

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 и применение тематических моделей**
 - Сравнение тематических профилей
 - Оценивание качества тематических моделей
 - Применение тематических моделей

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:

$p(w|t, y)$, $p(t|y)$, $p(y|t)$ — изменения тематики 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 «почти во всех» документах, 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

Вероятностная модель документа d — это вероятность реализации цепочки слов документа $W_d = \{w_1, \dots, w_{n_d}\}$:

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

$p(w|d)$ — неизвестное мультиномиальное распределение на W ;
 n_{dw} — число вхождений слова w в документ d ;
 n_d — длина документа d .

Принцип максимума правдоподобия:

$$\ln \prod_{d \in D} p(W_d) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \rightarrow \max_{\{p(w|d)\}}$$

при ограничении типа равенства

$$\sum_{w \in W} p(w|d) = 1.$$

Униграммная модель порождения текста

Лагранжиан \mathcal{L} — сумма независимых слагаемых по документам:

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

$$\frac{\partial \mathcal{L}}{\partial p(w|d)} = \frac{\partial \mathcal{L}(d)}{\partial p(w|d)} = n_{dw} \frac{1}{p(w|d)} - \lambda_d = 0.$$

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

Модель смеси униграмм [Nigam и др., 2000]

- 1 Документ $W_d = \{w_1, \dots, w_{n_d}\}$ — это смесь униграмм:

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

- 2 Документы, относящиеся к разным темам $t \in T$, описываются разными униграммными моделями:

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

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

Тематическая модель смеси униграмм:

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

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:

- Все слова документа относятся к одной и той же теме t :

$$p(W_d|t) \equiv p(w_1, \dots, w_{n_d}|t) = p(w_1|t) \cdots p(w_{n_d}|t)$$

Обучение модели смеси униграмм

Принцип максимума правдоподобия:

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

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

Лагранжиан:

$$\mathcal{L} = \sum_{d \in D} \ln \underbrace{\sum_{t \in T} p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}_{p(d)} -$$

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

Оценка максимального правдоподобия для $p(t)$

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

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

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 максимального правдоподобия 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 (оценки максимума апостериорной вероятности)

- 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: оценки ML и MAP

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

- 1 Метод *максимума правдоподобия* (ML, maximum likelihood):

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

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

- 2 Метод *максимума апостериорной вероятности* (MAP, maximum a posteriori probability):

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

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

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

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

Основные недостатки униграммной модели

Предположение, что для каждого документа $d = \{w_1, \dots, w_{n_d}\}$

$$p(W_d|t) \equiv p(w_1, \dots, w_{n_d}|t) = p(w_1|t) \cdots p(w_{n_d}|t)$$

фактически означает, что

- все слова документа относятся к одной и той же теме t ;
- внутри темы t слова появляются независимо.

Это довольно сильные ограничения. Они снимаются в последующих вероятностных тематических моделях.

Вероятностная тематическая модель порождения текстов

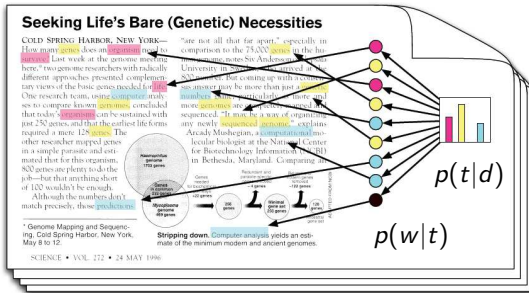
- 1: задать распределение $p(w|t)$ для каждой темы $t \in T$;
- 2: задать распределение $p(t|d)$ для каждого документа $d \in D$;
- 3: **для всех** документов $d \in D$
- 4: **для всех** слов w в документе d
- 5: выбрать тему t из $p(t|d)$;
- 6: выбрать слово w из $p(w|t)$;

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
...	



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

- **Вероятностная модель коллекции документов:**

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

(теперь каждое слово связано со своей темой).

- **Гипотеза условной независимости:** $p(w|d, t) = p(w|t)$
(распределения слов связаны с темами, а не с документами)

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

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(d, w) \rightarrow \max$$

при ограничениях нормировки

$$\sum_{t \in T} p(t) = 1; \quad \sum_{d \in D} p(d|t) = 1; \quad \sum_{w \in W} p(w|t) = 1.$$

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
 $p(w_1, \dots, w_{n_d}|t) = p(w_1|t) \cdots p(w_{n_d}|t)$.
- 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 Оценка профиля 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:

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

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

$$p(\theta|\alpha) = \text{Dir}(\theta; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_t^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \quad \sum_t \theta_t = 1.$$

Тематическая модель документа d как цепочки слов

$W_d = \{w_1, \dots, w_{n_d}\}$ — непрерывная смесь распределений $p(d|\theta_d)$:

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

Почему именно Дирихле?

Пусть профили документов $\theta_d = (p(t|d))_{t \in T}$ выбираются из Dir:

$$p(\theta_d | \alpha) = \text{Dir}(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad d \in D,$$

затем темы слов в документах $d \in D$ выбираются из θ_d :

$$X_d = \{t_1, \dots, t_{n_d}\} \sim \theta_d.$$

Тогда вероятность встретить каждую из тем t ровно n_{td} раз подчиняется мультиномиальному распределению:

$$p(X_d | \theta_d) = \text{Mult}(n_{1d}, \dots, n_{Td} | \theta_d) = \frac{n_d!}{\prod_t n_{td}!} \prod_t \theta_{td}^{n_{td}},$$

при этом апостериорное распределение — снова Дирихле!, т. е. **распределение Дирихле — сопряжённое к мультиномиальному**:

$$p(\theta_d | X_d, \alpha) = \frac{p(X_d | \theta_d) p(\theta_d | \alpha)}{\int p(X_d | \theta) p(\theta | \alpha) d\theta} = \text{Dir}(\theta_d; \alpha'), \quad \alpha'_t = \alpha_t + n_{td}.$$

Свойство сопряженности упрощает оценивание параметров

Апостериорный профиль документа d после того, как стала известна выборка тем $X_d = \{t_1, \dots, t_{n_d}\} \sim \theta_d$:

$$\begin{aligned} p(t|d, X_d, \alpha) &= \int p(t|d)p(\theta_d|X_d) d\theta_d = \\ &= \int \theta_{td} \text{Dir}(\theta_d, \alpha') d\theta_d = \frac{n_{td} + \alpha_t}{\sum_{t'} (n_{t'd} + \alpha_{t'})} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}, \end{aligned}$$

n_{td} — сколько раз слово документа d было отнесено к теме t ,
 n_d — длина документа в словах.

Это совпадает с оценкой максимума апостериорной вероятности для профиля $p(t|d)$ документа d

и совпадает с оценкой максимума правдоподобия при $\alpha_t \equiv 0$.

Аналогично, для распределений слов $p(w|t)$ в темах

Вероятностная тематическая модель:

$$p(d, w) = \sum_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}} p(d).$$

Гипотеза: $\phi_t = (\phi_{wt})_{w \in W} \in \mathbb{R}^{|W|} \sim \text{Dir}(\phi; \beta)$.

Документ порождается двумя распределениями $p(t|d)$, $p(w|t)$

$$X_d = \{(t_i, w_i)\}_{i=1}^{n_d}, \quad t_i \sim \theta_d, \quad w_i \sim \phi_{t_i}.$$

Сопряжённое апостериорное распределение — снова Дирихле:

$$p(\phi_t | X_d, \beta) = \frac{p(X_d | \phi_t) p(\phi_t | \beta)}{\int p(X_d | \phi) p(\phi | \beta) d\phi} = \text{Dir}(\phi_t; \beta'), \quad \beta'_w = \beta_w + n_{wt}.$$

Апостериорное распределение слов в теме t :

$$p(w|t, X_d, \beta) = \int \phi_{tw} \text{Dir}(\phi_t, \beta') d\phi_t = \frac{n_{wt} + \beta_w}{n_t + \beta_0}.$$

Сэмплирование Гиббса [Griffiths, Steyvers, 2004]

Тематическая модель вхождения слова w в документ d :

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

Апостериорные (ненормированные) оценки $p(w|t)$ и $p(t|d)$:

$$\tilde{p}(t|d, w) = \frac{n_{wt} + \beta_w}{n_t + \beta_0} \cdot \frac{n_{td} + \alpha_t}{n_d + \alpha_0},$$

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

n_{td} — сколько раз слово документа d было отнесено к теме t ;

n_t — сколько раз слово было отнесено к теме t ;

n_d — длина документа в словах.

Идея сэмплирования: для всех слов в коллекции (d, w) выберем тему $t \sim \tilde{p}(t|d, w)$, обновим счётчики n_{wt} , n_{td} , n_t , n_d , уточним распределение $\tilde{p}(t|d, w)$, и так много раз в цикле.

Алгоритм сэмплирования Гиббса

Вход: коллекция $X = \{(d, w) : d \in D, w \in d\}$; параметры α , β ;

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

-
- 1: $n_{wt} := \beta$; $n_{td} := \alpha$ для всех $d \in D$, $w \in W$, $t \in T$;
 - 2: $n_t := \beta|W|$; $n_d := \alpha|T|$ для всех $d \in D$, $t \in T$;
 - 3: **для** $i := 1, \dots, M$ — генерация M сэмплов
 - 4: **для** всех документов $d \in D$ и всех слов $w \in d$
 - 5: $\tilde{p}(t|d, w) := (n_{wt}/n_t) \cdot (n_{td}/n_d)$ для всех $t \in T$;
 - 6: **если** $i \geq 2$ **то** $t := t_{dw}$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
 - 7: выбрать t из ненормированного распределения $\tilde{p}(t|d, w)$;
 - 8: $t_{dw} := t$; $++n_{wt}$; $++n_{td}$; $++n_t$; $++n_d$;
 - 9: $\hat{p}(w|t) := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
 $\hat{p}(t|d) := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Сэмплирование Гиббса: особенности, преимущества, недостатки

- Алгоритм прост и эффективен, время $O(LM)$, где L — длина коллекции (в словах), M — число сэмплов;
- Добавление нового документа (возможно в онлайн!) — тот же алгоритм, но без инициализации счётчиков.
- Сэмплов иногда достаточно десятка, иногда нужны сотни...

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

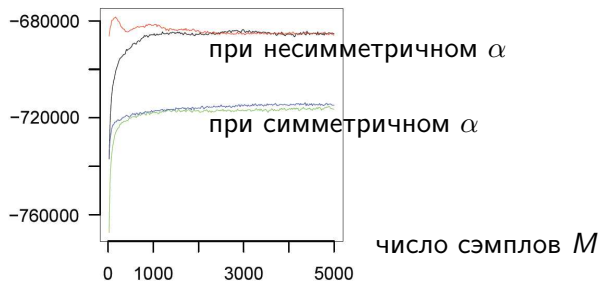
Проблема выбора гиперпараметров α и β

Стандартная рекомендация [2004]: $\alpha_t = 50/|T|$, $\beta_t = 0.01$.

Выводы по результатам более тонкого исследования [2009]:

- $p(t|d) \sim \text{Dir}(\theta; \alpha)$, оптимизировать $\alpha = (\alpha_1, \dots, \alpha_T)$.
- $p(w|t) \sim \text{Dir}(\phi; \beta)$, взять симметричное $\beta_1 = \dots = \beta_T \ll 1$.

правдоподобие



Hanna Wallach, David Mimno, Andrew McCallum.

Rethinking LDA: why priors matter. Neural Information Processing Systems, 2009.

Оптимизация гиперпараметра α

Обоснованность (evidence) модели на коллекции $X = (X_d)_{d \in D}$:

$$\begin{aligned} P(X|\alpha) &= \int P(X|\theta)p(\theta|\alpha) d\theta = \\ &= \prod_{d \in D} \frac{\Gamma(\alpha_0)}{\Gamma(n_d + \alpha_0)} \prod_{t \in T} \frac{\Gamma(n_{td} + \alpha_t)}{\Gamma(\alpha_t)} \rightarrow \max_{\alpha} \end{aligned}$$

Метод неподвижной точки [Minka, 2003] — итерационный процесс, встраиваемый между последовательными сэмплами:

$$\alpha_t := \alpha_t \frac{\sum_d \psi(n_{td} + \alpha_t) - \psi(\alpha_t)}{\sum_d \psi(n_d + \alpha_0) - \psi(\alpha_0)},$$

где $\psi(z) = (\ln \Gamma(z))' = \Gamma'(z)/\Gamma(z)$ — дигамма-функция.

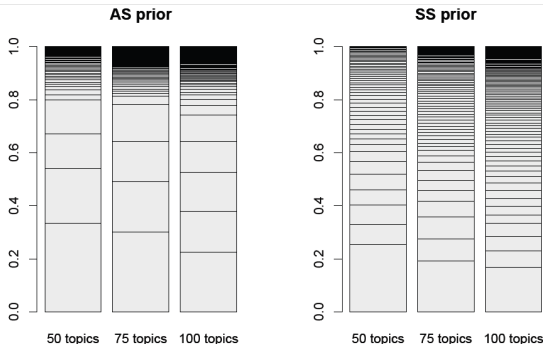
Более быстрые и точные методы оптимизации α :

Hanna Wallach. Structured Topic Models for Language.

PhD thesis, University of Cambridge, 2008.

Преимущество оптимизации гиперпараметра α

- Правдоподобие существенно выше.
- Сходимость быстрее, сэмплов нужно намного меньше.
- Меньшая чувствительность к избыточному $|T|$.
- Меньшее дробление тематики (это хорошо или плохо?):



Преимущества LDA

- Модель порождения документа d общая для всей коллекции D , а не отдельная для каждого $d \in D$.
- Профиль нового документа $p(t|d)$ оценивается по той же модели, что и для всех документов обучающей коллекции.
- Число параметров $|T| + |T| \cdot |W|$ не зависит от $|D|$.
- При $p(w|t) \sim \text{Dir}(\beta)$ число параметров $|T| + |W|$ или $|T| + 1$.
- Существуют эффективные методы оценивания профилей $p(t|d)$, распределений $p(w|t)$ и гиперпараметров α, β .

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/>

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:

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

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):

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

- На обучающей коллекции D : насколько нарушается гипотеза условной независимости $p(w|d, t) = p(w|t)$

$$\text{KL} \left(\hat{p}(d, w|t), \hat{p}(d|t) \cdot \hat{p}(w|t) \right) = \sum_{d, w} \frac{n_{dwt}}{n_t} \log \frac{n_{dwt} \cdot n_t}{n_{td} \cdot n_{wt}}$$

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

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 связанных сущностей: журналов, конференций, организаций
- Модели, учитывающие связь слов внутри документа
- Модели связей 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>