Вероятностные тематические модели
Лекция 2. Постановка задачи, оптимизация и регуляризация

К. В. Воронцов
k.vorontsov@iai.msu.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 3 марта 2025

Вероятностные тематические модели (BTM): о чём этот курс

Probabilistic Topic Modeling (PTM) — область автоматической обработки текстов (Natural Language Processing, NLP)

Курс о том, как

- выявлять тематику документов в текстовых коллекциях
- строить и упрощать прикладные математические теории
- искать тексты по тематике, а не по ключевым словам
- создавать технологии текстовой аналитики для
  — поиска и систематизации научных текстов
  — социо-гуманитарных исследований

Пререквизиты (какие знания потребуются)

- теория вероятности (в основном на конечных множествах)
- машинное обучение (базовые понятия и методология)
- линейная алгебра, методы оптимизации (самые азы)
- язык Python

Содержание

Пусть

- W — конечное множество *термов* (слов, терминов)
- D — конечное множество текстовых *документов*
- T — конечное множество *тем* (topics)
- каждый терм w в документе d связан с некоторой темой t
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен (bag of docs)
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- *гипотеза условной независимости*: $p(w|d,t) = p(w|t)$

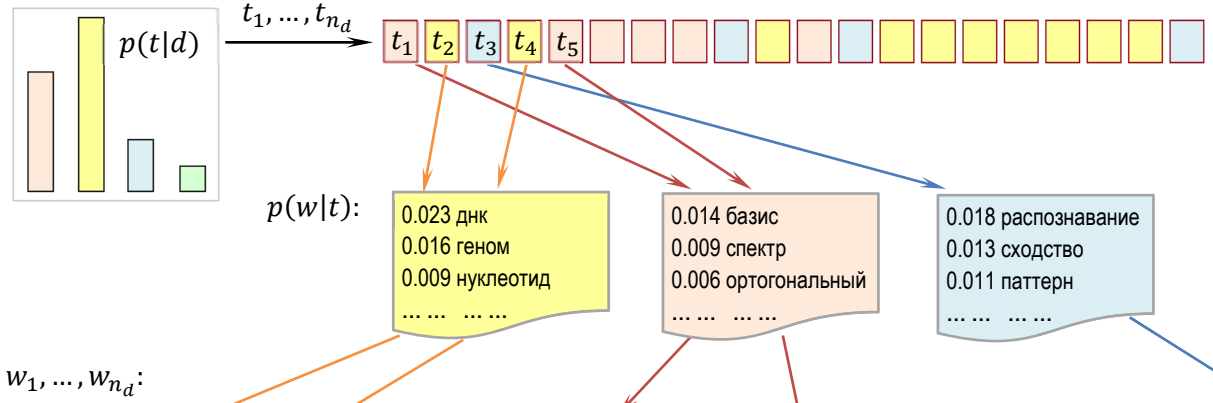
Тематическая модель, по формуле полной вероятности:

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

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

Вероятностная тематическая модель коллекции документов D описывает появление термов w по темам t в документах d:

$$p(w|d) = \sum_t p(w|t)\, p(t|d)$$



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 термов w по темам t в документах d:

$$p(w|d) = \sum_t p(w|t)\, p(t|d)$$

Вход: распределение $p(w|t)$ для каждой темы $t \in T$;
распределение $p(t|d)$ для каждого документа $d \in D$;
Выход: коллекция документов;

для всех $d \in D$
  для всех позиций i в документе d
    сгенерировать тему t_i из $p(t|d)$;
    сгенерировать терм w_i из $p(w|t_i)$;

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Обратная задача: восстановление $p(w|t)$ и $p(t|d)$ по коллекции

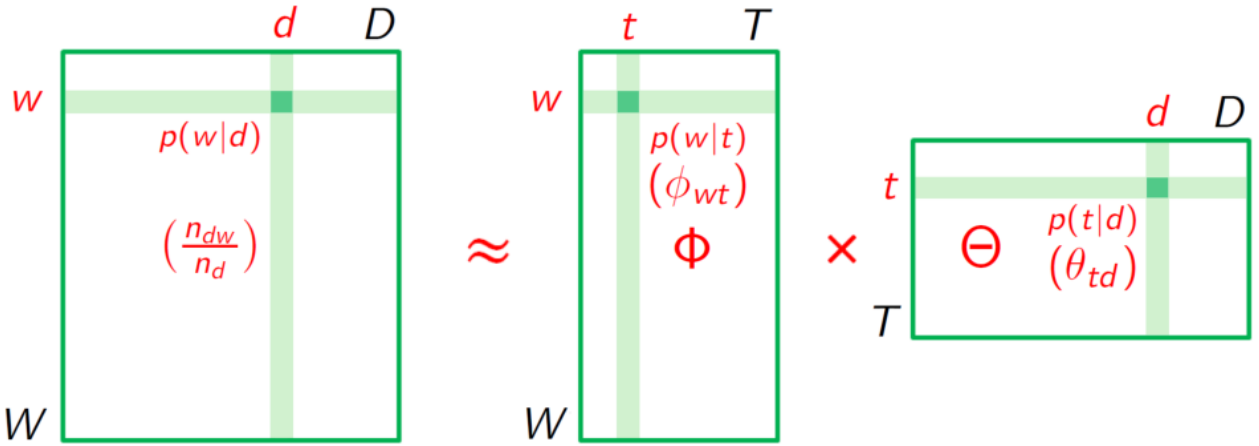
Дано: коллекция текстовых документов

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

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

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

Это задача стохастического матричного разложения:



Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Система обозначений для частот — счётчиков числа термов

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 терма w в теме t

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

$n_t = \sum_{d,w} n_{dwt}$ — частота термов темы t в коллекции

Наблюдаемые частоты, не зависящие от t:

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

$n_w = \sum_d n_{dw}$ — частота терма w в коллекции

$n_d = \sum_w n_{dw}$ — длина документа d

$n = \sum_{d,w} n_{dw}$ — длина коллекции

Частотные оценки условных вероятностей

Имеем ли мы формальное право записывать такие равенства:

- $p(w|d) = \frac{n_{dw}}{n_d}$ — распределение термов в документе d
- $p(t|d) = \frac{n_{td}}{n_d}$ — искомое распределение тем в документе d
- $p(w|t) = \frac{n_{wt}}{n_t}$ — искомое распределение термов в теме t

ДА, но только в ограниченной вероятностной модели текста, при предположении, что $(d_i, w_i, t_i)_{i=1}^n$ — фиксированная последовательность элементарных событий с вероятностями $\frac{1}{n}$

При общем предположении $(d_i, w_i, t_i) \overset{\text{i.i.d.}}{\sim} p(d, w, t)$ это лишь *приближённые частотные оценки условных вероятностей*
(i.i.d. — independent identically distributed)

Элементарное решение обратной задачи

Выразим $p(t|d,w)$ через ϕ_{wt}, θ_{td} по формуле Байеса:

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

Частотные оценки условных вероятностей $\phi_{wt} = \frac{n_{wt}}{n_t}$, $\theta_{td} = \frac{n_{td}}{n_d}$, $p(t|d,w) = \frac{n_{dwt}}{n_{dw}}$ приводят к системе уравнений для ϕ_{wt} и θ_{td}:

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

Численное решение — методом простых итераций

Почему это работает? Какой критерий оптимизируется?

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
Первые 10 слов и их частоты $p(w|t)$ 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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 2. Биграммная модель научных конференций

Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
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

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Свойство интерпретируемости тематических моделей

Тематическая модель даёт тематические векторы:

- $p(t|d) = \dfrac{n_{td}}{n_d} = \theta_{td}$ для каждого документа d
- $p(t|w) = \dfrac{n_{wt}}{n_w} = \dfrac{n_{wt}}{n_t}\dfrac{n_t}{n_w} = \phi_{wt}\dfrac{n_t}{n_w}$ для каждого терма w
- $p(t|d,w) = \dfrac{n_{dwt}}{n_{dw}}$ для каждого локального контекста (d,w)

Интерпретируемость тематических векторов:

- каждая тема t описывается *семантическим ядром* — частотным словарём слов $\left\{w\colon p(w|t) \gg p(w)\right\}$
- тема может «рассказать о себе» словами или фразами
- любой объект x с вектором $p(t|x)$ описывается частотным словарём слов $\left\{w\colon p(w|x) = \sum\limits_{t\in T} p(w|t)p(t|x) \gg p(w)\right\}$

Цели и не-цели тематического моделирования

Цели:

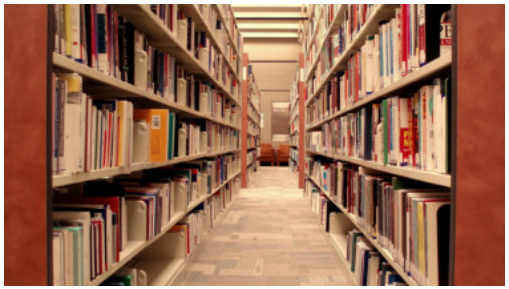
- Выявлять кластерную тематическую структуру текстовой коллекции, сколько в ней тем и о чём они
- Получать *интерпретируемые* тематические векторные представления (эмбединги) слов $p(t|w)$, $p(t|d,w)$, документов $p(t|d)$, фрагментов $p(t|s)$, объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

Не-цели:

- Угадывать слова по контексту (TM слабы как модели языка)
- Понимать смысл текста
- Генерировать связный текст

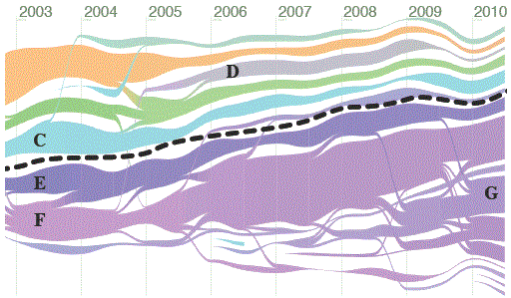
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

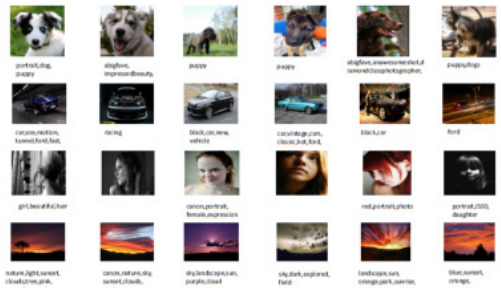


поиск тематических сообществ в соцсетях

выявление и отслеживание цепочек новостей

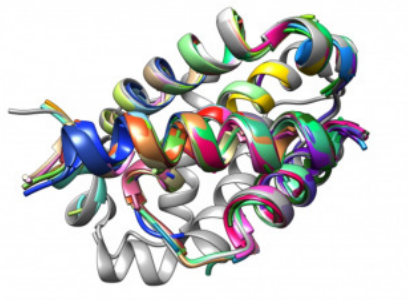


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

поиск паттернов в задачах биоинформатики



J.Boyd-Graber, Yuening Hu, D.Mimno. Applications of Topic Models. 2017.

H.Jelodar et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. 2019.

Принцип максимума правдоподобия

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

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

Максимизация логарифма правдоподобия

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

эквивалентна максимизации функционала

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

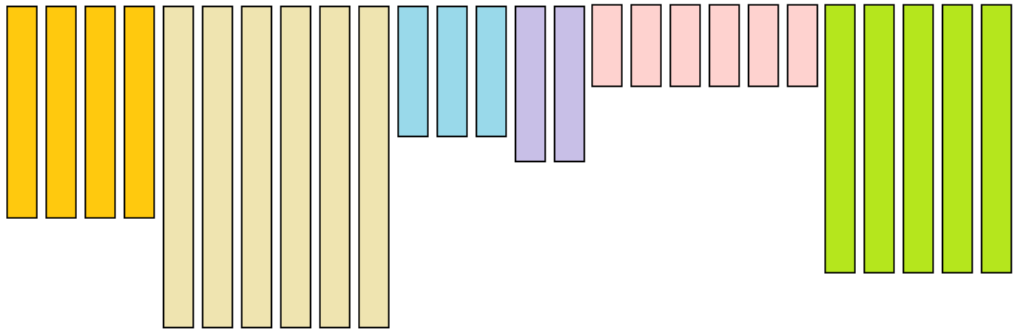
при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \underset{i \in I}{\text{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \dfrac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Векторы $\omega_j = 0$ отбрасываются как вырожденные решения
- Итерации похожи на градиентную оптимизацию:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}},$$

но учитывают ограничения и не требуют подбора шага η

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Доказательство леммы о максимизации на симплексах

Задача: $f(\Omega) \to \max\limits_{\Omega}$; $\sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J.$

Функция Лагранжа:

$$\mathscr{L}(\Omega; \mu, \lambda) = -f(\Omega) + \sum_{j \in J} \lambda_j \Big(\sum_{i \in I_j} \omega_{ij} - 1 \Big) - \sum_{j \in J} \sum_{i \in I_j} \mu_{ij} \omega_{ij}.$$

Условия Каруша–Куна–Таккера для вектора ω_j:

$$\frac{\partial f(\Omega)}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}, \quad \mu_{ij} \omega_{ij} = 0, \quad \mu_{ij} \geqslant 0.$$

Умножим обе части первого равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij} \frac{\partial f(\Omega)}{\partial \omega_{ij}} = \omega_{ij} \lambda_j.$$

Согласно условию леммы $\exists i$: $A_{ij} > 0$. Значит, $\lambda_j > 0$.

Если $\frac{\partial f(\Omega)}{\partial \omega_{ij}} < 0$ для некоторого i, то $\mu_{ij} > 0 \;\Rightarrow\; \omega_{ij} = 0$.

Тогда $\omega_{ij} \lambda_j = (A_{ij})_+$; $\lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname{norm}\limits_i (A_{ij})$. ■

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)
- $\exists \lambda > 0 \;\; f(\Omega^{t+1}) - f(\Omega^t) \geqslant \lambda H(\Omega^t)$ (монотонный рост f)

Тогда $\left|\omega_{ij}^{t+1} - \omega_{ij}^{t}\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей. Труды Института математики и механики УрО РАН. 2020.

Открытая проблема: неудобное четвёртое условие

Определение. $H(\Omega^t)$ есть линейное приближение приращения функции f в окрестности точки Ω^t:

$$f(\Omega^{t+1}) - f(\Omega^t) = H(\Omega^t) + o(\Delta\Omega^t)$$

Лемма. Квадратичное представление функции $H(\Omega)$:

$$H(\Omega) = \frac{1}{2}\sum_{j\in J}\sum_{i,k\in I_j}\left(\frac{\partial f(\Omega)}{\partial \omega_{ij}} - \frac{\partial f(\Omega)}{\partial \omega_{kj}}\right)^2 \omega_{ij}\omega_{kj}$$

Следовательно, $H(\Omega^t) \geqslant 0$.

$f(\Omega^{t+1}) - f(\Omega^t) \approx H(\Omega^t)$ — согласно определению;

$f(\Omega^{t+1}) - f(\Omega^t) \geqslant \lambda H(\Omega^t)$, начиная с некоторой итерации t при некотором $\lambda > 0$ — хотелось бы получить это как результат, а не вводить как предположение. Доказать это пока не удалось.

A.M.Ostrowski. Solution of equations and systems of equations. New York, 1966.

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена по Адамару*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Задача матричного разложения *некорректно поставлена*:
если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta), \quad \text{rank} S = |T|$
- $L(\Phi', \Theta') \approx L(\Phi, \Theta)$

Регуляризация или стабилизация — доопределение решения добавлением второго оптимизационного критерия.

А.Н.Тихонов
(1906–1993)

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

Дифференцируя, выделим вспомогательную переменную p_{tdw}:

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) =$$

$$= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\phi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \quad \blacksquare$$

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ)

Тема t вырождена, если для всех термов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leqslant 0.$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$; это означает, что тема исключается из модели (происходит отбор тем)

Документ d вырожден, если для всех тем $t \in T$

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leqslant 0.$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$; это означает, что модель не в состоянии описать данный документ

Модель вероятностного латентного семантического анализа

PLSA — Probabilistic Latent Semantic Analysis [Хофманн, 1999]:

- $R(\Phi, \Theta) = 0$ — нет никакой регуляризации

Получаем то самое «элементарное решение обратной задачи»

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w \in W}{\text{norm}}\left(\sum_{d \in D} n_{dw} p_{tdw}\right) \\ & \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} n_{dw} p_{tdw}\right) \end{cases}$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Модель латентного размещения Дирихле

LDA — Latent Dirichlet Allocation [Блэй, Ын, Джордан, 2001]:

- $R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W} \beta_w \ln \phi_{wt} + \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td}$

распределения ϕ_t близки к заданному распределению β
распределения θ_d близки к заданному распределению α

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw} p_{tdw} + \beta_w\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw} p_{tdw} + \alpha_t\right) \end{cases}$$

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. NIPS-2001. JMLR 2003.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайновый параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

Этапы исследования при решении практических задач

- Установка и настройка инструментария (BigARTM)
- Понимание задачи, «чего хочет заказчик»
- Получение коллекции, перевод в удобный формат
- Предварительная обработка (токенизация) текстов
- Реализация базовой модели (обычно PLSA)
- Измерение качества тематической модели
- Добавление данных, регуляризаторов, модальностей
- Оптимизация коэффициентов регуляризации
- Оптимизация весов / оценка полезности модальностей
- Оптимизация числа тем
- Интерпретация и визуализация тем
- Прикладное использование тематической модели

Методы предварительной обработки текста

- Удаление чисел, не-слов и «прочей грязи»
- Устранение переносов (когда текст был в pdf)
- Исправление опечаток (для пользовательских данных)
- Лемматизация (для русского языка)
- Стемминг (для английского языка)
- Удаление слишком редких слов (если «мешок слов»)
- Удаление стоп-слов (если не строить фоновые темы)
- Автоматическое выделение терминов (ATE)
- Выделение именованных сущностей (NER)
- Сокращение словаря (Vocabulary Reduction)

Извлечение объектов и фактов из текстов в Яндексе. Лекция для Малого ШАДа, 2013. `https://habr.com/ru/company/yandex/blog/205198`

`https://nlpub.ru/Обработка_текста`

Какими будут задания по курсу

Задача-минимум: научиться решать задачи NLP с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	$5X$
обзор по NeuralTM	$5X$
интеграция ARTM в pyTorch	$5X$
участие в одном из проектов	$10X$
работа над открытой проблемой	$10X$

где X — оценка за вид деятельности по 5-балльной шкале.

Итоговая оценка: $\min\big(10, \lfloor \text{score}/5 \rfloor\big)$ по 10-балльной шкале.

Теоретическое задание №1

Два упражнения на принцип максимума правдоподобия:

- Униграммная модель документов: $p(w|d) = \xi_{dw}$
  Найти параметры модели ξ_{dw}.
- Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d
  Найти параметры модели ξ_w.

Подсказка: применить условия KKT или «основную лемму».

Третье упражнение в продолжение — более творческое:

- Предложите модель, определяющую роли слов в текстах:
  — тематические слова
  — специфичные слова документа (шум)
  — слова общей лексики (фон)

Подсказка 1: ввести распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Примеры датасетов для заданий по курсу

- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Научные статьи: eLibrary, Semantic Scholar, arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- TechCrunch (английский)
- Данные социальных сетей: VK, Twitter, Telegram,...
- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Мастерская знаний» для научного поиска
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus.
  — задача: показать пользователю тематику подборки
  — понадобится автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем
  — конечная цель: ускорить понимание предметной области
- «Тематизатор» для социо-гуманитарных исследований
  — пользователь задаёт грубый фильтр текстового потока
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме
  — конечная цель: q&q аналитика проблемной среды

Открытые проблемы тематического моделирования

1. Проблема несбалансированности тем в коллекции
2. Обеспечение 100%-й интерпретируемости тем
3. Тематические модели внимания последовательного текста
4. Обнаружение новых тем или трендов в потоке текстов
5. Автоматическое именование и аннотирование тем
6. Обзор подходов в нейросетевых тематических моделях
7. Обеспечение полноты и устойчивости множества тем
8. Автоматический подбор гиперпараметров, AutoML
9. Оптимизация гиперпараметров в потоковом режиме
10. Проблема несбалансированности текстов по длине
11. Бережное слияние моделей нескольких коллекций
12. Гиперграфовые тематические модели в RecSys

Резюме

- Основная лемма о максимизации на единичных симплексах
- Вероятностная тематическая модель (PTM) — это:
  — мягкая кластеризация документов по кластерам-темам
  — стохастическое матричное разложение
  — вероятностные эмбединги текстов и слов
- Задача некорректно поставлена, её решение не единственно
- ARTM — для построения моделей с заданными свойствами
- BigARTM — открытая реализация http://bigartm.org
- Что дальше в этом курсе:
  — изучаем много разных моделей и регуляризаторов
  — применяем для решения практических задач
  — участвуем в прикладных проектах
  — создаём альтернативную реализацию в pyTorch