Вероятностные тематические модели
Лекция 9. Байесовский вывод для тематического моделирования

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • весна 2016

Содержание

Вероятностная порождающая модель

D — конечное множество документов
W — конечное множество терминов
T — конечное множество тем
$D \times W \times T$ — вероятностное пространство
$p(d, w, t)$ — распределение в этом пространстве

Исходные данные — наблюдаемые переменные:
$X = (d_i, w_i)_{i=1}^n = (w_{di})_{D \times n_d} = (n_{dw})_{D \times W}$

Скрытые переменные, объясняющие появление данных:
$Z = (t_i)_{i=1}^n$

Параметры вероятностной модели $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$:
$\Omega = (\Phi, \Theta), \quad \phi_{wt} = p(w|t), \quad \theta_{td} = p(t|d)$

Задача: по X найти Ω

Принцип максимума правдоподобия

Пусть скрытые переменные Z известны: $\ln p(X, Z|\Omega) \to \max_{\Omega}$.

Гипотеза независимости элементов выборки $(d_i, w_i, t_i)_{i=1}^n$:

$$p(X, Z|\Omega) = \prod_{i=1}^{n} p(d_i, w_i, t_i|\Omega) = \prod_{d,w,t} p(d, w, t|\Omega)^{n_{dwt}} = \prod_{d,w,t} (\phi_{wt}\theta_{td}p_d)^{n_{dwt}}$$

где $n_{dwt} = \sum_{i=1}^{n} [d_i = d]\,[w_i = w]\,[t_i = t]$, $\quad p_d = p(d) = \frac{n_d}{n}$.

Максимизация логарифма правдоподобия

$$\sum_{d,w,t} n_{dwt} \ln\left(\phi_{wt}\theta_{td}\right) \to \max_{\Phi,\Theta}$$

при ограничениях нормировки и неотрицательности

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

Решение задачи максимизации правдоподобия

Чтобы применить условия ККТ, выписываем лагранжиан:

$$\mathscr{L}(\Phi,\Theta) = \sum_{d,w,t} n_{dwt} \ln \phi_{wt} - \sum_t \lambda_t \Big(\sum_w \phi_{wt} - 1\Big) + \\ + \sum_{d,w,t} n_{dwt} \ln \theta_{td} - \sum_d \mu_d \Big(\sum_t \theta_{td} - 1\Big)$$

Условия ККТ для стационарной точки лагранжиана:

$$\frac{\partial \mathscr{L}}{\partial \phi_{wt}} = \underbrace{\textstyle\sum_d n_{dwt}}_{n_{wt}} \frac{1}{\phi_{wt}} - \lambda_t = 0 \qquad \frac{\partial \mathscr{L}}{\partial \theta_{td}} = \underbrace{\textstyle\sum_w n_{dwt}}_{n_{td}} \frac{1}{\theta_{td}} - \mu_d = 0$$

$$n_{wt} = \lambda_t \phi_{wt} \qquad n_{td} = \mu_d \theta_{td}$$

$$n_t = \lambda_t \qquad n_d = \mu_d$$

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

Промежуточный итог

Если значения скрытых переменных $Z = (t_1, \ldots, t_n)$ известны, то решением задачи максимизации правдоподобия

$$\ln p(X, Z|\Omega) \;\to\; \max_{\Omega}$$

являются частотные оценки условных вероятностей

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

$$\theta_{td} = \frac{n_{td}}{n_d} = \underset{t \in T}{\operatorname{norm}}(n_{td}), \qquad n_{td} = \sum_{t \in T} n_{dwt}.$$

Ну... мы это уже давно знали :)

Тем не менее, это простое решение нам ещё не раз пригодится.
Теперь перейдём к случаю, когда Z не известны.

Максимизация неполного правдоподобия

Проблема — возникает сумма под логарифмом:

$$\ln p(X|\Omega) = \ln \sum_Z p(X, Z|\Omega) \;\to\; \max_\Omega$$

Формула условной вероятности:

$$p(X, Z|\Omega) = p(Z|X, \Omega)p(X|\Omega) \;\Rightarrow\; p(X|\Omega) = \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$$

Для произвольного распределения $q(Z)$

$$\ln p(X|\Omega) = \sum_Z q(Z) \ln p(X|\Omega) = \sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)} =$$

$$= \underbrace{\sum_Z q(Z) \ln p(X, Z|\Omega) - \sum_Z q(Z) \ln q(Z)}_{L(q,\Omega) \text{ — нижняя оценка } \ln p(X|\Omega)} + \underbrace{\sum_Z q(Z) \ln \frac{q(Z)}{p(Z|X,\Omega)}}_{\mathrm{KL}\left(q(Z)\middle\|p(Z|X,\Omega)\right)\geqslant 0}$$

Идея EM-алгоритма. Задача E-шага

Максимизировать нижнюю оценку $L(q,\Omega)$ то по q, то по Ω:

$$\text{E-шаг:}\quad L(q,\Omega) \to \max_q$$
$$\text{M-шаг:}\quad L(q,\Omega) \to \max_\Omega$$

Задача E-шага.

Подставим $p(X,Z|\Omega) = p(Z|X,\Omega)p(X|\Omega)$ в формулу $L(q,\Omega)$:

$$\sum_Z q(Z)\ln p(Z|X,\Omega) + \underbrace{\textstyle\sum_Z q(Z)}_{=1}\underbrace{\ln p(X|\Omega)}_{\text{const по } q} - \sum_Z q(Z)\ln q(Z) \to \max_q$$

$$\mathrm{KL}\big(q(Z)\,\big\|\,p(Z|X,\Omega)\big) \to \min_q$$

Утв. 1. $q(Z) = p(Z|X,\Omega)$ — точное решение задачи E-шага.

Утв. 2. $L(q,\Omega)$ — точная нижняя оценка $\ln p(X|\Omega)$.

EM-алгоритм. Обоснование сходимости

Мы вывели EM-алгоритм для Z и Ω общего вида:

$$\text{E-шаг:} \quad q(Z) = p(Z|X,\Omega)$$

$$\text{M-шаг:} \quad \sum_Z q(Z) \ln p(X,Z|\Omega) \;\to\; \max_\Omega$$

и доказали его *сходимость в слабом смысле:*

- на каждом шаге правдоподобие $\ln p(X|\Omega)$ увеличивается;
- не гарантируется достижение max с заданной точностью;
- не гарантируется глобальная сходимость, так как задача в общем случае многоэкстремальная (на практике важен выбор начального приближения).

N.B. Если скрытая переменная Z не дискретна, а непрерывна, то суммирование $\sum_Z$ заменяется интегрированием $\int_Z$.

Максимизация регуляризованного правдоподобия

$D \times W \times T \times \{\Omega\}$ — вероятностное пространство
$p(\Omega)$ — априорное распределение параметров модели

Принцип максимума апостериорной вероятности:

$$\ln p(X, \Omega) = \ln p(X|\Omega) + \underbrace{\ln p(\Omega)}_{R(\Omega)} \to \max_{\Omega}$$

Регуляризатор $R(\Omega)$ может даже и не иметь вероятностной интерпретации, тем не менее, все выкладки остаются в силе!

$$\text{E-шаг:} \quad q(Z) = p(Z|X, \Omega)$$

$$\text{M-шаг:} \quad \sum_{Z} q(Z) \ln p(X, Z|\Omega) + R(\Omega) \to \max_{\Omega}$$

Регуляризаторы используются для формализации дополнительных требований к вероятностной модели.

Регуляризованный EM-алгоритм для тематической модели

Напоминание: $\Omega = (\Phi, \Theta)$, $X = (d_i, w_i)_{i=1}^n$, $Z = (t_i)_{i=1}^n$.

E-шаг: в силу независимости элементов выборки

$$q(Z) = p(Z|X,\Omega) = \prod_{i=1}^{n} p(t_i|d_i,w_i) = \prod_{i=1}^{n} \underset{t_i}{\text{norm}}\left(\phi_{w_i t_i}\theta_{t_i d_i}\right)$$

M-шаг:

$$\sum_{Z\in T^n} q(Z) \ln p(X,Z|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{(t_1,\ldots,t_n)\in T^n} \prod_{k=1}^{n} p(t_k|d_k,w_k) \sum_{i=1}^{n} \ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n} \sum_{t_1\in T} \cdots \sum_{t_n\in T} \prod_{k=1}^{n} p(t_k|d_k,w_k) \ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n} \sum_{t\in T} p(t|d_i,w_i) \ln p(d_i,w_i,t|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

Регуляризованный EM-алгоритм для тематической модели

... продолжаем вывод формулы M-шага:

$$\sum_{i=1}^{n} \sum_{t \in T} p(t|d_i, w_i) \ln p(d_i, w_i, t|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{d \in D} \sum_{w \in W} \sum_{t \in T} \underbrace{n_{dw} p(t|d, w)}_{\text{обозначим } n_{dwt}} \ln\left(\phi_{wt} \theta_{td}\right) + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}$$

Такую задачу мы уже решали, только без регуляризатора.

Чтобы применить условия KKT, выписываем лагранжиан:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w,t} n_{dwt} \ln \phi_{wt} - \sum_{t} \lambda_t \left(\sum_{w} \phi_{wt} - 1 \right) +$$

$$+ \sum_{d,w,t} n_{dwt} \ln \theta_{td} - \sum_{d} \mu_d \left(\sum_{t} \theta_{td} - 1 \right) + R(\Phi, \Theta)$$

Регуляризованный EM-алгоритм для тематической модели

Условия ККТ для стационарной точки лагранжиана:

$$\underbrace{\sum_d n_{dwt}}_{n_{wt}} \frac{1}{\phi_{wt}} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t \qquad \underbrace{\sum_w n_{dwt}}_{n_{td}} \frac{1}{\theta_{td}} + \frac{\partial R}{\partial \theta_{td}} = \mu_d$$

$$\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+ = \lambda_t \phi_{wt} \qquad \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+ = \mu_d \theta_{td}$$

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \qquad \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$$

Ещё раз вывели формулы ARTM, теперь из общего EM-алгоритма.

Частные случаи:

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

LDA: $R(\Phi, \Theta) = \ln \prod_{t\in T} \text{Dir}(\phi_t|\beta) \prod_{d\in D} \text{Dir}(\theta_d|\alpha).$

Промежуточный итог

Мы узнали более общий вариант ЕМ-алгоритма:

- также снабжённый возможностью регуляризации,
- для которого имеется доказательство слабой сходимости,
- применяемый для построения более сложных моделей,
- используемый также в методах байесовского вывода.

Далее мы рассматриваем *байесовский вывод*:

- Он даёт апостериорные распределения $p(\Omega|X)$,
  хотя в BTM используются только точечные оценки Ω.
- Он намного более громоздкий по сравнению с ARTM,
  хотя в литературе именно он в основном и используется.
- Он претендует на то, чтобы оценивать меньше параметров,
  хотя на деле оценивает те же Φ и Θ, плюс гиперпараметры.

Тематическая модель LDA (Latent Dirichlet Allocation)

Априорное распределение с гиперпараметрами $\alpha \in \mathbb{R}^T$, $\beta \in \mathbb{R}^W$:

$$p(\Phi, \Theta|\alpha, \beta) = \prod_{t \in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d \in D} \mathrm{Dir}(\theta_d|\alpha)$$

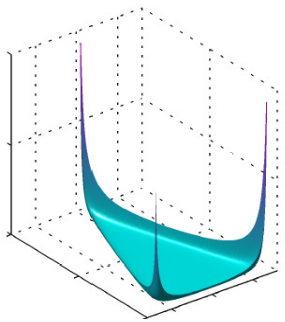
Распределения Дирихле:

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

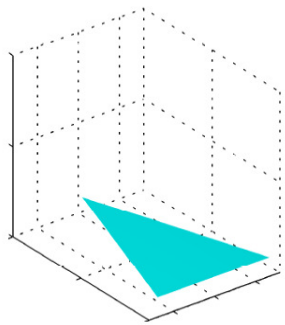
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

Пример:

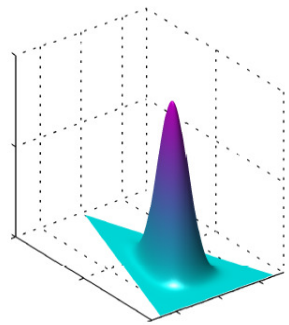
$\mathrm{Dir}(\theta|\alpha)$, $|T| = 3$,
$\theta, \alpha \in \mathbb{R}^3$



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



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



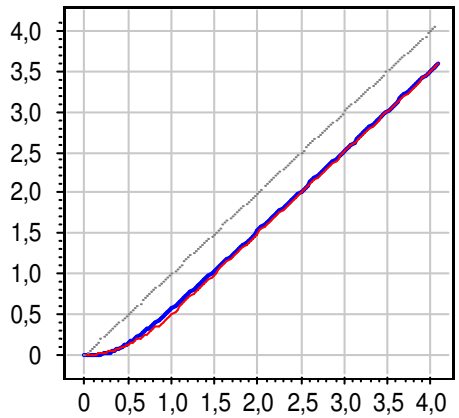
$\alpha_1 = \alpha_2 = \alpha_3 = 10$

Свойства распределения Дирихле

- Математическое ожидание:
  $\mathrm{E}\theta_t = \int \theta_t \mathrm{Dir}(\theta;\alpha)\,d\theta = \frac{\alpha_t}{\alpha_0} = \underset{t}{\mathrm{norm}}(\alpha_t)$
- Мода: $\hat{\theta}_t = \frac{\alpha_t - 1}{\alpha_0 - T} = \underset{t}{\mathrm{norm}}(\alpha_t - 1)$
- Дисперсия: $\mathrm{D}\theta_t = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0+1)}$
- Математическое ожидание логарифма:
  $\mathrm{E}\ln\theta_t = \int \ln\theta_t \mathrm{Dir}(\theta;\alpha)\,d\theta = \psi(\alpha_t) - \psi(\alpha_0)$,
  где $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ — дигамма-функция.

Простая, но очень точная аппроксимация экспоненты от дигамма-функции:

$$E(x) = \exp(\psi(x)) = \begin{cases} \frac{x^2}{2}, & 0 \leqslant x \leqslant 1 \\ x - \frac{1}{2}, & 1 \leqslant x \end{cases}$$



Основные идеи Variational Bayesian inference

EM-алгоритм для скрытых переменных Z и параметров Ω:

$$\text{E-шаг:} \quad \mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X,\Omega)\big) \;\to\; \min_q$$

$$\text{M-шаг:} \quad \sum_Z q(Z) \ln p(X,Z|\Omega) \;\to\; \max_\Omega$$

Идея 1. Если $q(Z) = p(Z|X,\Omega)$ не вычисляемо, то ищем $q(Z)$ в виде факторизации по группам переменных Z_j, $j \in J$:

$$q(Z) = \prod_{j \in J} q_j(Z_j)$$

Идея 2. Если теперь $p(\Phi,\Theta|\alpha,\beta) = \prod_t \mathrm{Dir}(\phi_t|\beta) \prod_d \mathrm{Dir}(\theta_d|\alpha)$, то параметры вероятностной модели — (α,β) вместо (Φ,Θ), и тогда скрытые переменные — (Z,Φ,Θ) вместо Z.

Основная теорема вариационного байесовского вывода

Теорема. Решение задачи $\mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X,\Omega)\big) \to \min_q$ в семействе факторизационных распределений $q(Z) = \prod_j q_j(Z_j)$ удовлетворяет системе уравнений

$$\ln q_j(Z_j) = \mathrm{E}_{q\setminus j} \ln p(X, Z|\Omega) + \mathrm{const},$$

где $\mathrm{E}_{q\setminus j}$ — матожидание по всем переменным кроме Z_j, const — нормировочный множитель распределения q_j.

Для решения этой системы используют метод простой итерации.

Идея доказательства: расписываем $\mathrm{KL}(\cdot\|\cdot)$ и сводим задачу к

$$\sum_{Z_j} q_j(Z_j) \underbrace{\sum_{Z\setminus Z_j} \prod_{i\neq j} q_i(Z_i) \ln p(X, Z|\Omega)}_{\mathrm{E}_{q\setminus j} \ln p(X,Z|\Omega)} - \sum_{Z_j} q_j(Z_j) \ln q_j(Z_j) \;\to\; \min_q$$

Доказательство

1. В оптимизационной задаче можно перекидывать X через условную черту:

$$\sum_Z q(Z) \ln \frac{p(Z|X,\Omega)}{q(Z)} \to \max_q \Leftrightarrow \sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{q(Z)} - \sum_Z q(Z) \ln p(X|\Omega) \to \max_q$$

2. Будем минимизировать KL-дивергенцию поочерёдно по всем Z_j.
Применим факторизацию и вынесем слагаемое с $q_j(Z_j)$ вперёд:

$$\sum_{Z_j} q_j(Z_j) \underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i) \ln p(X,Z|\Omega)}_{\mathrm{E}_{q \setminus j} \ln p(X,Z|\Omega)} - \sum_{Z_j} q_j(Z_j) \underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i) \sum_{k \in J} \ln q_k(Z_k)}_{\ln q_j(Z_j) + \mathrm{const}} \to \max_{q_j}$$

3. Почему вторую фигурную скобку можно заменить на $\ln q_j(Z_j)$:

$$\underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i) \sum_{k \neq j} \ln q_k(Z_k)}_{\text{не зависит от } q_j} + \underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i)}_{1} \ln q_j(Z_j)$$

4. Введём $r(Z_j) \propto \exp\big(\mathrm{E}_{q \setminus j} \ln p(X,Z|\Omega)\big)$, тогда $\mathrm{KL}\big(q_j(Z_j) \big\| r(Z_j)\big) \to \min_{q_j}$
5. Точное решение данной задачи $q_j(Z_j) = r(Z_j)$, следовательно,

$$\ln q_j(Z_j) = \mathrm{E}_{q \setminus j} \ln p(X,Z|\Omega) + \mathrm{const}.$$

■

Снова максимизация неполного правдоподобия

(Z, Φ, Θ) — теперь это скрытые переменные,
$\Omega = (\alpha, \beta)$ — а это параметры вероятностной модели.

Задача максимизации неполного правдоподобия:

$$p(X|\alpha,\beta) = \sum_Z \int_\Phi \int_\Theta p(X, Z, \Phi, \Theta|\alpha,\beta) d\Phi d\Theta \;\to\; \max_{\alpha,\beta}$$

EM-алгоритм для решения данной задачи имеет вид:

$$\text{E-шаг:} \quad \mathrm{KL}\big(q(Z,\Phi,\Theta) \,\big\|\, p(Z,\Phi,\Theta|X,\alpha,\beta)\big) \;\to\; \min_q$$

$$\text{M-шаг:} \quad \sum_Z \int_\Phi \int_\Theta q(Z,\Phi,\Theta) \ln p(X,Z,\Phi,\Theta|\alpha,\beta) d\Phi d\Theta \;\to\; \max_{\alpha,\beta}$$

Теперь всё происходит на E-шаге (аналог старых E и M шагов)
Новый M-шаг часто опускают, фиксируя α, β без оптимизации

Вариационная аппроксимация для тематической модели LDA

Найти $q(Z, \Phi, \Theta)$ в общем виде не удаётся, поэтому ищем в семействе факторизованных распределений:

$$q(Z, \Phi, \Theta) = \prod_{i=1}^{n} q_i(t_i) \prod_{t \in T} q_t(\phi_t) \prod_{d \in D} q_d(\theta_d) \equiv \prod_{j \in J} q_j(Z, \Phi, \Theta),$$

$J = \{1, \ldots, n\} \sqcup T \sqcup D$ — индексы всех скрытых переменных.

Основная теорема вариационного байесовского вывода

Решение задачи E-шага удовлетворяет системе уравнений

$$\ln q_j = \mathsf{E}_{q \setminus j} \ln p(X, Z, \Phi, \Theta | \alpha, \beta) + \text{const},$$

где $\mathsf{E}_{q \setminus j}$ — матожидание по всем переменным кроме j-й,
const — нормировочный множитель распределения q_j.

Для решения этой системы используют метод простой итерации.

Расписываем логарифм $p(X, Z, \Phi, \Theta|\alpha, \beta)$

Применяя *основную теорему*, переводим независящие от скрытых переменных множители в const:

$$\ln p(X, Z, \Phi, \Theta|\alpha, \beta) =$$

$$= \ln \prod_{i=1}^{n} p(d_i, w_i, t_i|\Phi, \Theta) \prod_{t \in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d \in D} \mathrm{Dir}(\theta_d|\alpha) =$$

$$= \sum_{i=1}^{n} \ln \phi_{w_i t_i} \theta_{t_i d_i} + \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td} + \text{const}.$$

Теперь надо брать матожидания $\mathsf{E}_{q \setminus j}$ от этой величины по всем распределениям $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$, кроме j-го.

Замечание, сильно упрощающее выкладки:
если слагаемое S не зависит от j-й переменной, то $\mathsf{E}_{q_j} S = \text{const}$.

Распределения скрытых переменных $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$

Распределение скрытой переменной $\phi_t \in \mathbb{R}^W$:

$$\ln q_t(\phi_t) = \sum_{i=1}^{n} \mathsf{E}_{q_i(t_i)}[t_i = t] \ln \phi_{w_i t_i} + \sum_{w \in W} (\beta_w - 1) \ln \phi_{wt} + \text{const} =$$

$$= \sum_{i=1}^{n} \sum_{w \in W} [w_i = w] q_i(t) \ln \phi_{wt} + \sum_{w \in W} (\beta_w - 1) \ln \phi_{wt} + \text{const} =$$

$$= \sum_{w \in W} \Bigg(\underbrace{\sum_{i=1}^{n} [w_i = w] q_i(t)}_{n_{wt}} + \beta_w - 1 \Bigg) \ln \phi_{wt} + \text{const} =$$

$$= \ln \mathrm{Dir}(\phi_t | \tilde{\beta}_t).$$

Это распределение Дирихле с параметрами $\tilde{\beta}_{wt} = n_{wt} + \beta_w$,
n_{wt} — оценка числа генераций термина w из темы t.

Распределения скрытых переменных $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$

Распределение скрытой переменной $\theta_d \in \mathbb{R}^T$:

$$\begin{aligned}
\ln q_d(\theta_d) &= \sum_{i=1}^{n} \mathsf{E}_{q_i(t_i)}[d_i = d] \ln \theta_{t_i d_i} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \text{const} = \\
&= \sum_{i=1}^{n} [d_i = d] \sum_{t \in T} q_i(t) \ln \theta_{td} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \text{const} = \\
&= \sum_{t \in T} \Bigg(\underbrace{\sum_{i=1}^{n} [d_i = d] q_i(t)}_{n_{td}} + \alpha_t - 1 \Bigg) \ln \theta_{td} + \text{const} = \\
&= \ln \mathrm{Dir}(\theta_d | \tilde{\alpha}_d).
\end{aligned}$$

Это распределение Дирихле с параметрами $\tilde{\alpha}_{td} = n_{td} + \alpha_t$,
n_{td} — оценка числа терминов темы t в документе d.

Распределения скрытых переменных $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$

Распределение скрытой переменной $t_i \in \mathbb{R}$:

$$\begin{aligned}\ln q_i(t) &= \mathsf{E}_{q\setminus i}\big(\ln \phi_{w_i t_i} + \ln \theta_{t_i d_i}\big) + \text{const} = \\ &= \mathsf{E}_{q_t(\phi_t)} \ln \phi_{w_i t} + \mathsf{E}_{q_d(\theta_d)} \ln \theta_{t_i d} + \text{const} = \\ &= \psi\big(n_{w_i t} + \beta_{w_i}\big) - \psi\Big(\textstyle\sum_w \big(n_{wt} + \beta_w\big)\Big) + \\ &\quad + \psi\big(n_{td_i} + \alpha_t\big) - \psi\Big(\textstyle\sum_t \big(n_{td_i} + \alpha_t\big)\Big) + \text{const}\end{aligned}$$

Воспользуемся приближением $\exp(\psi(x)) \approx x - \frac{1}{2}$:

$$q_i(t) = \underset{t\in T}{\text{norm}} \left(\frac{n_{w_i t} + \beta_{w_i} - \frac{1}{2}}{\sum_w (n_{wt} + \beta_w) - \frac{1}{2}} \cdot \frac{n_{td_i} + \alpha_t - \frac{1}{2}}{\sum_t (n_{td_i} + \alpha_t) - \frac{1}{2}} \right)$$

Похоже на обычную формулу E-шага $p(t|d_i, w_i) = \underset{t\in T}{\text{norm}}(\phi_{w_i t}\theta_{td_i})$

Собираем всё воедино

В итерационном процессе чередуются два шага:

1) распределение терминов (d_i, w_i) по темам, $E(x) = \exp(\psi(x))$:

$$q_i(t) = \underset{t \in T}{\text{norm}} \left(\frac{E\left(n_{w_i t} + \beta_{w_i}\right)}{E\left(\sum_w (n_{wt} + \beta_w)\right)} \cdot \frac{E\left(n_{td_i} + \alpha_t\right)}{E\left(\sum_t (n_{td_i} + \alpha_t)\right)} \right)$$

2) аккумулирование счётчиков n_{wt} и n_{td}:

$$n_{wt} = \sum_{i=1}^{n} [w_i = w]\, q_i(t) \qquad n_{td} = \sum_{i=1}^{n} [d_i = d]\, q_i(t)$$

Точечные оценки параметров по матожиданию или моде:

$$\mathrm{E}\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w\right) \qquad \mathrm{E}\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t\right)$$

$$\hat{\phi}_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w - 1\right) \qquad \hat{\theta}_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t - 1\right)$$

Промежуточный итог

- Формулы для MAP и VB очень похожи [Asuncion]:
  — при $n_{wt}, n_{td} \gg 1$ различия неощутимы,
  — при $n_{wt}, n_{td} \lesssim 1$ тема t незначима для w или d.
- Распределение Дирихле, как в роли регуляризатора, так и в роли сопряжённого к мультиномиальному в байесовском выводе, приводит к схожим алгоритмам.
- Можно добавить M-шаг для оптимизации α, β [Wallach].
- Несколько смущает разнообразие оценок... какая лучше?

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.

Hanna Wallach, David Mimno, Andrew McCallum. Rethinking LDA: why priors matter. Neural Information Processing Systems, 2009.

Основные идеи Gibbs Sampling

$\Omega = (\alpha, \beta)$ — параметры вероятностной модели.
Поочерёдное выполнение двух шагов (аналог EM-алгоритма):

$$\text{E-шаг:} \quad Z \sim p(Z|X, \Omega)$$
$$\text{M-шаг:} \quad (\Phi, \Theta) \sim p(\Phi, \Theta|X, Z, \Omega)$$

Основная теорема о сходимости сэмплирования Гиббса

Процесс сэмплирования одномерных случайных величин

$$t_i^{(k+1)} \sim p(t_i|X, Z_{\setminus i}, \Omega) = \frac{p(X, Z|\Omega)}{p(X, Z_{\setminus i}|\Omega)},$$

где k — номер итерации, $Z_{\setminus i} = (t_1^{(k+1)}, \ldots, t_{i-1}^{(k+1)}, t_{i+1}^{(k)}, \ldots, t_n^{(k)})$,
сходится к многомерному распределению $Z \sim p(Z|X, \Omega)$

Распределение Дирихле — сопряжённое к мультиномиальному

$\theta \sim p(\theta|\alpha)$ — априорное распределение Дирихле,
$X = (t_1, \ldots, t_n) \sim p(t|\theta)$ — мультиномиальные данные; тогда
$p(\theta|X, \alpha)$ — апостериорное распределение — тоже Дирихле.

Вывод апостериорного распределения Φ, Θ при известных Z:

$$p(\Phi, \Theta|X, Z, \alpha, \beta) \propto p(\Phi, \Theta, X, Z|\alpha, \beta) \propto p(X, Z|\Phi, \Theta) p(\Phi, \Theta|\alpha, \beta)$$

$$\propto \prod_{d,w,t} (\phi_{wt}\theta_{td})^{n_{dwt}} \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha)$$

$$\propto \prod_{t\in T}\prod_{d,w} \phi_{wt}^{n_{dwt}} \phi_{wt}^{\beta_w - 1} \prod_{d\in D}\prod_{w,t} \theta_{td}^{n_{dwt}} \theta_{td}^{\alpha_t - 1}$$

$$\propto \prod_{t\in T}\prod_{w} \phi_{wt}^{n_{wt} + \beta_w - 1} \prod_{d\in D}\prod_{t} \theta_{td}^{n_{td} + \alpha_t - 1}$$

$$\propto \prod_{t\in T} \mathrm{Dir}(\phi_t|\tilde{\beta}_t) \prod_{d\in D} \mathrm{Dir}(\theta_d|\tilde{\alpha}_d), \quad \tilde{\beta}_{wt} = n_{wt} + \beta_w, \quad \tilde{\alpha}_{td} = n_{td} + \alpha_t.$$

Распределение $p(X, Z|\alpha, \beta)$ для схемы сэмплирования Гиббса

Подынтегральное распределение мы только что вывели, но теперь будем аккуратнее с нормировочными множителями:

$$p(X, Z|\alpha, \beta) = \int_\Phi \int_\Theta p(X, Z|\Phi, \Theta) p(\Phi, \Theta|\alpha, \beta) d\Phi d\Theta =$$

$$= \int_\Phi \int_\Theta \prod_{d,w,t} (\phi_{wt}\theta_{td})^{n_{dwt}} \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha) d\Phi d\Theta =$$

$$= \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \underbrace{\int_{\phi_t} \prod_w \phi_{wt}^{\tilde\beta_{wt}-1} d\phi_t}_{\propto \mathrm{Dir}(\phi_t|\tilde\beta_t)} \prod_{d\in D} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \underbrace{\int_{\theta_d} \prod_t \theta_{td}^{\tilde\alpha_{td}-1} d\theta_d}_{\propto \mathrm{Dir}(\theta_d|\tilde\alpha_d)} =$$

$$= \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde\beta_{wt})}{\Gamma(\sum_w \tilde\beta_{wt})} \prod_{d\in D} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde\alpha_{td})}{\Gamma(\sum_t \tilde\alpha_{td})}$$

Распределение $p(X, Z_{\setminus i}|\alpha, \beta)$ для схемы сэмплирования Гиббса

Итак, мы только что получили распределение

$$p(X, Z|\alpha, \beta) =$$
$$= \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde{\beta}_{wt})}{\Gamma(\sum_w \tilde{\beta}_{wt})} \prod_{d\in D} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde{\alpha}_{td})}{\Gamma(\sum_t \tilde{\alpha}_{td})}$$

Распределение $p(X, Z_{\setminus i}|\alpha, \beta)$ отличается от него лишь тем, что оно построено по выборке без одной i-й точки (d_i, w_i, t_i):

$$p(X, Z_{\setminus i}|\alpha, \beta) =$$
$$= \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde{\beta}_{wt} - \delta^i_{wt})}{\Gamma(\sum_w (\tilde{\beta}_{wt} - \delta^i_{wt}))} \prod_{d\in D} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde{\alpha}_{td} - \delta^i_{td})}{\Gamma(\sum_t (\tilde{\alpha}_{td} - \delta^i_{td}))}$$

где $\delta^i_{wt} = [w = w_i][t = t_i]$, $\delta^i_{td} = [t = t_i][d = d_i]$

Ещё чуть-чуть... осталось только поделить одно на другое

Для сэмплирования Гиббса нужно одномерное распределение

$$p(t_i|X, Z_{\setminus i}, \alpha, \beta) = \frac{p(X, Z|\alpha, \beta)}{p(X, Z_{\setminus i}|\alpha, \beta)} =$$

В числителе и знаменателе сократятся все множители кроме i-х:

$$= \frac{\Gamma(n_{w_i t_i} + \beta_{w_i})\,\Gamma\left(\sum_w (n_{wt_i} + \beta_w) - 1\right)}{\Gamma(n_{w_i t_i} + \beta_{w_i} - 1)\,\Gamma\left(\sum_w (n_{wt_i} + \beta_w)\right)} \frac{\Gamma(n_{t_i d_i} + \alpha_{t_i})\,\Gamma\left(\sum_t (n_{td_i} + \alpha_t) - 1\right)}{\Gamma(n_{t_i d_i} + \alpha_{t_i} - 1)\,\Gamma\left(\sum_t (n_{td_i} + \alpha_t)\right)}$$

Воспользуемся свойством гамма-функции $\frac{\Gamma(x)}{\Gamma(x-1)} = x - 1$:

$$p(t|X, Z_{\setminus i}, \alpha, \beta) = \underset{t \in T}{\text{norm}} \left(\frac{n_{w_i t} + \beta_{w_i} - 1}{\sum_w (n_{wt} + \beta_w) - 1} \cdot \frac{n_{td_i} + \alpha_t - 1}{\sum_t (n_{td_i} + \alpha_t) - 1} \right)$$

Похоже на обычную формулу E-шага $p(t|d_i, w_i) = \underset{t \in T}{\text{norm}}(\phi_{w_i t} \theta_{td_i})$

Собираем всё воедино

Выделены отличия от вариационного алгоритма

1) для каждого (d_i, w_i), $i = 1, \ldots, n$ сэмплирование темы t_i:

$$t_i \sim p_i(t) = \underset{t \in T}{\text{norm}} \left(\frac{n_{w_i t} + \beta_{w_i} - 1}{\sum_w (n_{wt} + \beta_w) - 1} \cdot \frac{n_{td_i} + \alpha_t - 1}{\sum_t (n_{td_i} + \alpha_t) - 1} \right)$$

2) аккумулирование счётчиков n_{wt} и n_{td}:

$$n_{wt} = \sum_{i=1}^{n} [w_i = w][t_i = t] \qquad n_{td} = \sum_{i=1}^{n} [d_i = d][t_i = t]$$

Параметры Φ, Θ сэмплируются или оцениваются точечно:

$$(\phi_{wt})_w \sim \text{Dir}\left(\phi \,\middle|\, (n_{wt} + \beta_w)_w\right) \qquad (\theta_{td})_t \sim \text{Dir}\left(\theta \,\middle|\, (n_{td} + \alpha_t)_t\right)$$

$$\text{E}\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w\right) \qquad \text{E}\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t\right)$$

$$\hat{\phi}_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w - 1\right) \qquad \hat{\theta}_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t - 1\right)$$

Алгоритм сэмплирования Гиббса

Вход: коллекция D, число тем $|T|$, параметры α, β;
Выход: распределения Φ и Θ;

$n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $k := 1, \ldots, k_{\max}$
  для всех документов $d \in D$ и терминов $w = w_1, \ldots, w_{n_d} \in d$
    если $k \geqslant 2$ **то** $t := t_{dw}$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
    $p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\frac{n_{wt}+\beta_w}{n_t+\beta_0} \cdot \frac{n_{td}+\alpha_t}{n_d+\alpha_0}\right)$ для всех $t \in T$;
    сэмплировать одну тему t из распределения $p(t|d,w)$;
    $t_{dw} := t$; $++n_{wt}$; $++n_{td}$; $++n_t$; $++n_d$;

$\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
$\theta_{td} := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Griffiths T., Steyvers M. Finding scientific topics. 2004.

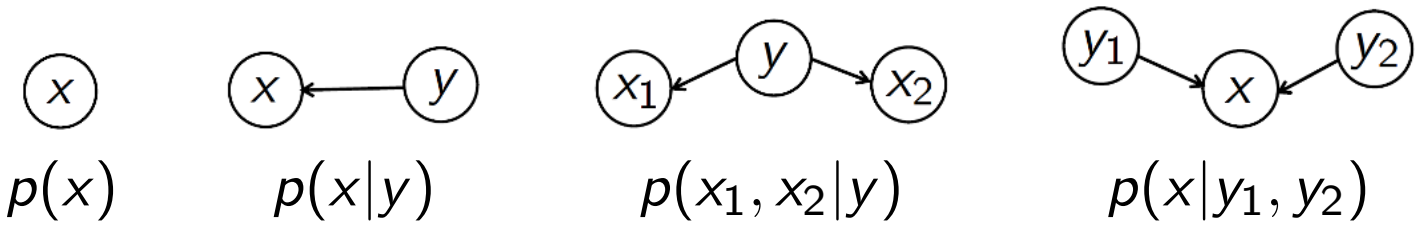
Промежуточный итог

- Похожий алгоритм получится в ARTM, если на E-шаге вместо $p(t|d,w)$ брать $\hat{p}(t|d,w) = [t = t_i]$, $t_i \sim p(t|d,w)$.
- Формулы для MAP, VB и GS очень похожи [Asuncion]:
  — при $n_{wt}, n_{td} \gg 1$ различия неощутимы,
  — при $n_{wt}, n_{td} \lesssim 1$ тема t незначима для w или d.
- Различия в алгоритмах кажутся второстепенными...
  стоило ли преодолевать столько технических трудностей,
  чтобы получать каждый раз почти одно и то же?
- Без сопряжённых априорных распределений ещё труднее!
- В методах VB и GS нет удобных механизмов регуляризации,
  т.к. нет, собственно, и задачи оптимизации по (Φ, Θ)

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.

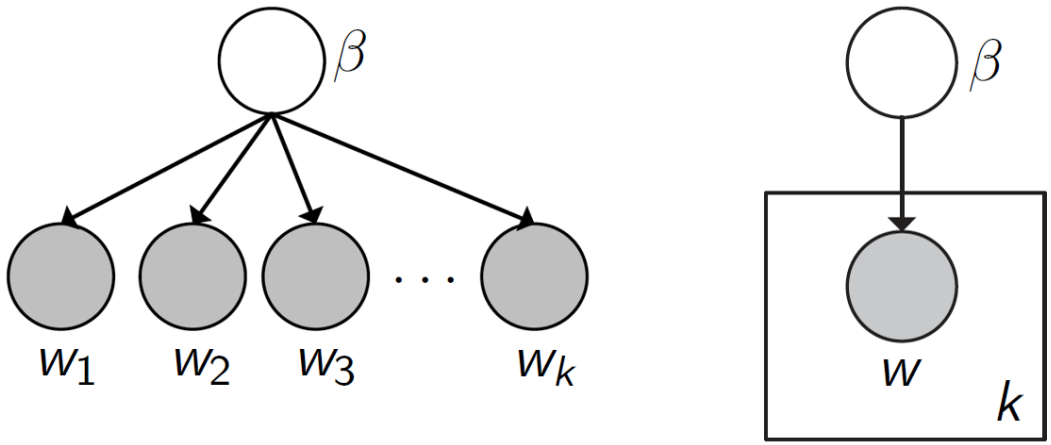
Графический язык плоской нотации (plate notation)

Графическое представление условных зависимостей



$p(x)$ $p(x|y)$ $p(x_1, x_2|y)$ $p(x|y_1, y_2)$

Графическое представление выборки $w_1, \ldots, w_k$,
порождаемой дискретным распределением $\beta_w = p(w)$

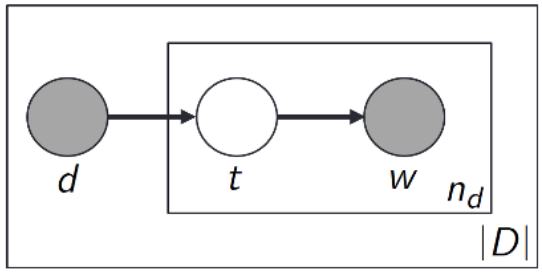


Плоская нотация: модели PLSA и LDA

Модель PLSA

каждый $d \in D$ порождает скрытые темы:
$t_i \sim p(t|d), \quad i = 1, \ldots, n_d$;

каждая тема t_i порождает слово:
$w_i \sim p(w|t_i), \quad i = 1, \ldots, n_d$.

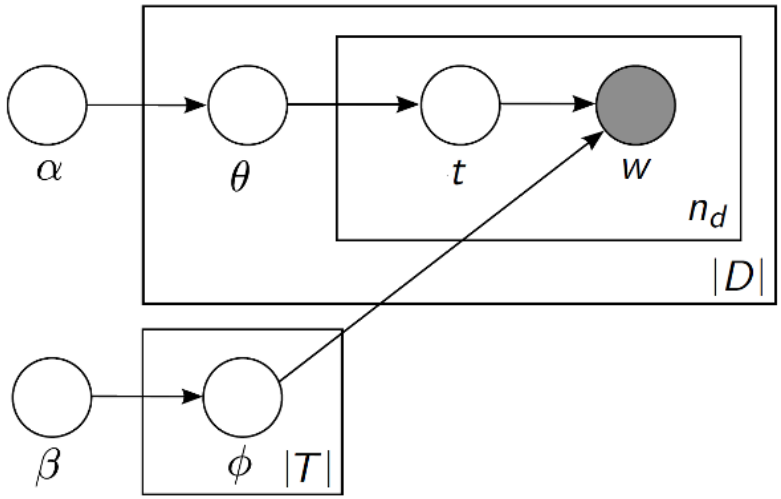


Модель LDA

α порождает векторы документов:
$\theta_d \sim \mathrm{Dir}(\theta|\alpha), \quad d \in D$;

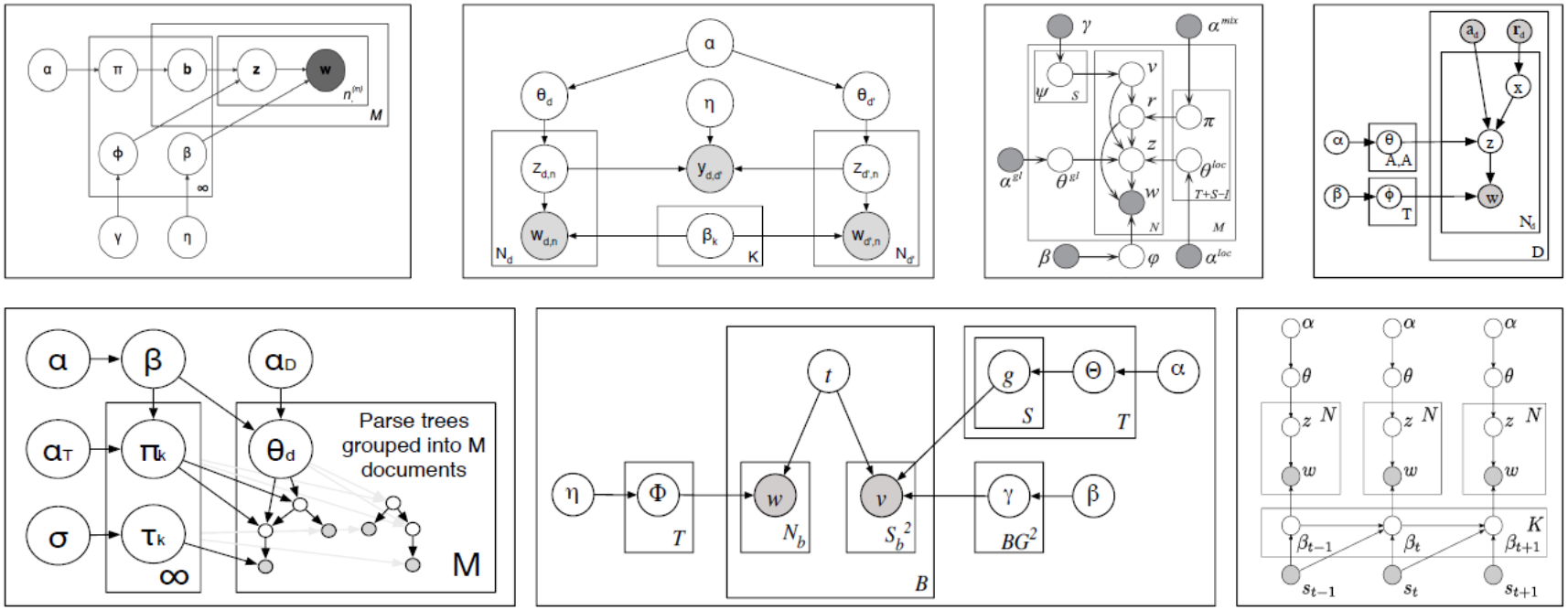
β порождает векторы тем:
$\phi_t \sim \mathrm{Dir}(\phi|\beta), \quad t \in T$;

далее как в PLSA.



Плоская нотация: представление структуры модели

Большое структурное разнообразие тематических моделей:



David Blei. Probabilistic topic models // Communications of the ACM, 2012.

Обсуждение «Stop using Plate Notation»

Единственное достоинство и куча недостатков:

- + хорошо запоминающийся наглядный образ модели
- – множественность вариантов отображения одной модели
- – неполнота и неоднозначность интерпретации
- – не ясен переход от картинки к модели и алгоритму
- – во многих статьях этот переход скрыт или скомкан
- – иллюзия понятности и практическая бесполезность

Один из комментариев:

Every now and then the topic comes up as to why algorithms and procedures are explained in obtuse forms across the entirety of the paper it is described in, usually we just conclude that it would look too simple if it were explained any other way.

Rob Zinkov. Stop using Plate Notation. 2013-07-28.
`http://zinkov.com/posts/2013-07-28-stop-using-plates`

Язык псевдокода порождающего процесса (generative story)

Пример: вероятностная порождающая модель PLSA

Вход: $p(w|t)$ для всех $t \in T$, $p(t|d)$ для всех $d \in D$;
Выход: коллекция документов;

для всех документов $d \in D$
  для всех позиций слов $i = 1, \ldots, n_d$ в документе d
    выбрать тему t_i из $p(t|d)$;
    выбрать слово w_i из $p(w|t_i)$;

- $+$ легко понимать модель, описание недвусмысленно
- $-$ не ясен переход от модели к алгоритму
- $-$ во многих статьях этот переход скрыт или скомкан

Язык задач оптимизации: модель → критерий → алгоритм

1. Вероятностная модель порождения данных:

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

2. Постановка задачи оптимизации (ARTM):

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

3. Алгоритм решения — итерационный процесс:

$$\text{E-шаг:} \quad \begin{cases} p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} \end{cases}$$

Математический язык оптимизационных задач

Один недостаток и куча достоинств:

- $-$ плохо понятен специалистам гуманитарных профессий
- $+$ ясность и однозначность записи модели
- $+$ ясность критерия оптимальности
- $+$ ясность вариантов выбора методов оптимизации
- $+$ переход от модели к алгоритму — это теорема
- $+$ возможность стандартизации кода (BigARTM)

Резюме

- MAP (максимизация апостериорной вероятности) полностью совместима с ARTM.
- *Байесовский вывод* оценивает не (Φ, Θ), а $p(\Phi, \Theta|X)$. Так ли это необходимо в тематическом моделировании?
- Уникальная математическая задача для каждой модели. Решение плохо унифицируется до общих формул и кода.
- В нём нет удобных механизмов регуляризации, нет понимания проблемы неединственности решения, т.к. нет, собственно, и задачи оптимизации по (Φ, Θ).
- Тем не менее, итерационный процесс в методах VB и GS получается предельно похожим на тот, что даёт MAP.
- Мы рассмотрели лишь самую простую модель (LDA), где сопряжённость распределения Дирихле упрощает вывод. Без сопряжённых распределений вывод ещё сложнее!