Вероятностные тематические модели
Лекция 2. Аддитивная регуляризация тематических моделей (ARTM)

К. В. Воронцов
vokov@forecsys.ru

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

ВМК МГУ • весна 2017

Содержание

1. **Теория ARTM**
   - EM-алгоритм для ARTM
   - Мультимодальные тематические модели
   - Оффлайновый и онлайновый EM-алгоритм
2. **Модель латентного размещения Дирихле LDA**
   - Распределение Дирихле
   - Максимизация апостериорной вероятности
   - Мифы про LDA
3. **Эксперименты**
   - Неустойчивость на синтетических данных
   - Неустойчивость на реальных данных
   - Задания

Напоминание. Задача тематического моделирования

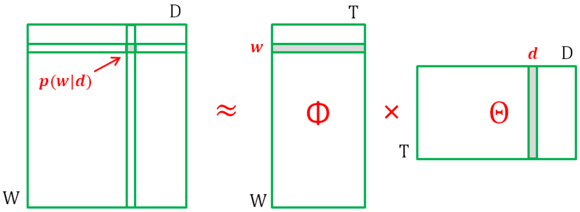
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Вероятностная тематическая модель:

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

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

Это задача стохастического матричного разложения:



Напоминание. PLSA (Probabilistic Latent Semantic Analysis)

Задача: найти максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

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

$$\begin{array}{ll} \text{E-шаг:} & \\ & \\ \text{M-шаг:} & \\ & \end{array} \begin{cases} n_{dwt} = n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}; \\ \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d\in D} n_{dwt}; \quad n_t = \sum\limits_{w} n_{wt} \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w\in d} n_{dwt}; \quad n_d = \sum\limits_{t} n_{td} \end{cases}$$

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар (1865–1963)

Задача стохастического матричного разложения *некорректно поставлена*, так как имеется бесконечное множество решений:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для невырожденных $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Регуляризация — стандартный приём, введение новых ограничений или критериев, доопределяющих решение.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl}
\text{E-шаг:} & \left\{\begin{array}{l}
p_{tdw} \equiv p(t|d,w) = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\[2ex]
\phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\Bigr), \quad n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw} \\[2ex]
\theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Bigr), \quad n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw}
\end{array}\right. \\
\text{M-шаг:} &
\end{array}$$

где $\operatorname*{norm}_{t\in T}(x_t) = \dfrac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$ — операция нормирования вектора.

Условия невырожденности решения

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

Тема t невырождена, если хотя бы для одного термина $w \in W$

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

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

Документ d невырожден, если хотя бы для одной темы $t \in T$

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

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

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для ϕ_{wt} (для θ_{td} всё аналогично):

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \mu_{wt}; \quad \mu_{wt} \geqslant 0; \quad \mu_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Если $\lambda_t \leqslant 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w.
4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$:

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

5. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w\in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

6. Подставим λ_t из (5) в (4), получим требуемое. ■

Комбинирование регуляризаторов

Максимизация log правдоподобия с k регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

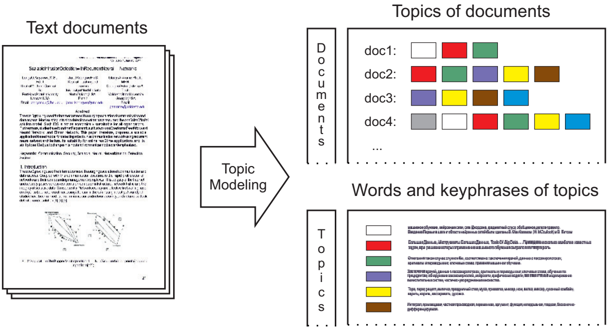
где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl} \text{Е-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\sum_{i=1}^{k}\tau_i \frac{\partial R_i}{\partial \phi_{wt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\sum_{i=1}^{k}\tau_i \frac{\partial R_i}{\partial \theta_{td}}\Big) \end{array} \right. \\ \text{М-шаг:} & \end{array}$$

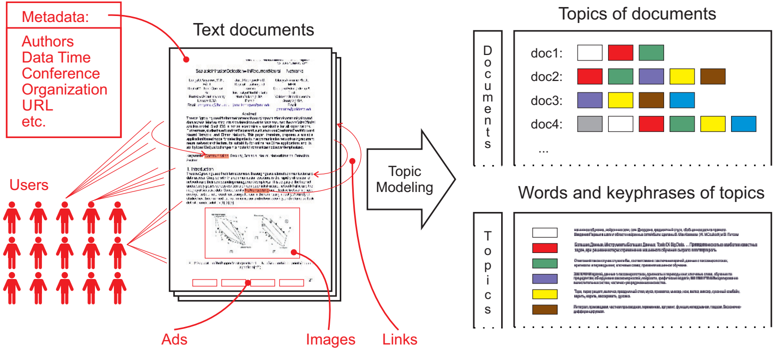
Тематика (тематический вектор, тематический профиль)

Тематика документа: $p(t|d) = \theta_{td}$
Тематика термина: $p(t|w) = p(w|t)p(t)/p(w) = \phi_{wt} n_t / n_w$



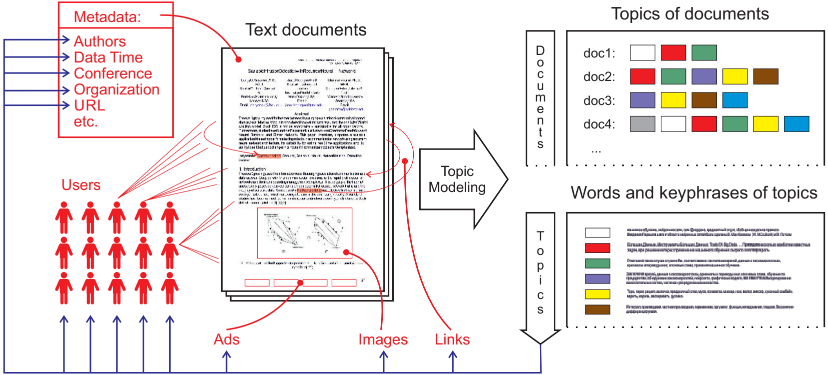
Мультимодальная тематическая модель

Документ — универсальный контейнер не только терминов, но и токенов других *модальностей*: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{ссылка})$, $p(t|\text{баннер})$, $p(t|\text{изображение})$, $p(t|\text{пользователь})$



Мультимодальная тематическая модель

Документы могут содержать *токены* разных модальностей.
Каждая *модальность* $m \in M$ описывается своим словарём W^m.
Каждая тема имеет своё распределение $\phi_{wt} = p(w|t)$ на W^m.



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$
$W = W^1 \sqcup \dots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{ll} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \underset{t \in T}{\mathrm{norm}} \bigl(\phi_{wt} \theta_{td}\bigr) \\ \phi_{wt} = \underset{w \in W^m}{\mathrm{norm}} \Bigl(\tau_m \sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \Bigr) \\ \theta_{td} = \underset{t \in T}{\mathrm{norm}} \Bigl(\sum_{m \in M} \tau_m \sum_{w \in W^m} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \Bigr) \end{array} \right. \\ \text{M-шаг:} & \end{array}$$

Оффлайновый EM-алгоритм для ARTM

Идея: E-шаг встраивается внутрь M-шага

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

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \dots, i_{\max}$
- $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
- **для всех** $d \in D$, $m \in M$, $w \in d \cap W^m$
  - $p_{tdw} = \mathop{\mathrm{norm}}_{t \in T} \bigl(\phi_{wt}\theta_{td}\bigr)$ для всех $t \in T$;
  - $n_{wt}, n_{td} \mathrel{+}= \tau_m n_{dw} p_{tdw}$ для всех $t \in T$;
- $\phi_{wt} := \mathop{\mathrm{norm}}_{w \in W^m} \Bigl(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\Bigr)$ для всех $m \in M$, $w \in W^m$, $t \in T$;
- $\theta_{td} := \mathop{\mathrm{norm}}_{t \in T} \Bigl(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\Bigr)$ для всех $d \in D$, $t \in T$;

Онлайновый EM-алгоритм для ARTM

Идея: коллекция D разбивается на пакеты D_b, $b = 1, \dots, B$

Вход: коллекция D, $\delta \equiv \texttt{decay\_weight}$, $\alpha \equiv \texttt{apply\_weight}$;
Выход: матрица Φ;

инициализировать ϕ_{wt} для всех $w \in W$, $t \in T$;
$n_{wt} := 0$, $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;
для всех пакетов D_b, $b = 1, \dots, B$
- $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \textbf{ProcessBatch}(D_b, \Phi)$;
- **если** пора обновить матрицу Φ **то**
  - $n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
  - $\phi_{wt} := \operatorname*{norm}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $m \in M$, $w \in W^m$, $t \in T$;
  - $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

ProcessBatch: обработка пакета в онлайновом EM-алгоритме

Идея: обработать пакет, не меняя матрицу Φ

Вход: пакет D_b, матрица $\Phi = (\phi_{wt})$;
Выход: матрица $(\tilde{n}_{wt})$;

$\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех $d \in D_b$
- инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
- **повторять**
  - $p_{tdw} := \underset{t \in T}{\mathrm{norm}}\bigl(\phi_{wt}\theta_{td}\bigr)$ для всех $w \in d,\ t \in T$;
  - пост-обработка матрицы $(p_{tdw})_{T \times n_d}$ при необходимости;
  - $\theta_{td} := \underset{t \in T}{\mathrm{norm}}\left(\sum\limits_{m \in M} \tau_m \sum\limits_{w \in W^m} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
- **пока** θ_d не сойдётся;
- $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_m n_{dw} p_{tdw}$ для всех $m \in M,\ w \in W^m,\ t \in T$;

Сравнение оффлайнового и онлайнового алгоритмов

Оффлайн EM-алгоритм:

1. многократное итерирование по коллекции
2. однократный проход по документу
3. хранение матрицы Θ
4. обновление Φ в конце каждого прохода по коллекции
5. применяется при обработке небольших коллекций

Онлайн EM-алгоритм:

1. однократный проход по коллекции
2. многократное итерирование по каждому документу
3. нет необходимости хранить матрицу Θ
4. обновление Φ через заданное число пакетов
5. применяется при потоковой обработке больших коллекций

Гипотеза об априорных распределениях Дирихле

Вероятностная тематическая модель: $p(w|d) = \sum\limits_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$

1. Пусть $\theta_d = (\theta_{td})_{t \in T} \in \mathbb{R}^{|T|}$ — случайные векторы из распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:

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

2. Пусть $\phi_t = (\phi_{wt})_{w \in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:

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

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

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:

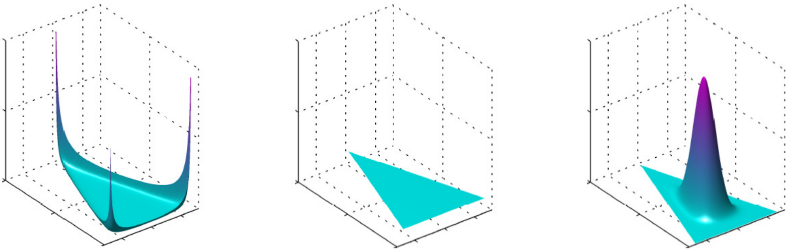
Вход: векторы гиперпараметров β, α;
Выход: коллекция документов;

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 слов $i = 1, \dots, n_d$ в документе d
    выбрать тему t_i из $p(t|d) \equiv \theta_{td}$;
    выбрать слово w_i из $p(w|t_i) \equiv \phi_{wt}$;

Почему именно распределение Дирихле?

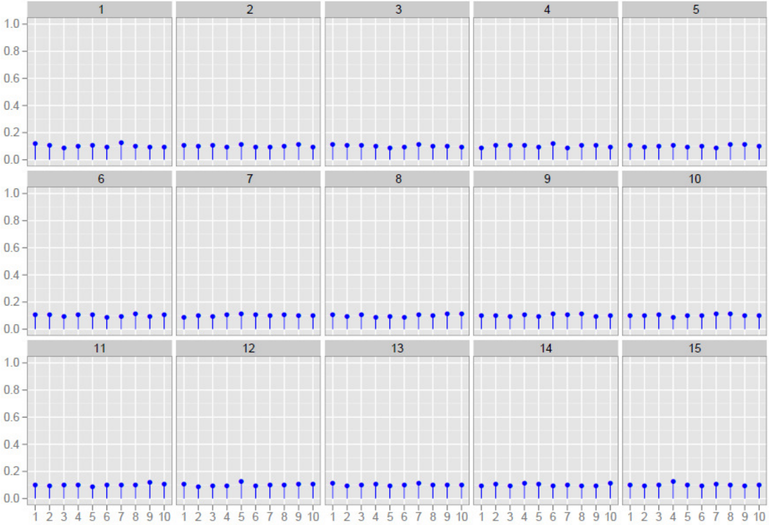
– Может порождать сглаженные или разреженные векторы
– Имеет параметры, управляющие степенью разреженности
– Неплохо описывает кластерные структуры на симплексе
– Является сопряжённым к мультиномиальному распределению

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

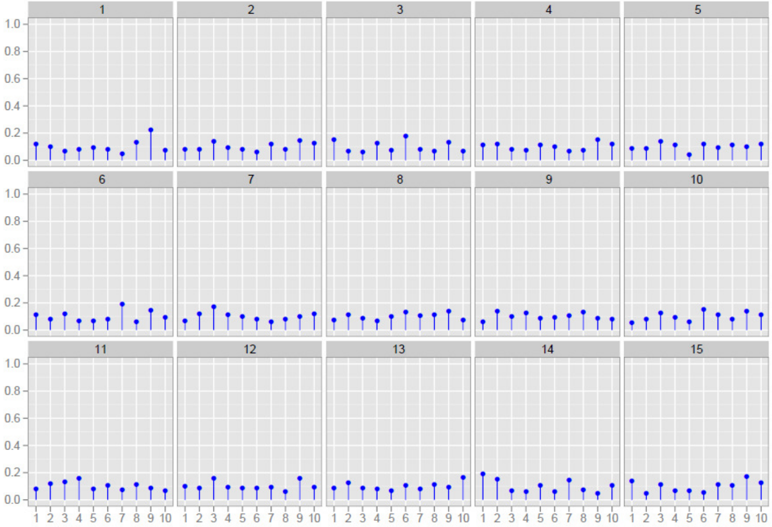


$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ $\qquad$ $\alpha_1 = \alpha_2 = \alpha_3 = 1$ $\qquad$ $\alpha_1 = \alpha_2 = \alpha_3 = 10$

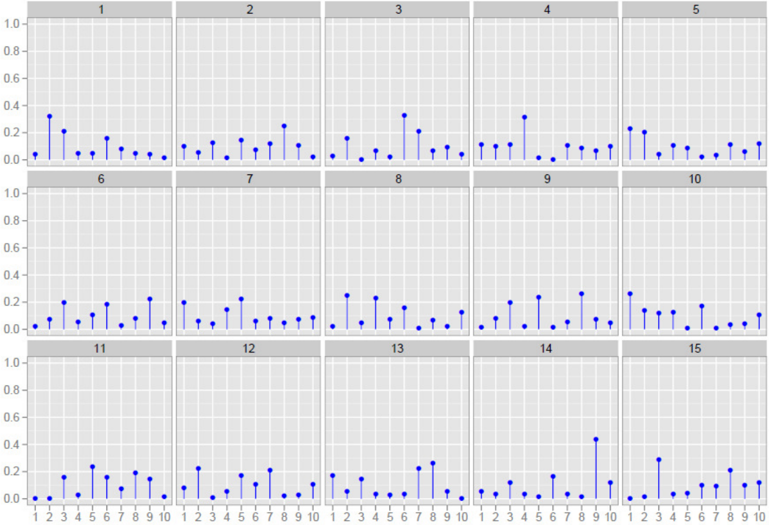
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 100$, 10 тем, 15 документов



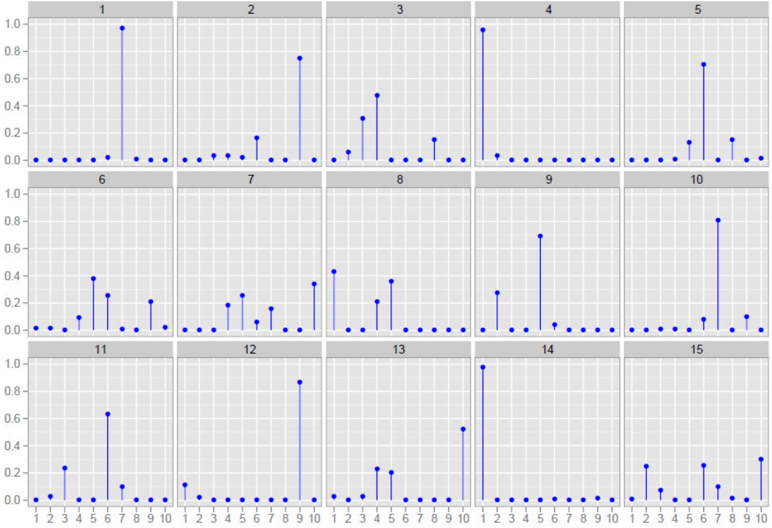
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 10$, 10 тем, 15 документов



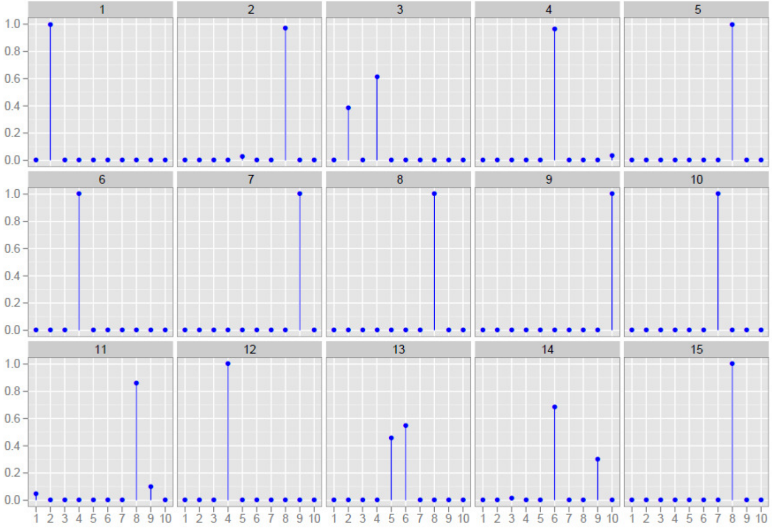
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 1$, 10 тем, 15 документов



Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 0.1$, 10 тем, 15 документов



Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 0.01$, 10 тем, 15 документов



Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

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

Принцип MAP (maximum a posteriori probability)

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

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

Максимум апостериорной вероятности:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td}}_{\ln\text{ правдоподобия } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}}_{\text{критерий регуляризации } R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T} \big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w\in W} \Big(\sum\limits_{d\in D} n_{dw} p_{tdw} + \beta_w - 1 \Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T} \Big(\sum\limits_{w\in d} n_{dw} p_{tdw} + \alpha_t - 1 \Big) \end{array} \right. \\ \text{M-шаг:} & \end{array}$$

Мифы про LDA

- LDA существенно меньше переобучается, чем PLSA
- LDA строит более разреженную тематическую модель
- LDA имеет меньше параметров по сравнению с PLSA
- LDA == тематическое моделирование

На самом деле,

- LDA отличается от PLSA только на малых n_{wt}, n_{td}
- LDA имеет больше параметров по сравнению с PLSA
- LDA не имеет убедительных лингвистических обоснований
- LDA не решает проблему неединственности разложения
- LDA — слабый и неинтересный регуляризатор

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.

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:

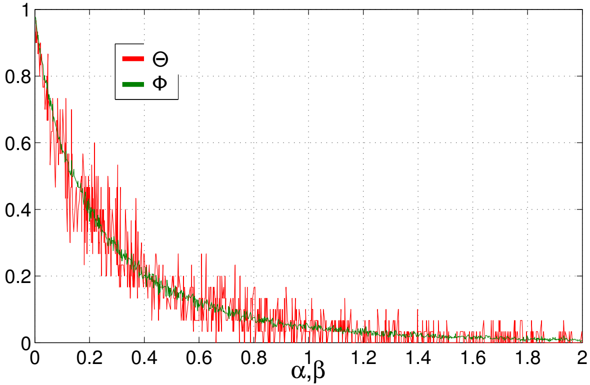
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

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

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

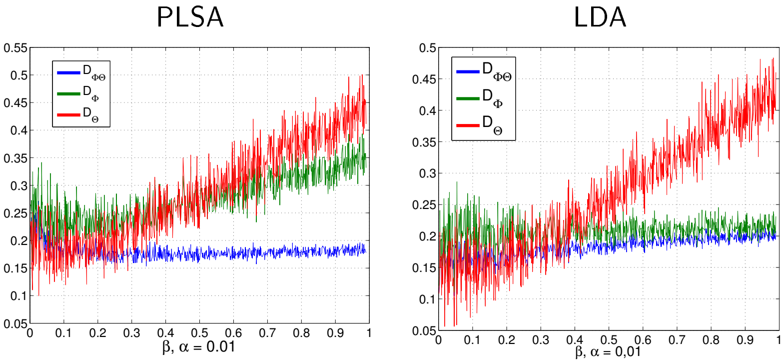
Разреженность векторов, порождаемых распределением Dir

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



Неустойчивость восстановления матриц Φ и Θ

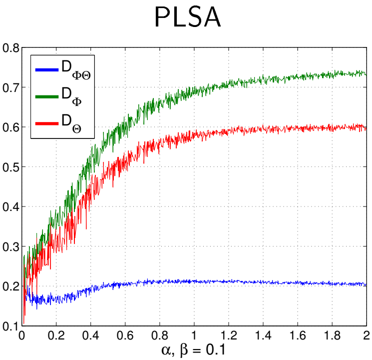
Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ_0 при фиксированном $\alpha = 0.01$

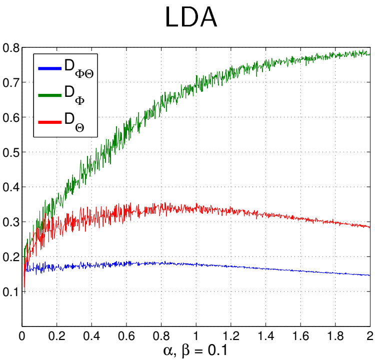


Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования // Магистерская диссертация. МФТИ, 2013.

Неустойчивость восстановления матриц Φ и Θ

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





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

Цель эксперимента

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 терминов в темах (word ratio):

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

Доля существенных тем в документах (document ratio):

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

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 K → 8 K.

Дивергенция Кульбака–Лейблера между темами t и s:

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

Нормированная KL-близость пар тем t и s:

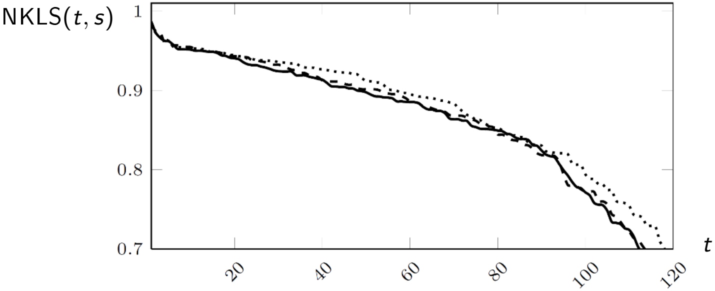
$$\mathrm{NKLS}(t,s) = \left(1 - \frac{\mathrm{KL}(t,s)}{\max\limits_{t',s'} \mathrm{KL}(t',s')}\right)$$

При $\mathrm{NKLS}(t,s) > 0.9$ 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-GS 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

1. Матрицы Φ, Θ устойчиво восстанавливаются только при сильной разреженности Φ_0, Θ_0 (более 90% нулей)
2. Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных Φ_0, Θ_0
3. В разных запусках с использованием случайных начальных приближений или сэмплирования EM-алгоритм находит существенно различающиеся наборы тем

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.

Резюме

- Задача тематического моделирования некорректно поставлена, её решение не единственно и не устойчиво.
- Регуляризация — стандартный приём решения таких задач.
- Подход ARTM позволяет комбинировать регуляризаторы и строить тематические модели с требуемыми свойствами.
- Реализация — в проекте с открытым кодом BigARTM.
- Модель LDA — слишком слабый регуляризатор, не решает проблему неединственности и неустойчивости.
- Модель LDA лучше описывает вероятности редких слов, но для выявления тематики они как раз и не важны.
- Регуляризаторы и модальности — в следующих лекциях.

Прикладные проекты и исследовательские задачи

- Медиапланирование. Поток новостей. Иерархия тем.
- Контакт-центр. Обучение тем. Сегментация. Сценарии.
- Новостной мониторинг. Обучение тем. Поиск проблем.
- Финансовые транзакции. Трёхматричная модель.
- Агрегатор научно-популярного контента в Рунете.
- Разведочный поиск по Хабру. Улучшение модели.
- Исследование устойчивости.
- Исследование онлайнового алгоритма.
- Исследование инициализаций матрицы Φ.
- Адаптивная оптимизация коэффициентов регуляризации.
- Распознавание новых тем в потоке новостей.
- Тематическая сегментация.
- Суммаризация тем.
- Автоматическое именование тем.
- SyntaxNet. Автоматическое выделение терминов.