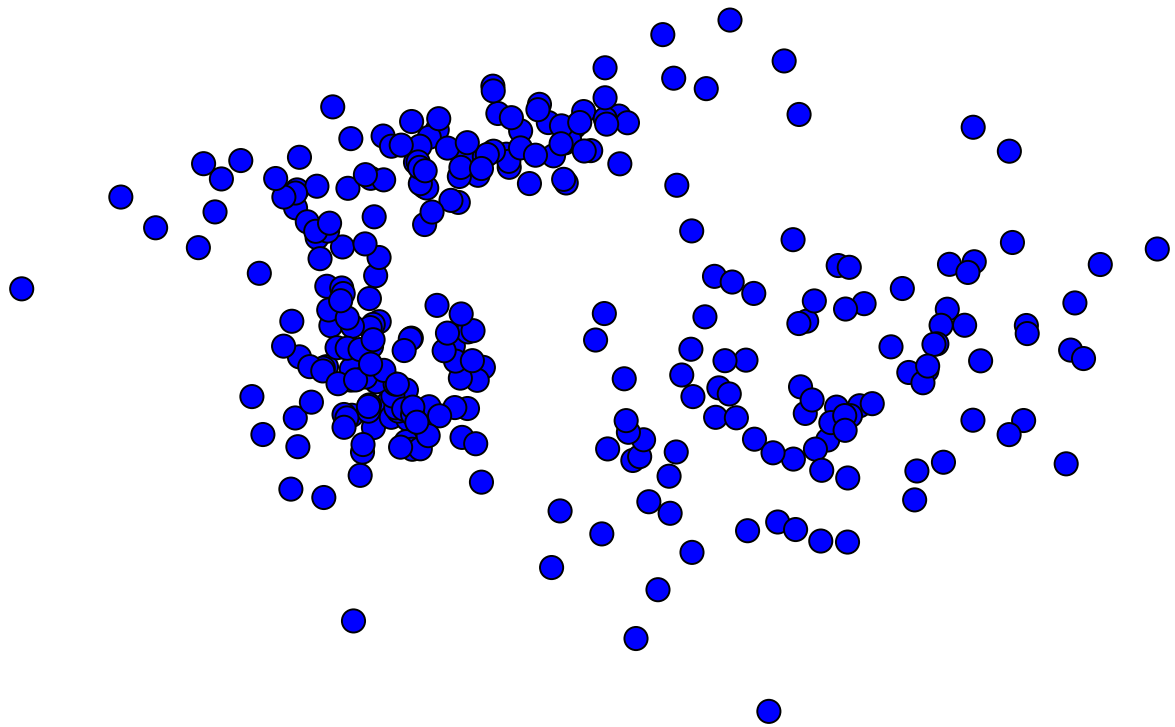
- Каждый кластер состоит из *похожих* объектов
- Объекты в разных кластерах сильно различаются

1. Нет чёткой постановки задачи $\Rightarrow$ нет правильного решения
2. Число кластеров обычно не известно заранее, но часто задаётся исследователем
3. Кластеризация может быть *жёсткой* (одному объекту один кластер) или *мягкой* (объект принадлежит кластеру с некоторой вероятностью)

Модельная задача: сколько кластеров?

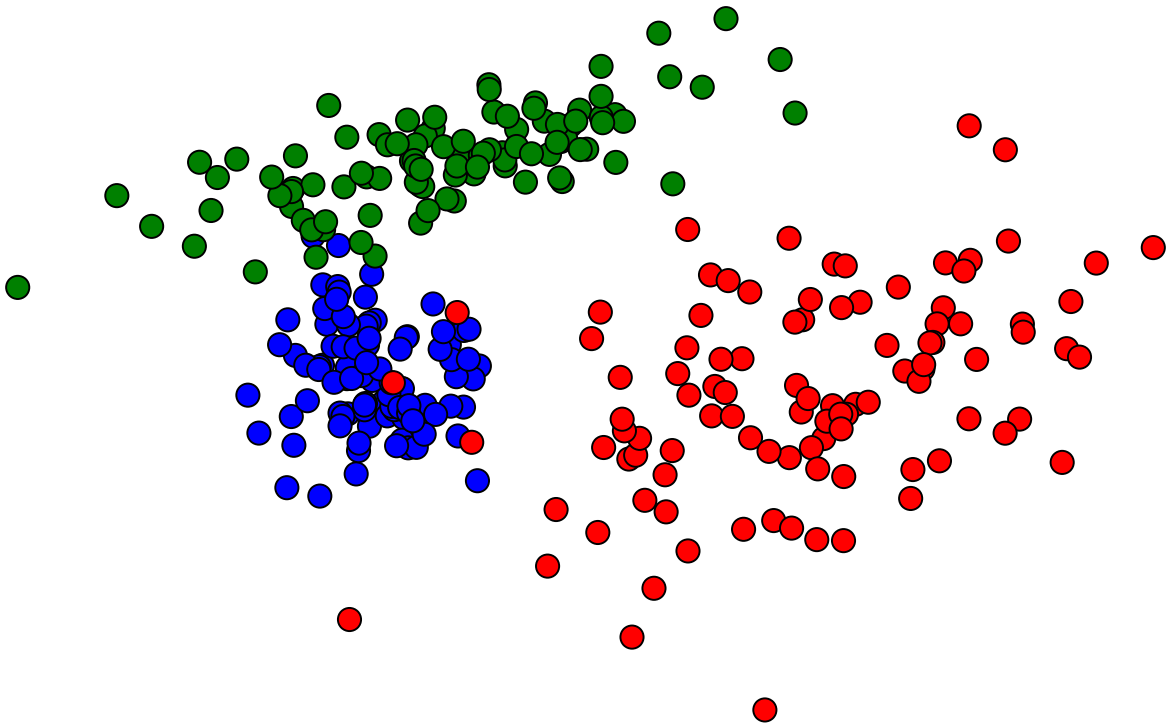


Модельная задача: сколько кластеров?



Нет правильного ответа

Модельная задача: истинные данные



Смесь гауссиан для кластеризации

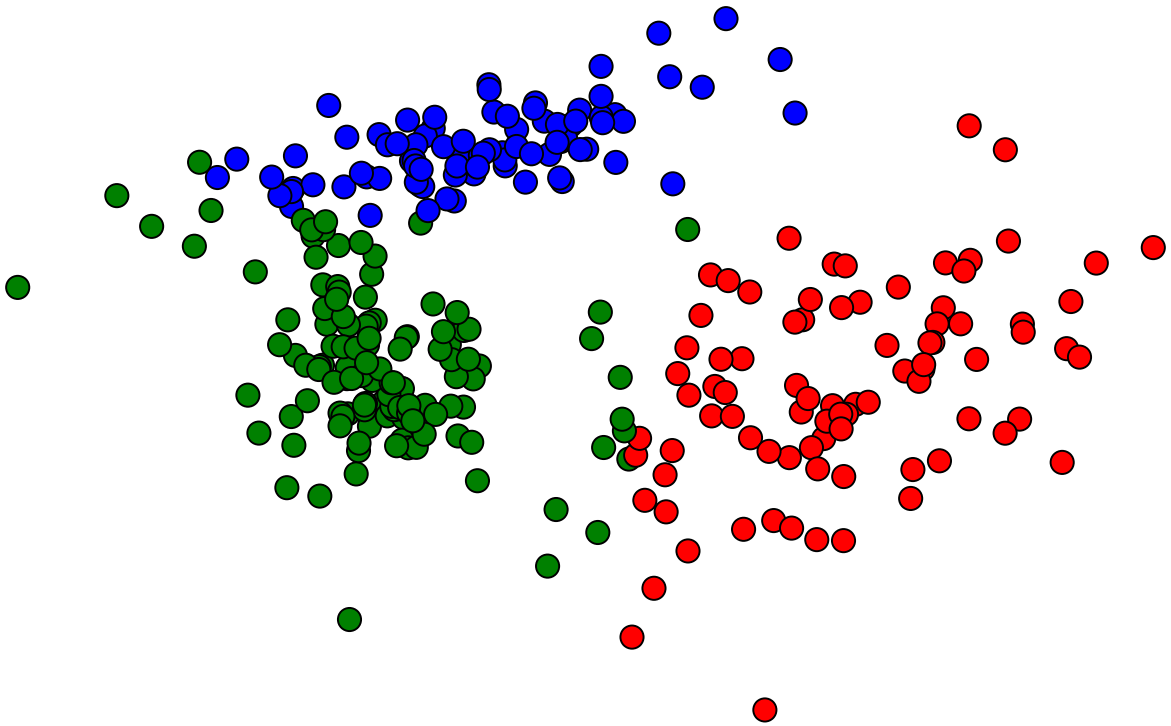
EM-алгоритм для смеси гауссиан можно использовать для решения задачи мягкой кластеризации.

Скрытые переменные:
$T = t_1, t_2, \ldots t_K$ – номера компонент смеси, каждая компонента соответствует кластеру

$p(t_i = j | x_i, \Theta)$ — вероятность принадлежности объекта i кластеру j

Какая форма у кластеров?

Результат кластеризации (5 итераций)



Постановка задачи вычитания фона

Дано:

$X = \{\hat{x}_1, \dots \hat{x}_N\}$ — видеопоследовательность (камера статична)

На каждом изображении присутствуют фон (слабо изменяется во времени) и объекты (сильно изменяются во времени)

Найти:

Для каждого изображения x_i выделить пиксели фона

Гауссиана для описания фона

Выберем последовательность $[i, j]$ пикселей: $x_1^{ij}, \dots x_N^{ij} \in \mathbb{R}^3$
$\forall i, j$ последовательность $x_1^{ij}, \dots x_N^{ij}$ слабо меняется во времени

Опишем последовательность трёхмерной гауссианой:

$$p(x^{ij}) \sim \mathcal{N}(\mu^{ij}, \Sigma^{ij}), \qquad \mu^{ij} \in \mathbb{R}^3, \Sigma^{ij} \in \mathbb{R}^{3\times 3}$$

Как получить μ и Σ?

Гауссиана для описания фона

Выберем последовательность $[i, j]$ пикселей: $x_1^{ij}, \dots x_N^{ij} \in \mathbb{R}^3$
$\forall i, j$ последовательность $x_1^{ij}, \dots x_N^{ij}$ слабо меняется во времени

Опишем последовательность трёхмерной гауссианой:

$$p(x^{ij}) \sim \mathcal{N}(\mu^{ij}, \Sigma^{ij}), \qquad \mu^{ij} \in \mathbb{R}^3, \Sigma^{ij} \in \mathbb{R}^{3\times 3}$$

Как получить μ и Σ?

Оценка максимального правдоподобия:

$$\sum_{n=1}^{N_{train}} \ln \mathcal{N}(x_n^{ij} | \mu^{ij}, \Sigma^{ij}) \to \max_{\mu, \Sigma}$$

Как отделять фон от объектов?

Пороговое правило:

$$x_n^{ij} \text{ — фон } \Leftrightarrow |p(x_n^{ij}) - \mu^{ij}| \leq kq$$

Как отделять фон от объектов?

Пороговое правило:

$$x_n^{ij} \text{ — фон } \Leftrightarrow |p(x_n^{ij}) - \mu^{ij}| \leq kq$$

Проблемы модели: фон не настолько статичен, чтобы описать его нормальным распределением

Способы улучшить модель:

1. Использовать более сложное семейство распределений
   Вместо гауссианы — смесь гауссиан
   EM-алгоритм для вычисления параметров смеси
2. Адаптивно обновлять параметры при поступлении новых изображений
3. При предсказании учитывать соседние пиксели

Адаптивная модель

Учитываем постепенное изменение фона
(для краткости $\mu = \mu^{ij}$, $\Sigma = \Sigma^{ij}$):

Вычислить μ_n, σ_n^2 по первым n объектам.

для $t = n+1, \ldots$

I_t — яркости пикселя

если $|(I_t - \mu_t)| \leq \sigma_t k$ **то**

$$\mu_{t+1} = \rho I_t + (1-\rho)\mu_t$$
$$\Sigma_{t+1} = \rho(I_t - \mu_{t+1})(I_t - \mu_{t+1})^T + (1-\rho)\Sigma_t$$

иначе

$$\mu_{t+1} = \mu_t$$
$$\Sigma_{t+1} = \Sigma_t$$

Тематическое моделирование

Тематическое моделирование (*Topic Modeling*) — приложение машинного обучения к анализу текстов.

Тема (неформально) — набор терминов часто встречающихся вместе в документах, семантически однородное множество документов и т.д.

Тема (формально) задаётся распределениями:

- распределение $p(w|t)$ над терминами $w \in W$
- распределение $p(t|d)$ над темами $t \in T$, где $d \in D$

Цели тематического моделирования:

- выявить структуру текстовой коллекции документов
- построить представления для документов

Пример темы

Топ слова	Топ документы
вода	В школах Авдеевки установили резервуары для питьевой воды
павел	В Авдеевке заканчивается питьевая вода
авдеевка	В Авдеевке установили резервуары с водой и стеклят окн
фильтровальный	Жители Авдеевки будут набирать питьевую воду в специальных емкостях
станция	Авдеевка осталась без воды - надежда на колодцы, скважины и развозку
донецкий	Донецкая фильтровальная станция возобновила работу
водоснабжение	В Авдеевке снова попытаются разминировать фильтровальную станцию
подача	П.Жебривский-В Авдеевке осталось воды на две подачи
ремонтный	Авдеевка с водой из запасов

Пример темы

Коллекция статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Основные положения модели

- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- каждое слово w в документе d связано с некоторой темой t
- d, w — наблюдаемые переменные, темы t — скрытые
- $D \times W \times T$ — дискретное вероятностное пространство
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

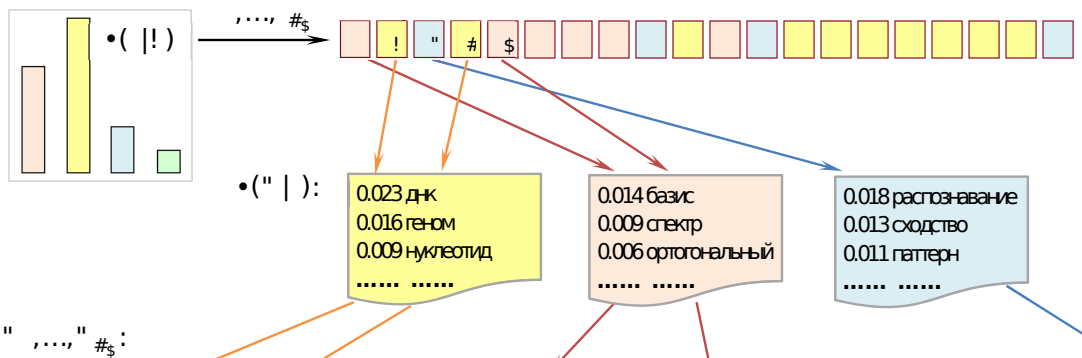
Вероятность появления слова w в документе d:

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

Прямая задача: порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t:

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



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Прямая задача: порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t:

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

Input: распределение $p(w|t)$ для каждой темы $t \in T$;
распределение $p(t|d)$ для каждого документа $d \in D$;
длины документов n_d

Output: коллекция документов;

forall $d \in D$ **do**
  forall *позиций* $i = 1, \ldots, n_d$ *в документе* d **do**
    выбрать тему t_i из $p(t|d)$;
    выбрать термин w_i из $p(w|t_i)$;

Тематическая модель PLSA

n_{wd} — сколько раз слово w встретилось в документе d

Максимизация правдоподобия модели:

$$\sum_{d\in D}\sum_{w\in d} \ln p(w,d) = \sum_{d\in D}\sum_{w\in d} \ln p(w|d)p(d) =$$

$$= \sum_{d\in D}\sum_{w\in d} \ln p(w|d) + const \hat{=} \sum_{d\in D}\sum_{w\in d} \ln \sum_{t\in T} \phi_{wt}\theta_{td} =$$

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

при условии:

$$\sum_{w\in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0, \quad \sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geq 0$$

Вывод PLSA с помощью вероятностного EM

Если есть время, выведем на доске :)

Тематическое моделирование как матричное разложение

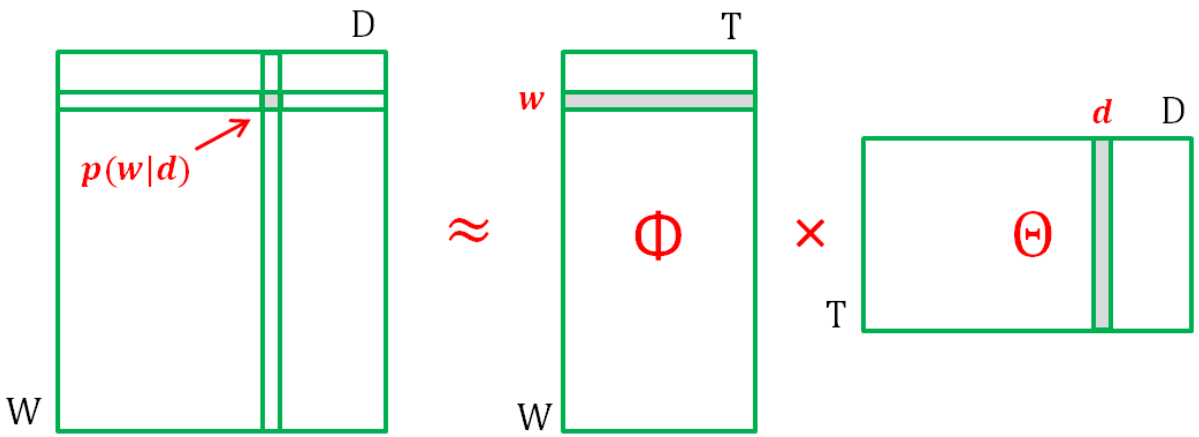
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели
$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

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

Это задача стохастического матричного разложения:



Тематическое моделирование как матричное разложение

Минимизация взвешенных KL-дивергенций между частотными оценками и модельными:

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

$$\Leftrightarrow -\sum_{d\in D} n_d \sum_{w\in W} \frac{n_{dw}}{n_d} \left(\ln p(w|d) - \ln \frac{n_{dw}}{n_d} \right) =$$

$$= \sum_{d\in D} n_d KL(\tilde{p}(w|d) \parallel p(w|d)) \to \min_{\Phi,\Theta} \quad (1)$$

Тематическое моделирование как система уравнений

Ненаблюдаемые частоты, зависящие от t:

$n_{dwt} = \sum_{i=1}^{n} [d_i = d]\,[w_i = w]\,[t_i = t]$ — частота (d, w, t) в коллекции

$n_{wt} = \sum_d n_{dwt}$ — частота термина w в теме t

$n_{td} = \sum_w n_{dwt}$ — частота терминов темы t в документе d

Наблюдаемые частоты, не зависящие от t:

$n_{dw} = \sum_t n_{dwt}$ — частота термина w в документе d

$n_d = \sum_{w,t} n_{dwt}$ — длина документа d

По частотным оценкам:

$$\phi_{wt} = \frac{n_{wt}}{n_w} \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

Тематическое моделирование как система уравнений

Выразим n_{dwt} через ϕ_{wt}, θ_{td} по формуле Байеса:

$$\frac{n_{dwt}}{n_{dw}} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

Получим систему уравнений относительно параметров модели ϕ_{wt}, θ_{td} и вспомогательных переменных n_{dwt}:

$$\begin{cases} n_{dwt} = n_{dw}\dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}, \quad d \in D,\ w \in W,\ t \in T; \\ \phi_{wt} \equiv \dfrac{n_{wt}}{n_t} = \dfrac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, \quad w \in W,\ t \in T; \\ \theta_{td} \equiv \dfrac{n_{td}}{n_d} = \dfrac{\sum_w n_{dwt}}{\sum_{t,w} n_{dwt}}, \quad d \in D,\ t \in T. \end{cases}$$

Решаем методом простых итераций (решение совпадает с ЕМ)

Усложнение алгоритма ТМ

Два пути — задание априорных распределений и регуляризация

- задание априорных распределений

  вместо $\ln p(X|\Theta)$ оптимизируем $\ln p(X|\Theta)p(\Theta|\alpha)$

- регуляризация

  вместо $\ln p(X|\Theta)$ оптимизируем $\ln p(X|\Theta) + R(\Theta)$

  $R(\Theta)$ не обязан иметь вероятностный смысл

Регуляризованные тематические модели

При некоторых ограничениях на R формулы почти не отличаются:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \; + \; R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм:

$$\begin{cases} p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt}\theta_{td}\right) & \text{E-шаг:} \\ n_{dwt} = n_{dw} p_{tdw} & \\ \phi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \phi_{wt} \dfrac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum\limits_{d \in D} n_{dwt} & \text{M-шаг:} \\ \theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum\limits_{w \in d} n_{dwt} & \end{cases}$$

где $\underset{t \in T}{\operatorname{norm}}(x_t) = \dfrac{\max\{x_t, 0\}}{\sum\limits_{s \in T} \max\{x_s, 0\}}$ — операция нормировки вектора.

ЕМ-алгоритм для ТМ

Идея: Е-шаг встраивается внутрь М-шага,
чтобы не хранить трёхмерный массив значений n_{dwt}.

Input: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Output: матрицы терминов тем Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
forall *итераций* $i = 1, \ldots, i_{\max}$ **do**
  $n_{wt}, n_{td} := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
  forall *документов* $d \in D$ *и всех слов* $w \in d$ **do**
    $n_{tdw} := n_{dw} \operatorname{norm}_{t \in T}\left(\phi_{wt}\theta_{td}\right)$ для всех $t \in T$;
    $n_{wt} \mathrel{+}= n_{tdw}$; $n_{td} \mathrel{+}= n_{tdw}$ для всех $t \in T$;
  $\phi_{wt} := \operatorname{norm}_{w \in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := \operatorname{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $d \in D$, $t \in T$;

Примеры регуляризации

- Дальность от заданного распределения β_w или β_t

$$\sum_{t\in T} KL(\beta_w||p(w|t)) \to \max_{\Phi} \qquad \sum_{d\in D} KL(\beta_t||p(t|d)) \to \max_{\Theta}$$

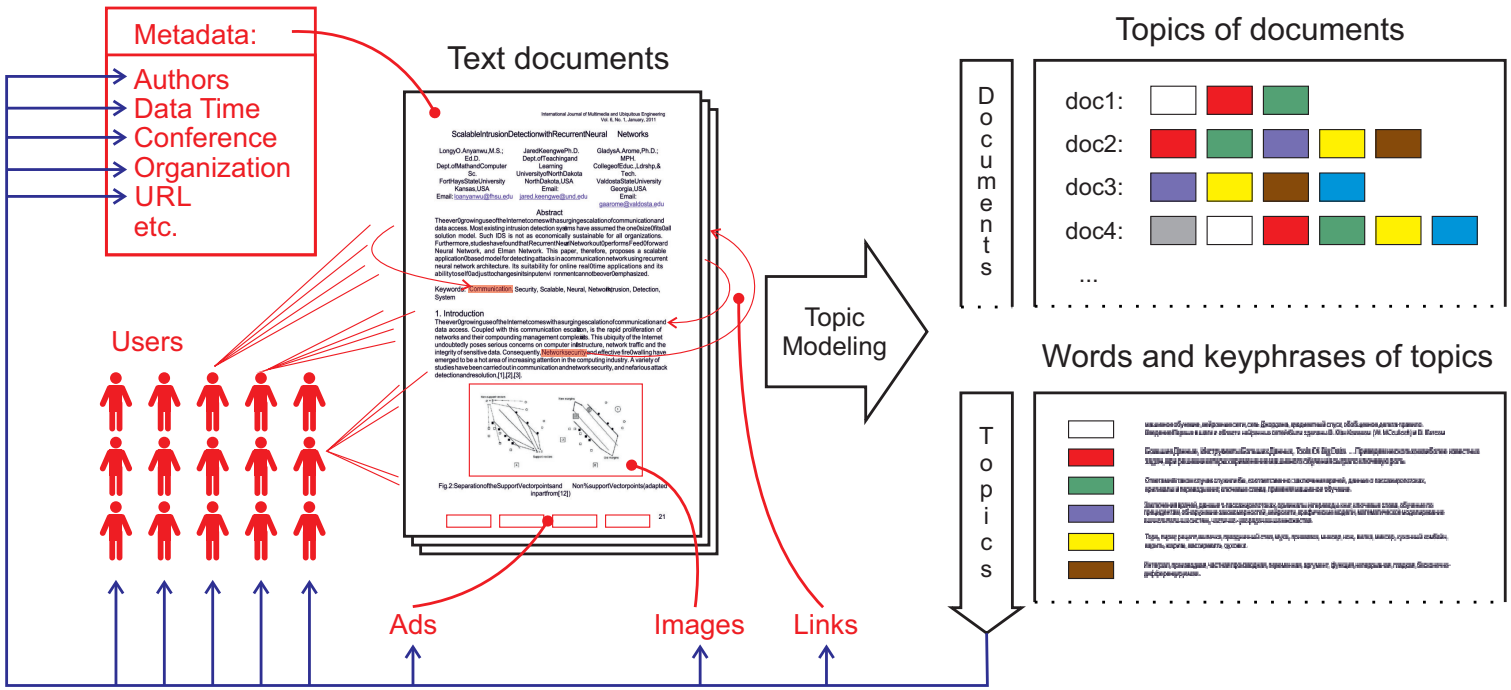
- Близость к данному распределению β_w или β_t

$$-\sum_{t\in T} KL(\beta_w||p(w|t)) \to \max_{\Phi} \qquad -\sum_{d\in D} KL(\beta_t||p(t|d)) \to \max_{\Theta}$$

- Понижение корреляции между разными темами

$$-\frac{1}{2}\sum_{t\in T}\sum_{s\in T, s\neq t}\sum_{w\in W} \phi_{wt}\phi_{ws} \to \max_{\Phi}$$

Мультимодальная тематическая модель

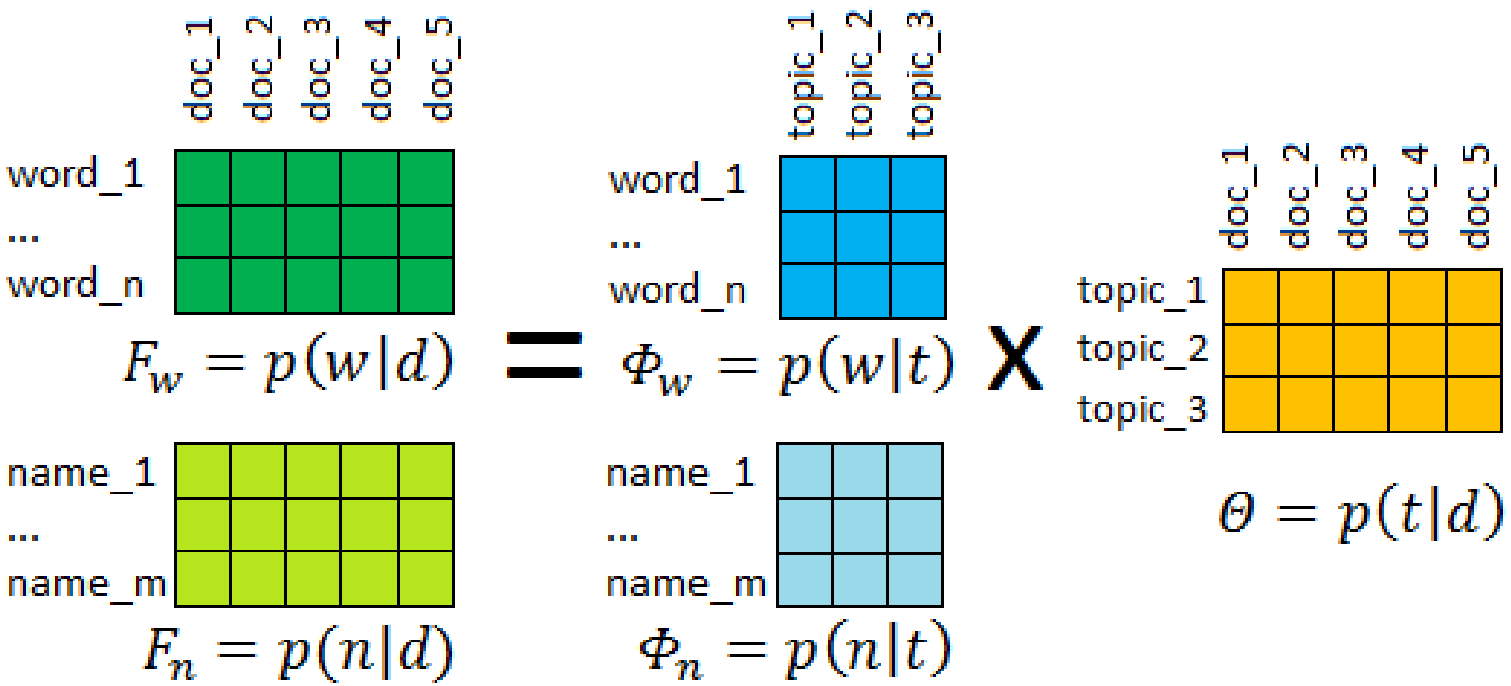


Мультимодальная TM строит распределения тем на терминах $p(w|t)$, авторах $p(a|t)$, метках времени $p(y|t)$, связанных документах $p(d'|t)$, рекламных баннерах $p(b|t)$, пользователях $p(u|t)$, и объединяет все эти модальности в одно тематическую модель.

Пример

Пусть у нас есть две модальности:

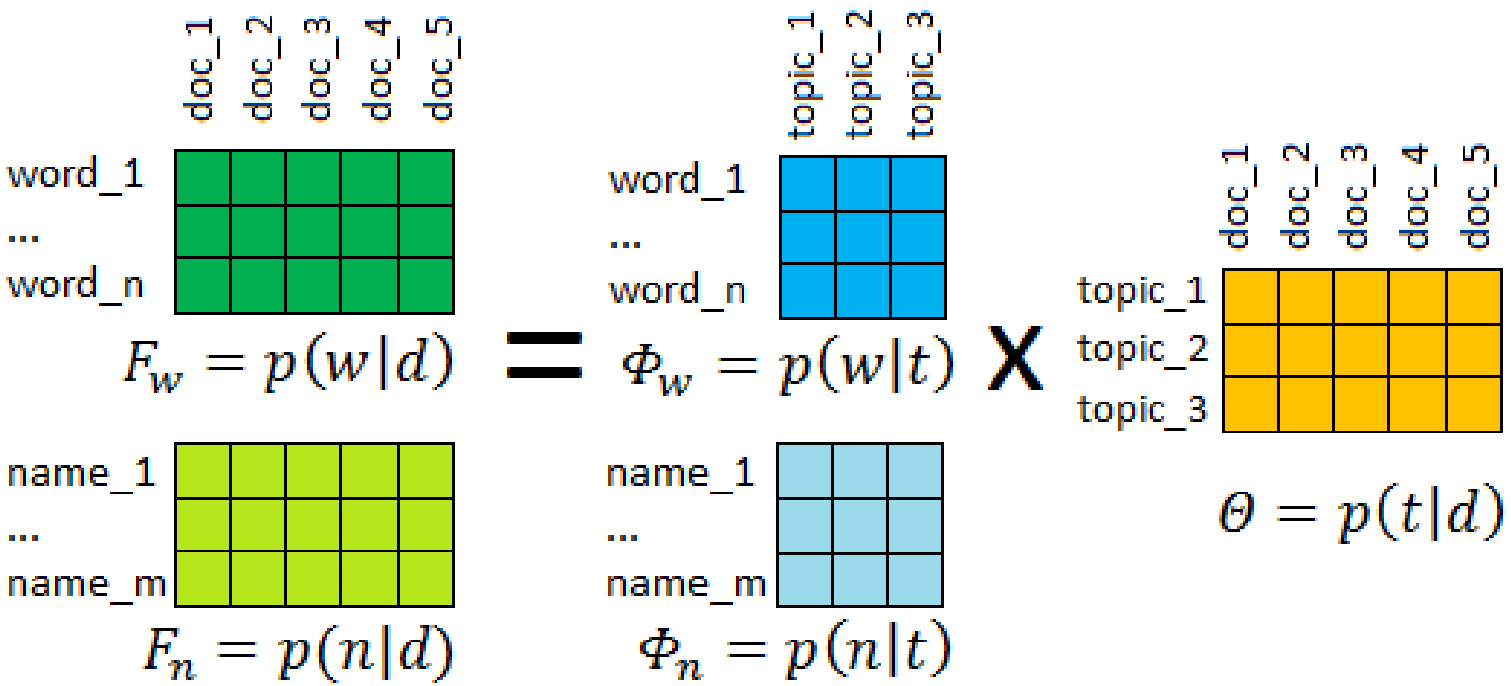
- обычные слова;
- слова-имена авторов



Пример

Пусть у нас есть две модальности:

- обычные слова;
- слова-имена авторов



M-ARTM и EM-алгоритм

W^m — словарь терминов m-й модальности, $m \in M$

Максимизация логарифма мультимодального регуляризированного правдоподобия:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм:

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right), \\ & n_{dwt} = n_{dw} p_{tdw} \\ \text{M-шаг:} & \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt} \dfrac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum\limits_{d \in D} \lambda_{m(w)} n_{dwt} \\ & \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum\limits_{w \in d} \lambda_{m(w)} n_{dwt} \end{cases}$$

Мультиязычные модели

Хотим строить единую модель сразу для нескольких языков
Единая модель == одни и те же темы описывают слова в разных языках

W^ℓ — словарь языка ℓ из множества языков L.

Что у нас для этого есть:

- *Parallel texts* — точный перевод (с выравниванием предложений),
  пример: EuroParl, протоколы европарламента, 21 язык.
- *Comparable* — не перевод, а пересказ на другом языке,
  пример: Википедия.
- *Словари* — все переводы слова w из языка l в язык m

Примеры тем

(вероятности в %)

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Использование словарей

Все переводы слова w из языка L в M:

$$\Pi_{LM}(w) = \left\{ u \;\middle|\; u \in W^M,\ u \text{ — перевод } w \right\}$$

Регуляризация (1) — сближение $p(t|w)$ для comparable текстов:

$$-\sum_{L,M} \sum_{w \in W^L} \sum_{u \in \Pi_{LM}(w)} KL\left(p(t|u) \,\|\, p(t|w)\right) \to \max_{\Phi}$$

Регуляризация (2) — сближение $p(t|d)$ одного языка и частотной оценки $\frac{n_{td}}{n_d}$ для другого языка:

$$-\sum_{L,M} \sum_{w \in W^L} \sum_{u \in \Pi_{LM}(w)} KL\left(\frac{n_{tu}}{n_u} \,\|\, p(t|w)\right) \to \max_{\Phi}$$

$$p(t|w) = \frac{p(w|t)p(t)}{p(w)} = \frac{\phi_{wt} n_t N}{n_w N} = \frac{\phi_{wt} n_t}{n_w}$$

Использование параллельных корпусов

Пары comparable текстов для языков L и M:
$\Pi_{LM} = \left\{ (d_l, d_m) \mid d_l \in L, d_m \in M, d_l \text{ и } d_m \text{ — comparable} \right\}$

Регуляризация (1) — сближение $p(t|d)$ для comparable текстов:

$$-\sum_{L,M} \sum_{(d_l,d_m)\in\Pi_{LM}} KL\left(p(t|d_l) \,\|\, p(t|d_m)\right) \to \max_{\Theta}$$

Регуляризация (2) — сближение $p(t|d)$ одного языка и частотной оценки $\frac{n_{td}}{n_d}$ для другого языка:

$$-\sum_{L,M} \sum_{(d_l,d_m)\in\Pi_{LM}} KL\left(\frac{n_{td_l}}{n_{td_l}} \,\|\, p(t|d_m)\right) \to \max_{\Theta}$$

Использование параллельных корпусов: модальности

Пары comparable текстов для языков L и M:
$\Pi_{LM} = \left\{ (d_l, d_m) \mid d_l \in L, d_m \in M, d_l \text{ и } d_m \text{ — comparable} \right\}$

Пусть каждая пара документов из Π_{LM} представляет собой один документ с двумя модальностями (язык L и язык M)

Тогда, каждому документу из пары (d_l, d_m) соответствует одно представление

Применение TM на практике

Когда полезно использовать:

- Когда нужно узнать что-то о коллекции в целом
- Кластеризация документов
- Тематический поиск
- Задачи с большим числом модальностей
- Один из факторов в рекоммендациях

Заключение

- EM-алгоритм != разделение смеси гауссиан
- EM-алгоритм может использоваться в большом числе задач (в любых, где есть скрытые переменные)
- Иногда, формулы EM-алгоритма можно вывести более простым путём
- С помощью восстановления плотности можно решать задачу индентификации
- С помощью EM-алгоритма можно строить тематические модели
- Тематические модели легко расширяются с помощью регуляризации и модальностей
- Регуляризованные мультимодальные тематические модели могут использоваться для задач мультиязычного поиска