Вероятностные тематические модели
Лекция 6.
Классика тематических моделей: PLSA, LDA и EM-алгоритм

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)»

МФТИ – ФИЦ ИУ РАН • 24 октября 2024

Содержание

Напоминание. Задача тематического моделирования

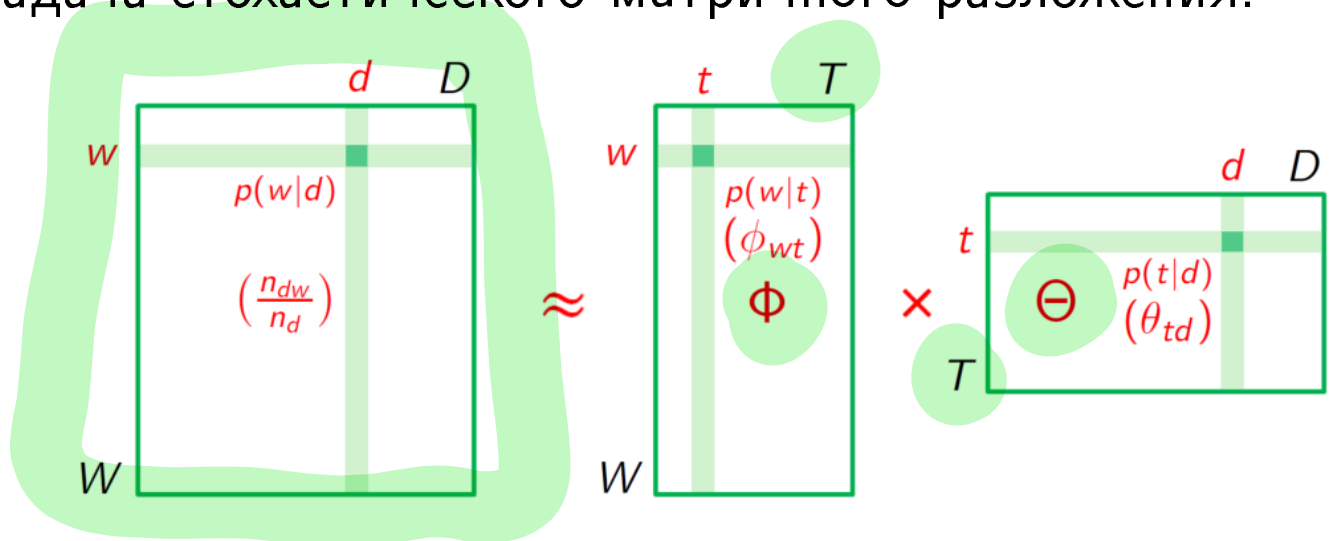
Дано: коллекция текстовых документов

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

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

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

Это задача стохастического матричного разложения:



Напоминание. PLSA (Probabilistic Latent Semantic Analysis)

Критерий — максимум логарифмированного правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Недостатки PLSA (и необходимость его регуляризации)

1. Большая размерность пространства параметров
2. Якобы из-за этого сильное переобучение
3. Якобы невозможность моделирования новых документов
4. Нет управления *разреженностью* Φ и Θ, т.к.
   (в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в финале $\phi_{wt} = 0$),
   (в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в финале $\theta_{td} = 0$)
5. Неединственность и неустойчивость решения:
   если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ — тоже решение
6. Темы не всегда интерпретируемы
7. Нет выделения нетематических (фоновых) слов
8. Не ясно, как учитывать дополнительную информацию

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

Гипотеза об априорных распределениях Дирихле

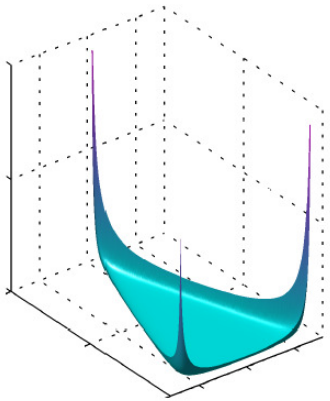
Гипотеза: вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_w > 0;$$

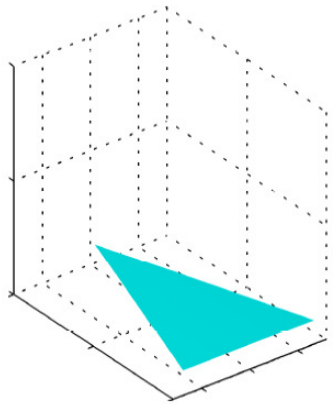
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

Пример:

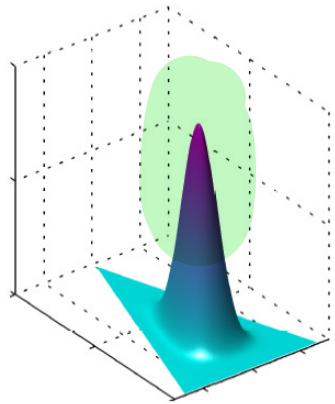
$\mathrm{Dir}(\theta|\alpha)$,
$|T| = 3$,
$\theta, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$

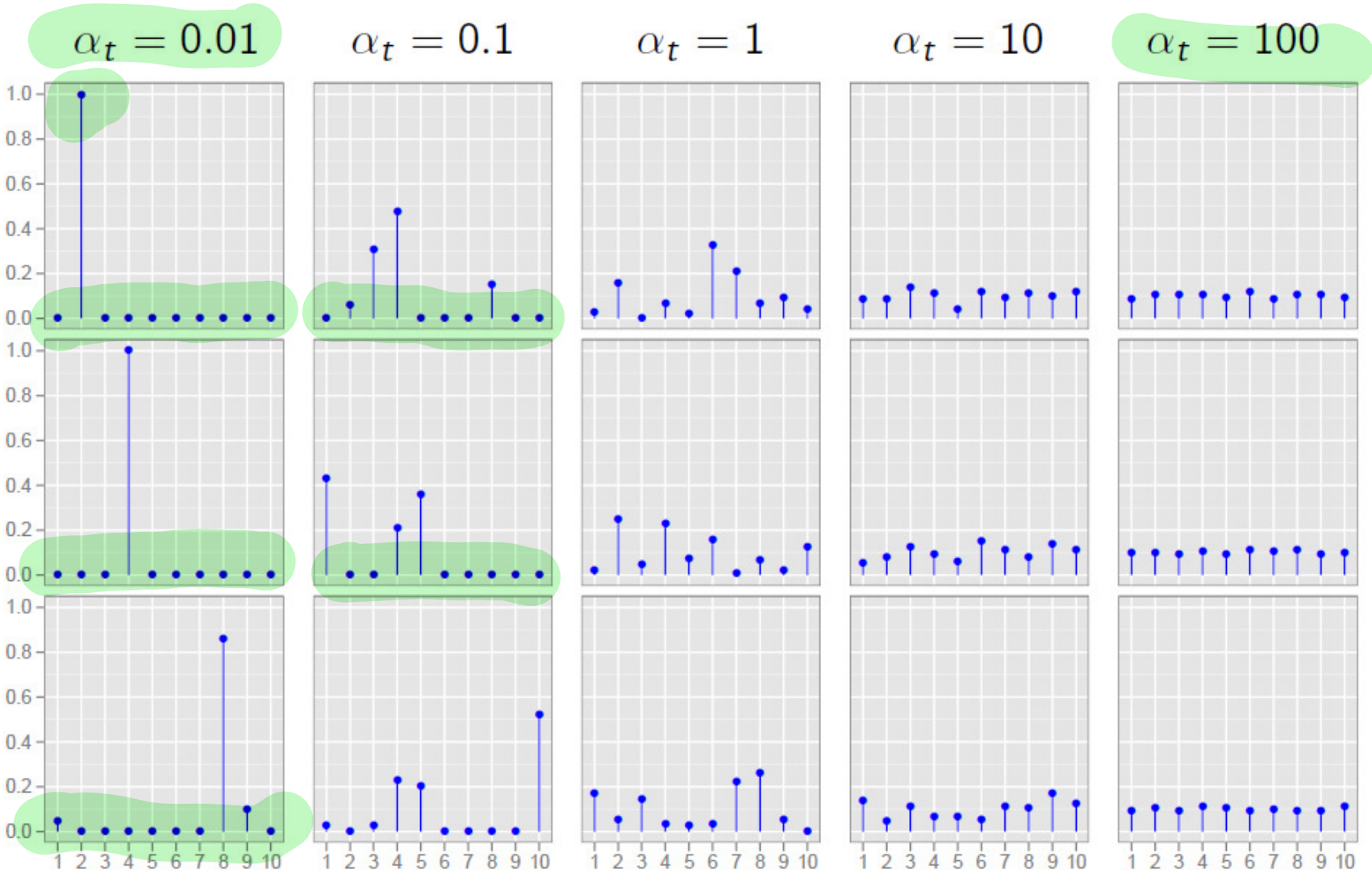


$\alpha_1 = \alpha_2 = \alpha_3 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = 10$

Пример. Выборки из трёх 10-мерных векторов $\theta \sim \mathrm{Dir}(\theta|\alpha)$



Вероятностная модель порождения текста

Тематическая модель LDA (Latent Dirichlet Allocation):

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, \quad \phi_t \sim \mathrm{Dir}(\phi|\beta), \quad \theta_d \sim \mathrm{Dir}(\theta|\alpha).$$

Процесс порождения документов $d = \{w_1 \dots w_{n_d}\}$ коллекции D:

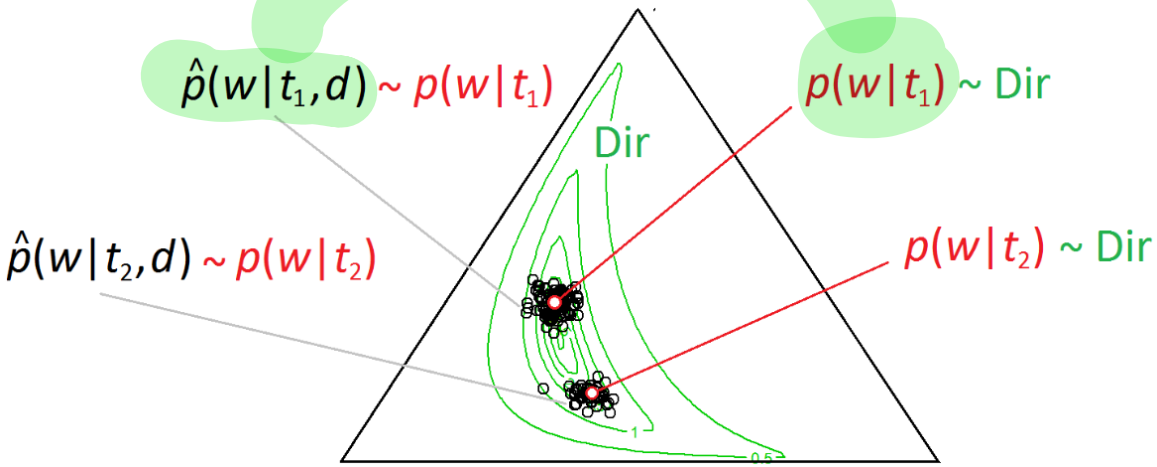
Вход: векторы гиперпараметров β, α;
Выход: коллекция документов;

выбрать вектор ϕ_t из $\mathrm{Dir}(\phi|\beta)$ для каждой темы $t \in T$;
выбрать вектор θ_d из $\mathrm{Dir}(\theta|\alpha)$ для каждого документа $d \in D$;
для всех документов $d \in D$
  для всех позиций термов $i = 1, \dots, n_d$ в документе d
    выбрать тему t_i из $p(t|d) \equiv \theta_{td}$;
    выбрать терм w_i из $p(w|t_i) \equiv \phi_{wt_i}$;

Почему именно распределение Дирихле?

- оно способно порождать разреженные векторы;
- имеет параметры, управляющие степенью разреженности;
- описывает кластерные структуры на симплексе (см. рис.);
- является сопряжённым с мультиномиальным распределением, что сильно упрощает байесовский вывод (в след. лекции).

Распределение $\mathrm{Dir}(\phi|\alpha)$ порождает векторы тем $\phi_t = p(w|t)$, которые порождают мультиномиальные распределения $\hat{p}(w|t,d)$.



Формула Байеса для апостериорного распределения

Введём более общие обозначения X, Ω, γ:

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*

$\Omega = (\Phi, \Theta)$ — параметры порождающей модели $p(X|\Omega)$

$\gamma = (\beta, \alpha)$ — гиперпараметры *априорного распределения* $p(\Omega|\gamma)$

Задача: зная X, моделируя $p(X|\Omega)$, найти Ω.

Формула Байеса даёт *апостериорное распределение* $p(\Omega|X,\gamma)$,
где символ $\propto$ означает «равно с точностью до нормировки»:

$$p(\Omega|X,\gamma) = \frac{p(\Omega,X|\gamma)}{p(X|\gamma)} \propto p(\Omega,X|\gamma) \propto p(X|\Omega)\,p(\Omega|\gamma)$$

Далее есть два пути:

- *Максимизация правдоподобия:* $\Omega = \arg\max_\Omega \ln p(\Omega|X,\gamma)$
- *Байесовский вывод:* вычисление распределения $p(\Omega|X,\gamma)$

Максимизация апостериорной вероятности для модели LDA

Максимизация *совместного правдоподобия* данных и модели, называется также *Maximum a Posteriori (MAP) estimation*:

$$\ln p(X|\Omega)\,p(\Omega|\gamma) = \ln \prod_{i=1}^{n} p(d_i, w_i|\Phi,\Theta)\,p(\Phi|\beta)\,p(\Theta|\alpha) =$$

$$= \ln \prod_{d\in D}\prod_{w\in d} p(d,w|\Phi,\Theta)^{n_{dw}} \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha) \to \max_{\Phi,\Theta}$$

Это задача максимизации регуляризованного log-правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{t,w} \ln \phi_{wt}^{\beta_w - 1} + \sum_{d,t} \ln \theta_{td}^{\alpha_t - 1} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

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

Регуляризованный EM-алгоритм для модели LDA в ARTM

Максимизация апостериорной вероятности эквивалентна регуляризатору логарифма априорного распределения:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\ln \text{ правдоподобия}} + \underbrace{\sum_{t,w}(\beta_w - 1)\ln \phi_{wt} + \sum_{d,t}(\alpha_t - 1)\ln \theta_{td}}_{\text{регуляризатор } R(\Phi,\Theta) = \ln p(\Phi,\Theta|\alpha,\beta)} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \beta_w - 1\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw} p_{tdw} + \alpha_t - 1\right) \end{cases}$$

Общая вероятностная модель со скрытыми переменными

Вернёмся к общим обозначениям X, Ω, γ, добавив Z:

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*

$Z = (t_i)_{i=1}^n$ — *скрытые переменные*

$\Omega = (\Phi, \Theta)$ — параметры порождающей модели $p(X|\Omega)$

$\gamma = (\beta, \alpha)$ — гиперпараметры *априорного распределения* $p(\Omega|\gamma)$

Задача: зная X, моделируя $p(X, Z|\Omega)$, найти Ω.

Апостериорное распределение:

$$p(\Omega|X,\gamma) \propto p(X|\Omega)\, p(\Omega|\gamma) = \sum_Z p(X, Z|\Omega)\, p(\Omega|\gamma)$$

Принцип максимума апостериорной вероятности:

$$\ln p(X|\Omega) + \underbrace{\ln p(\Omega|\gamma)}_{R(\Omega)} \to \max_\Omega$$

$R(\Omega)$ может и не иметь вероятностной интерпретации.

Общий EM-алгоритм для задачи со скрытыми переменными

Теорема. Точка Ω локального максимума регуляризованного *маргинализованного правдоподобия* (Marginal log-Likelihood)

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega \qquad \text{(RML)}$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

$$\text{E-шаг:} \quad q(Z) = p(Z|X, \Omega);$$

$$\text{M-шаг:} \quad \sum_Z q(Z) \ln p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega .$$

Это общий вид EM-алгоритма, используемый не только в тематическом моделировании.

A.P.Dempster, N.M.Laird, D.B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. 1977.

Доказательство теоремы

Необходимые условия локального экстремума:

$$\frac{\partial}{\partial\Omega}\left(\ln\sum_z p(X,Z|\Omega)+R(\Omega)\right)=\frac{1}{p(X|\Omega)}\sum_z\frac{\partial p(X,Z|\Omega)}{\partial\Omega}+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

По формуле условной вероятности $p(X|\Omega)=\frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$, подставляем:

$$\sum_z\frac{p(Z|X,\Omega)}{p(X,Z|\Omega)}\frac{\partial p(X,Z|\Omega)}{\partial\Omega}+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

$$\sum_z\underbrace{p(Z|X,\Omega)}_{q(Z)}\frac{\partial}{\partial\Omega}\ln p(X,Z|\Omega)+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

Это необходимые условия локального экстремума задачи M-шага.
Выделение уравнений E-шага $q(Z)=p(Z|X,\Omega)$ относительно дополнительных переменных $q(Z)$ приводит к EM-алгоритму.

■

Ещё более общий EM-алгоритм и его сходимость

Теорема. Значение маргинализованного правдоподобия

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega \qquad \text{(RML)}$$

не убывает на каждом шаге итерационного процесса

$$\text{E-шаг:} \quad \mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X,\Omega)\big) \;\to\; \min_q;$$

$$\text{M-шаг:} \quad \sum_Z q(Z) \ln p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega.$$

$q(Z) = p(Z|X,\Omega)$ является точным решением задачи E-шага.

Минимизация KL на E-шаге используется в тех случаях, когда $p(Z|X,\Omega)$ не удаётся вычислить в явном виде.

Сходимость *в слабом смысле*: глобальный max не гарантируется.

Доказательство теоремы

По формуле условной вероятности $p(X|\Omega) = \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$.

Для произвольного распределения $q(Z)$

$$\ln p(X|\Omega) = \sum_Z q(Z) \ln p(X|\Omega) = \sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)} =$$

$$= \underbrace{\sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{q(Z)}}_{L(q,\Omega)} + \underbrace{\sum_Z q(Z) \ln \frac{q(Z)}{p(Z|X,\Omega)}}_{\mathrm{KL}\left(q(Z)\middle\|p(Z|X,\Omega)\right)\geqslant 0}$$

Максимизируем достижимую нижнюю оценку RML то по q, то по Ω:

E-шаг: $L(q,\Omega) + \cancel{R(\Omega)} \to \max\limits_q \Leftrightarrow \mathrm{KL}\left(q(Z) \,\|\, p(Z|X,\Omega)\right) \to \min\limits_q$

M-шаг: $L(q,\Omega) + R(\Omega) \to \max\limits_\Omega \Leftrightarrow \sum\limits_Z q(Z) \ln p(X,Z|\Omega) + R(\Omega) \to \max\limits_\Omega$

На каждом шаге значение функционала может только возрастать.

■

Регуляризованный EM-алгоритм для тематической модели

Напоминание: $X = (d_i, w_i)_{i=1}^n$, $Z = (t_i)_{i=1}^n$, $\Omega = (\Phi, \Theta)$.

Лемма. Точка (Φ, Θ) локального максимума RML (регуляризованного маргинализованного log-правдоподобия)

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi, \Theta)$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

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

M-шаг: $\sum_{d,w,t} n_{dw} p(t|d,w) \ln\left(\phi_{wt}\theta_{td}\right) + R(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}$

Доказательство леммы

E-шаг: в силу независимости элементов выборки

$$q(Z) = p(Z|X,\Omega) = \prod_{i=1}^{n} p(t_i|d_i,w_i) = \prod_{i=1}^{n} \operatorname*{norm}_{t_i \in T}\left(\phi_{w_i t_i}\theta_{t_i d_i}\right)$$

M-шаг: подставим $q(Z)$ и $p(X,Z|\Omega)$ в общую формулу M-шага:

$$\sum_{Z\in T^n} q(Z)\ln p(X,Z|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{t_1\in T}\cdots\sum_{t_n\in T}\prod_{k=1}^{n} p(t_k|d_k,w_k)\sum_{i=1}^{n}\ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n}\sum_{t_1\in T}\cdots\sum_{t_n\in T}\prod_{k=1}^{n} p(t_k|d_k,w_k)\ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n}\sum_{t\in T} p(t|d_i,w_i)\ln p(d_i,w_i,t|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

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

■

Вывод формул M-шага ARTM, теперь из общего EM-алгоритма

Оптимизационная задача M-шага:

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

Применим лемму о максимизации на единичных симплексах:

$$\phi_{wt}=\underset{w\in W}{\text{norm}}\left(\phi_{wt}\frac{\partial f}{\partial\phi_{wt}}\right)=\underset{w\in W}{\text{norm}}\left(\underbrace{\sum_{d\in D} n_{dw}p(t|d,w)}_{n_{wt}}+\phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\right)$$

$$\theta_{td}=\underset{t\in T}{\text{norm}}\left(\theta_{td}\frac{\partial f}{\partial\theta_{td}}\right)=\underset{t\in T}{\text{norm}}\left(\underbrace{\sum_{w\in d} n_{dw}p(t|d,w)}_{n_{td}}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)$$

Таким образом, снова получили формулы ARTM

Резюме по теории

Модель латентного размещения Дирихле LDA

- LDA проще вводить через KL-дивергенцию, как регуляризатор сглаживания/разреживания,
- тогда заодно снимаются ограничения $\beta_w > 0$, $\alpha_t > 0$
- Распределение Дирихле играет особую роль в байесовских методах тематического моделирования
- ARTM — это более простая альтернатива байесовским методам, но в статьях по тематическому моделированию они преобладают, поэтому в них надо уметь разбираться
- Мы рассмотрим байесовские методы в следующей лекции

Общий вариант EM-алгоритма

- снабжён возможностью регуляризации $R(\Omega)$
- имеет обоснование слабой сходимости

Способны ли PLSA и LDA восстановить истинные темы?

Матрицы Φ_0 и Θ_0 порождаются распределением Дирихле.
Синтетическая коллекция порождается матрицами Φ_0 и Θ_0.
Размеры: $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Цель — сравнить восстановленные распределения $p(i|j)$
с исходными синтетическими распределениями $p_0(i|j)$
по среднему расстоянию Хеллингера:

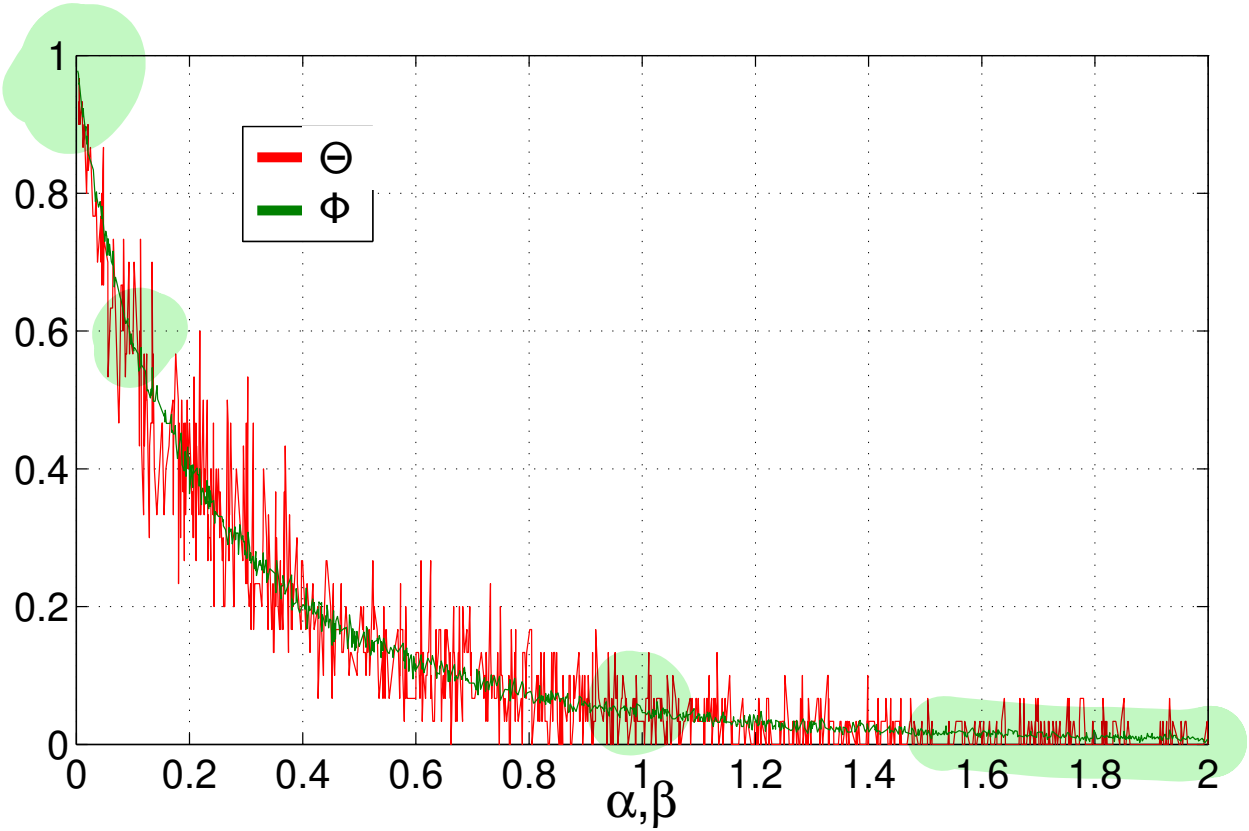
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

как для самих матриц Φ и Θ, так и для их произведения:

$$D_\Phi = H(\Phi, \Phi_0);$$
$$D_\Theta = H(\Theta, \Theta_0);$$
$$D_{\Phi\Theta} = H(\Phi\Theta, \Phi_0\Theta_0).$$

Разреженность векторов, порождаемых распределением Dir

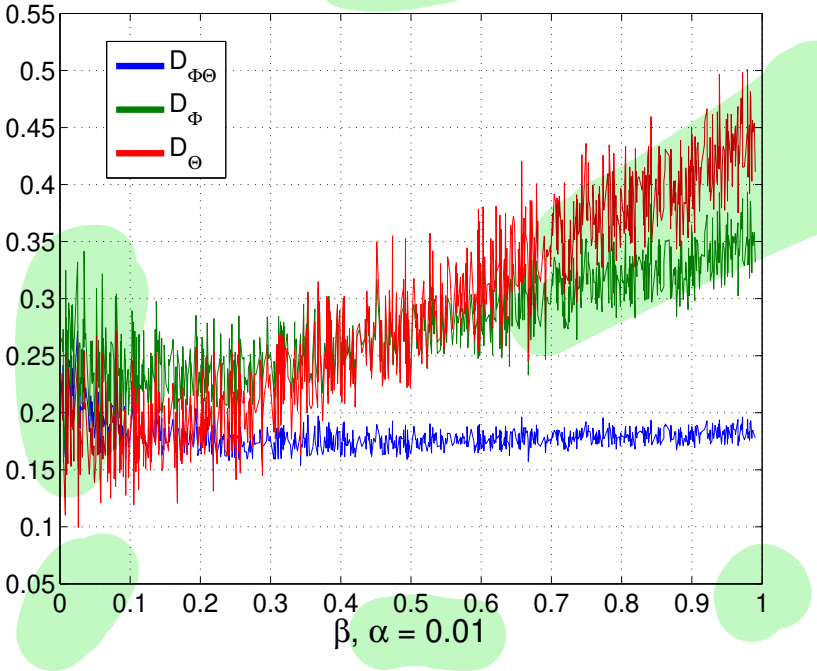
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \text{Dir}(\alpha)$ и $\phi_t^0 \sim \text{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



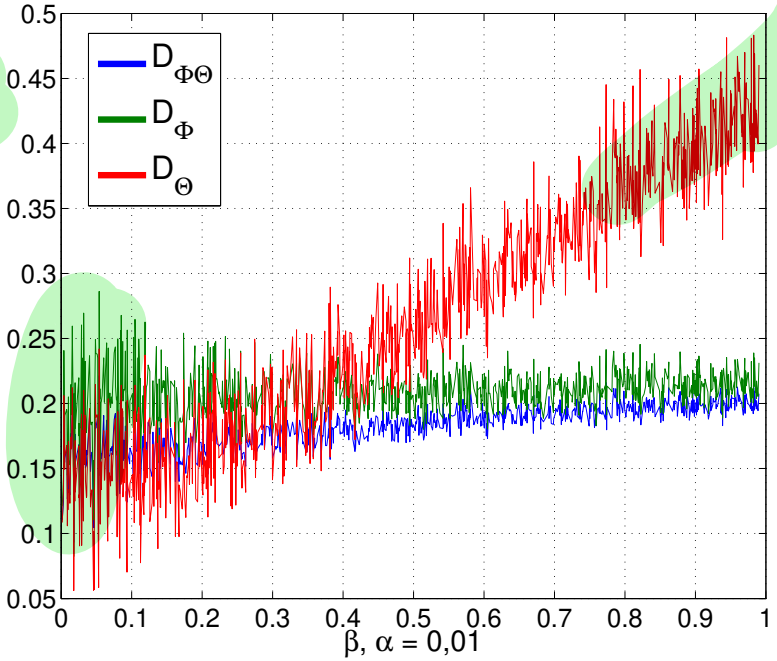
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ_0 при фиксированном $\alpha = 0.01$

PLSA



LDA

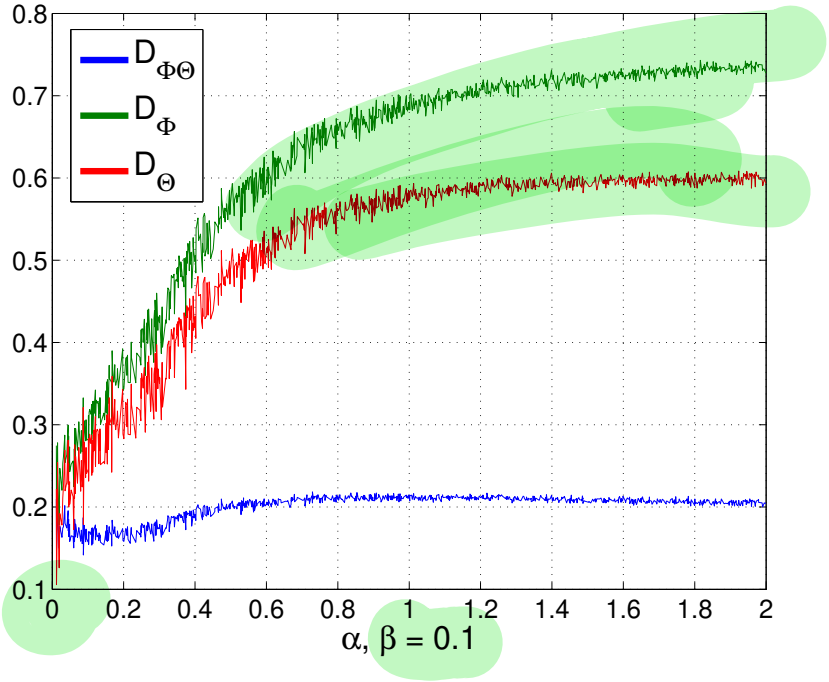


Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

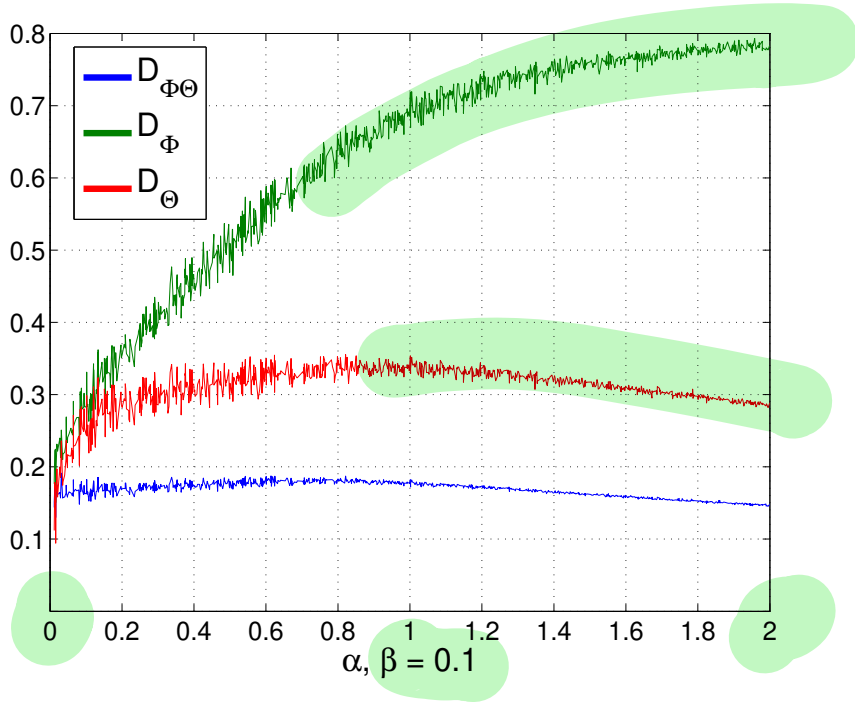
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0 при фиксированном $\beta = 0.1$

PLSA



LDA



Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

Второй эксперимент — на реальных данных

Посты ЖЖ: $|D| = 300\,\text{K}$, $|W| = 154\,\text{K}$, $n = 35\,\text{M}$, $|T| = 120$.
LDA: симметричное распределение Дирихле, $\beta = 0.1$, $\alpha = 0.5$.

Цель эксперимента — оценить различность тем, получаемых в нескольких запусках алгоритма LDA Gibbs Sampling.

Проблема «проклятия размерности»:
длинные хвосты мешают сравнивать распределения.

Доля существенных слов в темах (word ratio):

$$WR = \frac{1}{|W|}\frac{1}{|T|}\sum_{w\in W}\sum_{t\in T}\left[\phi_{wt} > \frac{1}{|W|}\right] \quad \text{(в эксперименте } \sim 3.5\%\text{)}$$

Доля существенных тем в документах (document ratio):

$$DR = \frac{1}{|D|}\frac{1}{|T|}\sum_{d\in D}\sum_{t\in T}\left[\theta_{td} > \frac{1}{|T|}\right] \quad \text{(в эксперименте } \sim 11.5\%\text{)}$$

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Методика эксперимента

Оставлены слова w, имеющие $\phi_{wt} > \frac{1}{|W|}$ хотя бы в одной теме

Сокращение словаря (vocabulary reduction): $154\,\text{K} \to 8\,\text{K}$.

Дивергенция Кульбака–Лейблера между темами t и s:

$$\text{KL}(t,s) = \sum_{w \in W} p(w|t) \ln \frac{p(w|t)}{p(w|s)}$$

Нормированная KL-близость пар тем t и s:

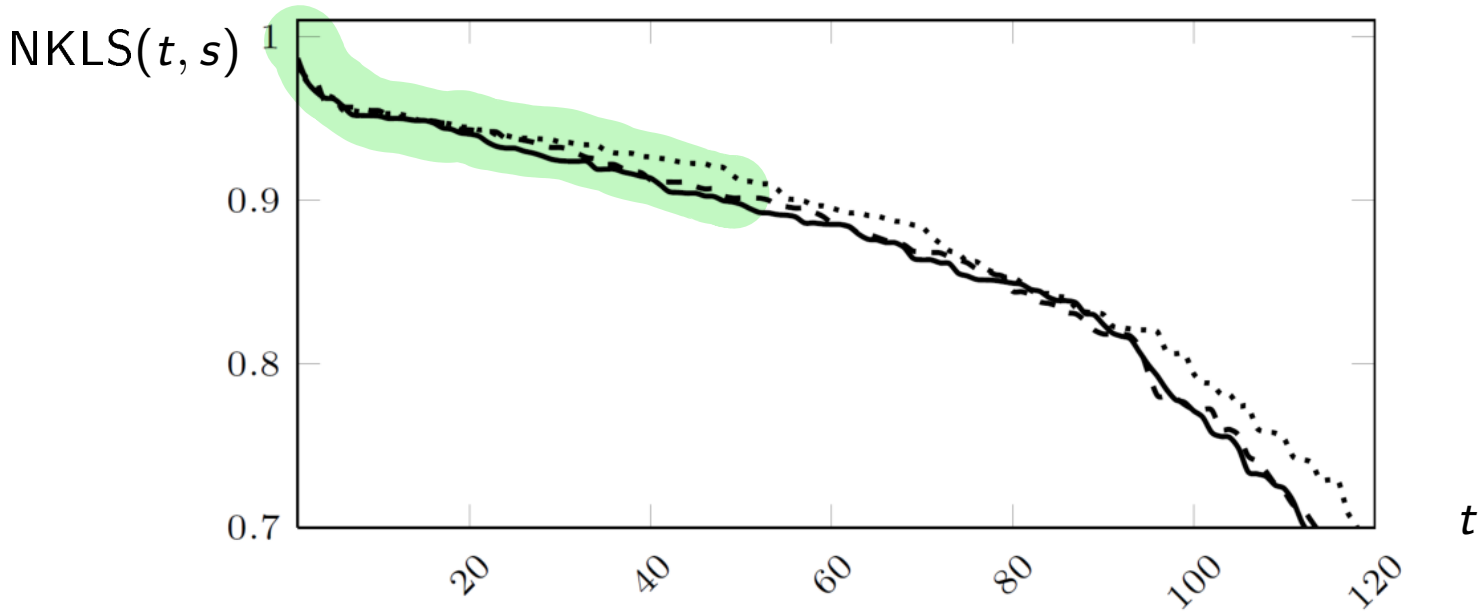
$$\text{NKLS}(t,s) = \left(1 - \frac{\text{KL}(t,s)}{\max\limits_{t',s'} \text{KL}(t',s')}\right)$$

При $\text{NKLS}(t,s) > 0.9$ в темах совпадают 30–50 топовых слов, и эксперты-социологи признают такие темы одинаковыми.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Неустойчивость LDA в разных запусках

Результат эксперимента: нормированная KL-близость NKLS между темой t и ближайшей к ней s в другом запуске.



1. Менее 50% тем воспроизводятся от запуска к запуску.
2. Плохо воспроизводятся как мусорные темы, так и хорошие.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Выводы из экспериментов

- Матрицы Φ, Θ устойчиво восстанавливаются только при сильной разреженности Φ_0, Θ_0 (более 90% нулей)
- Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных Φ_0, Θ_0
- В разных запусках со случайной инициализацией или сэмплированием строятся существенно различные темы
- PLSA не переобучается, а лишь хуже моделирует малые вероятности редких слов, которые не интересны.
- Распределение Дирихле — слишком слабый регуляризатор

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models. Machine Learning. Springer, 2015.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Робастная тематическая модель

Гипотеза: каждое слово в документе (d, w) является
— либо тематическим, связанным с какой-то темой t,
— либо специфичным для данного документа (шум),
— либо общеупотребительным (фон).

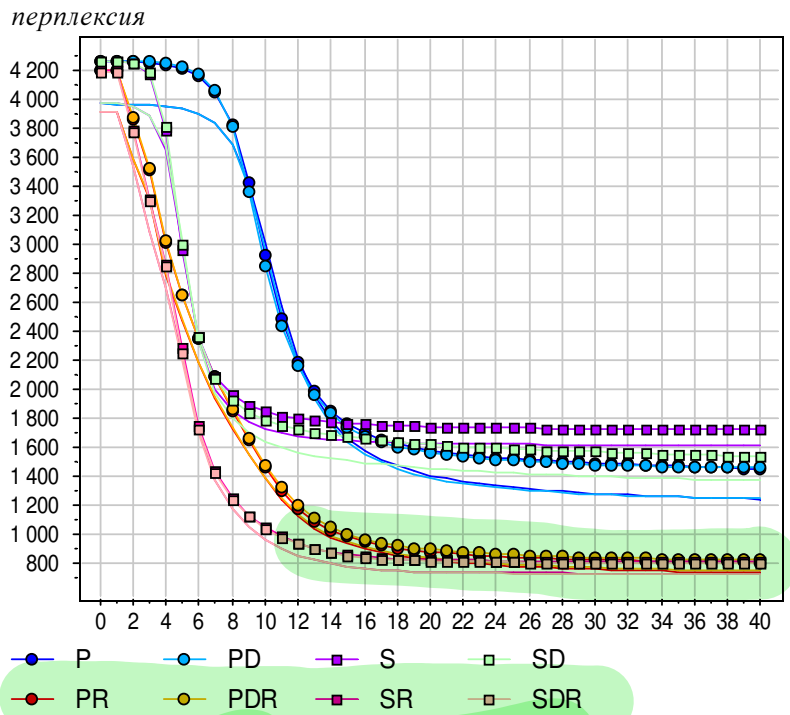
Модель вероятностной смеси тематической, шумовой и фоновой компонент SWB (Special Words with Background):

$$p(w|d) = \gamma\pi_{dw} + \varepsilon\pi_w + (1-\gamma-\varepsilon)\sum_{t\in T}\phi_{wt}\theta_{td}$$

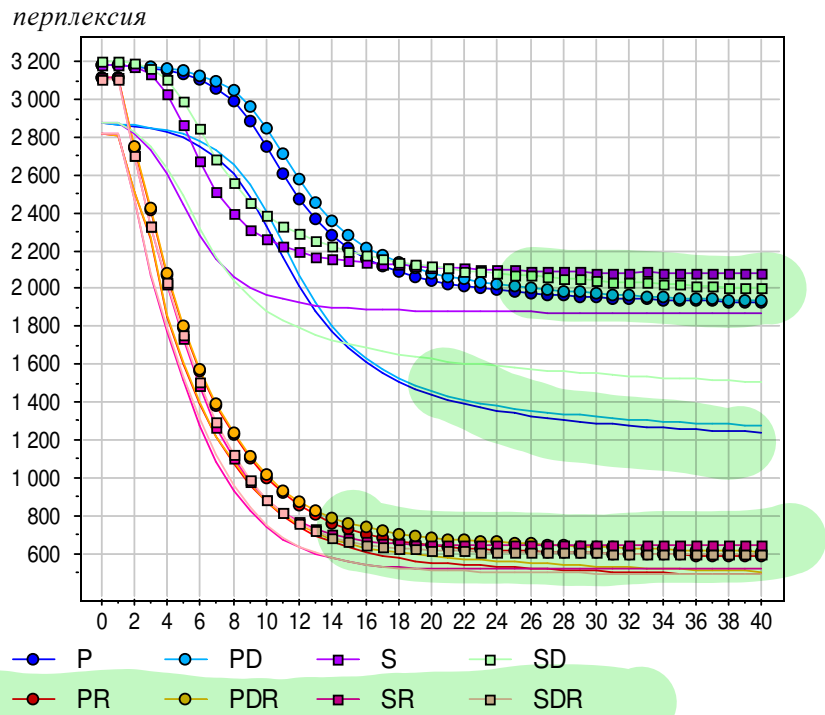
$\pi_{dw} \equiv p_{\text{ш}}(w|d)$ — *шумовая* компонента, γ — параметр;
$\pi_w \equiv p_{\text{ф}}(w)$ — *фоновая* компонента, ε — параметр.

Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model. NIPS, 2006
Потапенко А. А., Воронцов К. В. Модификации EM-алгоритма для вероятностного тематического моделирования. JMLDA, 2013

Эксперименты с робастными PLSA и LDA



Коллекция RuDis



Коллекция NIPS

Обозначения: P – PLSA, D – LDA ($\alpha_t = 0.5$, $\beta_w = 0.01$)
S – сэмплирование темы из $p(t|d, w)$ для каждого d, w
R – робастность (шум $\gamma = 0.3$, фон $\varepsilon = 0.01$)

A.Potapenko, K.Vorontsov. Robust PLSA performs better than LDA. ECIR-2013.

Выводы

1. Переобучение проявляется только на редких словах
2. LDA точнее моделирует вероятности редких слов
3. Но они как раз наименее интересны для описания тем
4. Робастные PLSA и LDA почти одинаковы по перплексии и почти не переобучаются
5. Робастный PLSA лучше, чем обычный LDA [1]
6. PLSA и LDA почти одинаковы на больших коллекциях [2,3,4]
7. Перплексия — не вполне адекватная мера качества

---

1. *Potapenko A. A., Vorontsov K. V.* Robust PLSA performs better than LDA. 2013
2. *Tomonari Masada, Senya Kiyasu, Sueharu Miyahara.* Comparing LDA with PLSI as a dimensionality reduction method in document clustering. 2008
3. *Yue Lu, Qiaozhu Mei, ChengXiang Zhai.* Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. 2011
4. *Yonghui Wu, Yuxin Ding, Xiaolong Wang, Jun Xu.* A comparative study of topic models for topic clustering of Chinese web news. 2010

Вместо резюме: мифы про LDA

- LDA существенно меньше переобучается, чем PLSA
- LDA строит разреженные тематические модели
- LDA имеет меньше параметров по сравнению с PLSA
- LDA == тематическое моделирование

На самом деле,

- LDA и PLSA почти не отличаются на больших данных
- LDA не максимизирует разреженность моделей
- LDA имеет больше параметров по сравнению с PLSA
- LDA не решает проблему неединственности разложения
- LDA не имеет убедительных лингвистических обоснований
- LDA — самый примитивный способ регуляризации

Тем не менее («парадокс тематического моделирования»)

- LDA — самая известная и часто используемая модель

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.