

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)»

ноябрь 2011

Содержание

- 1** **Задача тематического моделирования**
 - Постановка задачи
 - Основные вероятностные гипотезы
 - Униграммная модель документа
- 2** **Базовые тематические модели**
 - Модель смеси униграмм
 - Вероятностный латентный семантический анализ
 - Латентное размещение Дирихле
- 3** **Оценивание, использование, разновидности т.м.**
 - Оценивание качества тематических моделей
 - Сравнение тематических профилей
 - Обобщения и модификации

Определения и обозначения

Дано:

W — словарь, множество слов (терминов);

D — множество текстовых документов;

каждый $d \in D$ — это последовательность слов из W .

Найти:

T — множество скрытых (латентных) тем;

$p(w|t)$ — распределение на W , задающее тему $t \in T$;

$p(t|d)$ — *тематический профиль* документа, для всех $d \in D$.

Дополнительно:

$p(w|t, y)$ — изменение темы по годам y ;

$p(t|y)$, $p(y|t)$ — распределения тем по годам;

$p(t|a)$, $p(t|a, y)$ — *тематический профиль* автора a ;

$p(t|x)$, $p(t|x, y)$ — *тематический профиль* объекта x , связанного с документами (журнала, конференции, организации, страны);

Цели тематического моделирования (topic modeling)

- Тематический поиск (запрос — тема или документ)
- Рубрикация и классификации текстов
- Поиск экспертов
- Прослеживание фронта исследований
- Аннотация документов
- Суммаризация множества документов

Типичные приложения:

- Анализ коллекций научных статей
- Анализ новостных потоков
- Рубрикация коллекций изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики

Стандартные гипотезы тематического моделирования

- 1 Порядок документов в коллекции не важен
- 2 Порядок слов в документе не важен (bag of words)
- 3 Слова, встречающиеся в большинстве документов, не важны
- 4 Слово в разных формах — это одно и то же слово

Предварительная обработка текстов:

- Приведение всех слов к нормальной форме (стемминг или лемматизация)
- Выделение терминов (term extraction) и выделение словосочетаний (key phrase extraction); (сводятся к задачам классификации или ранжирования)
- Удаление стоп-слов $w \in W$: $\#\{d : w \in d\} \geq \alpha|D|$,
 $\alpha \sim 0.05 \dots 0.5$

Основная вероятностная гипотеза:

коллекция документов — i.i.d. выборка $\{(d, w) : d \in D, w \in d\}$.

Униграммная модель порождения текста

Дополнительная вероятностная гипотеза:

появления слов w в документе d — независимые события.

Вероятностная модель документа:

$$p(d) = \prod_{w \in d} p(w|d)^{n_{dw}},$$

$p(w|d)$ — неизвестное мультиномиальное распределение на W ;

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

Принцип максимума правдоподобия:

$$\ln \prod_{d \in D} \prod_{w \in d} p(w|d)^{n_{dw}} \rightarrow \max_{\{p(w|d)\}} \text{ при } \sum_{w \in W} p(w|d) = 1.$$

Лагранжиан:

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) - \sum_{d \in D} \lambda_d \left(\sum_{w \in W} p(w|d) - 1 \right).$$

Униграммная модель порождения текста

Лагранжиан \mathcal{L} распадается на независимые слагаемые $\mathcal{L}(d)$:

$$\mathcal{L}(d) = \sum_{w \in d} n_{dw} \ln p(w|d) - \lambda_d \left(\sum_{w \in W} p(w|d) - 1 \right).$$

$$\frac{\partial \mathcal{L}}{\partial p(w|d)} = \frac{\partial \mathcal{L}(d)}{\partial p(w|d)} = \frac{n_{dw}}{p(w|d)} - \lambda_d = 0,$$

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

Умножим обе части равенства на $p(w|d)$ и просуммируем по w :

$$\sum_{w \in W} \lambda_d p(w|d) = \sum_{w \in W} n_{dw} \Rightarrow \lambda_d = n_d,$$

где n_d — длина документа d .

Получим (тривиальную) оценку максимума правдоподобия:

$$p(w|d) = \frac{n_{dw}}{n_d}.$$

Униграммная модель порождения текста

Недостатки униграммной модели:

- тематика не выявляется
- число $|W| \cdot |D|$ оцениваемых параметров $p(w|d)$ линейно зависит от $|D|$ — числа документов в коллекции
- зависимости между документами не учитываются

Эти недостатки устраняются в модели смеси униграмм:

Nigam, McCallum, Thrun, Mitchell.

Text classification from labeled and unlabeled documents using EM.

Journal of Machine Learning, 2000, 39(2–3): 103–134

Вероятностные тематические модели порождения текста

темы $t \in T$
 $p(w|t)$

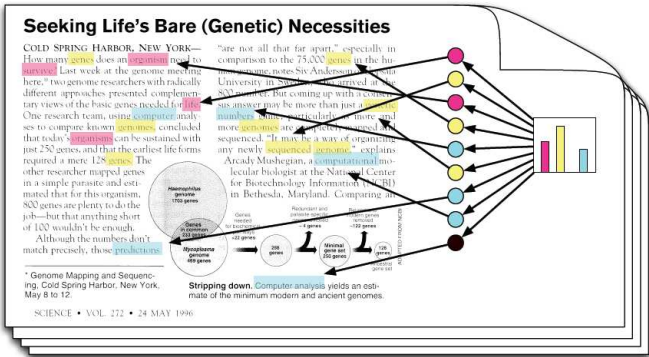
gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

документы $d \in D$



слова

из $p(w|t)$

темы

из $p(t|d)$

Модель смеси униграмм [Nigam и др., 2000]

- 1 Будем описывать униграммной моделью не весь документ d , а только ту его часть, которая относится к теме $t \in T$:

$$p(d|t) = \prod_{w \in d} p(w|t, d)^{n_{dw}}.$$

- 2 Гипотеза условной независимости: $p(w|t, d) = p(w|t)$ (распределения слов связаны с темами, а не с документами)
- 3 Документ — это смесь униграмм:

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

Тематическая модель смеси униграмм:

$$p(d) = \sum_{t \in T} p(t) \prod_{w \in d} p(w|t)^{n_{dw}}.$$

Преимущества и недостатки модели смеси униграмм

Преимущества:

- Модель позволяет выявлять тематику
- *Тематический профиль* каждого документа выражается через параметры модели $p(t)$, $p(w|t)$ по формуле Байеса:

$$p(t|d) = \frac{p(t)p(d|t)}{p(d)} = \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}}$$

- Число параметров $|T| + |W| \cdot |T|$ не зависит от $|D|$

Недостатки:

- Наивное байесовское предположение о независимости слов:

$$p(w_1, \dots, w_n|t) = p(w_1|t), \dots, p(w_n|t)$$

Обучение модели смеси униграмм

Принцип максимума правдоподобия:

$$\ln \prod_{d \in D} p(d) = \ln \prod_{d \in D} \sum_{t \in T} p(t) \prod_{w \in d} p(w|t)^{n_{dw}} \rightarrow \max_{\{p(t), p(w|t)\}};$$

$$\text{при } \sum_{w \in W} p(w|t) = 1, t \in T; \quad \sum_{t \in T} p(t) = 1.$$

Лагранжиан:

$$\begin{aligned} \mathcal{L} = & \sum_{d \in D} \ln \underbrace{\sum_{t \in T} p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}_{p(d)} - \\ & - \sum_{t \in T} \lambda_t \left(\sum_{w \in W} p(w|t) - 1 \right) - \mu \left(\sum_{t \in T} p(t) - 1 \right). \end{aligned}$$

Оценка МП для $p(t)$

$$\mathcal{L} = \sum_d \ln \underbrace{\sum_t p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}_{p(d)} - \sum_t \lambda_t \left(\sum_w p(w|t) - 1 \right) - \mu \left(\sum_t p(t) - 1 \right).$$

$$\frac{\partial \mathcal{L}}{\partial p(t)} = \sum_{d \in D} \frac{1}{p(d)} p(d|t) - \mu = 0.$$

Умножим обе части равенства на $p(t)$ и просуммируем по t :

$$\mu \sum_{t \in T} p(t) = \sum_{d \in D} \sum_{t \in T} \frac{p(d|t)p(t)}{p(d)} \Rightarrow \mu = |D|.$$

Если умножить, но не суммировать:

$$\mu p(t) = \sum_{d \in D} \frac{p(d|t)p(t)}{p(d)} = \sum_{d \in D} p(t|d) \Rightarrow p(t) = \frac{\sum_{d \in D} p(t|d)}{|D|}.$$

Оценка МП для $p(w|t)$

$$\mathcal{L} = \sum_d \ln \underbrace{\sum_t p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}_{p(d)} - \sum_t \lambda_t \left(\sum_w p(w|t) - 1 \right) - \mu \left(\sum_t p(t) - 1 \right).$$

$$\frac{\partial \mathcal{L}}{\partial p(w|t)} = \sum_{d \in D} \frac{1}{p(d)} n_{dw} \frac{p(d|t)}{p(w|t)} - \lambda_t = 0.$$

Умножим обе части равенства на $p(t)p(w|t)$, просуммируем по w :

$$\lambda_t p(t) \sum_w p(w|t) = \sum_{d,w} n_{dw} \frac{p(d|t)p(t)}{p(d)} \Rightarrow \lambda_t p(t) = \sum_{d,w} n_{dw} p(t|d).$$

Если умножить, но не суммировать:

$$\lambda_t p(t) p(w|t) = \sum_{d \in D} n_{dw} p(t|d) \Rightarrow p(w|t) = \frac{\sum_{d \in D} n_{dw} p(t|d)}{\sum_{d \in D} \sum_{u \in W} n_{du} p(t|d)}.$$

EM-алгоритм (оценки максимума правдоподобия)

- Инициализировать $p(t|d)$;
- M-шаг: оценить параметры модели

$$p(t) := \frac{\sum_{d \in D} p(t|d)}{|D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{\sum_{d \in D} p(t|d) n_{dw}}{\sum_{d \in D} p(t|d) \sum_{u \in d} n_{du}} \quad \text{для всех } w \in W, d \in D;$$

- E-шаг: вычислить скрытые профили документов

$$p(t|d) := \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D, t \in T;$$

EM-алгоритм (байесовские оценки, maximum a posteriori)

- Инициализировать $p(t|d)$;
- M-шаг: оценить параметры модели

$$p(t) := \frac{1 + \sum_{d \in D} p(t|d)}{|T| + |D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{1 + \sum_{d \in D} p(t|d)n_{dw}}{|W| + \sum_{d \in D} p(t|d) \sum_{u \in d} n_{du}} \quad \text{для всех } w \in W, d \in D;$$

- E-шаг: вычислить скрытые профили документов

$$p(t|d) := \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D, t \in T;$$

Лирическое отступление: чем отличаются ML и MAP?

Задача:

оценить параметр θ распределения $p(x|\theta)$ по выборке $\{x_i\}_{i=1}^n$.

Метод максимума правдоподобия (ML, maximum likelihood)

$$\prod_{i=1}^n p(x_i|\theta) \rightarrow \max_{\theta};$$

$$\sum_{i=1}^n \log p(x_i|\theta) \rightarrow \max_{\theta};$$

Метод максимума апостериорной вероятности (MAP, maximum a posteriori probability)

$$\prod_{i=1}^n p(x_i|\theta) p(\theta) \rightarrow \max_{\theta};$$

$$\sum_{i=1}^n \log p(x_i|\theta) + \log p(\theta) \rightarrow \max_{\theta};$$

Инициализация $p(t|d)$ и частичное обучение

1. Использование априорной классификации документов
 $y_{dt} = [\text{документ } d \text{ относится к теме } t]$:

$$p(t|d) := \frac{y_{dt}}{\sum_{s \in T} y_{ds}}.$$

2. Если размечена только часть коллекции $D^\ell \subset D$, то
 - на первом M-шаге суммируем только по $d \in D^\ell$,
 - на первом E-шаге вычисляем $p(t|d)$ для всех $d \in D$;
 - далее EM-алгоритм выполняется как обычно, на E-шаге вычисляем $p(t|d)$ для всех $d \in D$ либо $D \setminus D^\ell$.
3. Если $|D^\ell| \ll |D|$, то вес неразмеченных документов $\lambda < 1$:

$$\Lambda_d = \begin{cases} 1, & d \in D^\ell; \\ \lambda, & d \notin D^\ell. \end{cases}$$

EM-алгоритм (с частичным обучением)

- Инициализировать $p(t|d)$ для всех $d \in D^\ell$, $t \in T$;
- M-шаг: оценить параметры модели

$$p(t) := \frac{1 + \sum_{d \in D} \Lambda_d p(t|d)}{|T| + |D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{1 + \sum_{d \in D} \Lambda_d p(t|d) n_{dw}}{|W| + \sum_{d \in D} \Lambda_d p(t|d) \sum_{u \in d} n_{du}} \quad \text{для всех } w \in W, d \in D;$$

- E-шаг: вычислить скрытые профили документов

$$p(t|d) := \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D, t \in T;$$

Замечание

В обзорах часто называется недостаток модели смеси униграмм: «каждый документ относится только к одной теме».

Это действительно так при использовании другой техники вывода формул М-шага (через неравенство Йенсена), однако мы получили те же формулы без этого предположения!

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.

Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(имеется русский перевод)

Вероятностный латентный семантический анализ Probabilistic Latent Semantic Analysis, PLSA [Hofmann, 1999]

- Вероятностная модель документа:

$$p(d) = \prod_{w \in d} p(d, w)^{n_{dw}}$$

- Гипотеза условной независимости: $p(w|t, d) = p(w|t)$
(распределения слов связаны с темами, а не с документами)
- Модель смеси распределений для пар (d, w) :

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

Задача максимизации правдоподобия по $p(t)$, $p(d|t)$, $p(w|t)$:

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(d, w) \rightarrow \max$$
$$\sum_{t \in T} p(t) = 1; \quad \sum_{d \in D} p(d|t) = 1; \quad \sum_{w \in W} p(w|t) = 1.$$

Особенности, преимущества, недостатки модели PLSA

- Симметричность модели относительно $d \Leftrightarrow w$:

$$\begin{aligned} p(d, w) &= \sum_{t \in T} p(t) p(d|t) p(w|t); \\ &= \sum_{t \in T} p(d) p(t|d) p(w|t); \\ &= \sum_{t \in T} p(w) p(t|w) p(d|t); \end{aligned}$$

- Тематические профили вычисляются по формуле Байеса:

$$p(t|d) = \frac{p(d|t) p(t)}{\sum_{s \in T} p(d|s) p(s)}; \quad p(t|w) = \frac{p(w|t) p(t)}{\sum_{s \in T} p(w|s) p(s)}.$$

- Нет наивного байесовского предположения
 $p(w_1, \dots, w_n|t) = p(w_1|t), \dots, p(w_n|t)$
- Число параметров $|D||T| + |W||T|$, возможно переобучение!

Максимизация правдоподобия: EM-алгоритм

Сформировать начальные приближения $p(t)$, $p(d|t)$, $p(w|t)$;
Повторять итерации до сходимости:

- **E-шаг:** скрытые переменные H по формуле Байеса:

$$H(t|d, w) = \frac{p(t)p(d|t)p(w|t)}{p(d, w)};$$

- **M-шаг:** аналитическое решение задачи $\mathcal{L} \rightarrow \max$:

$$p(t) = \frac{S(t)}{S}; \quad S(t) = \sum_{d,w} n_{dw} H(t|d, w); \quad S = \sum_{d,w} n_{dw};$$

$$p(d|t) = \frac{1}{S(t)} \sum_w n_{dw} H(t|d, w);$$

$$p(w|t) = \frac{1}{S(t)} \sum_d n_{dw} H(t|d, w).$$

Вывод формул M-шага

Распишем Лагранжиан:

$$\mathcal{L} = \sum_{d,w} n_{dw} \ln p(d, w) - \nu \left(\sum_{t \in T} p(t) - 1 \right) -$$

$$- \sum_{t \in T} \lambda_t \left(\sum_{d \in D} p(d|t) - 1 \right) - \sum_{t \in T} \mu_t \left(\sum_{w \in W} p(w|t) - 1 \right).$$

$$\frac{\partial \mathcal{L}}{\partial p(t)} = \sum_{d,w} n_{dw} \frac{p(d|t)p(w|t)}{p(d, w)} - \nu = 0;$$

$$\sum_{d,w} n_{dw} \frac{p(d|t)p(w|t)p(t)}{p(d, w)} = \nu p(t) \Rightarrow \nu = \sum_{d,w} n_{dw} = S;$$

$$p(t) = \frac{1}{S} \sum_{d,w} n_{dw} H(t|d, w) = \frac{S(t)}{S};$$

Вывод формул M-шага (продолжение)

$$\frac{\partial \mathcal{L}}{\partial p(d|t)} = \sum_w n_{dw} \frac{p(t)p(w|t)}{p(d, w)} - \lambda_t = 0;$$

$$\sum_w n_{dw} \frac{p(t)p(w|t)p(d|t)}{p(d, w)} = \lambda_t p(d|t) \Rightarrow \lambda_t = \sum_{d,w} n_{dw} H(t|d, w);$$

$$p(d|t) = \frac{\sum_w n_{dw} H(t|d, w)}{\sum_{d,w} n_{dw} H(t|d, w)}.$$

$$\frac{\partial \mathcal{L}}{\partial p(w|t)} = \sum_d n_{dw} \frac{p(t)p(d|t)}{p(d, w)} - \mu_t = 0;$$

$$\sum_d n_{dw} \frac{p(t)p(d|t)p(w|t)}{p(d, w)} = \mu_t p(w|t) \Rightarrow \mu_t = \sum_{d,w} n_{dw} H(t|d, w);$$

$$p(w|t) = \frac{\sum_d n_{dw} H(t|d, w)}{\sum_{d,w} n_{dw} H(t|d, w)}.$$

Замечания о методе PLSA

- 1 Как быстро оценить профиль нового документа (folding-in):

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

где $p(w|d) = n_{dw}/n_d$ — оценка униграммной модели.

- 2 Меры по уменьшению переобучения:
 - обнуление незначимых компонент профилей $p(t|d)$, $p(t|w)$ (сокращение числа параметров модели);
 - обнуление незначимых скрытых переменных $H(t|d, w)$;
 - аккуратное формирование начального приближения;
 - ранний останов (оптимизация числа итераций);
 - симметризованный EM-алгоритм, в котором профили $p(t|w)$ и $p(t|d)$ уточняются по очереди.

Латентное размещение Дирихле

Latent Dirichlet Allocation [David Blei, 2003]

Вероятностная тематическая модель:

$$p(d) = \prod_{w \in d} p(d, w)^{n_{dw}}; \quad p(d, w) = \sum_{t \in T} p(w|t) \underbrace{p(t|d)}_{\theta_{dt}} p(d).$$

Обозначим $\theta_{dt} = p(t|d)$.

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

$$p(\theta|\alpha) = \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_t^{\alpha_t - 1}.$$

Тогда тематическая модель примет вид:

$$p(d) = \int p(\theta_d|\alpha) \prod_{w \in d} \left(\sum_{t \in T} p(w|t) \theta_{dt} \right)^{n_{dw}} d\theta_d$$

Лирическое отступление: почему именно Дирихле?

Рассмотрим формулу Байеса:

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

Хотим, чтобы:

- априорное распределение $p(t)$ и апостериорное $p(t|d)$ были из одного параметрического семейства;
- $p(d|t)$ было мультиномиальным распределением.

Этим требованиям удовлетворяет параметрическое распределение, сопряжённое с мультиномиальным — распределение Дирихле:

$$p(t), p(t|d) \sim \text{Dir}(\alpha), \quad \alpha \in \mathbb{R}^{|T|}.$$

Оценивание параметров и вывод профилей

- Задача оценивания параметров в модели LDA:

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

где α — параметр Дирихле, $\beta = (p(w|t))_{|W| \times |T|}$.

- Тематический профиль документа d :

$$p(t|d) = \frac{p(t, d)}{p(d)} = \frac{\int p(\theta_d | \alpha) \prod_{w \in d} (p(w|t)\theta_{dt})^{n_{dw}} d\theta_d}{\int p(\theta_d | \alpha) \prod_{w \in d} \left(\sum_{t \in T} p(w|t)\theta_{dt} \right)^{n_{dw}} d\theta_d}.$$

Методы оценивания параметров α , β и профилей $p(t|d)$:

- Сэмплирование Гиббса (Gibbs sampling).
- Вариационный вывод (variational inference).

Особенности, преимущества, недостатки LDA

- Модель порождения документов d общая для всей коллекции D , а не отдельная для каждого $d \in D$.
- Число параметров $|T| + |T| \cdot |W|$ не зависит от $|D|$.
- Профиль нового документа $p(t|d)$ оценивается по той же модели, что и для всех документов обучающей коллекции.
- Несмотря на устрашающий вид модели, существуют эффективные методы оценивания параметров α , β и профилей $p(t|d)$.

David Blei, Andrew Ng, Michael Jordan.

Latent Dirichlet allocation.

Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Эффективная реализация — в проекте Vowpal Wabbit:

<http://hunch.net/~vw/>

Сэмплирование Гиббса [Griffiths, Steyvers, 2004]

Тематическая модель вхождения слова w в документ d :

$$p(t|d, w) = \frac{p(d)}{p(d, w)} p(w|t) p(t|d) \propto \frac{p(w, t)}{p(t)} \cdot \frac{p(d, t)}{p(d)}.$$

Оценка максимума правдоподобия:

$$\hat{p}(t|d, w) = \frac{C_{wt}}{\sum_{w' \in W} C_{w't}} \cdot \frac{C_{dt}}{\sum_{t' \in T} C_{dt'}}$$

C_{wt} — сколько раз слово w было отнесено к теме t ;

C_{dt} — сколько раз слово документа d было отнесено к теме t ;

Обобщение — байесовская оценка (maximum a posteriori):

$$\hat{p}(t|d, w) = \frac{\beta + C_{wt}}{\beta |W| + \sum_{w' \in W} C_{w't}} \cdot \frac{\alpha + C_{dt}}{\alpha |T| + \sum_{t' \in T} C_{dt'}}$$

Рекомендуемые параметры $\alpha = 50/|T|$, $\beta = 0.01$.

Сэмплирование Гиббса: основные идеи

- 1 возьмём за основу байесовскую эмпирическую оценку:

$$\hat{p}(t|d, w) = \frac{\beta + C_{wt}}{\beta|W| + \sum_{w' \in W} C_{w't}} \cdot \frac{\alpha + C_{dt}}{\alpha|T| + \sum_{t' \in T} C_{dt'}}$$

- 2 для каждой пары (d, w) в коллекции выберем тему t из распределения $\hat{p}(t|d, w)$ и обновим счётчики C_{wt} , C_{dt} .
- 3 пробежим много раз, чтобы оценка стала устойчивой.
- 4 инициализация $C_{dt} := \alpha$ эквивалентна предположению об априорном распределении Дирихле $p(t) \sim \text{Dir}(\alpha, \dots, \alpha)$.
- 5 инициализация $C_{wt} := \beta$ эквивалентна предположению об априорном распределении Дирихле $p(w) \sim \text{Dir}(\beta, \dots, \beta)$.

Сэмплирование Гиббса: Алгоритм

Вход: Коллекция документов $\{(d, w): d \in D, w \in d\}$;
параметры α, β ;

Выход: Оценки $\hat{p}(w|t), \hat{p}(t|d)$;

-
- 1: $C_{wt} := \beta; C_{dt} := \alpha$ для всех $d \in D, w \in W, t \in T$;
 - 2: $C_t := \beta|W|; C_d := \alpha|T|$ для всех $d \in D, t \in T$;
 - 3: **для** $i := 1, \dots, N$ — число итераций сэмплирования
 - 4: **для** всех документов $d \in D$ и всех слов $w \in d$
 - 5: $\hat{p}(t|d, w) := (C_{wt}/C_t) \cdot (C_{dt}/C_d)$ для всех $t \in T$;
 - 6: выбрать t из распределения $\hat{p}(t|d, w)$;
 - 7: $C_{wt} := C_{wt} + 1; C_t := C_t + 1$;
 $C_{dt} := C_{dt} + 1; C_d := C_d + 1$;
 - 8: $\hat{p}(w|t) := C_{wt}/C_t$ для всех $w \in W, t \in T$;
 $\hat{p}(t|d) := C_{dt}/C_d$ для всех $d \in D, t \in T$;

Сэмплирование Гиббса: особенности, преимущества, недостатки

- Алгоритм прост и эффективен, время $O(NL)$, где L — длина коллекции (в словах), N — число сэмплов;
- Добавление нового документа (возможно в онлайн!) — тот же алгоритм, но без инициализации счётчиков.

Griffiths, Steyvers. Finding scientific topics // Proceedings of the National Academy of Sciences. USA, 2004. — Vol. 101. — Pp. 5228–5235.

- Более современные вариационные методы ещё быстрее (CVB — Collapsed Variational Bayesian);

Teh, Newman, Wellingm. A collapsed variational Bayesian inference algorithm for Latent Dirichlet Allocation // Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2006

Методы оценивания качества тематических моделей

- На размеченной тестовой коллекции D' :
 - число ошибок классификации (чем меньше, тем лучше).
- На неразмеченной тестовой коллекции D' :
 - perplexity, неопределённость (чем меньше, тем лучше):

$$\text{perplexity} = \exp \left(- \frac{\sum_{d \in D'} \ln p(d)}{\sum_{d \in D'} \sum_{w \in d} n_{dw}} \right).$$

- Posterior Predictive Checking

David Mimno, David Blei. Bayesian Checking for Topic Models // Empirical Methods in Natural Language Processing, 2011.

Методы сравнения тематических профилей

При классификации и поиске документов необходимо оценивать сходство документов по профилям $p(t|d)$, $p(t|d')$

- Среднеквадратичное отклонение не подходит!
- Расстояние Кульбака-Лейблера:

$$\text{KL}(d, d') = \sum_{t \in T} p(t|d) \log \frac{p(t|d)}{p(t|d')}$$

(не симметрично, поэтому также не рекомендуется).

- Симметризованное расстояние Кульбака-Лейблера:

$$\text{SymKL}(d, d') = \frac{1}{2}\text{KL}(d, d') + \frac{1}{2}\text{KL}(d', d).$$

- Симметризованное расстояние Йенсена-Шеннона:

$$\text{SymJS}(d, d') = \frac{1}{2}\text{KL}(d, \bar{d}) + \frac{1}{2}\text{KL}(d', \bar{d}), \quad \bar{d} = \frac{1}{2}(d + d').$$

Обобщения и модификации тематических моделей

- Иерархические модели, в том числе с адаптивной иерархией
- Темпоральные модели, учитывающие годы публикаций
- Author-topic models — пытаются приписать распределение авторов $p(a|w)$ каждому слову документа
- Entity-topic models — оценивают тематику авторов, журналов, конференций, организаций, стран
- Модели, учитывающие связь слов внутри документа
- Модели, связи между документами (ссылки, цитирование)

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.

Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(имеется русский перевод)

Topic Modeling Bibliography:

<http://www.cs.princeton.edu/~mimno/topics.html>