Аддитивная регуляризация вероятностных тематических моделей

Константин Воронцов

Вычислительный центр им. А. А. Дородницына РАН
Московский Физико-Технический Институт

Математические методы распознавания образов
ММРО-16 • Казань • 7–11 октября 2013

Содержание

Задача определения тематики коллекции документов

Тема — это набор терминов, неслучайно часто совместно встречающихся в относительно узком подмножестве документов.

Дано:

W — словарь, множество слов (терминов)

D — множество (коллекция, корпус) текстовых документов

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Постановки задач:

— найти, какими терминами определяется каждая тема

— найти, к каким темам относится каждый документ

— определить число статистически различимых тем

— восстановить иерархию тем

— построить динамику развития тем во времени

— найти тематику связанных с документами объектов

Цели тематического моделирования (topic modeling)

- Тематический поиск документов и объектов по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация текстовых документов

Типичные приложения:

- Поиск научной информации
- Поиск экспертов (expert search), рецензентов, проектов
- Выявление трендов и фронта исследований
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендательные сервисы (коллаборативная фильтрация)
- Аннотация генома и другие задачи биоинформатики

Вероятностная формализация постановки задачи

Базовые предположения:

- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция D — выборка троек $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

Вероятностная модель порождения документа d:

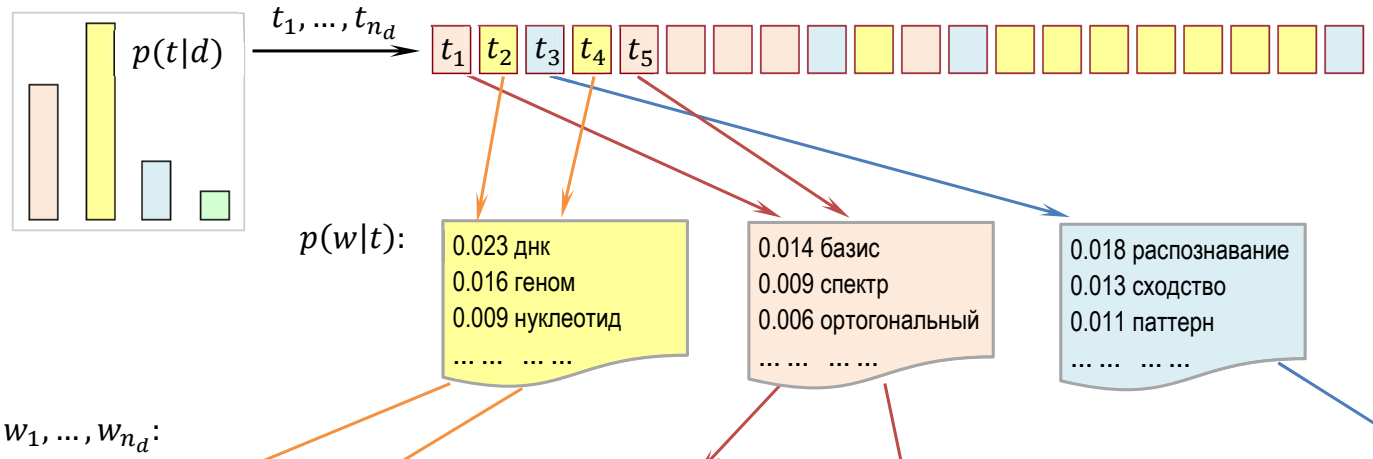
$$p(w|d) = \sum_{t \in T} p(w|d,t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

Дано $p(w|d) \equiv n_{dw}/n_d$, **найти:**

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темеах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

Вероятностная модель порождения документа d

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Задача максимизации правдоподобия

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

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

Интерпретация: найти стохастическое матричное разложение

$$\|F - \Phi\Theta\|_{\mathrm{KL}} \to \min_{\Phi,\Theta}$$

$F = \big(\hat{p}(w|d)\big)_{W\times D}$ — известная матрица исходных данных,
$\Phi = \big(\phi_{wt}\big)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$,
$\Theta = \big(\theta_{td}\big)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

EM-алгоритм, вероятностный латентный семантический анализ PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

E-шаг. Выразим $p(t|d,w)$ через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}$$

$n_{dwt} = n_{dw}p(t|d,w)$ — оценка числа троек (d,w,t) в коллекции

M-шаг. Частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d\in D} n_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w\in d} n_{dwt}}, \qquad \theta_{td} = \frac{n_{dt}}{n_d} \equiv \frac{\sum\limits_{w\in d} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dwt}},$$

или краткая запись:

$$\phi_{wt} \propto n_{wt} \qquad \theta_{td} \propto n_{dt}$$

Недостатки классического PLSA

1. PLSA переобучается, т.к. число параметров ϕ_{wt} и θ_{td} слишком велико, $|D| \cdot |T| + |W| \cdot |T|$
2. PLSA не позволяет управлять *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 $\theta_{td} = 0$)
3. PLSA не позволяет управлять *разреженностью* $p(t|d, w)$
4. PLSA вынужден хранить 3D-матрицу $p(t|d, w)$
5. PLSA медленно сходится на больших коллекциях, т.к. Φ и Θ обновляются после каждого прохода коллекции
6. PLSA неверно оценивает вероятность новых слов:
   если $n_w = 0$, то $\hat{p}(w|t) = 0$ для всех $t \in T$

Латентное размещение Дирихле
LDA — Latent Dirichlet Allocation [David Blei, 2003]

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

Гипотеза об априорных распределениях Дирихле:

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

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

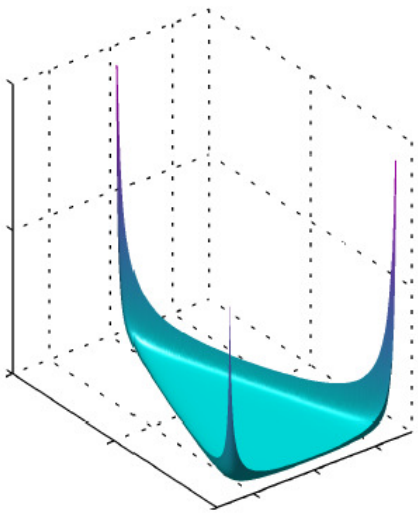
- $\phi_t = (\phi_{wt})_{w \in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:

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

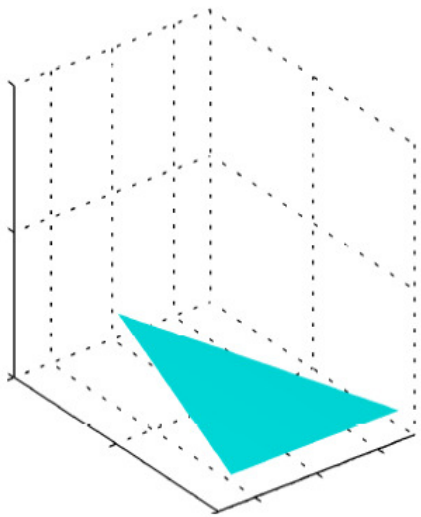
Почему именно распределение Дирихле?

– Является сопряжённым к мультиномиальному распределению
– Порождает как сглаженные, так и разреженные векторы
– Неплохо описывает кластерные структуры на симплексе

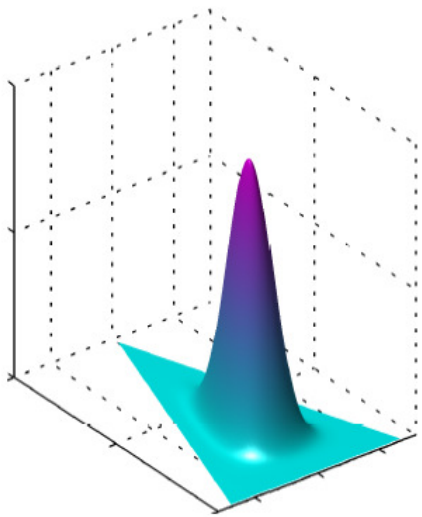
Пример. $\mathrm{Dir}(\theta|\alpha)$ при $|T| = 3$, $\theta, \alpha \in \mathbb{R}^3$:



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



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



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

Главное отличие LDA от PLSA

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- в LDA — сглаженные байесовские оценки:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

Байесовский вывод алгоритма сэмплирования Гиббса:

Yi Wang. Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.

Недостатки LDA

1. слабые лингвистические обоснования, «особая роль» распределения Дирихле
2. сглаживание вместо разреживания
3. байесовский вывод требует интегрирования по пространству параметров модели, которое только в базовом варианте LDA элементарно
4. построение композитных и многофункциональных моделей — громоздкая математическая задача
5. практика показывает, что на достаточно больших данных нет значимых различий между LDA и PLSA

Эксперимент на модельных данных

Модельные коллекции порождаются заданными матрицами Φ_0 и Θ_0 при $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Отклонение восстановленных распределений $p(i|j)$ от исходных модельных распределений $p_0(i|j)$ измеряются средним расстоянием Хеллингера:

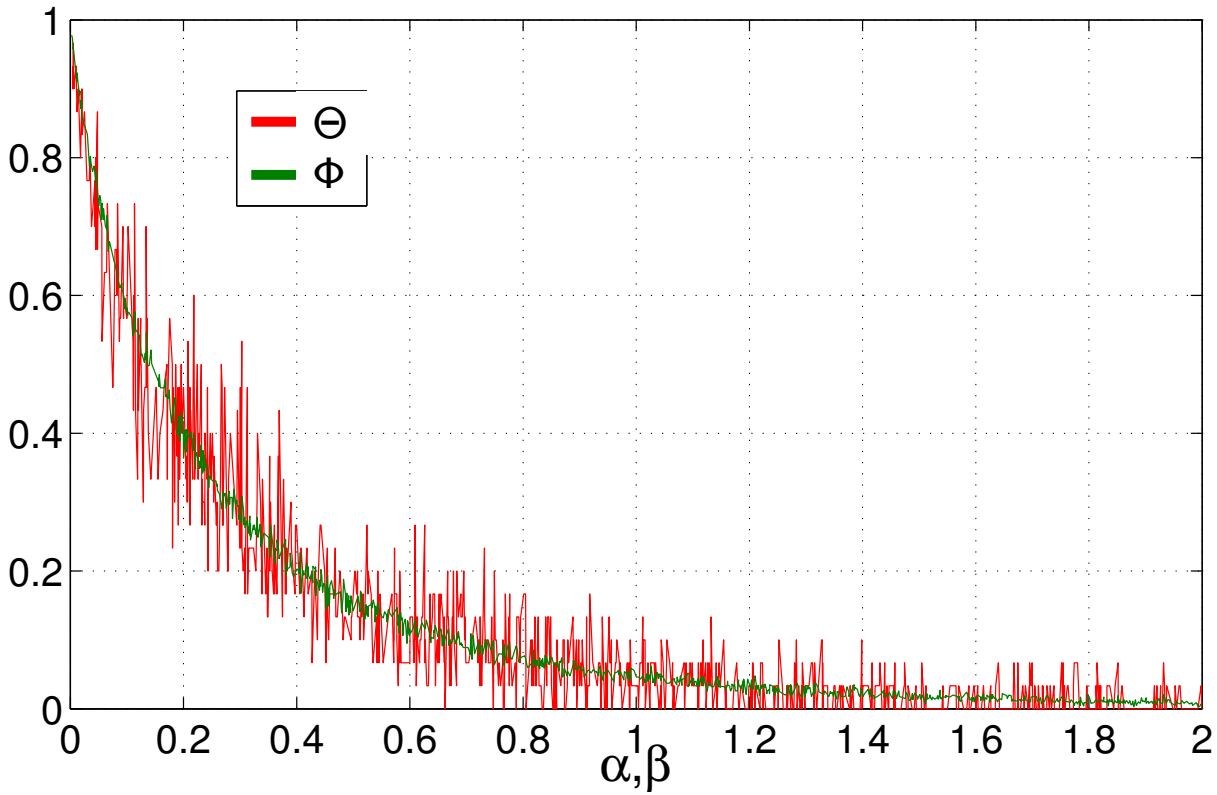
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

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

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

Генерация модельных данных различной степени разреженности

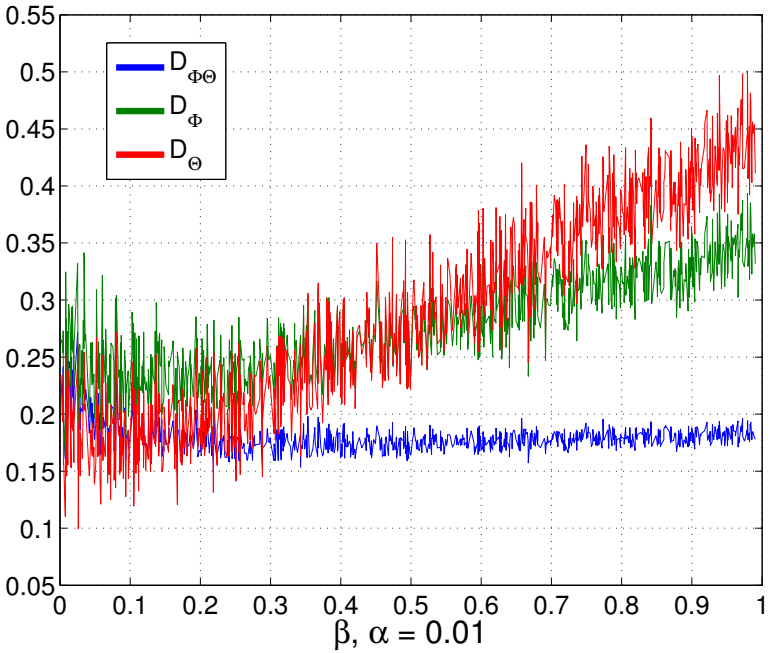
Зависимость разреженности (доли нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



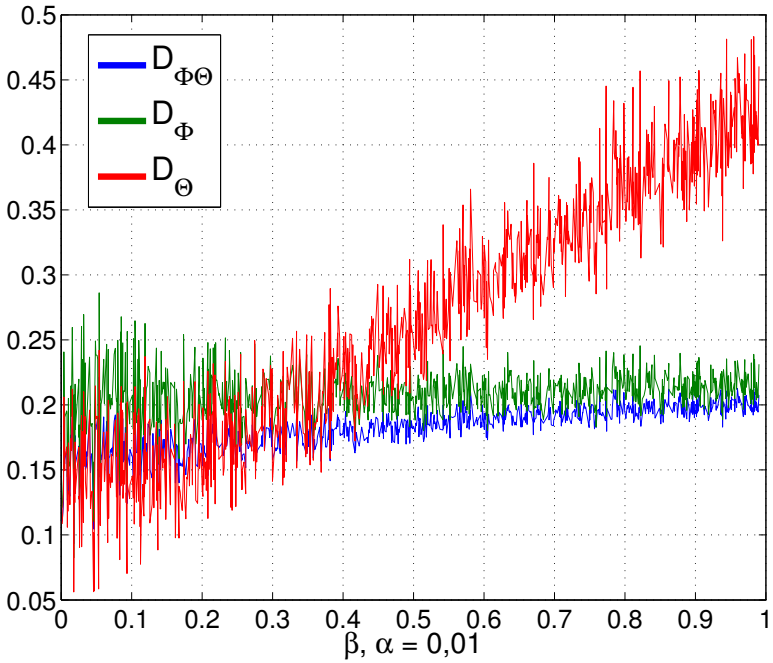
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ_0

PLSA



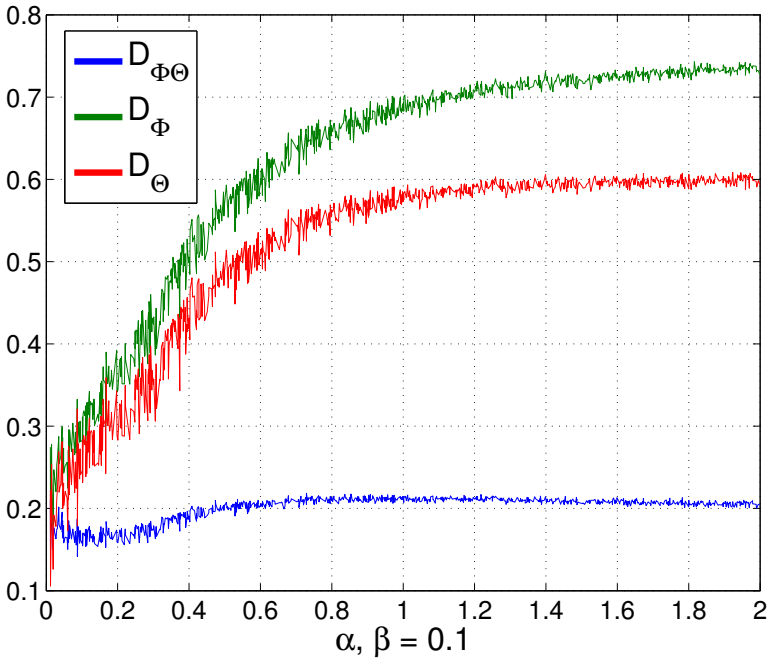
LDA



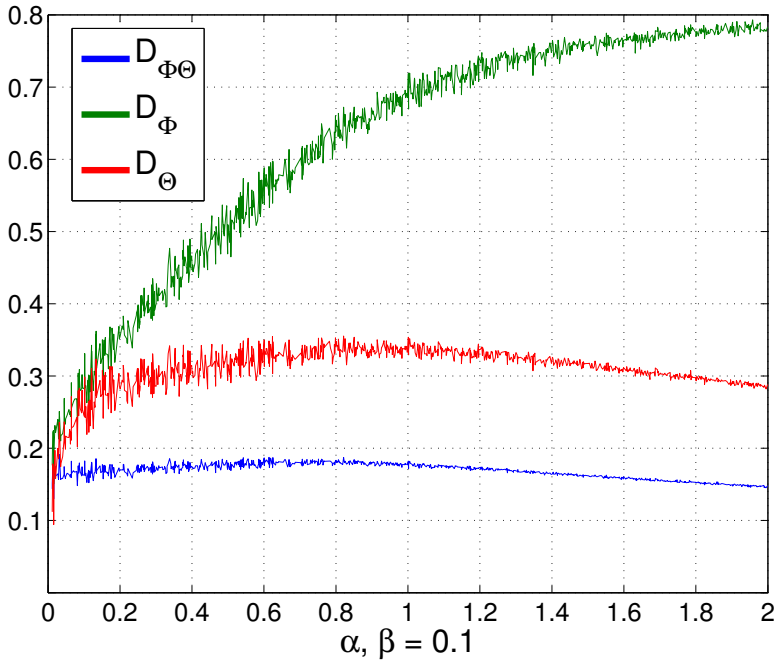
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0

PLSA



LDA



Выводы

1. Произведение $\Phi\Theta$ восстанавливается устойчиво, точность восстановления не зависит от разреженности исходных модельных данных Φ_0, Θ_0
2. Матрицы Φ, Θ восстанавливаются неустойчиво, результат сильно зависит от случайной инициализации, если разреженность (доля нулей) в Φ_0, Θ_0 менее 80%
3. Методы PLSA и LDA одинаково неустойчивы (сглаживание не спасает от неединственности)

Реализация экспериментов:
Виталий Глушаченков. Магистерская диссертация. МФТИ, 2013.
Михаил Колупаев. Курсовая работа. ВШЭ, 2013.

Причина неустойчивости тематических моделей

Задача стохастического матричного разложения:

$$\hat{F} \approx F = \Phi\Theta$$

$\hat{F} = \left(n_{dw}/n_d\right)_{W\times D}$ — известная матрица исходных данных;
$F = \left(p(w|d)\right)_{W\times D}$ — матрица тематической модели;
$\Phi = \left(\phi_{wt}\right)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$;
$\Theta = \left(\theta_{td}\right)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.
Все матрицы неотрицательные, с нормированными столбцами.

Проблема неединственности матричного разложения:

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

для невырожденных $S_{T\times T}$.

Какое из множества разложений лучше выбрать?

Обоснование EM-алгоритма PLSA

Теорема

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

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

и фиксированных значениях $p(t|d,w)$ достигается при

$$\phi_{wt} \propto n_{wt} \equiv \sum_{d\in D} n_{dw} p(t|d,w) \qquad \theta_{td} \propto n_{dt} \equiv \sum_{w\in W} n_{dw} p(t|d,w)$$

Обоснование регуляризованного EM-алгоритма PLSA

Теорема

Максимум регуляризованного правдоподобия

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

при ограничениях неотрицательности и нормировки

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

и фиксированных значениях $p(t|d,w)$ достигается, когда

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+ \qquad \theta_{td} \propto \left(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+$$

Дивергенция Кульбака–Лейблера

Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

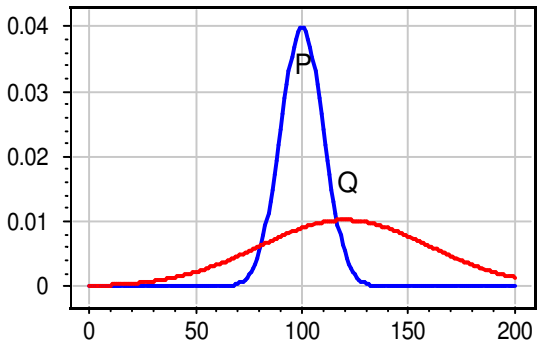
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

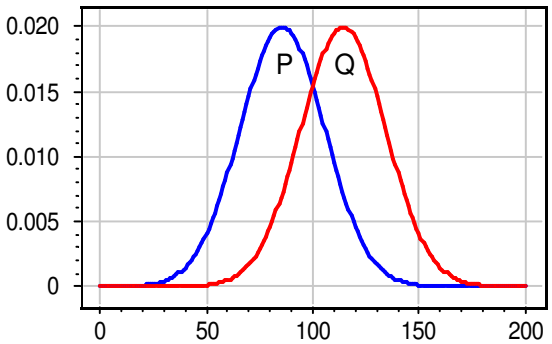
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

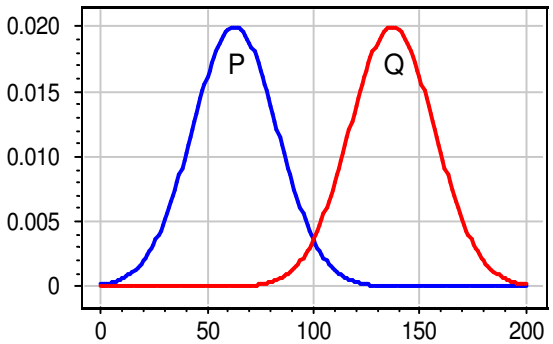
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$



$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Регуляризатор №1: Сглаживание LDA

Гипотеза сглаженности:

распределения ϕ_{wt} близки к заданным распределениям β_w

распределения θ_{td} близки к заданным распределениям 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 этих 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:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w, \qquad \theta_{td} \propto n_{dt} + \alpha_0\alpha_t.$$

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

Регуляризатор №1: Сглаживание LDA

Выводы:

- Найдено альтернативное обоснование LDA:
  оказывается, это всего лишь притягивание столбцов Φ, Θ
  к заданным распределениям
- Формулы M-шага LDA получены без байесовского вывода:
  — без предположения об априорном распределении
  — без интегрирования по пространству параметров модели
  — без требования сопряжённости
- Распределение Дирихле утрачивает «особую роль»,
  это лишь один из многих возможных регуляризаторов

Регуляризатор №2: Частичное обучение

Пусть известно, что
1) документы $d \in D_0$ относятся к темам $T_d \subset T$,
2) к темам $t \in T_0$ относятся термины $W_t \subset W$.

ϕ^0_{wt} — распределение, равномерное на W_t
θ^0_{td} — распределение, равномерное на T_d

Максимизируем сумму этих регуляризаторов:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi^0_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \ln \theta_{td} \to \max$$

Подставляем, получаем обобщение LDA:

$$\theta_{td} \propto n_{dt} + \beta_0 \theta^0_{td} \qquad \phi_{wt} \propto n_{wt} + \alpha_0 \phi^0_{wt}$$

Nigam K., McCallum A., Thrun S., Mitchell T. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, no. 2–3.

Регуляризатор №2: Частичное обучение (новое обобщение)

Гипотеза: вместо логарифма можно взять любую другую монотонно возрастающую функцию μ

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T_0}\sum_{w\in W_t} \phi^0_{wt}\mu(\phi_{wt}) + \alpha_0 \sum_{d\in D_0}\sum_{t\in T_d} \theta^0_{td}\mu(\theta_{td}) \to \max$$

Подставляем, получаем ещё одно обобщение LDA:

$$\theta_{td} \propto n_{dt} + \beta_0\theta^0_{td}\theta_{td}\mu'(\theta_{td}) \qquad \phi_{wt} \propto n_{wt} + \alpha_0\phi^0_{wt}\phi_{wt}\mu'(\phi_{wt})$$

При $\mu(z) = z$ максимизируется сумма ковариаций $\text{cov}(\theta^0_d, \theta_d)$. Если θ^0_{td} равномерно на T_d, то ковариация не накладывает ограничений на распределение θ_{td} между темами из T_d.

Регуляризатор №3: Разреживание

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей.

Чем сильнее разрежено распределение, тем ниже его энтропия.
Максимальной энтропией обладает равномерное распределение.

Поэтому максимизируем дивергенцию между равномерным распределением и искомыми распределениями ϕ_{wt}, θ_{td}:

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

Подставляем, получаем «анти-LDA»:

$$\phi_{wt} \propto \left(n_{wt} - \beta\right)_+, \qquad \theta_{td} \propto \left(n_{dt} - \alpha\right)_+.$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Регуляризатор №4: Антикорреляция

Гипотеза некоррелированности тем:
чем различнее темы, тем лучше они интерпретируются.

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:

$$\phi_{wt} \propto \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right)_+ .$$

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

Регуляризатор №5: Максимизация когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, v \in W$.

Пусть C_{uv} — оценка когерентности.

$$R(\Phi, \Theta) = \tau \sum_{t \in T} \sum_{(u,v)} C_{uv} n_{ut} \ln \phi_{vt} \to \max,$$

Подставляем, получаем ещё один вариант сглаживания: векторы ϕ_{wt} притягиваются к эмпирическим оценкам распределений $p(w|t)$, вычисляемым по когерентным словам:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} n_{ut}.$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Регуляризатор №6: Связи между документами

Гипотеза: чем больше n_{dc} — число ссылок из d на c, тем более близки тематики документов d и c.

Минимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Phi, \Theta) = \tau \sum_{d,c \in D} n_{dc} \text{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} \propto \hat{n}_{dt} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Регуляризатор №7: Классификация

Пусть C — множество классов документов (категории, пользователи, авторы, ссылки, годы, конференции,...)

Гипотеза:
классификация документа d объясняется его темами:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}$$

Минимизируем дивергенцию между моделью $p(c|d)$ и «эмпирической частотой» классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct}\theta_{td} \to \max$$

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

Регуляризатор №7: Классификация (EM-алгоритм)

E-шаг. По формуле Байеса:

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}} \qquad p(t|d,c) = \frac{\psi_{ct}\theta_{td}}{\sum_{s\in T}\psi_{cs}\theta_{sd}}$$

M-шаг. Максимизация регуляризованного правдоподобия:

$$\phi_{wt} \propto n_{wt} \qquad n_{wt} = \sum_{d\in D} n_{dw} p(t|d,w)$$

$$\theta_{td} \propto n_{dt} + \tau m_{dt} \qquad n_{dt} = \sum_{w\in W} n_{dw} p(t|d,w) \qquad m_{dt} = \sum_{c\in C} m_{dc} p(t|d,c)$$

$$\psi_{ct} \propto m_{ct} \qquad m_{ct} = \sum_{d\in D} m_{dc} p(t|d,c)$$

Регуляризатор №8: Динамическая тематическая модель

Пусть классы C — это годы публикации

Гипотеза:
тематика меняется медленно, поэтому вероятности ψ_{ct} в последовательные годы $(c-1, c)$ должны быть близки:

$$R_2(\Psi) = -\frac{\tau_2}{2} \sum_{c \in C} \sum_{t \in T} (\psi_{ct} - \psi_{c-1,t})^2 \to \max.$$

Сглаживание–разреживание:
если значение ψ_{ct} меньше полусуммы соседних вероятностей $\psi_{c-1,t}$, $\psi_{c+1,t}$, то оно увеличивается, иначе — уменьшается:

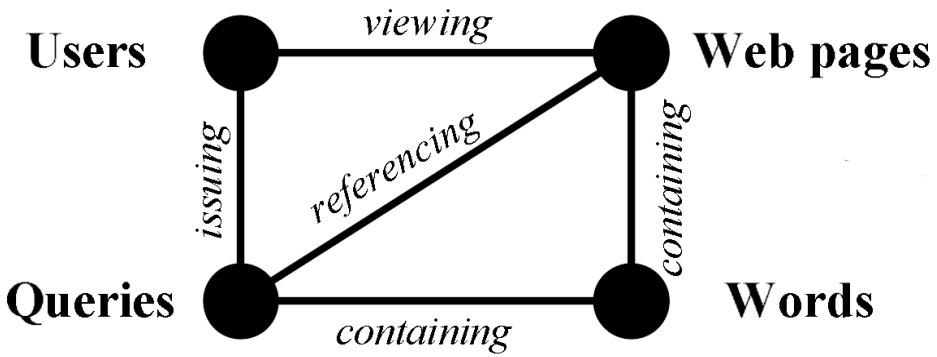
$$\psi_{ct} \propto \tau_1 m_{ct} + \tau_2 \psi_{ct} (\psi_{c-1,t} + \psi_{c+1,t} - 2\psi_{ct}).$$

Мультимодальные ТМ: коллаборативная фильтрация

Тематическое моделирование с классификацией документов:
документы D, термины W, классы C

Коллаборативная фильтрация:
предметы D с их описаниями, термины W, пользователи U

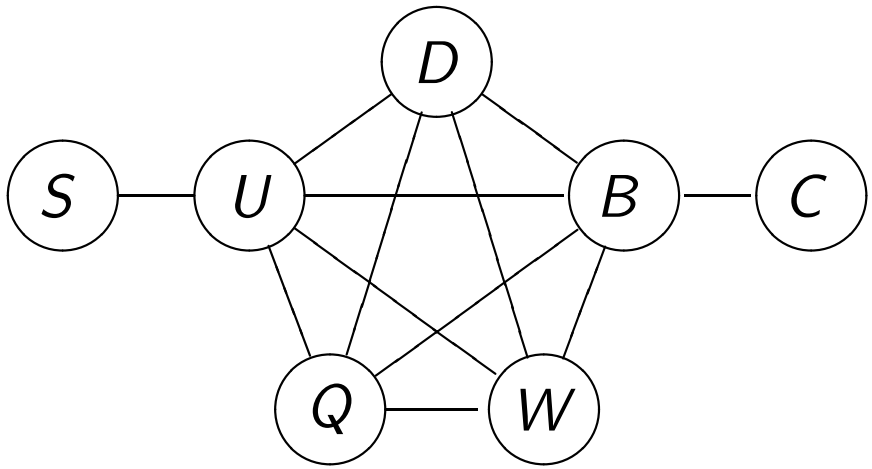
Персонализация поиска:
сайты D, термины W, пользователи U, запросы Q



Wang X., Sun J. T., Chen Z., Zhai C. X. Latent semantic analysis for multiple-type interrelated data objects // SIGIR'06, Pp. 236–243

Мультимодальные ТМ: рекламная сеть поисковой системы

Персонализация показов рекламы:
сайты D, термины W, пользователи U, запросы Q, баннеры B, социально-демографические классы пользователей S, рекламные кампании C



Объекты x всех типов получают тематические профили $p(t|x)$, учитывающие всевозможные взаимодействия между объектами

Подбор траекторий регуляризации

Пусть задана линейная комбинация регуляризаторов:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$$

Задача: выбрать вектор коэффициентов $\tau = (\tau_i)_{i=1}^{n}$

Ближайшие аналоги:

- Построение «Regularization Path» в задачах регрессии с двумя регуляризаторами L_1 и L_2 (Elastic Net)
- Постепенное разреживание тематической модели (в докладе Анны Потапенко)

Идея построения траектории в пространстве коэффициентов τ:
1) достичь сходимости нерегуляризованного PLSA,
2) усиливать регуляризаторы постепенно, в определённом порядке.

Открытые проблемы и задачи

Математические:

1. Доказательство сходимости регуляризованного PLSA-EM
2. Разреживание $p(t|d,w)$: сэмплирование—максимизация
3. Устойчивое определение числа тем без HDP

Экспериментальные:

1. Регуляризаторы, улучшающие интерпретируемость тем
2. Многофункциональные и композитные модели, в частности, мультимодальные тематические модели
3. Подбор траекторий регуляризации

Технические:

1. Реализация библиотеки регуляризаторов
2. Распределённая параллельная реализация (Big Data)

Воронцов Константин Вячеславович
voron@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование