Методы машинного обучения. Байесовские латентные модели и тематическое моделирование

Воронцов Константин Вячеславович
www.MachineLearning.ru/wiki?title=User:Vokov
вопросы к лектору: voron@forecsys.ru

материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-22-23
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 7 марта 2023

Содержание

Напоминание. Вероятностные модели порождения данных

Дано:

$X = (x_1, \ldots, x_\ell)$ — исходные данные, *наблюдаемые переменные*

Найти:

$p(X|\Omega)$ — модель порождения данных, с параметром Ω

$p(X|\Omega) = \prod_{i=1}^{\ell} p(x_i|\Omega)$ — в случае простой (i.i.d.) выборки

Критерии максимизации:

— *правдоподобия* (Maximum Likelihood Estimate, MLE):

$$\ln p(X|\Omega) = \sum_{i=1}^{\ell} \ln p(x_i|\Omega) \;\to\; \max_{\Omega}$$

— *апостериорной вероятности* (Maximum a Posteriori, MAP):

$$\ln p(X,\Omega) = \ln p(X|\Omega)p(\Omega|\gamma) = \sum_{i=1}^{\ell} \ln p(x_i|\Omega) + \ln p(\Omega|\gamma) \;\to\; \max_{\Omega}$$

где $R(\Omega) = \ln p(\Omega|\gamma)$ играет роль регуляризатора.

Порождающая модель с латентными переменными

Дано:
$X = (x_1, \dots, x_\ell)$ — исходные данные, *наблюдаемые переменные*
$Z = (z_1, \dots, z_m)$ — *латентные (скрытые) переменные*

Найти:
$p(X, Z|\Omega)$ — модель совместного порождения наблюдаемых данных и скрытых переменных, с параметром Ω

Критерий:
максимум правдоподобия (Maximum Likelihood Estimate, MLE):

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

Для дискретных переменных Z вместо интеграла $\int_Z$ сумма $\sum_Z$
Договоримся далее dZ опускать

Bishop C. M. Pattern recognition and machine learning. Springer, 2006.

Пример. Задача разделения смеси распределений

Порождающая модель смеси k вероятностных распределений:

$$p(x) = \sum_{j=1}^{k} p(x,j) = \sum_{j=1}^{k} p(j)p(x|j) = \sum_{j=1}^{k} w_j \varphi(x,\theta_j)$$

$X = (x_1, \dots, x_\ell)$ — исходные данные, *наблюдаемые переменные*

$Z = (j_1, \dots, j_\ell)$ — компоненты смеси j_i, порождающие объекты x_i

$\Omega = \bigl(w_j, \theta_j\bigr)_{j=1}^{k}$ — параметры смеси, $w_j \geqslant 0$, $\sum_{j=1}^{k} w_j = 1$

Порождающая модель наблюдаемых и скрытых i.i.d. данных:

$$p(X,Z|\Omega) = \prod_{i=1}^{\ell} p(x_i, j_i) = \prod_{i=1}^{\ell} p(j_i)p(x_i|\theta_{j_i}) = \prod_{i=1}^{\ell} w_{j_i}\varphi(x_i,\theta_{j_i})$$

Задача разделения смеси — максимизация log правдоподобия:

$$\ln p(X|\Omega) = \ln \prod_{i=1}^{\ell} p(x_i) = \sum_{i=1}^{\ell} \ln \sum_{j=1}^{k} w_j \varphi(x_i,\theta_j) \;\to\; \max_{\Omega}$$

Байесовская порождающая модель с латентными переменными

$X = (x_1, \dots, x_\ell)$ — исходные данные, *наблюдаемые переменные*
$Z = (z_1, \dots, z_m)$ — *латентные (скрытые) переменные*
$p(X, Z|\Omega)$ — модель наблюдаемых и скрытых переменных
$p(\Omega|\gamma)$ — *априорное распределение* с гиперпараметрами γ

Задача: по X найти Ω.

Апостериорное распределение, по формуле Байеса:

$$p(\Omega|X, \gamma) \propto p(X|\Omega)\, p(\Omega|\gamma) = \int_Z p(X, Z|\Omega)\, p(\Omega|\gamma)$$

Принцип максимума апостериорной вероятности:

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

$R(\Omega) = \ln p(\Omega|\gamma)$ — байесовский регуляризатор, хотя
$R(\Omega)$ может и не иметь вероятностной интерпретации.

Общий EM-алгоритм для задачи со скрытыми переменными

Теорема. Точка Ω локального максимума регуляризованного *маргинализованного правдоподобия* (Marginal log-Likelihood)

$$\ln \int_Z p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega \qquad \text{(RML)}$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

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

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

Общий EM-алгоритм используется не только для разделения смесей, но и в анализе сигналов, изображений, текстов и др.

A.P.Dempster, N.M.Laird, D.B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. 1977.

Доказательство теоремы

Необходимые условия локального экстремума:

$$\frac{\partial}{\partial\Omega}\left(\ln\int_Z p(X,Z|\Omega)+R(\Omega)\right)=\frac{1}{p(X|\Omega)}\int_Z\frac{\partial p(X,Z|\Omega)}{\partial\Omega}+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

По формуле условной вероятности $p(X|\Omega)=\frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$, подставляем:

$$\int_Z\frac{p(Z|X,\Omega)}{p(X,Z|\Omega)}\frac{\partial p(X,Z|\Omega)}{\partial\Omega}+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

$$\int_Z\underbrace{p(Z|X,\Omega)}_{q(Z)}\frac{\partial}{\partial\Omega}\ln p(X,Z|\Omega)+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

Это уравнение совпадает с необходимым условием локального экстремума для задачи M-шага, при этом $q(Z)$ рассматривается как переменная, не зависящая от Ω.

■

Ещё более общий EM-алгоритм и его сходимость

Теорема. Значение маргинализованного правдоподобия

$$\ln \int_Z p(X, Z|\Omega) + R(\Omega) \;\to\; \max_{\Omega} \qquad \text{(RML)}$$

не убывает на каждом шаге итерационного процесса

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

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

$q(Z) = p(Z|X, \Omega)$ является точным решением задачи E-шага.

Минимизация KL-дивергенции на E-шаге используется в случаях, когда не удаётся вычислить $p(Z|X, \Omega)$ в явном виде.

Сходимость *в слабом смысле*: глобальный max не гарантируется.

Доказательство теоремы

По формуле условной вероятности $p(X|\Omega) = \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$.
Для произвольного распределения $q(Z)$

$$\ln p(X|\Omega) = \int_Z q(Z) \ln p(X|\Omega) = \int_Z q(Z) \ln \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)} =$$
$$= \underbrace{\int_Z q(Z) \ln \frac{p(X,Z|\Omega)}{q(Z)}}_{L(q,\Omega)} + \underbrace{\int_Z q(Z) \ln \frac{q(Z)}{p(Z|X,\Omega)}}_{\mathrm{KL}\left(q(Z)\middle\|p(Z|X,\Omega)\right)\geqslant 0}$$

Максимизируем достижимую нижнюю оценку RML то по q, то по Ω:

$$\text{E-шаг: } L(q,\Omega) + \cancel{R(\Omega)} \to \max_q \;\Leftrightarrow\; \mathrm{KL}\big(q(Z) \,\|\, p(Z|X,\Omega)\big) \to \min_q$$
$$\text{M-шаг: } L(q,\Omega) + R(\Omega) \to \max_\Omega \;\Leftrightarrow\; \int_Z q(Z) \ln p(X,Z|\Omega) + R(\Omega) \to \max_\Omega$$

На каждом шаге значение функционала может только возрастать, откуда и следует сходимость в слабом смысле. ■

EM-алгоритм для разделения смеси распределений

$X = (x_1, \dots, x_\ell)$ — исходные данные, *наблюдаемые переменные*

$Z = (j_1, \dots, j_\ell)$ — компоненты смеси j_i, порождающие объекты x_i

$$p(X, Z|\Omega) = \prod_{i=1}^{\ell} p(x_i, j_i|\Omega) = \prod_{i=1}^{\ell} p(j_i)p(x_i|\theta_{j_i}) = \prod_{i=1}^{\ell} w_{j_i}\varphi(x_i, \theta_{j_i})$$

E-шаг: в силу независимости элементов выборки и формулы Байеса

$$q(Z) = p(Z|X, \Omega) = \prod_{i=1}^{\ell} p(j_i|x_i, \Omega); \qquad p(j|x, \Omega) = \frac{p(x,j|\Omega)}{p(x|\Omega)} = \frac{w_j\varphi(x,\theta_j)}{\sum_t w_t\varphi(x,\theta_t)}$$

M-шаг: подставим $q(Z)$ и $p(X, Z|\Omega)$ в общую формулу M-шага:

$$\sum_Z q(Z) \ln p(X, Z|\Omega) + R(\Omega) =$$

$$= \sum_{j_1=1}^{k} \cdots \sum_{j_\ell=1}^{k} \prod_{t=1}^{\ell} p(j_t|x_t, \Omega) \sum_{i=1}^{\ell} \ln p(x_i, j_i|\Omega) + R(\Omega) =$$

$$= \sum_{i=1}^{\ell} \sum_{j=1}^{k} \underbrace{p(j|x_i, \Omega)}_{g_{ij}} \ln \underbrace{p(x_i, j|\Omega)}_{w_j\varphi(x_i,\theta_j)} + R(\Omega) \to \max_{\Omega}, \quad \Omega = (w_j, \theta_j)_{j=1}^{k}$$

EM-алгоритм для разделения смеси распределений

M-шаг распадается на $2k$ подзадач по w_j, θ_j, $j = 1, \dots, k$:

$$\sum_{j=1}^{k} \sum_{i=1}^{\ell} \Bigl(g_{ij} \ln w_j + g_{ij} \ln \varphi(x_i, \theta_j) \Bigr) + R(\Omega) \;\to\; \max_{\Omega}$$

Необходимые условия локального экстремума:

$$\begin{cases} \dfrac{\partial}{\partial \theta_j} \left(\sum\limits_{i=1}^{\ell} g_{ij} \ln \varphi(x_i, \theta_j) + R(\Omega) \right) = 0; \\ \dfrac{\partial}{\partial w_j} \left(\sum\limits_{i=1}^{\ell} g_{ij} \ln w_j + R(\Omega) + \Bigl(1 - \sum\limits_{j=1}^{k} w_j \Bigr) \lambda_j - w_j \mu_j \right) = 0; \\ w_j \geqslant 0, \quad \sum_{j=1}^{k} w_j = 1, \quad w_j \mu_j = 0, \quad \mu_j \geqslant 0 \end{cases}$$

Относительно w_j решение аналитическое, из условий ККТ:

$$w_j = \operatorname*{norm}_{j} \left(\sum_{i=1}^{\ell} g_{ij} + w_j \frac{\partial R}{\partial w_j} \right), \; j = 1, \dots, k.$$

EM-алгоритм для разделения смеси распределений

Задача разделения смеси: $(X, Z) = (x_i, j_i)_{i=1}^{\ell}$, $\Omega = (w_j, \theta_j)_{j=1}^{k}$

Теорема. Точка Ω локального максимума (RML)

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) = \sum_{i=1}^{\ell} \ln \sum_{j=1}^{k} w_j \varphi(x_i, \theta_j) + R(\Omega)$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

$$\text{E-шаг:} \quad g_{ij} \equiv p(j|x_i) = \operatorname*{norm}_j \bigl(w_j \varphi(x_i, \theta_j) \bigr), \quad i = 1..\ell,\ j = 1..k;$$

$$\text{M-шаг:} \quad \theta_j = \arg\max_\theta \Bigl(\sum_{i=1}^{\ell} g_{ij} \ln \varphi(x_i, \theta) + R(\Omega) \Bigr),\ j = 1..k;$$

$$w_j = \operatorname*{norm}_j \Bigl(\sum_{i=1}^{\ell} g_{ij} + w_j \frac{\partial R}{\partial w_j} \Bigr),\ j = 1..k.$$

Регуляризация в EM-алгоритме для разделения смеси

- При $R(\Omega) = 0$ это уже знакомый нам EM-алгоритм для разделения смеси вероятностных распределений
- При разделении смеси n-мерных гауссиан $\mathcal{N}(x; \mu_j, \Sigma_j)$ регуляризация ковариационных матриц: $\Sigma_j + \tau I_n$
- $R(w) = \tau \sum_j p_j \ln w_j$ — регуляризатор сглаживания, приближает веса w_j к априорному распределению p_j:

$$w_j = \operatorname*{norm}_j \left(\sum_{i=1}^{\ell} g_{ij} + \tau p_j \right)$$

- $R(w) = -\tau \sum_j \ln w_j$ — регуляризатор разреживания, удаляет компоненты с весами $w_j < \tau$:

$$w_j = \operatorname*{norm}_j \left(\sum_{i=1}^{\ell} g_{ij} - \tau \right)$$

Преимущества общего EM-алгоритма

- Широкий класс задач связан с выявлением латентных структур, порождающих наблюдаемые данные:
  — разделение смеси вероятностных распределений;
  — «мягкая» или «жёсткая» кластеризация;
  — вероятностное тематическое моделирование;
  — сегментация временных рядов, сигналов, изображений;
  — восстановление пропущенных данных;
- Готовый рецепт для вывода вычислительных формул E и M шагов по известной порождающей модели $p(X, Z|\Omega)$
- Можно добавлять какие угодно регуляризаторы $R(\Omega)$, причём не обязательно вероятностные вида $\ln p(\Omega|\gamma)$
- Гарантируется сходимость в слабом смысле
- Это обучение без учителя, не требующее разметки

Постановка задачи тематического моделирования

Дано: коллекция текстовых документов

- W — конечное множество (словарь) термов (слов)
- D — конечное множество документов
- T — конечное множество тем
- $X = (d_i, w_i)_{i=1}^{n}$ — данные, *наблюдаемые переменные*
- $Z = (t_i)_{i=1}^{n}$ — *латентные (скрытые) переменные*, $t_i \in T$
- n_{dw} — частота терма $w \in W$ в документе $d \in D$
- n_d — длина документа $d \in D$
- n — суммарная длина всех документов коллекции

Найти: *вероятностную тематическую модель* (BTM)

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

где $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ — параметры модели

Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

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

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \underbrace{\cancel{p(d)}}_{\text{const}} \to \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

Несколько интерпретаций постановки задачи BTM

1. **Вероятностная языковая модель** $p(w|d)$

2. **Мягкая би-кластеризация** по кластерам-темам $t \in T$ как документов: $p(t|d)$, так и термов: $p(t|w) = p(w|t)\frac{p(t)}{p(w)}$

3. **Векторные представления (эмбединги)** — вероятностные, интерпретируемые, разреженные: $p(t|d), p(t|w), p(t|d, w), \ldots$

4. **Автокодировщик** документов в тематические эмбединги:

кодировщик: $f_\Phi: \left(\hat{p}(w|d) = \frac{n_{dw}}{n_d}\right) \;\to\; \left(p(t|d) = \theta_d\right)$
декодировщик: $g_\Phi: \theta_d \to \Phi\theta_d$
задача реконструкции: $\sum\limits_d n_d \mathrm{KL}\left(\hat{p}(w|d) \;\|\; \langle\phi_w, \theta_d\rangle\right) \;\to\; \min\limits_{\Phi,\Theta}$

5. **Матричное разложение** — низкоранговое, неотрицательное (стохастическое), приближённое: $\left(\frac{n_{dw}}{n_d}\right) \approx \Phi\Theta$

Свойство интерпретируемости тематических моделей

Тематические векторные представления (эмбединги) текста:

- $p(t|d) = \theta_{td}$ для документа d
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ для терма w
- $p(t|d,w)$ для локального контекста (d,w)
- $p(t|x)$ для нетекстового объекта x

Интерпретируемость тематических векторов:

- каждая тема t описывается *семантическим ядром* — частотным словарём слов $\bigl\{w\colon p(w|t) > \gamma p(w)\bigr\}$, встречающихся в данной теме в γ раз чаще обычного
- любой объект x с вектором $p(t|x)$ описывается частотным словарём слов $\Bigl\{w\colon p(w|x) = \sum\limits_{t\in T} p(w|t)p(t|x) > \gamma p(w)\Bigr\}$

Цели и не-цели тематического моделирования

Цели:

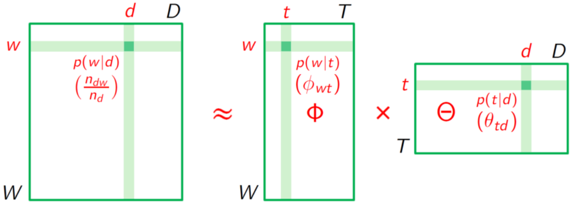
- Выяснить тематическую кластерную структуру текстовой коллекции, сколько в ней тем и какие они
- Получать интерпретируемые тематические векторные представления (эмбединги) документов, фрагментов, слов $p(t|d)$, $p(t|w)$, $p(t|d,w)$ и нетекстовых объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

Не-цели:

- Угадывать следующие слова (ТМ — слабые модели языка)
- Генерировать связный текст
- Понимать смысл текста

Некорректно поставленная задача матричного разложения

Низкоранговое стохастическое матричное разложение:



Если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\operatorname{rank} S = |T|$
- $L(\Phi', \Theta') = L(\Phi, \Theta)$ — линейно не зависимые решения
- $L(\Phi', \Theta') \geqslant L(\Phi, \Theta) - \varepsilon$ — приближённые решения

Регуляризация необходима для доопределения решения
Аддитивная регуляризация — сумма регуляризаторов

ВТМ как порождающая модель с латентными переменными

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*
$Z = (t_i)_{i=1}^n$ — *латентные (скрытые) переменные*
$\Omega = (\Phi, \Theta)$ — параметры порождающей модели $p(X|\Omega)$
$R(\Omega)$ — регуляризатор; не обязательно байесовский $\ln p(\Omega|\gamma)$

Задача максимизации правдоподобия — по X найти Ω:

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

Порождающая модель наблюдаемых и скрытых i.i.d. данных:

$$p(X, Z|\Omega) = \prod_{i=1}^{n} p(d_i, w_i, t_i) =$$
$$= \prod_{i=1}^{n} p(w_i|t_i)p(t_i|d_i)p(d_i) = \prod_{i=1}^{n} \phi_{w_i t_i}\theta_{t_i d_i} p(d_i)$$

Регуляризованный ЕМ-алгоритм для тематической модели

Наблюдаемые $X = (d_i, w_i)_{i=1}^n$, латентные $Z = (t_i)_{i=1}^n$

Теорема. Точка $\Omega = (\Phi, \Theta)$ локального максимума RML (регуляризованного маргинализованного log-правдоподобия)

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi, \Theta)$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

$$\text{Е-шаг:} \quad p(t|d, w) = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr), \quad \forall(d\in D, w\in d, t\in T)$$

$$\text{М-шаг:} \quad \sum_{d,w,t} n_{dw} p(t|d, w) \ln\bigl(\phi_{wt}\theta_{td}\bigr) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

Доказательство леммы

E-шаг: в силу независимости элементов выборки и формулы Байеса

$$q(Z) = p(Z|X,\Omega) = \prod_{i=1}^{n} p(t_i|d_i,w_i) = \prod_{i=1}^{n} \mathop{\mathrm{norm}}_{t_i\in T}\bigl(\phi_{w_i t_i}\theta_{t_i d_i}\bigr)$$

M-шаг: подставим $q(Z)$ и $p(X,Z|\Omega)$ в общую формулу M-шага:

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

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

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

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

$$\sum_{d\in D}\sum_{w\in W}\sum_{t\in T} n_{dw}\underbrace{p(t|d,w)}_{p_{tdw}}\ln\bigl(\phi_{wt}\theta_{td}\bigr) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

ARTM: аддитивная регуляризация тематических моделей

Задача М-шага декомпозируется на независимые подзадачи:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl} \text{Е-шаг:} & \left\{ \begin{array}{l} p_{tdw} \equiv p(t \mid d, w) = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad n_{wt} = \sum_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname*{norm}_{t\in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w\in d} n_{dw} p_{tdw} \end{array} \right. \\ \text{М-шаг:} & \end{array}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Два наиболее известных частных случая: модели PLSA и LDA

PLSA: probabilistic latent semantic analysis [Hofmann, 1999] (вероятностный латентный семантический анализ):

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

М-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt}\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td}\bigr).$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}.$$

М-шаг — частотные оценки с поправками $\beta_w > 0$, $\alpha_t > 0$:

$$\phi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt} + \beta_w - 1\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td} + \alpha_t - 1\bigr).$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

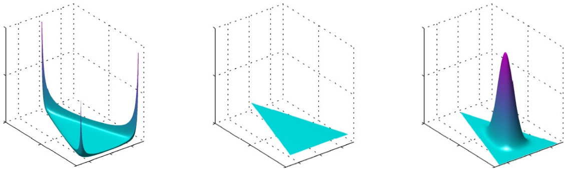
Распределение Дирихле

Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})$ и $\theta_d = (\theta_{td})$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

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

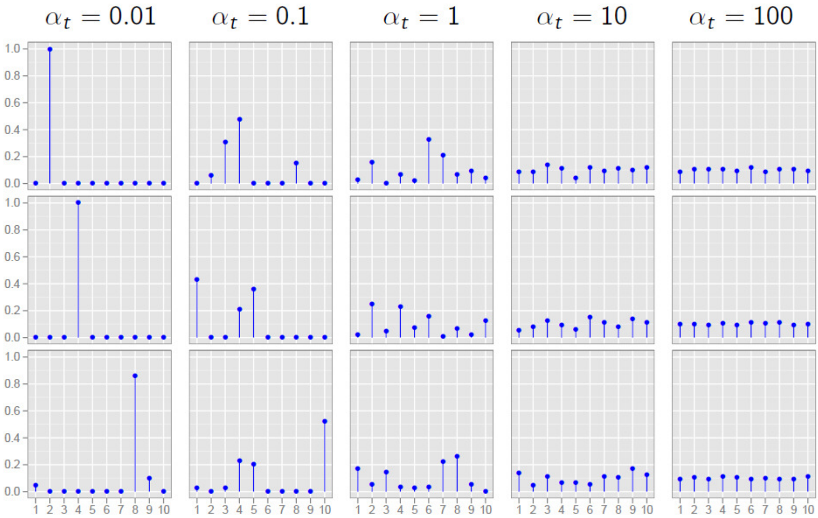
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

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



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ $\alpha_1 = \alpha_2 = \alpha_3 = 1$ $\alpha_1 = \alpha_2 = \alpha_3 = 10$

Пример. Выборки из трёх 10-мерных векторов $\theta \sim \mathrm{Dir}(\theta|\alpha)$



Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

$$\ln \prod_{d\in D} \prod_{w\in d} p(w,d|\Phi,\Theta)^{n_{dw}} \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha) \to \max_{\Phi,\Theta}$$

Регуляризатор — логарифм априорного распределения:

$$R(\Phi,\Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}$$

M-шаг — сглаженные или разреженные частотные оценки:

$$\phi_{wt} = \underset{w}{\mathrm{norm}}\bigl(n_{wt} + \beta_w - 1\bigr), \qquad \theta_{td} = \underset{t}{\mathrm{norm}}\bigl(n_{td} + \alpha_t - 1\bigr).$$

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,
при $0 < \beta_w < 1$, $0 < \alpha_t < 1$ — слабое разреживание,
при $\beta_w = 1$, $\alpha_t = 1$ априорное распределение равномерно, PLSA.

Обобщение LDA: регуляризатор сглаживания и разреживания

Общий вид регуляризатора сглаживания и разреживания:

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

где $\beta_0 > 0$, $\alpha_0 > 0$ — коэффициенты регуляризации,
β_{wt}, α_{td} — параметры, задаваемые пользователем:

- $\beta_{wt} > 0$, $\alpha_{td} > 0$ — сглаживание
- $\beta_{wt} < 0$, $\alpha_{td} < 0$ — разреживание

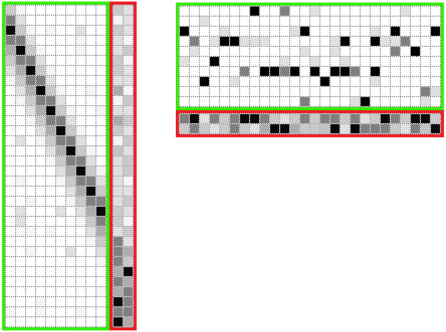
Возможные применения сглаживания и разреживания:

- задать фоновые темы с общей лексикой языка
- задать шумовую тему для нетематичных термов
- задать псевдо-документ с ключевыми термами темы
- скорректировать состав термов и документов темы

Разделение тем на предметные и фоновые

Предметные темы S содержат термины предметной области,
$p(w|t)$, $p(t|d)$, $t \in S$ — разреженные, существенно различные
Фоновые темы B содержат слова общей лексики,
$p(w|t)$, $p(t|d)$, $t \in B$ — существенно отличные от нуля

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



Регуляризатор декоррелирования тем

Цель: усилить различность тем; выделить в каждой теме лексическое ядро, отличающее её от других тем;
вывести слова общей лексики из предметных тем в фоновые.

Минимизируем ковариации между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t\in T} \sum_{s\in T\backslash t} \sum_{w\in W} \phi_{wt}\phi_{ws} \to \max.$$

Подставляем в формулы М-шага, получаем ещё один вариант разреживания — контрастирование строк матрицы Φ (малые вероятности ϕ_{wt} в строке становятся ещё меньше):

$$\phi_{wt} = \operatorname*{norm}_{w}\Bigl(n_{wt} - \tau\phi_{wt} \sum_{s\in T\backslash t} \phi_{ws}\Bigr).$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. 2010

Байесовская и классическая регуляризация

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (громоздкий, приближённый) ради точечной оценки Ω:

$$\mathrm{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\mathrm{Prior}(\Omega|\gamma)$$

$$\Omega := \arg\max_{\Omega} \mathrm{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg\max_{\Omega}\bigl(\ln p(X|\Omega) + \ln \mathrm{Prior}(\Omega|\gamma)\bigr)$$

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg\max_{\Omega}\Bigl(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Bigr)$$

Резюме

- Тематическое моделирование — «мягкая кластеризация», автокодировщик или стохастическое матричное разложение
- Стандартные методы — PLSA и LDA
- Нестандартные — огромное разнообразие регуляризаторов
- Аддитивная регуляризация позволяет комбинировать модели
- Обычно в ТМ используется байесовское обучение. **Почему оно не нужно в ТМ:** на практике используются не апостериорные распределения $p(\Omega|X, \gamma)$, а их точечные оценки Ω по максимуму правдоподобия
- В ARTM те же модели выводятся намного проще — с помощью Леммы о максимизации на симплексах,
- она применима для оптимизации любых моделей, параметры которых — неотрицательные нормированные векторы (дискретные вероятностные распределения)

Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. 2017–2023.
http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf