

Тематическое моделирование

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
<http://www.MachineLearning.ru/wiki>
«Машинное обучение (курс лекций, К.В.Воронцов)»

осень 2016

- 1 Вероятностное тематическое моделирование**
 - Цели, приложения, постановка задачи
 - Аддитивная регуляризация тематических моделей
 - Классические модели: PLSA и LDA
- 2 Регуляризация тематических моделей**
 - Мультимодальные тематические модели
 - Классификация и регрессия на текстах
 - Декоррелирование и отбор тем
- 3 Оценивание качества и эксперименты**
 - Внутренние (intrinsic) критерии качества
 - Внешние (extrinsic) критерии качества
 - Несколько примеров

Что такое «тема» в коллекции текстовых документов?

- *тема* — семантически однородный кластер текстов
- *тема* — специальная терминология предметной области
- *тема* — набор терминов (слов или словосочетаний), совместно встречающихся в документах

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность термина w в теме t ;
- *тематический профиль* документа — условное распределение $p(t|d)$ — вероятность темы t в документе d .

Когда автор писал термин w в документе d , он думал о теме t , и мы хотели бы выявить, о какой именно.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Цели и приложения тематического моделирования

- Выявить тематическую структуру коллекции текстов
- Найти сжатое семантическое описание каждого документа

Приложения:

- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов
- Разведочный информационный поиск (exploratory search)
- Аннотирование изображений, видео, музыки
- Анализ и агрегирование новостных потоков
- Поиск трендов, фронта исследований (research front)
- Поиск экспертов, рецензентов, подрядчиков (expert search)
- Рекомендательные системы
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Основные предположения

- Порядок слов в документе не важен (bag of words)
- Порядок документов в коллекции не важен (bag of docs)
- Каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- Коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

Предварительная обработка текстов:

- Лемматизация (русский) или стемминг (английский)
- Выделение терминов (term extraction)
- Выделение именованных сущностей (named entities)
- Удаление стоп-слов и слишком редких слов

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

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

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



w_1, \dots, w_{n_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

Дано: W — словарь терминов

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

$\hat{p}(w|d) = \frac{n_{dw}}{n_d}$ — частота термина w в документе d

n_d — длина документа d

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

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

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

Это задача стохастического матричного разложения.

Она *некорректно поставлена* — её решение не единственно:

$$\left(\frac{n_{dw}}{n_d} \right)_{W \times D} \approx \Phi_{W \times T} \cdot \Theta_{T \times D} = (\Phi S)(S^{-1} \Theta) = \Phi' \cdot \Theta'$$

Для регуляризации задачи нужны дополнительные критерии.

Принцип максимума правдоподобия

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

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

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)p(d) \rightarrow \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

ARTM: аддитивная регуляризация тематических моделей

Максимизация \log правдоподобия с регуляризатором R :

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Элементарная интерпретация EM-алгоритма

EM-алгоритм — это чередование E и M шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d, w)$ для всех t, d, w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

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

M-шаг: при $R = 0$ частотные оценки условных вероятностей, вычисляемые суммированием счётчика $n_{tdw} = n_{dw}p(t|d, w)$:

$$\begin{aligned} \phi_{wt} &= \frac{n_{wt}}{n_t}, & n_{wt} &= \sum_{d \in D} n_{tdw}, & n_t &= \sum_{w \in W} n_{wt}; \\ \theta_{td} &= \frac{n_{td}}{n_d}, & n_{td} &= \sum_{w \in D} n_{tdw}, & n_d &= \sum_{t \in T} n_{td}. \end{aligned}$$

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \rightarrow \min_x; \\ g_i(x) \leq 0, & i = 1, \dots, m; \\ h_j(x) = 0, & j = 1, \dots, k. \end{cases}$$

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

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, & \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \lambda_j h_j(x); \\ g_i(x) \leq 0; h_j(x) = 0; & \text{(исходные ограничения)} \\ \mu_i \geq 0; & \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; & \text{(условие дополняющей нежёсткости)} \end{cases}$$

Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для ϕ_{wt} , $w \in W$ (для θ_{td} всё аналогично):

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

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. Альтернатива: либо $\phi_{wt} = 0$ для всех w , либо $\lambda_t > 0$ и

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

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

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

5. Подставим λ_t из (4) в (3), получим требуемое. ■

Рациональный EM-алгоритм для тематического моделирования

Идея: E-шаг встраивается внутрь M-шага, чтобы не хранить трёхмерный массив значений n_{dwt} .

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 $i = 1, \dots, i_{\max}$

$n_{wt}, n_{td} := 0$ для всех $d \in D, w \in W, t \in T$;

для всех документов $d \in D$ и всех слов $w \in d$

$n_{tdw} := n_{dw} \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td})$ для всех $t \in T$;

$n_{wt} += n_{tdw}; n_{td} += n_{tdw}$ для всех $t \in T$;

$\phi_{wt} := \operatorname{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $w \in W, t \in T$;

$\theta_{td} := \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ для всех $d \in D, t \in T$;

Онлайновый EM-алгоритм

Вход: коллекция D , число тем $|T|$, параметры i_{\max} , j_{\max} , γ ;

Выход: матрицы терминов тем Θ и тем документов Φ ;

инициализировать $n_{wt} := 0$;

для всех $i = 1, \dots, i_{\max}$ (для больших коллекций $i_{\max} = 1$)

для всех документов $d \in D$

для всех $j = 1, \dots, j_{\max}$ (итерации по документу)

$$p_{tdw} := \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \text{ для всех } w \in D;$$

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

$$n_{wt} := \gamma n_{wt} + n_{dw} p_{tdw};$$

если пора обновить матрицу Φ **то**

$$\phi_{wt} := \operatorname{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

Классические модели PLSA и LDA

PLSA, Probabilistic Latent Semantic Analysis [Hofmann, 1999]
(вероятностный латентный семантический анализ)

$$R(\Phi, \Theta) = 0.$$

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt}), \quad \theta_{td} = \underset{t}{\text{norm}}(n_{td}).$$

LDA, Latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}.$$

M-шаг — сглаженные частотные оценки с параметрами β_w, α_t :

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt} + \beta_w), \quad \theta_{td} = \underset{t}{\text{norm}}(n_{td} + \alpha_t).$$

Hofmann T. Probabilistic Latent Semantic Indexing. SIGIR 1999.

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

Вероятностная байесовская интерпретация LDA [Blei, 2003]

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

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

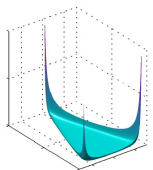
$$\text{Dir}(\theta_d | \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \quad \alpha_t > 0;$$

Пример:

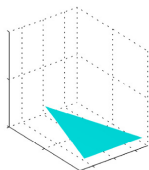
$$\text{Dir}(\theta | \alpha)$$

$$|T| = 3$$

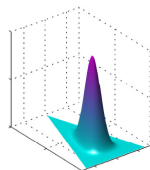
$$\theta, \alpha \in \mathbb{R}^3$$



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

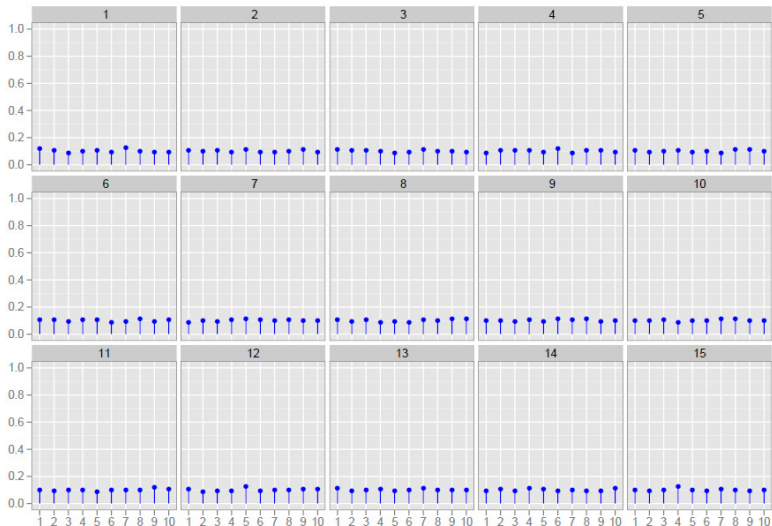


$$\alpha_1 = \alpha_2 = \alpha_3 = 1$$

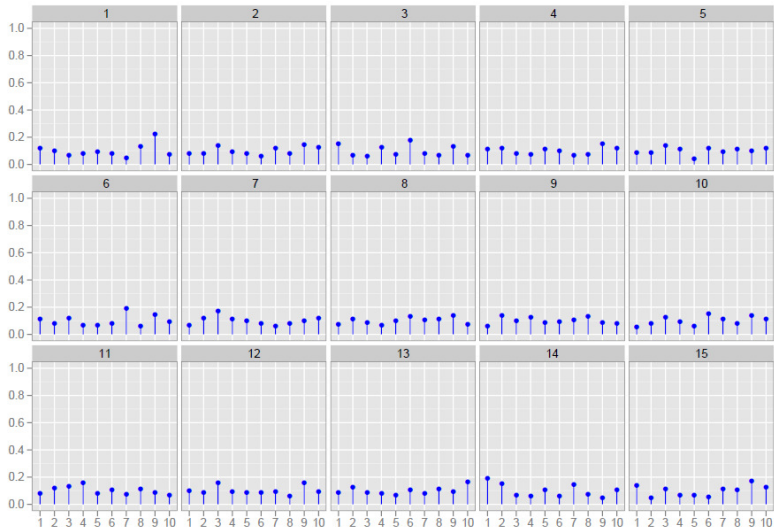


$$\alpha_1 = \alpha_2 = \alpha_3 = 10$$

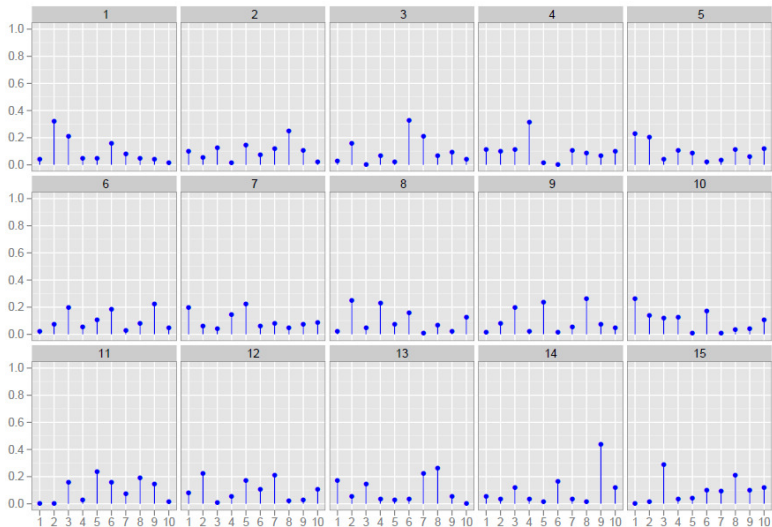
Распределение Дирихле при $\alpha_t \equiv 100$, 10 тем, 15 документов



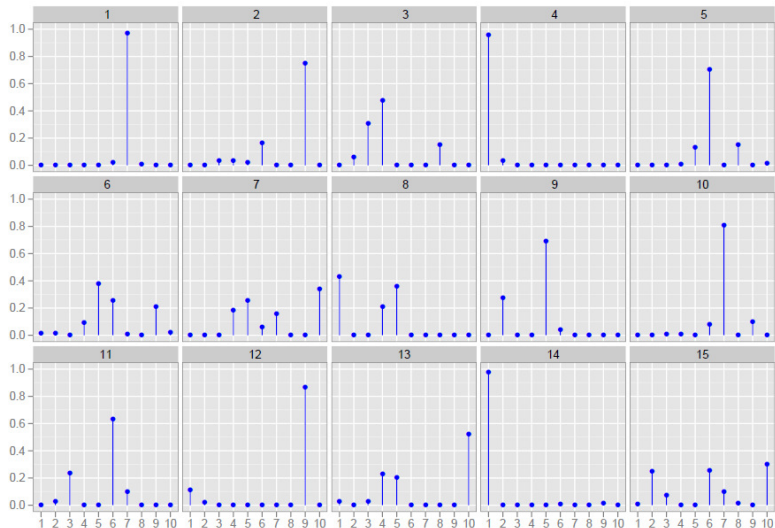
Распределение Дирихле при $\alpha_t \equiv 10$, 10 тем, 15 документов



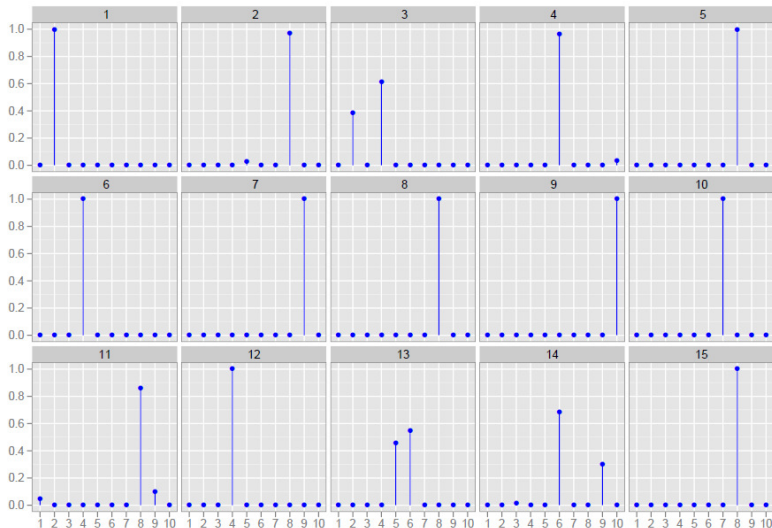
Распределение Дирихле при $\alpha_t \equiv 1$, 10 тем, 15 документов



Распределение Дирихле при $\alpha_t \equiv 0.1$, 10 тем, 15 документов



Распределение Дирихле при $\alpha_t \equiv 0.01$, 10 тем, 15 документов



Максимизация апостериорной вероятности для модели LDA

Регуляризатор — логарифм априорного распределения:

$$\begin{aligned} R(\Phi, \Theta) &= \ln \prod_{t \in T} \text{Dir}(\phi_t | \beta) \prod_{d \in D} \text{Dir}(\theta_d | \alpha) = \\ &= \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td} + \text{const} \end{aligned}$$

M-шаг — сглаженные или слабо разреженные оценки:

$$\phi_{wt} = \text{norm}_w(n_{wt} + \beta_w - 1), \quad \theta_{td} = \text{norm}_t(n_{td} + \alpha_t - 1).$$

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,

при $0 < \beta_w < 1$, $0 < \alpha_t < 1$ — слабое разреживание,

при $\beta_w = 1$, $\alpha_t = 1$ априорное распределение равномерно, PLSA.

Почему именно распределение Дирихле?

Плюсы:

- описывает широкий класс распределений на симплексе
- позволяет управлять разреженностью ϕ_{wt} и θ_{td}
- иногда уменьшает переобучение
- удобно для байесовского вывода, т. к. является сопряжённым к дискретному распределению

Минусы:

- не имеет лингвистических обоснований
- не даёт выигрыша против PLSA на больших коллекциях
- слишком слабый разреживатель
- слишком слабый регуляризатор

Обобщённая не-байесовская интерпретация LDA

Сглаживание распределений по KL-дивергенции:

приблизить $\phi_{wt} \equiv p(w|t)$ к заданным распределениям $\beta_t(w)$,
приблизить $\theta_{td} \equiv p(t|d)$ к заданным распределениям $\alpha_d(t)$:

$$\sum_{t \in T} \tau_t \text{KL}(\beta_t(w) \parallel \phi_{wt}) \rightarrow \min_{\Phi}; \quad \sum_{d \in D} \tau_d \text{KL}(\alpha_d(t) \parallel \theta_{td}) \rightarrow \min_{\Theta}.$$

Взвешенная сумма регуляризаторов:

$$R(\Phi, \Theta) = \sum_{t \in T} \tau_t \sum_{w \in W} \beta_t(w) \ln \phi_{wt} + \sum_{d \in D} \tau_d \sum_{t \in T} \alpha_d(t) \ln \theta_{td}.$$

Формулы M-шага:

$$\phi_{wt} = \text{norm}_w(n_{wt} + \tau_t \beta_t(w)), \quad \theta_{td} = \text{norm}_t(n_{td} + \tau_d \alpha_d(t)).$$

Преимущества не-байесовского обобщения LDA

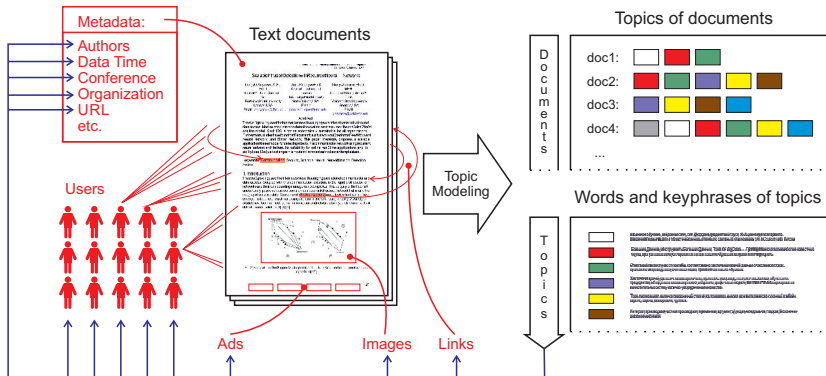
Формулы M-шага:

$$\phi_{wt} = \operatorname{norm}_w(n_{wt} + \beta_{wt}), \quad \theta_{td} = \operatorname{norm}_t(n_{td} + \alpha_{td}).$$

- максимизация KL ведёт к $\beta_{wt} < 0$, $\alpha_{td} < 0$ и разреживанию
- разреживание и сглаживание описывается общей формулой
- можно разбивать темы на два подмножества, $T = S \sqcup B$:
 S — разреженные *предметные* темы со специальной лексикой
 B — сглаженные *фоновые* темы с общей лексикой языка
- можно собирать предметные темы S , сглаживая их по словарю терминов W_0 : $\beta_{wt} = p(w)[w \in W_0][t \in S]$
- можно использовать *частичное обучение*:
 $\beta_{wt} > 0$ — сглаживание, термин w в «белом списке» темы t
 $\beta_{wt} < 0$ — разреживание, термин w в «чёрном списке» темы t
 $\alpha_{td} > 0$ — сглаживание, тема t в «белом списке» документа d
 $\alpha_{td} < 0$ — разреживание, тема t в «чёрном списке» документа d

ARTM легко обобщается на мультимодальные задачи

Выявление тематики документов $p(t|d)$ и терминов $p(t|w)$, а также модальностей: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{ссылка})$, $p(t|\text{баннер})$, $p(t|\text{элемент изображения})$, $p(t|\text{пользователь})$,...



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:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

Регуляризатор для классификации и категоризации текстов

Цель: построить тематическую модель классификации.

Y — множество классов;

$n_{dy} = [\text{документ } d \text{ относится к классу } y]$ — обучающие данные;

$p(y|d) = \sum_{t \in T} \phi_{yt} \theta_{td}$ — линейная модель классификации.

Регуляризатор — правдоподобие модальности классов:

$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{y \in Y} n_{dy} \ln \sum_{t \in T} \phi_{yt} \theta_{td} \rightarrow \max,$$

это тематическая модель с двумя модальностями, W и Y .

TM превосходит SVM в случае несбалансированных классов.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

Регуляризатор для задач регрессии

Цель: построить тематическую модель регрессии.

$y_d \in \mathbb{R}$ для всех документов $d \in D$ — обучающие данные.

$E(y|d) = \sum_{t \in T} v_t \theta_{td}$ — линейная модель регрессии, $v \in \mathbb{R}^{|T|}$.

Регуляризатор — среднеквадратичная ошибка (МНК):

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

Формулы М-шага:

$$\theta_{td} = \text{norm}_t \left(n_{td} + \tau \left(y_d - \sum_{t \in T} v_t \theta_{td} \right) \right);$$

$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

Sokolov E., Bogolubsky L. Topic Models Regularization and Initialization for Regression Problems // CIKM-2015 Workshop on Topic Models. ACM, pp. 21–27.

Регуляризатор декоррелирования тем

Цели: усилить различность тем; выделить в каждой теме лексическое ядро, отличающее её от других тем; вывести слова общей лексики из предметных тем в фоновые.

Минимизируем ковариации между вектор-столбцами ϕ_t :

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

Формулы М-шага (контрастирование строк матрицы Φ с эффектами разреживания и выделения слов общей лексики):

$$\phi_{wt} = \operatorname{norm}_w \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right).$$

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

Цели: избавиться от «мелких» незначимых тем;
удалить дублирующие, зависимые и расщеплённые темы.

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

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

Формулы M-шага (строки матрицы Θ целиком обнуляются для тем t , собравших слишком мало слов, $n_t < \tau$):

$$\theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right) \approx \operatorname{norm}_{t \in T} \left(n_{td} \left(1 - \frac{\tau}{n_t} \right) \right).$$

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

Правдоподобие и перплексия (perplexity)

Правдоподобие языковой модели $p(w|d)$ (чем выше, тем лучше):

$$\mathcal{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d), \quad p(w|d) = \sum_t \phi_{wt} \theta_{td}$$

Перплексия языковой модели $p(w|d)$ (чем меньше, тем лучше):

$$\mathcal{P}(D) = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)\right), \quad n = \sum_{d \in D} \sum_{w \in d} n_{dw}$$

Интерпретация перплексии:

- если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- мера различности или неопределённости слов в тексте
- коэффициент ветвления (branching factor) текста

Перплексия тестовой (отложенной) коллекции

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{1}{n''} \sum_{d \in D'} \sum_{w \in d''} n_{dw} \ln p(w|d)\right), \quad n'' = \sum_{d \in D'} \sum_{w \in d''} n_{dw}$$

$d = d' \sqcup d''$ — случайное разбиение тестового документа на две половины равной длины;

параметры ϕ_{wt} оцениваются по обучающей коллекции D ;

параметры θ_{td} оцениваются по первой половине d' ;

перплексия вычисляется по второй половине d'' .

Меры интерпретируемости тем и когерентность

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- Экспертные оценки:
 - интерпретируемость темы по балльной шкале;
 - каждую тему оценивают несколько экспертов.
- Метод интрузий (intrusion):
 - в список топовых слов внедряется лишнее слово;
 - измеряется доля ошибок экспертов его при определении.

Нужна автоматически вычисляемая мера интерпретируемости, коррелирующая с экспертными оценками.

Ею оказалась *когерентность* (согласованность, coherence).

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

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена между 15 метрикам и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	-0.20	0.19
	LCH	-0.31	-0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	-0.08	0.27
	WuP	0.41	0.26
	Wikipedia	RACO	0.62
MiW		0.68	0.70
DOCsim		0.59	0.60
PMI		0.74	0.77
Google	TITLES	0.51	
	LOGHITS	-0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

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

Когерентность как внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

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

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

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

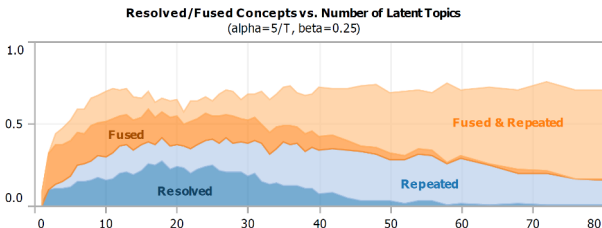
N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

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.

Внешние критерии качества

- Полнота и точность тематического поиска
- Качество ранжирования при тематическом поиске
- Качество категоризации документов
- Точность соответствия тем заданным *концептам* (число найденных и расщеплённых тем и концептов)



Chuang J., Gupta S., Manning C., Heer J. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. ICML-2013.

Пример тем. Мультиязычная модель Википедии

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

Дударенко М. А. Регуляризация многоязычных тематических моделей.
 Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Пример тем. Мультиязычная модель Википедии

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

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Биграммы радикально улучшают интерпретируемость тем

Коллекция 850 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

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

- Тематическое моделирование — это восстановление латентных тем по коллекции текстовых документов
- Задача сводится к стохастическому матричному разложению
- Стандартные методы — PLSA и LDA.
- Задача является некорректно поставленной, так как множество её решений в общем случае бесконечно
- Аддитивная регуляризация позволяет комбинировать модели и строить модели с заданными свойствами
- В отличие от классических задач машинного обучения, регуляризаторы до крайности разнообразны
- Не забывать о внешних критериях оценивания моделей