Вероятностные тематические модели коллекций текстовых документов

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)»

апрель 2014

Содержание

Понятие «латентной темы»

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.
- *Тема* — вероятностное распределение на терминах: $p(w|t)$ — вероятность встретить термин w в теме t.

Документ d состоит из наблюдаемых терминов $w_1, \ldots, w_{n_d}$, $p(w|d)$ — известная частота термина w в документе d.

Документ имеет ненаблюдаемый *семантический профиль*: $p(t|d)$ — неизвестная частота темы t в документе d.

Когда автор писал термин w в документ d, он думал о теме t. *Тематическая модель* пытается выявить латентные темы.

Цели и приложения тематического моделирования

- Выявить скрытую тематическую структуру коллекции текстов
- Выявить семантический профиль каждого документа

Приложения:

- Семантический поиск по текстовому запросу любой длины
- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов
- Поиск научной информации, трендов, фронта исследований
- Поиск специалистов (expert search), рецензентов, проектов
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендующие системы, коллаборативная фильтрация
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Основные предположения

1. Порядок документов в коллекции не важен
2. Порядок слов в документе не важен (bag of words)
3. Слова, встречающиеся «почти во всех» документах, не важны
4. Слово в разных формах — это одно и то же слово
5. Документ обычно относится к небольшому числу тем
6. Тема обычно определяется небольшим числом терминов

Предварительная обработка текстов:

- Приведение всех слов к нормальной форме (лемматизация или стемминг)
- Выделение терминов (key phrase extraction) (сводится к задаче классификации или ранжирования)
- Удаление стоп-слов и слишком редких слов

Вероятностная формализация постановки задачи

Формализация основных предположений:

- 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: $p(w|d,t) = p(w|t)$

Вероятностная модель порождения документа d:

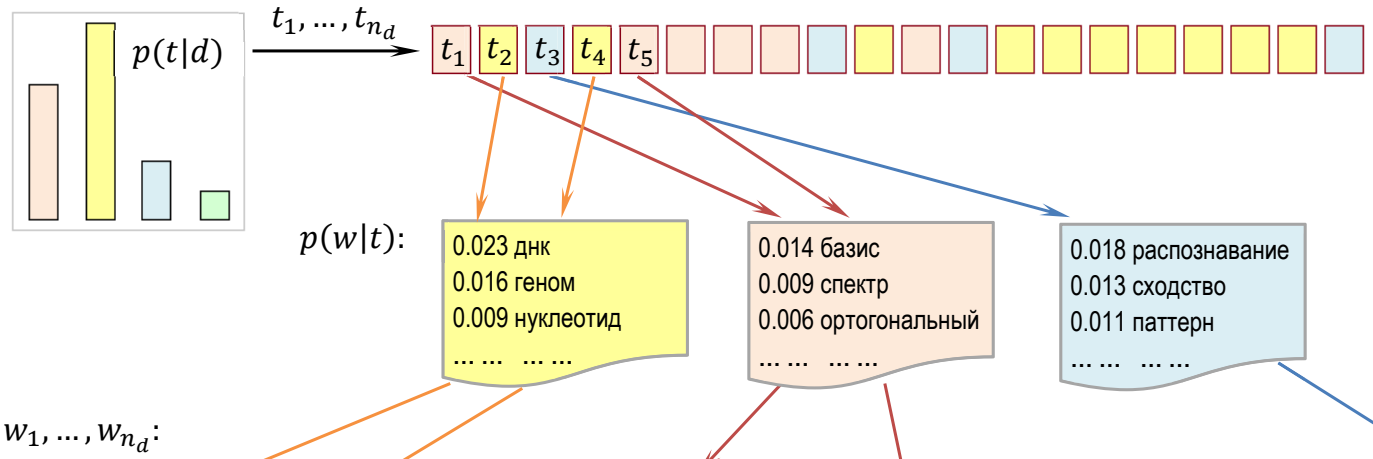
$$p(w|d) = \sum_{t \in T} p(w|d,t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

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

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

Вероятностная модель порождения документа

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).

Принцип максимума правдоподобия

Правдоподобие — это плотность распределения выборки D:

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

где n_{dw} — число вхождений термина w в документ d.

Пусть $p(w|d, \alpha)$ — параметрическая вероятностная модель документа d, зависящая от вектора параметров α.

Логарифм правдоподобия выборки D:

$$\log p(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d, \alpha) p(d) \to \max_{\alpha}.$$

Избавимся от $p(d)$, не влияющего на точку максимума:

$$\mathscr{L}(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d, \alpha) \to \max_{\alpha}.$$

Униграммные модели порождения текстовых документов

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{dw} \to \max_{\xi}, \quad \sum_{w\in W} \xi_{dw} = 1, \quad \xi_{dw} \geqslant 0.$$

$$\mathscr{L} = \sum_{d\in D}\Big(\sum_{w\in d} n_{dw} \ln \xi_{dw} - \lambda_d\Big(\sum_{w\in W} \xi_{dw} - 1\Big)\Big);$$

$$\frac{\partial \mathscr{L}}{\partial \xi_{dw}} = n_{dw}\frac{1}{\xi_{dw}} - \lambda_d = 0 \quad \Rightarrow \quad \lambda_d = n_d, \quad \xi_{dw} = \frac{n_{dw}}{n_d} \equiv \hat{p}(w|d).$$

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{w} \to \max_{\xi}, \quad \sum_{w\in W} \xi_{w} = 1, \quad \xi_{w} \geqslant 0.$$

$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{w} - \lambda\Big(\sum_{w\in W} \xi_{w} - 1\Big);$$

$$\frac{\partial \mathscr{L}}{\partial \xi_{w}} = n_{w}\frac{1}{\xi_{w}} - \lambda = 0 \quad \Rightarrow \quad \lambda = n, \quad \xi_{w} = \frac{n_{w}}{n} \equiv \hat{p}(w).$$

Вероятностная латентная семантическая модель PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

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

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

Интерпретация №1: минимизация суммарной (по $d \in D$) дивергенции Кульбака–Лейблера между тематическими моделями $p(w|d)$ и униграммными $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$:

$$\mathrm{KL}(\hat{p}\|p) = \sum_{d\in D} n_d \sum_{w\in d} \hat{p}(w|d) \ln \frac{\hat{p}(w|d)}{p(w|d)} \to \min.$$

Вероятностная латентная семантическая модель PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

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

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

Интерпретация №2: стохастическое матричное разложение

$$F \approx \Phi\Theta,$$

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

EM-алгоритм. Элементарная интерпретация

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: частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{dwt}, \qquad n_t = \sum_{w\in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w\in d} n_{dwt}, \qquad n_d = \sum_{t\in T} n_{td}.$$

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

EM-алгоритм. Вывод формулы M-шага для ϕ_{wt}

Лагранжиан задачи максимизации правдоподобия при ограничениях нормировки но без ограничений неотрицательности:

$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T}\lambda_t\left(\sum_{w\in W}\phi_{wt} - 1\right) - \sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right)$$

$$\frac{\partial \mathscr{L}}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} - \lambda_t = 0;$$

$$\sum_{d\in D} n_{dw}\frac{\theta_{td}\phi_{wt}}{p(w|d)} = \lambda_t\phi_{wt} \quad \Rightarrow \quad \lambda_t = \sum_{d\in D}\sum_{w\in W} n_{dw}p(t|d,w);$$

$$\phi_{wt} = \frac{\sum\limits_{d\in D} n_{dw}p(t|d,w)}{\sum\limits_{d\in D}\sum\limits_{w'\in d} n_{dw'}p(t|d,w')} \equiv \frac{n_{wt}}{n_t} \quad \text{для всех } w\in W,\ t\in T.$$

EM-алгоритм. Вывод формулы M-шага для θ_{td}

Лагранжиан задачи максимизации правдоподобия при ограничениях нормировки но без ограничений неотрицательности:

$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T}\lambda_t\left(\sum_{w\in W}\phi_{wt} - 1\right) - \sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right)$$

$$\frac{\partial \mathscr{L}}{\partial \theta_{td}} = \sum_{w\in d} n_{dw}\frac{\phi_{wt}}{p(w|d)} - \mu_d = 0;$$

$$\sum_{w\in d} n_{dw}\frac{\phi_{wt}\theta_{td}}{p(w|d)} = \mu_d\theta_{td} \quad \Rightarrow \quad \mu_d = \sum_{t\in T}\sum_{w\in d} n_{dw}p(t|d,w);$$

$$\theta_{td} = \frac{\sum\limits_{w\in d} n_{dw}p(t|d,w)}{\sum\limits_{w\in d} n_{dw}\sum\limits_{t'\in T} p(t'|d,w)} \equiv \frac{n_{td}}{n_d} \quad \text{для всех } d\in D,\ \ t\in T.$$

Недостатки EM-PLSA и способы их устранения

1. необходимость хранить 3D-матрицу $p(t|d,w)$,
   медленная сходимость на больших коллекциях
   — рациональный EM-алгоритм
2. неединственность и неустойчивость решения,
   на малых коллекциях возможно переобучение
   — регуляризации: сглаживание (LDA), разреживание,
   учёт дополнительной внешней информации
3. нет управления *разреженностью* Φ и Θ, т.к.
   (в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в финале $\phi_{wt} = 0$),
   (в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в финале $\theta_{td} = 0$)
   — регуляризации, постепенное разреживание
4. нет выделения нетематических слов
   — робастные модели, игнорирующие редкие слова

Рациональный EM-алгоритм

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 $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех документов $d \in D$ и всех слов $w \in d$
    $p(t|d,w) = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $t \in T$;
    $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dw} p(t|d,w)$ для всех $t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;

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

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, завышающие вероятности редких терминов и редких тем:

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

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.

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

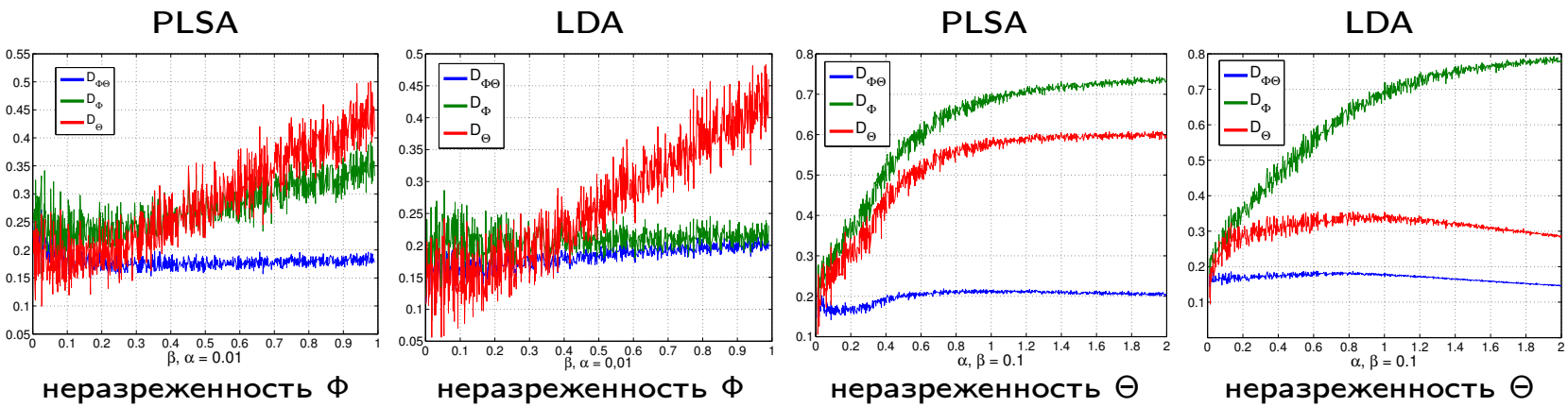
Задача построения BTM — некорректно поставленная

Неединственность стохастического матричного разложения:

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

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

Эксперимент. Произведение $\Phi\Theta$ восстанавливается устойчиво, матрица Φ и матрица Θ — только когда сильно разрежены:



Вывод 1: нужно вводить дополнительные требования.

Вывод 2: требований сглаживания в LDA не достаточно.

ARTM — аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё n критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$ — *регуляризаторов*.

Метод многокритериальной оптимизации — *скаляризация*.

Задача: максимизировать регуляризованное правдоподобие

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

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

где $\tau_i > 0$ — *коэффициенты регуляризации*.

Регуляризованный EM-алгоритм

Теорема

Если Φ, Θ — решение задачи максимизации регуляризованного правдоподобия, то оно удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} \quad n_{dwt} = n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}; \\ \text{M-шаг:} \\ \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \left(\sum\limits_{d \in D} n_{dwt} + \phi_{wt} \dfrac{\partial R}{\partial \phi_{wt}} \right)_+; \quad n_t = \sum\limits_{w \in W} n_{wt}; \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \left(\sum\limits_{w \in d} n_{dwt} + \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} \right)_+; \quad n_d = \sum\limits_{t \in T} n_{td} \end{cases}$$

При $R(\Phi, \Theta) = 0$ это формулы EM-алгоритма для PLSA.

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Доказательство Теоремы о регуляризации M-шага

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

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

2. Умножим обе части равенства на ϕ_{wt} и выделим n_{dwt}:

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

3. Учтём ограничение $\phi_{wt} \geqslant 0$ и предположение $\lambda_t > 0$:

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

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

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

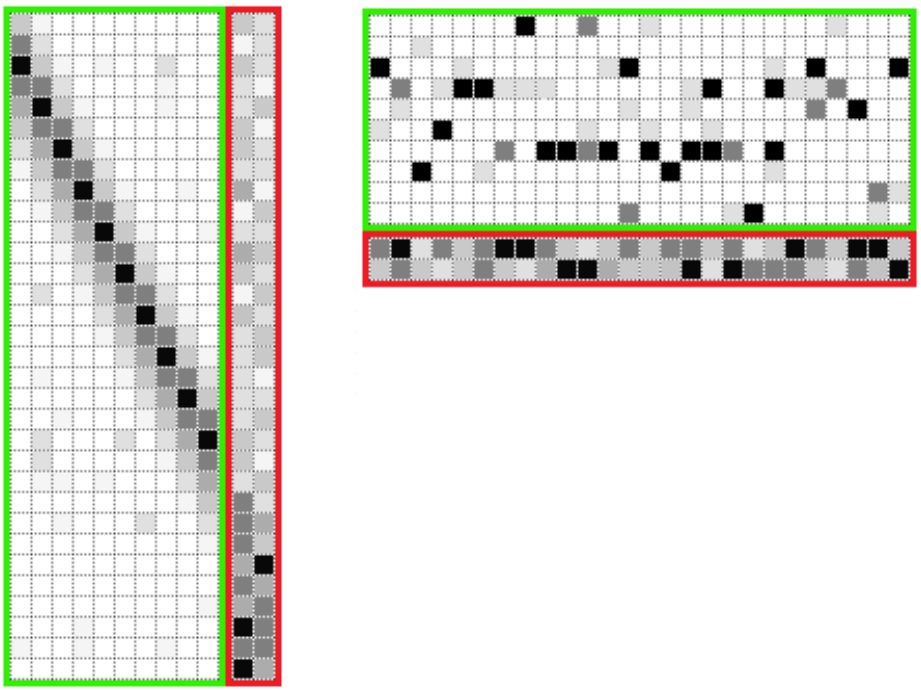
5. Подставим λ_t из (4) в (3), получим требуемое. ■

Требования интерпретируемости и гипотезы о структуре тем

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

Фоновые темы B содержат слова общей лексики, $p(w|t)$ и $p(t|d)$ не разреженные в этих темах

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Напоминания. Дивергенция Кульбака–Лейблера

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

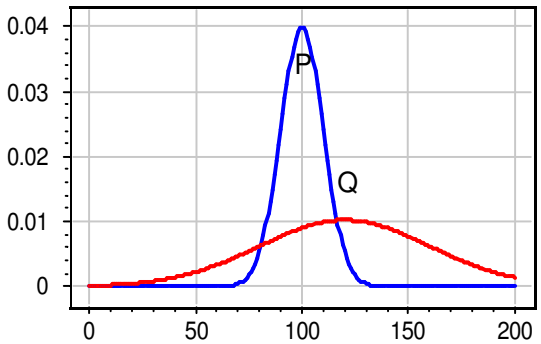
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

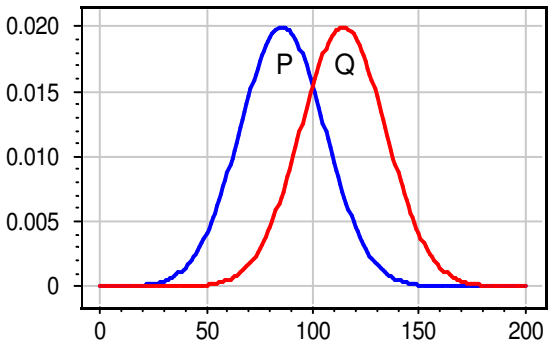
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

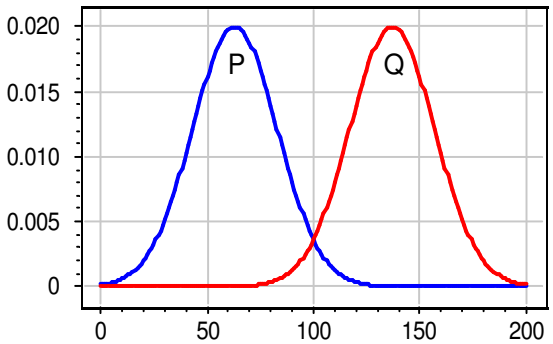
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



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

Регуляризатор сглаживания (почти совпадает с LDA)

Гипотеза сглаженности:

распределения ϕ_{wt} близки к заданным распределениям β_w

распределения θ_{td} близки к заданным распределениям α_t

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

Максимизируем сумму этих регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA, для всех $t \in B$:

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

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.

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

Максимизируем дивергенцию между распределениями β_w, α_t (равномерными?) и искомыми распределениями ϕ_{wt}, θ_{td}:

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

Подставляем, получаем «анти-LDA», для всех $t \in S$:

$$\phi_{wt} \propto \left(n_{wt} - \beta_0\beta_w\right)_+, \qquad \theta_{td} \propto \left(n_{td} - \alpha_0\alpha_t\right)_+.$$

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 декоррелирования тем

Гипотеза: в каждой теме должно быть своё лексическое ядро, отличающее 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:

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

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

Регуляризатор удаления незначимых тем

Гипотеза: если тема собрала 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:

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

Подставляем, получаем:

$$\theta_{td} \propto \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.$$

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

Разреживание + Сглаживание + Декорреляция + Отбор тем

M-шаг при комбинировании 6 регуляризаторов:

$$\phi_{wt} \propto \Big(n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{сглаживание} \\ \text{фоновых} \\ \text{тем}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{разреживание} \\ \text{предметных} \\ \text{тем}}} - \tau_3 \phi_{wt} \underbrace{\sum_{s \in S \setminus t} \phi_{ws}}_{\text{декорреляция}} \Big)_+$$

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

Траектория регуляризации (*regularization path*) в пространстве $\tau = (\tau_1, \ldots, \tau_6)$ подбирается экспериментально в ходе итераций.

Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Межд. конф. по компьютерной лингвистике Диалог-2014.

Резюме

- 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.
- Уточнение постановки задачи с помощью регуляризации приводит к многокритериальной оптимизации.
- Регуляризаторы тематических моделей разнообразны, аддитивная регуляризация позволяет их комбинировать, не сильно изменяя EM-алгоритм.