Вероятностные тематические модели
Лекция 4. Аддитивная регуляризация тематических моделей

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 2016

Содержание

Напоминание. Задача тематического моделирования

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} — сколько раз термин w встретился в документе d

n_d — длина документа d

Найти: модель $p(w|d) = \sum_t \phi_{wt}\theta_{td}$ с параметрами ϕ, θ:

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Критерий: максимум логарифма правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \to \max_{\phi,\theta},$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geqslant 0; \quad \sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_t \theta_{td} = 1$$

Напоминание. Модель LDA (Latent Dirichlet Allocation)

Максимум апостериорной вероятности

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \mathop{\mathrm{norm}}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \mathop{\mathrm{norm}}_{w\in W}\bigl(n_{wt} + \beta_w\bigr), \quad n_{wt} = \sum_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \mathop{\mathrm{norm}}_{t\in T}\bigl(n_{td} + \alpha_t\bigr), \quad n_{td} = \sum_{w\in d} n_{dw} p_{tdw} \end{array} \right. \\ \text{M-шаг:} & \end{array}$$

где $\mathop{\mathrm{norm}}_{t\in T} x_t = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация ln 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

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

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014. Т. 455., № 3. 268–271.

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}, $w \in W$ (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 ln 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

$$\begin{array}{rl} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \underset{t\in T}{\mathrm{norm}}\big(\phi_{wt}\theta_{td}\big) \\ \\ \phi_{wt} = \underset{w\in W}{\mathrm{norm}}\left(\sum\limits_{d\in D} n_{dw} p_{tdw} + \phi_{wt} \sum\limits_{i=1}^{k} \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \\ \theta_{td} = \underset{t\in T}{\mathrm{norm}}\left(\sum\limits_{w\in d} n_{dw} p_{tdw} + \theta_{td} \sum\limits_{i=1}^{k} \tau_i \frac{\partial R_i}{\partial \theta_{td}} \right) \end{array} \right. \\ \text{M-шаг:} & \end{array}$$

Рациональный ЕМ-алгоритм для ARTM

Идея: Е-шаг встраивается внутрь М-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** документов $d \in D$ и всех слов $w \in d$
  - $p_{tdw} = \operatorname{norm}\limits_{t \in T}\bigl(\phi_{wt}\theta_{td}\bigr)$ для всех $t \in T$;
  - $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dw}p_{tdw}$ для всех $t \in T$;
- $\phi_{wt} := \operatorname{norm}\limits_{w \in W}\Bigl(\sum\limits_{d \in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Bigr)$ для всех $w \in W$, $t \in T$;
- $\theta_{td} := \operatorname{norm}\limits_{t \in T}\Bigl(\sum\limits_{w \in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr)$ для всех $d \in D$, $t \in T$;

Онлайновый параллельный EM-алгоритм для ARTM

Вход: коллекция D, разбитая на пакеты D_b, $b = 1, \dots, B$;
коэффициент дисконтирования $\rho \in (0, 1]$;
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, $b = 1, \dots, B$
- $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) +$ **ProcessBatch**(D_b, Φ);
- **если** пора выполнить синхронизацию, **то**
  - $n_{wt} := \rho n_{wt} + \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
  - $\phi_{wt} := \underset{w \in W}{\mathrm{norm}} \Big(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \Big)$ для всех $w \in W$, $t \in T$;
  - $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

Онлайновый параллельный EM-алгоритм для ARTM

ProcessBatch обрабатывает пакет D_b при фиксированной 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**
  - $p_{tdw} := \mathop{\mathrm{norm}}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr)$ для всех $w \in d,\ t \in T$;
  - $\theta_{td} := \mathop{\mathrm{norm}}_{t\in T}\Bigl(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr)$ для всех $t \in T$;
- **пока** θ_d не сойдётся;
- $\tilde{n}_{wt} := \tilde{n}_{wt} + n_{dw}p_{tdw}$ для всех $w \in d,\ t \in T$;

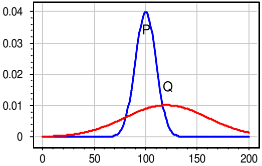
Напоминание. Дивергенция Кульбака–Лейблера

1. $\mathrm{KL}(P\|Q) \geqslant 0$; $\quad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q$;

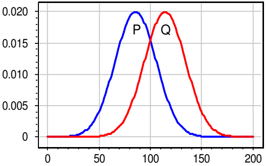
2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_{\alpha} \quad \Leftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_{\alpha}$$

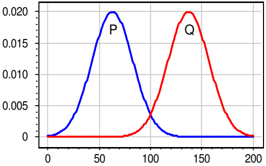
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P вложено в Q:



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 2.97$



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 0.44$



$\mathrm{KL}(P\|Q) = 2.97$
$\mathrm{KL}(Q\|P) = 2.97$

Регуляризатор сглаживания (переосмысление LDA)

Гипотеза сглаженности:
распределения ϕ_{wt} близки к заданному распределению β_w;
распределения θ_{td} близки к заданному распределению α_t.

$$\sum_{t\in T} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

Максимизируем сумму регуляризаторов:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T} \sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D} \sum_{t\in T} \alpha_t \ln \theta_{td} \to \max.$$

Подставляем, получаем формулы M-шага LDA:

$$\phi_{wt} = \mathop{\mathrm{norm}}_{w\in W}\bigl(n_{wt} + \beta_0\beta_w\bigr), \qquad \theta_{td} = \mathop{\mathrm{norm}}_{t\in T}\bigl(n_{td} + \alpha_0\alpha_t\bigr).$$

Этого вы не найдёте в *D.Blei, A.Ng, M.Jordan.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — Vol. 3. — Pp. 993–1022.

Регуляризатор разреживания (обобщение LDA)

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей;
распределения ϕ_{wt} далеки от заданного распределения β_w;
распределения θ_{td} далеки от заданного распределения α_t.

$$\sum_{t\in T} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \max_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \max_{\Theta}.$$

Максимизируем сумму регуляризаторов:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t\in T} \sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D} \sum_{t\in T} \alpha_t \ln \theta_{td} \to \max.$$

Подставляем, получаем «анти-LDA»:

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\bigl(n_{wt} - \beta_0\beta_w\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t\in T}\bigl(n_{td} - \alpha_0\alpha_t\bigr).$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010.

Объединение сглаживания и разреживания

Общий вид регуляризаторов сглаживания и разреживания:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \to \max,$$

где $\beta_0 > 0$, $\alpha_0 > 0$ — коэффициенты регуляризации,
β_{wt}, α_{td} — параметры, задаваемые пользователем:

- $\beta_{wt} > 0$, $\alpha_{td} > 0$ — сглаживание
- $\beta_{wt} < 0$, $\alpha_{td} < 0$ — разреживание

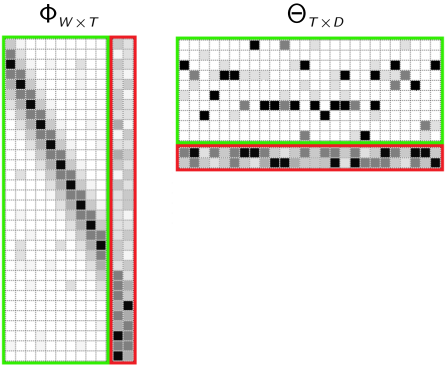
Частичное обучение (semi-supervised learning) темы t:

- $\beta_{wt} = [w \in W_t]$ — по словарю терминов $W_t \subset W$
- $\alpha_{td} = [d \in D_t]$ — по списку документов $D_t \subset D$

Разделение тем на предметные и фоновые

Предметные темы S содержат термины предметной области, $p(w|t)$, $p(t|d)$, $t \in S$ — разреженные, существенно различные

Фоновые темы B содержат слова общей лексики, $p(w|t)$, $p(t|d)$, $t \in B$ — существенно отличные от нуля



Регуляризатор декоррелирования тем

Цель — выделить *лексическое ядро* каждой темы, набор терминов, отличающий её от других тем.

Минимизируем ковариации между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max.$$

Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} = \operatorname*{norm}_{w \in W} \Bigl(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \Bigr).$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Регуляризатор для сокращения числа тем

Гипотеза: если в теме слишком мало слов, то она не нужна.

Разреживаем распределение $p(t) = \sum_d p(d)\theta_{td}$, максимизируя KL-дивергенцию между $p(t)$ и равномерным распределением:

$$R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \to \max.$$

Подставляем, получаем:

$$\theta_{td} = \operatorname*{norm}_{t \in T} \Bigl(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \Bigr)_+.$$

Эффект: строки матрицы Θ могут целиком обнуляться для тем t, собравших мало слов по коллекции, $n_t = \sum_d \sum_w n_{dwt}$.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // SLDS 2015.

Некоторые критерии качества тематической модели

Построение ВТМ — многокритериальная оптимизация.
Поэтому критериев для контроля качества модели тоже много.

- Перплексия контрольной коллекции: $\mathcal{P} = \exp(-\frac{1}{n}\mathscr{L})$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы: [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w\colon p(t|w) > 0.25\}$
  - чистота темы: $\sum\limits_{w\in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum\limits_{w\in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем: $|T|$
  - доля фоновых слов: $\frac{1}{n} \sum\limits_{d\in D} \sum\limits_{w\in d} \sum\limits_{t\in B} p(t|d,w)$

Оценки интерпретируемости: когерентность

Когерентность темы t

$$\mathrm{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j)$$

где w_i — i-й термин в порядке убывания ϕ_{wt}.

$\mathrm{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information),

N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Разреживание + Сглаживание + Декорреляция + Отбор тем

M-шаг при комбинировании 6 регуляризаторов:

$$\phi_{wt} = \operatorname*{norm}_{w}\Bigl(n_{wt} + \tau_1 \underbrace{\beta_w[t\in B]}_{\substack{\text{сглаживание}\\\text{фоновых}\\\text{тем}}} - \tau_2 \underbrace{\beta_w[t\in S]}_{\substack{\text{разреживание}\\\text{предметных}\\\text{тем}}} - \underbrace{\tau_3\, \phi_{wt} \sum_{s\in S\setminus t} \phi_{ws}}_{\text{декорреляция}}\Bigr)$$

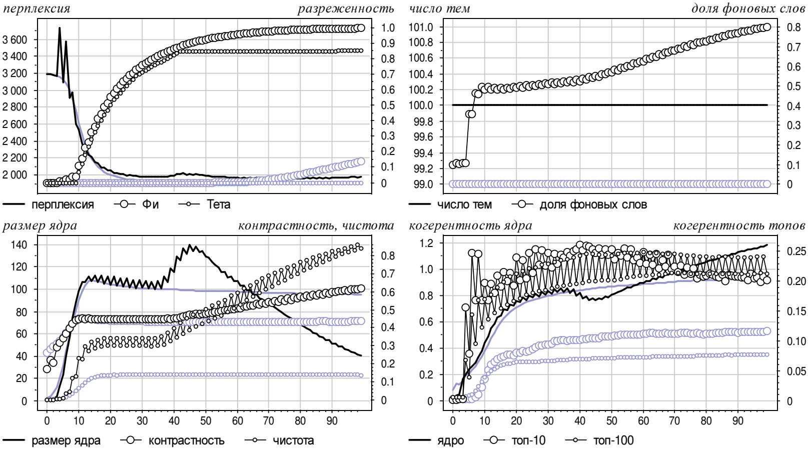
$$\theta_{td} = \operatorname*{norm}_{t}\Bigl(n_{td} + \tau_4 \underbrace{\alpha_t[t\in B]}_{\substack{\text{сглаживание}\\\text{фоновых}\\\text{тем}}} - \tau_5 \underbrace{\alpha_t[t\in S]}_{\substack{\text{разреживание}\\\text{предметных}\\\text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление}\\\text{малых тем}}}\Bigr)$$

Данные: NIPS (Neural Information Processing System)
$|D| = 1566$ статей, $n = 2.3\,\text{M}$, $|W| = 13\,\text{K}$,
контрольная коллекция: $|D'| = 174$.

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST'2014.

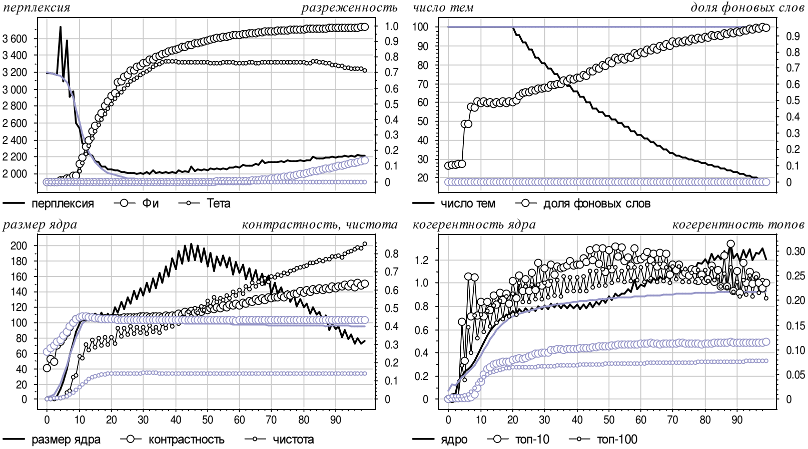
Разреживание, сглаживание, декорреляция

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Те же регуляризаторы, плюс отбор тем

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Выводы

Одновременное улучшение многих показателей:

- *разреженность* выросла от 0 до 95%–98%
- *когерентность тем* выросла от 0.1 до 0.3
- *чистота тем* выросла от 0.15 до 0.8
- *контрастность тем* выросла от 0.4 до 0.6
- почти без потери *перплексии* (правдоподобия) модели

Подобраны траектории регуляризации:

- разреживание включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декорреляцию включать сразу и как можно сильнее
- сокращение числа тем включать постепенно,
- никогда не совмещая с декорреляцией на одной итерации

Эксперименты с регуляризатором отбора тем

Коллекция статей NIPS (Neural Information Processing System)

- $|D| = 1566$ обучающих документов; $|D'| = 174$ тестовых
- $|W| = 13\,\text{K}$ — мощность словаря

Синтетическая коллекция:

- строим PLSA за 500 итераций, $|T_0| = 50$ тем на NIPS
- генерируем (n^0_{dw}) из полученных Φ и Θ:

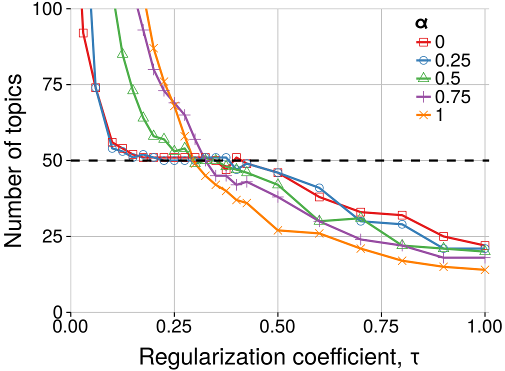
$$n^0_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

Параметрическое семейство полусинтетических данных:

- n^α_{dw} — смесь синтетических данных n^0_{dw} и реальных n_{dw}:

$$n^\alpha_{dw} = \alpha n_{dw} + (1 - \alpha) n^0_{dw}$$

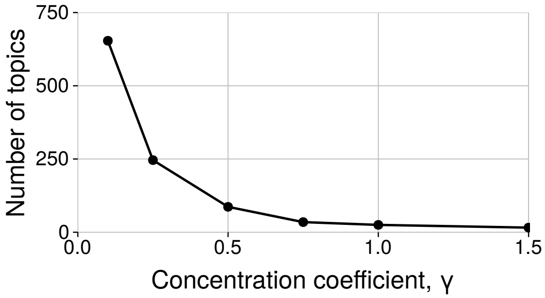
Попытка определения числа тем



- На синтетических данных надёжно находим $|T| = 50$,
- в широком интервале значений коэффициента τ;
- однако на реальных данных нет столь чёткого интервала.

Сравнение с байесовской тематической моделью HDP

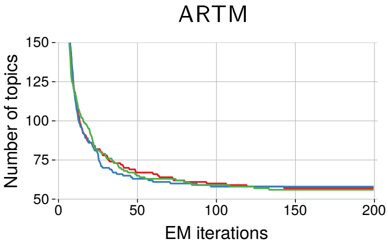
HDP (Hierarchical Dirichlet Process, Tech et. al, 2006) — «state-of-the-art» байесовский подход к определению числа тем

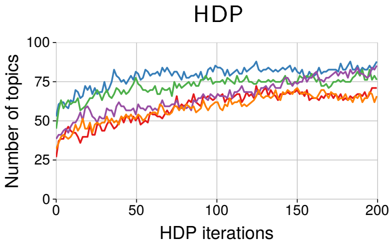


- *Коэффициент концентрации* γ в HDP влияет на $|T|$ так же сильно, как выбор коэффициента τ в ARTM.

Сравнение ARTM и HDP по устойчивости

Запуск ARTM и HDP много раз из случайных инициализаций:

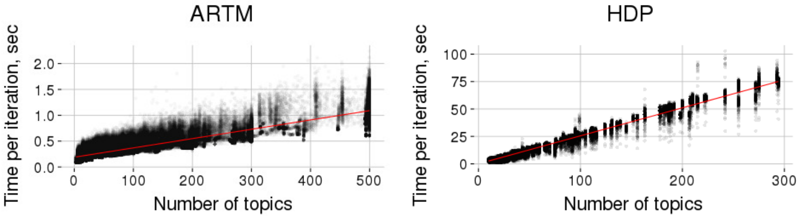




- HDP менее устойчив, причём в двух смыслах:
  - число тем сильнее флуктуирует от итерации к итерации;
  - результаты нескольких запусков различаются сильнее.
- «Рекомендуемые» значения параметров γ в HDP и τ в ARTM дают примерно равное число тем $|T| \approx 60$

Сравнение ARTM и HDP по времени вычислений

Сравнение времени одного прохода коллекции (sec)

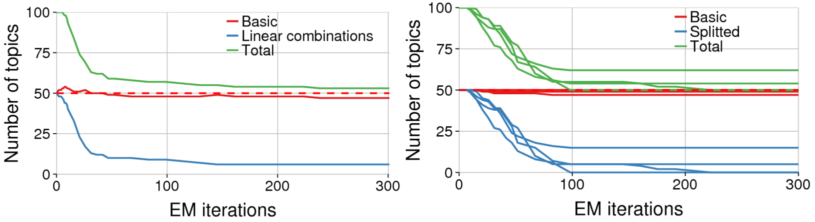


- ARTM в 100 раз быстрее!

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization // SLDS 2015, Royal Holloway, University of London, UK. pp. 193–202.

Удаление линейно зависимых и расщеплённых тем

Добавили 50 линейных комбинаций тем в модельную Φ.
Расщепили 50 тем, каждую на две подтемы в модельной Φ.



- Удаляются линейно зависимые и расщеплённые темы
- Остаются более различные темы исходной модели.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization // SLDS 2015, Royal Holloway, University of London, UK. pp. 193–202.

Резюме

- Задача тематического моделирования некорректно поставлена, её решение не единственно и не устойчиво.
- Регуляризация — стандартный приём решения таких задач.
- Подход ARTM позволяет комбинировать регуляризаторы, строить тематические модели с требуемыми свойствами, иметь общую для всех моделей реализацию.
- Модель LDA — частный случай сглаживающего регуляризатора.
- В реальных коллекциях «оптимального числа тем» нет.
- Регуляризатор отбора тем в ARTM проще, устойчивее и в 100 раз быстрее «штатного» байесовского метода HDP.