Оптимизация на единичных симплексах для обучения тематических моделей и нейронных сетей

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН • профессор ВМК МГУ,
руководитель лаборатории машинного обучения
и семантического анализа Института ИИ МГУ,
г.н.с. ФИЦ ИУ РАН, профессор МФТИ

Университет Сириус

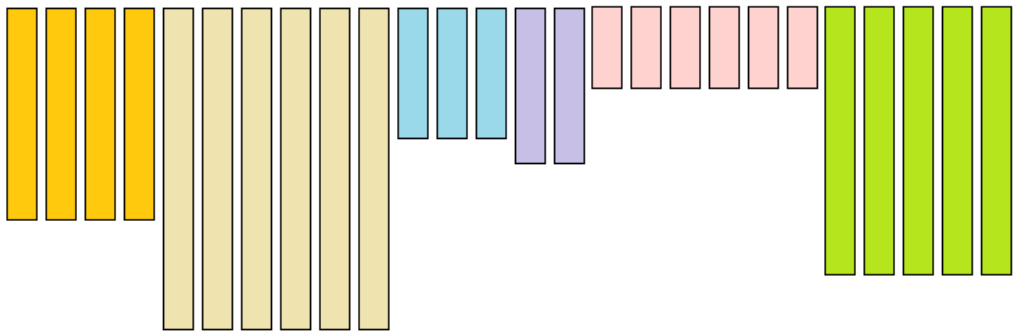
Научная школа
«Обратные некорректные задачи
и машинное обучение»
Сочи • 5 сентября 2023

Содержание

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 нашей задачи
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

Напоминание. 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.

Промежуточные итоги и направления исследований

- Метод похож на обычную градиентную оптимизацию, но не требует подбора градиентного шага η
- Ограничения неотрицательности и нормировки могут накладываться не на все векторы, а лишь на некоторые
- Операция norm может приводить к обнулению части координат, следовательно, к разреживанию векторов ω_j
- Приложение 1: вероятностное тематическое моделирование
- Приложение 2: неотрицательные матричные разложения
- Приложение 3: нейронные сети с неотрицательными весами

- **Открытая проблема:** упростить четвёртое условие в теореме сходимости (оно представляется избыточным)
- **Открытая проблема:** оценить скорость сходимости

Монотонная аппроксимация: постановка задачи

Дано: выборка $(x_i, y_i)_{i=1}^m$, $x_i \in \mathbb{R}^n$

Найти: предсказательную модель $a(x, w)$ с параметром w, обладающую свойством монотонности:

$$x \leqslant x' \;\Rightarrow\; a(x, w) \leqslant a(x', w)$$

Критерий: минимум эмпирического риска при зашумлённых данных (не обязательно $x_i \leqslant x_j \;\Rightarrow\; y_i \leqslant y_j$)

$$\sum_{i=1}^m \mathscr{L}\big(a(x_i, w), y_i\big) \to \min_w$$

Например, для монотонной регрессии (isotonic regression)

$$\sum_{i=1}^m \big(a(x_i, w) - y_i\big)^2 \to \min_w$$

Линейная модель $a(x, w) = \langle x, w \rangle$ монотонна $\Leftrightarrow$ $w \geqslant 0$

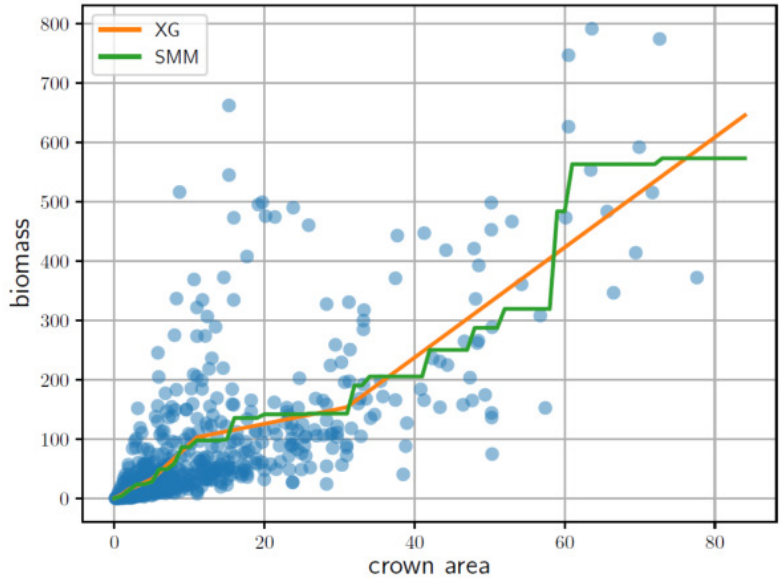
Ограничение монотонности: приложения и интерпретации

— учёт априорных знаний вида «чем больше значение признака, тем выше отклик»

— агрегирование нескольких моделей в ансамбль

— моделирование многомерных функций распределения

— синтез интерпретируемых векторных представлений в глубоких нейронных сетях



Пример.
Зависимость биомассы (и связанного углерода) от площади кроны деревьев

Christian Igel. Smooth monotonic networks. 1 Jun 2023.

J.-R. Cano et al. Monotonic classification: An overview on algorithms, performance measures and data sets. Neurocomputing, 2019.

Монотонная Min-Max нейронная сеть

Трёхслойная сеть с двумя слоями min и max пулинга:

$$a(x,w) = \min_{k\in K} \max_{h\in H_k} \big(\langle w_{kh}, x\rangle - b_{kh}\big), \quad w_{kh} \geqslant 0$$

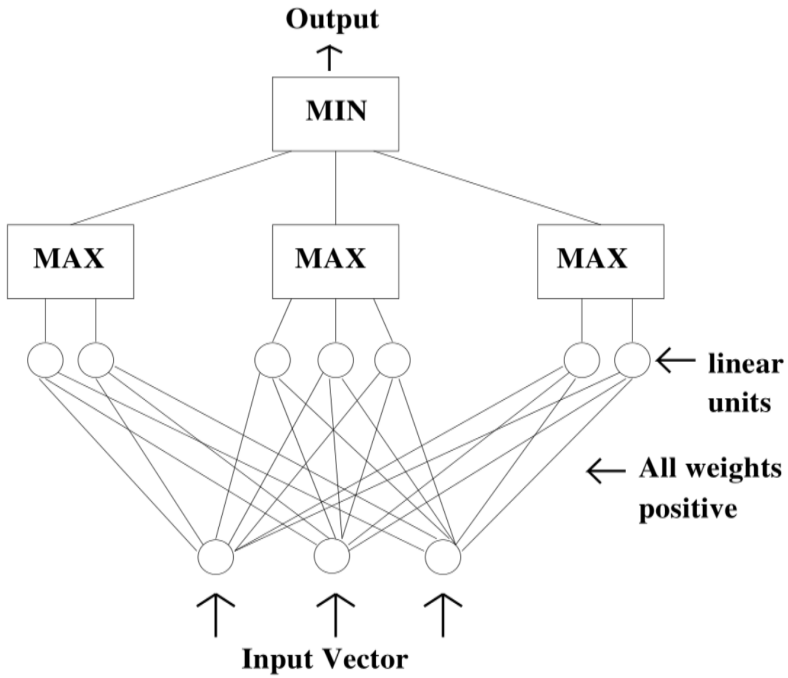
Репараметризация: $w_{kh} = \exp(z_{kh})$

Преимущества:

— можно использовать BackProp
— можно встраивать в DeepNN
— доказана аппроксимируемость

Недостатки:

— кусочно-линейная модель
— затухание градиентов
— недообучение, переусложнение
— чувствительность к инициализации



Joseph Sill. Monotonic networks. NeurIPS, 1997.

Сглаженная монотонная нейронная сеть (Smooth Min-Max)

Введём функцию LogSumExp с параметром β:

$$\underset{i\in I}{\mathrm{LSE}}{}_{\beta}(z_i) = \frac{1}{\beta}\ln\sum_{i\in I}\exp(\beta z_i) \to \begin{cases} \max_i(z_i), & \beta\to+\infty \\ \min_i(z_i), & \beta\to-\infty \end{cases}$$

Min-Max-сеть с заменой min и max на их сглаженные аналоги:

$$a(x,w) = \underset{k\in K}{\mathrm{LSE}}{}_{-\beta}\ \underset{h\in H_k}{\mathrm{LSE}}{}_{+\beta}\big(\langle w_{kh}, x\rangle - b_{kh}\big), \qquad w_{kh}\geqslant 0$$

Преимущества:

— гладкая аппроксимирующая монотонная функция
— доказаны асимптотические аппроксимационные свойства
— можно использовать BackProp, встраивать в DeepNN
— обобщающая способность существенно выше, чем у Min-Max

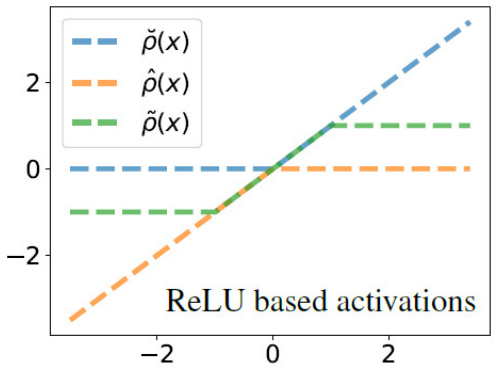
Christian Igel. Smooth Monotonic Networks. 1 Jun 2023

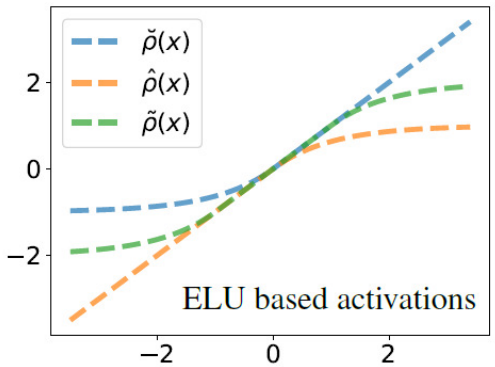
Ограниченная монотонная нейронная сеть (Constrained MNN)

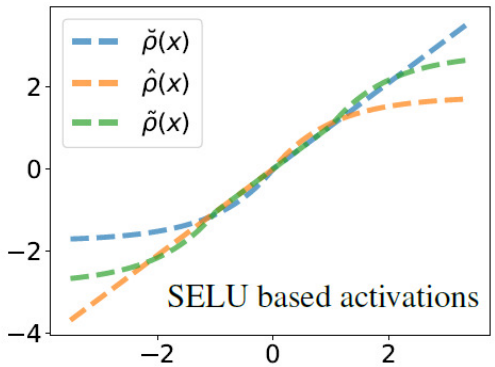
Двухслойная сеть, использующая три функции активации на скрытом слое: выпуклую, вогнутую и выпукло-вогнутую

$$a(x,w) = \sum_{k\in K} w_k^2 \sigma_k\big(\langle w_k^1, x\rangle - b_k\big), \quad w_k^L \geqslant 0$$

где $\sigma_k \in \{\check{\rho}, \hat{\rho}, \tilde{\rho}\}$





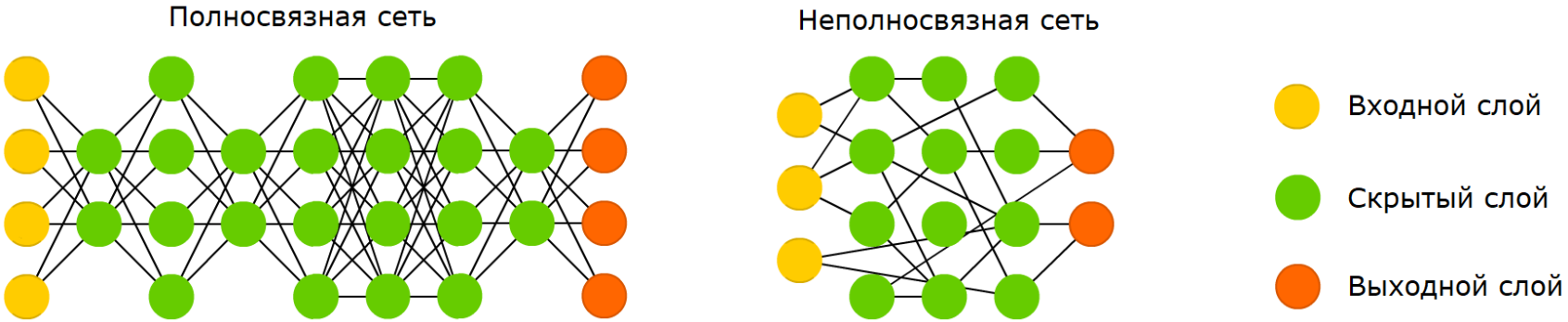


Преимущества те же, что у нейронной сети Smooth Min-Max
Кто из них лучше? Видимо, их ещё не успели сравнить.

D.Runje, S.M.Shankaranarayana. Constrained monotonic neural networks. 31 May 2023

Глубокие нейронные сети (Deep Neural Network, DNN)

Вычисление сети: $x^{\ell+1} = \sigma(W^\ell x^\ell)$, по слоям $\ell = 1, 2, \ldots, L$



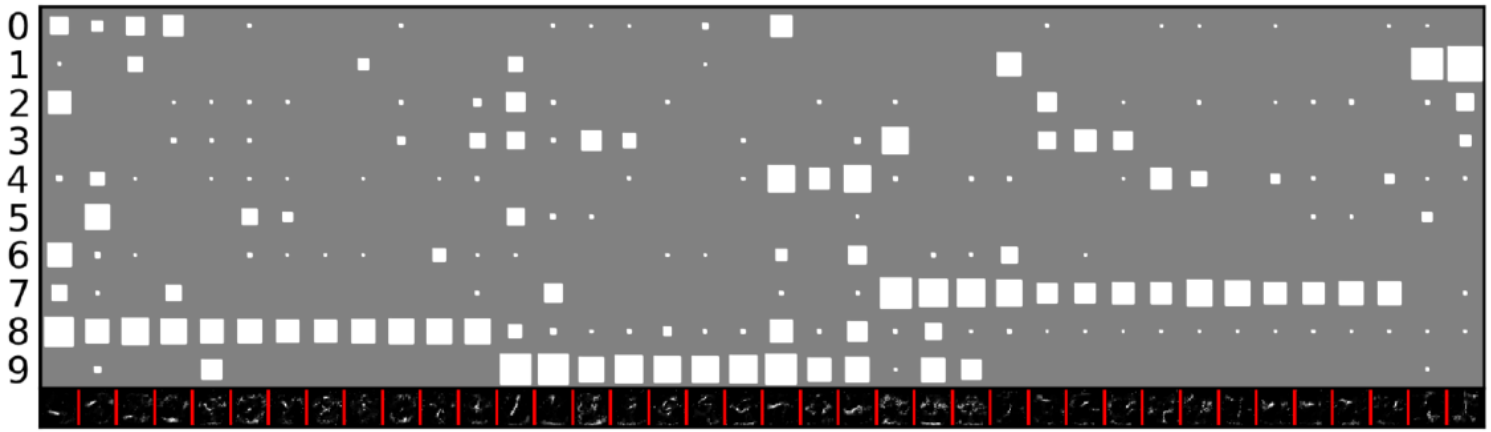
Достаточные условия монотонности DNN:
— неотрицательность коэффициентов $W^\ell \geqslant 0$
— монотонность функций активации $\sigma(z)$

Если предсказывать положительные значения $y(x)$ при $W^\ell \geqslant 0$, то векторы x^ℓ выучиваются детектировать части объекта x, при этом и веса W^ℓ, и векторы x^ℓ становятся разреженными

J.-R. Cano et al. Monotonic classification: An overview on algorithms, performance measures and data sets. Neurocomputing, 2019.

Глубокая монотонная нейронная сеть (Constrained MNN)

Двухслойная сеть для распознавания рукописных цифр MNIST
— первый слой выделяет информативные группы пикселей
— второй слой выделят группы, образующие цифры



Интерпретируемость и разреженность весов W^ℓ и векторов x^ℓ возникает также при обучении автокодировщиков без учителя

J.Chorowski, J.M.Zurada. Learning understandable neural networks with non-negative weight constraints. 2015

B.O.Ayinde, J.M.Zurada. Deep learning of nonnegativity-constrained autoencoders for enhanced understanding of data. 2018

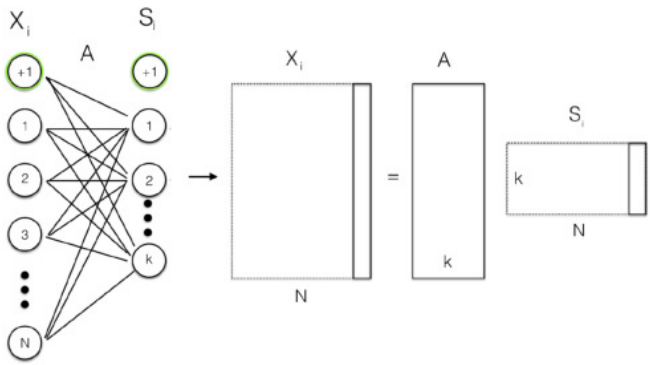
Неотрицательные матричные разложения и автокодировщики

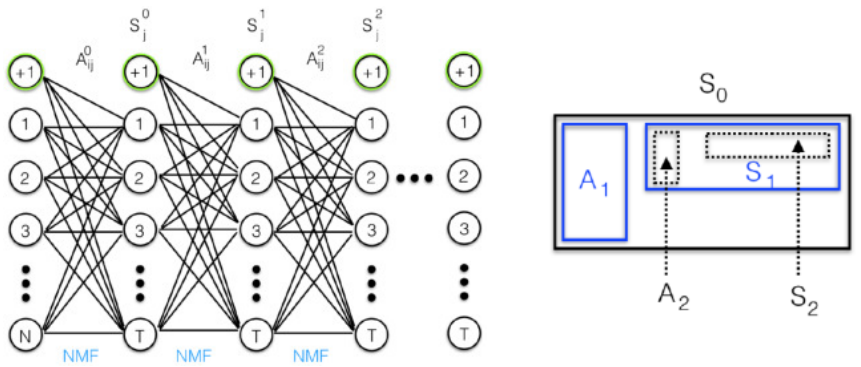
Дано: неотрицательная матрица данных X, $X_{ij} \geqslant 0$

Найти: неотрицательные матрицы A, S, $A_{ik} \geqslant 0$, $S_{kj} \geqslant 0$

Критерий: $\|X - AS\| \to \min_{A,S}$ — non-negative matrix factorization

Deep NMF — обучаемая иерархия векторных представлений:





J.Flenner, B.Hunter. A deep non-negative matrix factorization neural network. 2017

Zhikui Chen, Shan Jin, Runze Liu, Jianing Zhang. A Deep non-negative matrix factorization model for big data representation learning. 2021

T.Will et al. Neural nonnegative matrix factorization for hierarchical multilayer topic modeling. 2023

Промежуточные итоги и направления исследований

- Неотрицательность весов в нейросетевых моделях приводит к интерпретируемости и разреженности
- Известно, что введение нормировки при оптимизации повышает устойчивость обучения нейронных сетей
- Переход от нормированных векторов (при необходимости) к ненормированным — умножением на коэффициент
- Во всех перечисленных ситуациях может быть использована оптимизация на единичных симплексах
- Илья Дьяков, студент 2 курса ВМК МГУ, реализовал в pyTorch мультипликативный шаг с нормировкой

- **Открытая проблема:** находить применения этой технике, сравнивать со state-of-the-art на различных задачах

Тематическое моделирование: «о чём все эти тексты?»

Дано:

- коллекция текстовых документов D, словарь W
- n_{dw} — частота слов (термов) $w \in W$ в документе $d \in D$
- $|T|$ — сколько тем хотим найти в коллекции D

Найти:

- $p(w|t) = \phi_{wt}$ — вероятности слов w в каждой теме $t \in T$
- $p(t|d) = \theta_{td}$ — вероятности тем t в каждом документе d
- $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ — тематическую языковую модель

Критерий: правдоподобие предсказания слов w в документах d с дополнительными критериями-регуляризаторами $R_i(\Phi, \Theta)$:

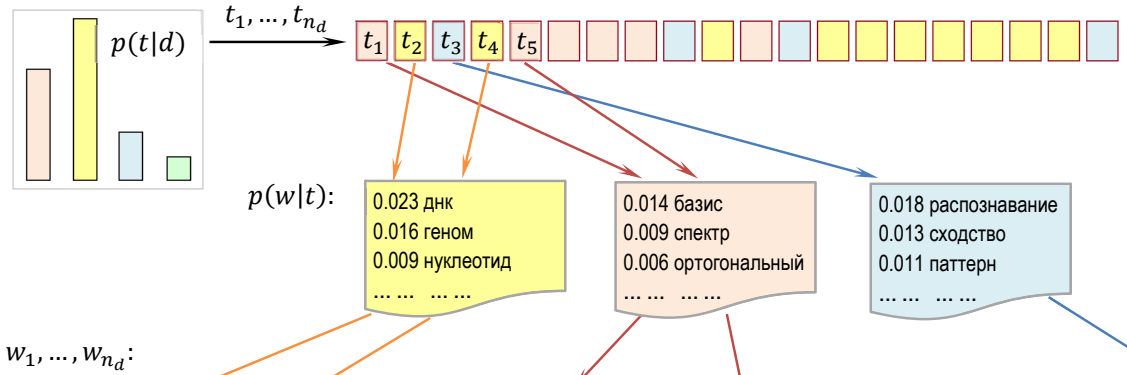
$$\sum_{d \in D} \sum_{w \in d} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

Почему это обратная задача, и почему именно такой критерий?

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

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

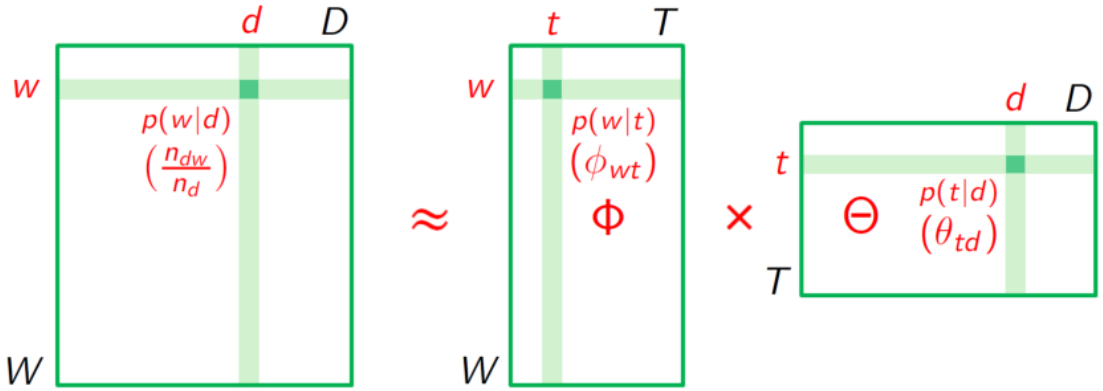
$$p(w|d) = \sum_{t \in T} p(w|t)\, p(t|d)$$



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).

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

1. Низкоранговое стохастическое **матричное разложение:**



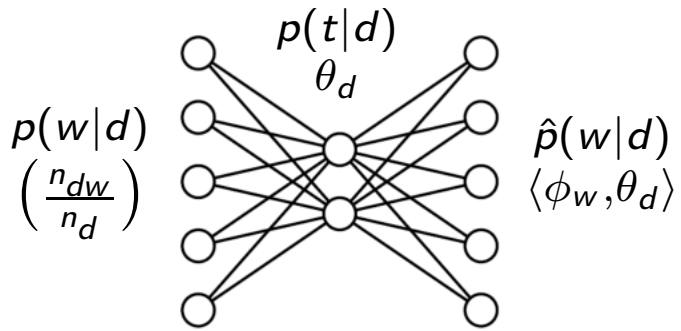
2. **Мягкая кластеризация** документов по кластерам-темам
3. **Автокодировщик** документов в тематические эмбединги:

кодировщик $f_\Phi : \frac{n_{dw}}{n_d} \to \theta_d$

декодировщик $g_\Phi : \theta_d \to \Phi\theta_d$

задача реконструкции текстов:

$$\sum_d \mathrm{KL}\left(\frac{n_{dw}}{n_d} \,\middle\|\, \langle \phi_w, \theta_d \rangle\right) \to \min_{\Phi,\Theta}$$



Свойство интерпретируемости тематических моделей

Тематическая модель формирует тематические векторы:

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

Интерпретируемость тематических векторов (эмбедингов):

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

Пример. Мультиязычная тематическая модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
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.

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

Пример. Мультиязычная тематическая модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
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.

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

Критерий 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 столбцов (такие матрицы Φ, Θ называются *стохастическими*)

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

Модель PLSA (Probabilistic Latent Semantic Analysis)

Максимизация log-правдоподобия для стохастических матриц:

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

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

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

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

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)

Модель LDA (Latent Dirichlet Allocation)

Максимизация log-правдоподобия + байесовская регуляризация с априорными распределениями Дирихле на столбцы Φ, Θ:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

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

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

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

Байесовская и классическая регуляризация

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (громоздкий, приближённый) ради точечной оценки Ω:

$$\text{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\text{Prior}(\Omega|\gamma)$$
$$\Omega := \arg\max_{\Omega} \text{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg\max_{\Omega} \big(\ln p(X|\Omega) + \ln \text{Prior}(\Omega|\gamma)\big)$$

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg\max_{\Omega} \Big(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Big)$$

Аддитивная Регуляризация Тематических Моделей (ARTM)

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

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

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

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

Доказательство (по лемме о максимизации на ед.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}:

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

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

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

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

Мультимодальная ARTM

W_m — словарь термов m-й модальности, $m \in M$

Максимизация суммы log-правдоподобий с регуляризацией:

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

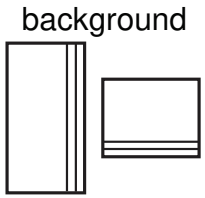
EM-алгоритм: метод простой итерации для системы уравнений

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

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}} \left(\sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}} \left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases}$$

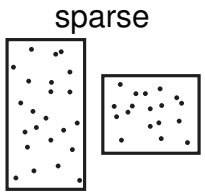
K.Vorontsov, O.Frei, M.Apishev et al. Non-Bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.

Регуляризаторы для улучшения интерпретируемости тем



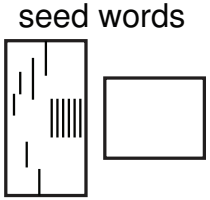
Сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$

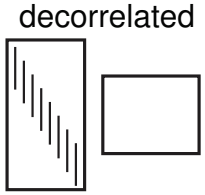


Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \phi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$

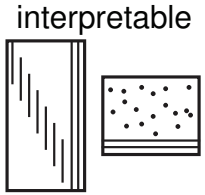


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

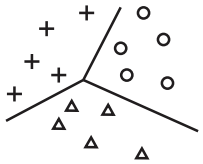
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

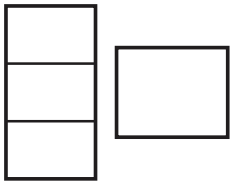
Регуляризаторы для мультимодальных тематических моделей

supervised



Модальности меток классов или категорий для задач классификации и категоризации текстов.

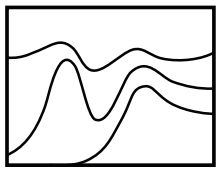
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\phi_{wt}$$

temporal



Темпоральные модели с модальностью времени i:

$$R(\Phi) = -\tau \sum_{i\in I} \sum_{t\in T} \left|\phi_{it} - \phi_{i-1,t}\right|.$$

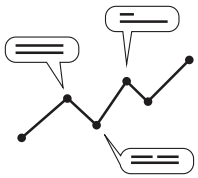
geospatial

Модальность геолокаций g с близостью $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}}\right)^2$$

Регуляризаторы для учёта дополнительной информации

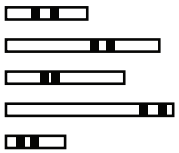
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2$$

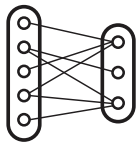
biterm



Связи сочетаемости слов (n_{uv} — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$

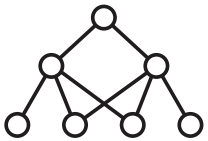
relational



Связи или ссылки между документами:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc}$$

hierarchy



Связи родительских тем t с дочерними подтемами s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}$$

Транзакционные данные

Выборка может содержать не только пары (d, w), но также тройки, четвёрки, ..., n-ки термов разных модальностей.

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g
- **Данные о пассажирских авиаперелётах:**
  (u, a, b, c) — перелёт клиента u из a в b авиакомпанией c

Задача: по наблюдаемой выборке рёбер гиперграфа найти латентные тематические векторные представления его вершин.

Гиперграфовая транзакционная ARTM

n_{kdx} — частота транзакции (d, x), $x \subset W$ типа k в выборке E_k

Максимизация суммы log-правдоподобий с регуляризацией:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$p_{tdx} = \operatorname*{norm}_{t \in T} \left(\theta_{td} \prod_{v \in x} \phi_{vt} \right)$$

M-шаг:
$$\phi_{vt} = \operatorname*{norm}_{v \in W^m} \left(\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} \left[v \in x \right] n_{kdx} p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right)$$

$$\theta_{td} = \operatorname*{norm}_{t \in T} \left(\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

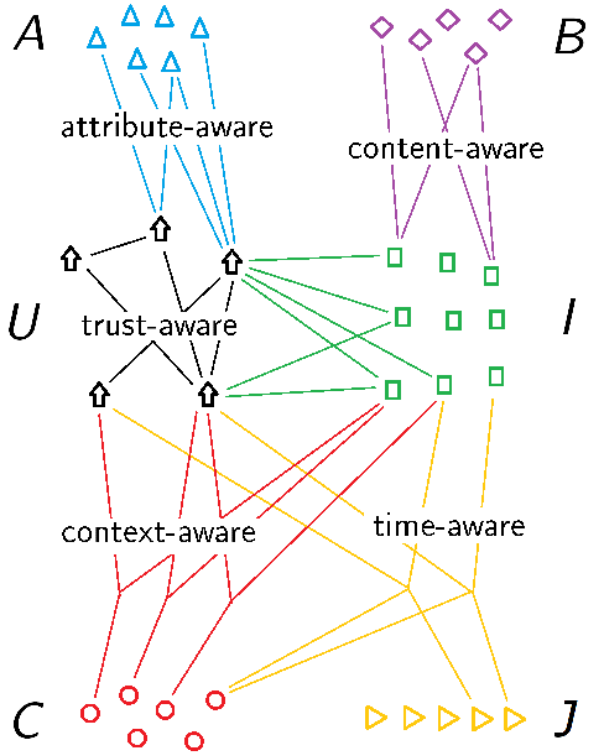
K.Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization // Springer Optimization and Its Applications. 2023

Транзакционные данные в рекомендательных системах

U — конечное множество (словарь) клиентов (users)
I — конечное множество (словарь) объектов (items)
A — словарь атрибутов клиентов (соцдем, регион, хобби...)
B — словарь свойств объектов (слова в текстовых объектах)
C — словарь ситуативных контекстов
J — словарь интервалов времени

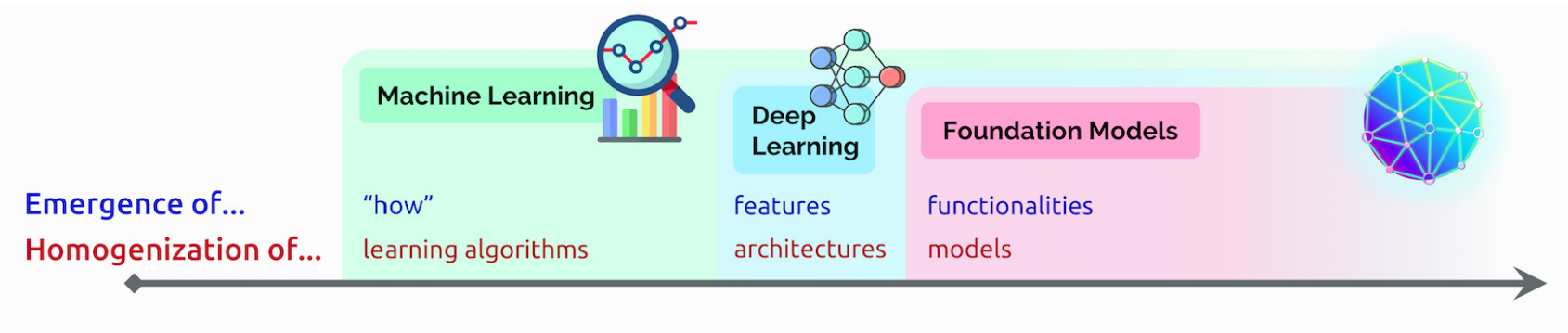
Возможные виды данных:

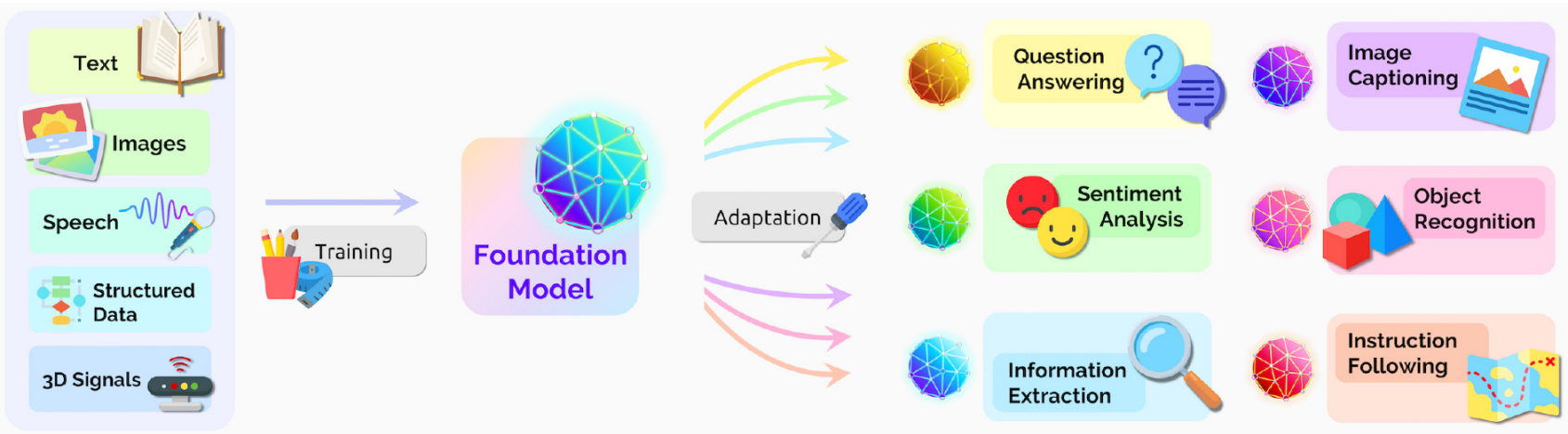
n_{ui} — клиент u выбрал объект i
n_{ua} — клиент u имеет атрибут a
n_{ib} — объект i имеет свойство b
n_{uv} — клиент u доверяет клиенту v
n_{uib} — клиент u отметил i тэгом b
n_{uic} — клиент u выбрал i в контексте c
n_{uicj} — u выбрал i в c в интервале j



Обучаемая векторизация данных — глобальный тренд AI/ML

Foundation Models — гомогенизация векторных моделей





R.Bommasani et al. (Center for Research on Foundation Models, Stanford University) On the opportunities and risks of foundation models // CoRR, 20 August 2021.

Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Тематика термов в документе $p(t|d, w_i)$ — матрица $T \times n_d$:

Регуляризация E-шага как постобработка матриц $p(t|d,w)$

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ & \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Набросок доказательства: три шага

1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

2. Введём вспомогательную функцию от переменных Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

$\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, значит

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_{d\in D} p_{tdw}Q_{tdw};\quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d} p_{tdw}Q_{tdw}.$$

3. Подставляем это в формулы M-шага:

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big);$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big).$$

Любая пост-обработка E-шага эквивалентна регуляризатору $R(\Pi)$

Итак, произвольному гладкому регуляризатору $R(\Pi, \Phi, \Theta)$ однозначно соответствует преобразование $p_{tdw} \to \tilde{p}_{tdw}$.
Верно и обратное:

Теорема. Если на k-й итерации EM-алгоритма для каждого (d, w): $n_{dw} > 0$ в формулах M-шага вместо вектора $(p^k_{tdw})_{t \in T}$ подставить вектор $(\tilde{p}^k_{tdw})_{t \in T}$, удовлетворяющий условию нормировки $\sum_t \tilde{p}^k_{tdw} = 1$, то это эквивалентно добавлению регуляризатора сглаживания–разреживания

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} (\tilde{p}^k_{tdw} - p^k_{tdw}) \ln p_{tdw}.$$

Общий вывод: пост-обработка E-шага позволяет учитывать порядок термов в документе в обход гипотезы «мешка слов».

Однопроходная тематизация текста

Дано: s — фрагмент текста, Φ — готовая тематическая модель
Найти: $p(t|s)$ — тематический вектор фрагмента текста

Проблемы:

- если текст короткий, то $p(t|s)$ может переобучаться
- текст s может быть внутри более объёмного (кон)текста
- $p(t|s)$ не согласуется с $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ отдельных термов

Наводящие соображения:

- первая итерация EM-алгоритма с инициализацией $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in W} n_{dw} p_{tdw}\Big) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}(\phi_{wt}\theta_{td}^0)$$

- формула полной вероятности:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(w|d)\,p(t|w) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}\big(\phi_{wt}p_t\big)$$

EM-алгоритм для ARTM без матрицы Θ

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}};$$

$$p'_{tdw} = p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}};$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right).$$

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КиМ, 2020.

Доказательство (по Лемме о максимизации на ед.симплексах)

Оптимизационная задача M-шага относительно Φ и $\Theta(\Phi)$:

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du}p_{sdu}\ln\big(\phi_{us}\theta_{sd}(\Phi)\big) + R\big(\Phi,\Theta(\Phi)\big) \to \max_{\Phi}$$

Применим Лемму к регуляризованному log-правдоподобию Q:

$$\phi_{wt}\frac{\partial Q}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw}p_{tdw} + \sum_{d,s,u} n_{du}p_{sdu}\frac{\phi_{wt}}{\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} + \phi_{wt}\sum_{d,s}\frac{\partial R}{\partial\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\left(p_{tdw} + \frac{1}{n_{dw}}\sum_{s\in T}\frac{\phi_{wt}}{\theta_{sd}}\underbrace{\left(\sum_{u\in d} n_{du}p_{sdu} + \theta_{sd}\frac{\partial R}{\partial\theta_{sd}}\right)}_{n_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}}\right) + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\underbrace{\left(p_{tdw} + \frac{1}{n_{dw}}\sum_{s\in T}\frac{n_{sd}}{\theta_{sd}}\phi_{wt}\frac{\partial\theta_{sd}}{\partial\phi_{wt}}\right)}_{p'_{tdw}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}.$$

■

EM-алгоритм для линейной тематизации документов

$$\theta_{td}(\Phi) = \sum_{w \in d} p_{wd} \operatorname*{norm}_{t \in T}(\phi_{wt} p_t) \quad \Rightarrow \quad \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} = p_{wd} \phi'_{tw} \left(\delta_{st} - \phi'_{sw}\right)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \operatorname*{norm}_{t \in T}\left(\phi_{wt} n_t\right); \qquad \theta_{td} = \sum_{w \in d} p_{wd} \phi'_{tw};$$

$$p_{tdw} = \operatorname*{norm}_{t \in T}\left(\phi_{wt} \theta_{td}\right); \qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw};$$

$$n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}};$$

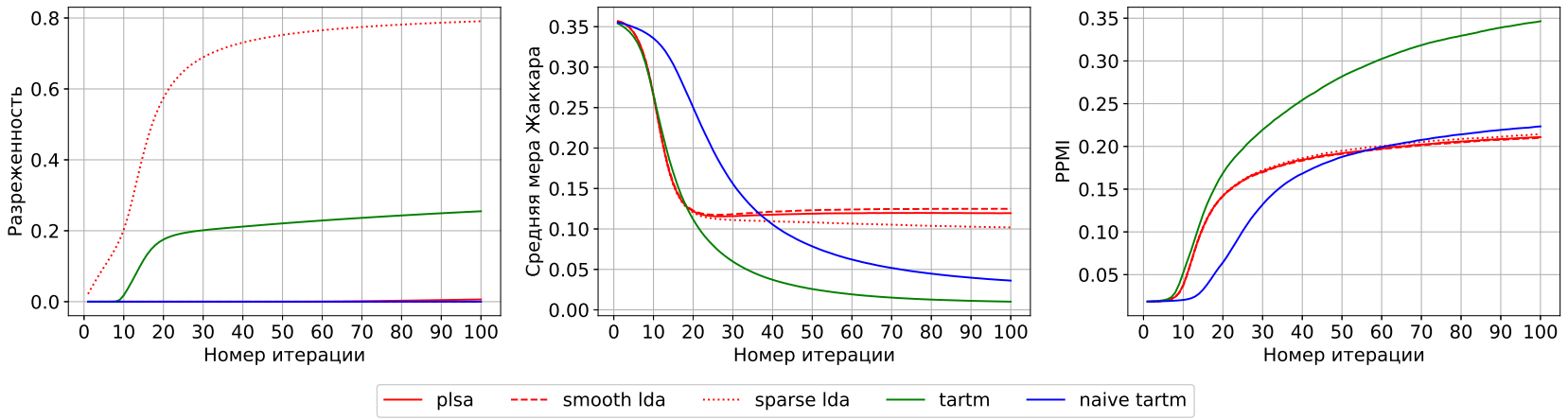
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}} \right);$$

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

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θless ARTM) — модифицированный EM-алгоритм
- naive TARTM — одна итерация обычного EM-алгоритма



- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность и когерентность тем

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

Упрощение EM-алгоритма для линейной тематизации

- Нет регуляризации по Θ, следовательно, $\frac{\partial R}{\partial \theta_{td}} = 0$
- Подстановка несмещённых оценок $\theta_{td} = \frac{n_{td}}{n_d}$, $\theta_{sd} = \frac{n_{sd}}{n_d}$ в формулу M-шага приводит к упрощению: $p'_{tdw} = p_{tdw}$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt} n_t\right); \qquad \theta_{td} = \sum_{w\in d} p_{wd}\phi'_{tw};$$

$$p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right); \qquad n_t = \sum_{d\in D}\sum_{w\in d} n_{dw} p_{tdw};$$

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right).$$

Это обычный EM-алгоритм, только с однопроходным E-шагом!
ОГО! И ТАК МОЖНО БЫЛО?!

Линейная тематизация: от документа к локальным контекстам

Тематизация документа $d = (w_1, \ldots, w_{n_d})$ за один проход:

$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i) = \frac{1}{n_d} \sum_{i=1}^{n_d} \phi'_{tw_i}$$

Тематизация *локального контекста* $C_i = (\ldots, w_i, \ldots)$ терма w_i:

$$\theta_{ti}(\Phi) \equiv p(t|i) = \frac{1}{|C_i|} \sum_{u \in C_i} p(t|u) = \frac{1}{|C_i|} \sum_{u \in C_i} \phi'_{tu}$$

Тематизация локального контекста с распределением весов:

$$\theta_{ti}(\Phi) \equiv p(t|i) = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i), \quad \sum_{u \in C_i} \alpha(u|i) = 1, \quad \alpha(u|i) \geqslant 0$$

Локализованная тематическая модель:

$$p(w|d,i) = \sum_{t \in T} p(w|t)\, p(t|i) = \sum_{t \in T} \phi_{wt} \sum_{u \in C_i} \phi'_{tu} \alpha(u|i)$$

EM-алгоритм с локализованным E-шагом

$w_1, \ldots, w_n$ — сквозная нумерация термов во всей коллекции
C_i — локальный контекст (окружение) терма w_i
$\alpha(u|i)$ — распределение важности термов $u \in C_i$ для терма w_i

- не нужна гипотеза «мешка слов»
- не нужно разбиение на документы

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt} n_t\right); \qquad \theta_{ti} = \sum_{u\in C_i} \phi'_{tu}\alpha(u|i);$$

$$p_{ti} \equiv p(t|C_i, w_i) = \underset{t\in T}{\text{norm}}\left(\phi_{w_i t}\theta_{ti}\right); \qquad n_t = \sum_{i=1}^{n} p_{ti};$$

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{i=1}^{n}[w_i = w]p_{ti} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right).$$

Быстрое вычисление двунаправленных векторов контекста

Два прохода по тексту — «слева направо» и «справа налево» для вычисления *экспоненциальных скользящих средних* (ЭСС):

$$\vec{p}(t|i) = \gamma_i\, p(t|w_i) + (1-\gamma_i)\, \vec{p}(t|i-1),\ \ i = 1,\ldots,n,\ \ \gamma_1 = 1$$
$$\overleftarrow{p}(t|i) = \gamma_i\, p(t|w_i) + (1-\gamma_i)\, \overleftarrow{p}(t|i+1),\ \ i = n,\ldots,1,\ \ \gamma_n = 1$$

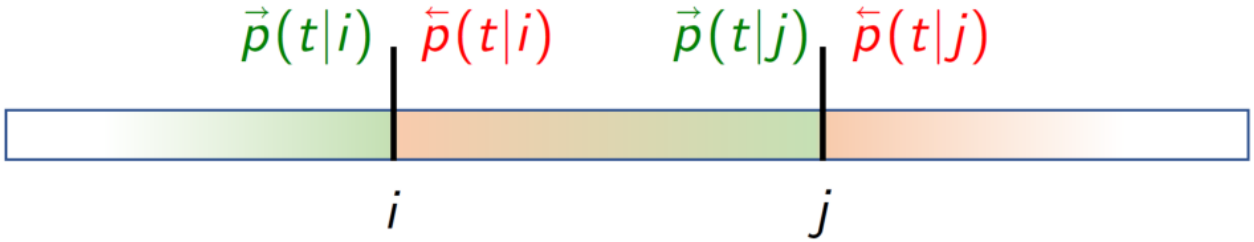
где γ_i — коэффициент сглаживания в позиции i

Основное свойство: если $\gamma_i = \gamma$, то $\alpha(w_k|i) = \gamma(1-\gamma)^{|i-k|}$

Несколько соображений, как распоряжаться выбором γ_i:

- $\gamma_i \approx \frac{1}{h}$, где h — ширина окна, размер контекста
- $\gamma_i = 1$, если надо забыть контекст, сменить документ
- $\gamma_i = 0$, если надо проигнорировать терм
- γ_i можно умножать на оценку важности терма

Использование двунаправленных векторов контекста



Двунаправленные тематические векторы определяют:

- $\vec{p}(t|i)$ — тематику левого контекста терма w_i
- $\overleftarrow{p}(t|i)$ — тематику правого контекста терма w_i
- $\frac{1}{2}\big(\vec{p}(t|i) + \overleftarrow{p}(t|i)\big)$ — тематику двустороннего контекста w_i
- $p(t|i\ldots j) = \frac{1}{2}\big(\overleftarrow{p}(t|i) + \vec{p}(t|j)\big)$ — тематику сегмента $[i\ldots j]$
- тематическую однородность сегмента $[i\ldots j]$:
  насколько распределения $\overleftarrow{p}(t|i)$ и $\vec{p}(t|j)$ схожи
- позиции i границ между сегментами:
  насколько распределения $\vec{p}(t|i)$ и $\overleftarrow{p}(t|i)$ не схожи
- короткие и длинные контексты при различных γ_i

Модель внимания Query–Key–Value

q — *вектор-запрос* для трансформации в вектор контекста z
$(k_1, \ldots, k_n)$ — *векторы-ключи*, чтобы сравнивать с q
$(v_1, \ldots, v_n)$ — *векторы-значения*, составляющие контекст

Модель внимания — это выпуклая комбинация векторов v_u:

$$z = \sum_u v_u \operatorname*{SoftMax}_u \langle k_u, q \rangle,$$

где $\langle k_u, q \rangle$ — *оценка релевантности* ключа k_u запросу q

Модель внутреннего внимания (самовнимания, self-attention):

$$z_i = \sum_u V x_u \operatorname*{SoftMax}_u \langle K x_u, Q x_i \rangle$$

трансформирует последовательность векторов $(x_1, \ldots, x_n)$
в выходную последовательность векторов контекста $(z_1, \ldots, z_n)$

Vaswani et al. Attention is all you need. 2017.
Dichao Hu. An Introductory Survey on Attention Mechanisms in NLP Problems. 2018.

Сравнение локализованного E-шага с моделью self-attention

Тематический вектор локального контекста на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \underset{t\in T}{\text{norm}}\big(\phi_{w_i t}\theta_{ti}\big) = \underset{t\in T}{\text{norm}}\Big(\sum_{u\in C_i} \phi'_{tu}\phi_{w_i t}\alpha(u|i)\Big)$$

Вектор контекста (эмбединг) на выходе модели внимания:

$$z_i = \sum_{u\in C_i} Vx_u\,\alpha(u|i) = \sum_{u\in C_i} Vx_u\,\underset{u\in C_i}{\text{SoftMax}}\langle Qx_i, Kx_u\rangle.$$

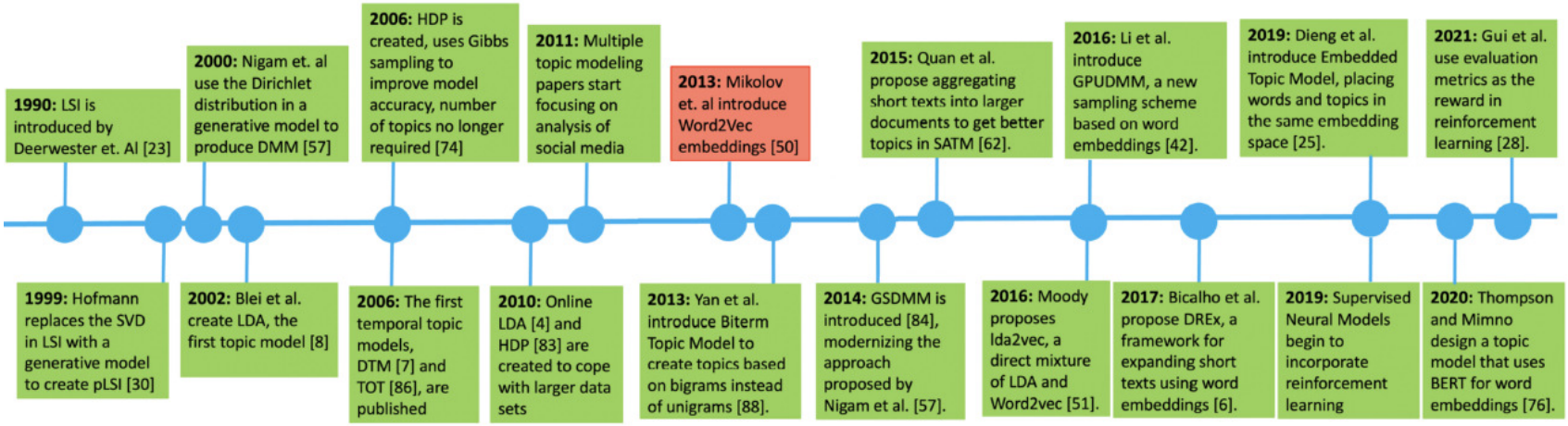
Сходство:

- вектор терма w_i трансформируется в вектор его контекста
- путём усреднения векторов ϕ'_u из контекста терма w_i,
- наиболее (семантически) схожих с вектором терма w_i.

Отличия:

- адамарово умножение вектора ϕ'_u на вектор-фильтр ϕ_{w_i};
- нет обучаемых параметров Q, K, V как у модели внимания;
- проецирование итогового вектора на единичный симплекс.

Эволюция тематического моделирования



Neural Topic Models — поток публикаций начиная с 2016

Как «объединить лучшее от двух миров»?

- **Neural:** качество, универсальность, генеративность
- **Topic:** скорость, интерпретируемость, простота

Что объединяет: векторизация, оптимизация, регуляризация, гомогенизация, локализация (контекст и внимание)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Промежуточные итоги и направления исследований

- В вероятностном тематическом моделировании (PTM) отказ от байесовской регуляризации в пользу обычной (классической, не-байесовской) сильно упрощает теорию
- Теперь PTM — это теория одной леммы

- **Открытая проблема:** построение интерпретируемых моделей внимания (что-то сделать неотрицательным?)
- **Открытая проблема:** объединение PTM и DeepNN не поверхностное (модели, архитектуры, чёрные ящики) а концептуальное (математика итерационных процессов)

Vorontsov K. V. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization. 2023 (принято к публикации)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022

He Zhao, Dinh Phung , Viet Huynh , Yuan Jin , Lan Du , Wray Buntine. Topic Modelling Meets Deep Neural Networks: A Survey. 2021

Открытые проблемы или чем заняться на досуге

- Упростить четвёртое условие в теореме сходимости (оно представляется избыточным)
- Оценить скорость сходимости в «основной лемме»
- Находить применения этой технике (реализована pyTorch), сравнивать со state-of-the-art на различных задачах
- Разработка интерпретируемых моделей внимания (что-то сделать неотрицательным?)
- Объединение тематических моделей и глубоких сетей не поверхностное (модели, архитектуры, чёрные ящики) а концептуальное (математика итерационных процессов)

Vorontsov K. V. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization. 2023 (принято к публикации)

Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. 2023? (принято к публикации в издательстве УРСС)