Методы машинного обучения. Тематическое моделирование: «мягкая» кластеризация текстов

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-21-22
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 22 марта 2022

Содержание

1. **Вероятностное тематическое моделирование**
   - Цели, приложения, постановка задачи
   - Метод оптимизации на единичных симплексах
   - Аддитивная регуляризация тематических моделей
2. **Примеры регуляризаторов**
   - Модели PLSA и LDA
   - Априорное распределение Дирихле
   - Регуляризатор декоррелирования тем
3. **EM-алгоритм для задач со скрытыми переменными**
   - Байесовская регуляризация
   - Теория EM-алгоритма
   - EM-алгоритм для тематического моделирования

Задача тематического моделирования

Дано: коллекция текстовых документов

- W — конечное множество термов (слов, токенов)
- D — конечное множество документов
- n_{dw} — частота терма w в документе d

Найти: вероятностную тематическую модель языка

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

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

Критерий: максимум логарифма правдоподобия

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

при ограничениях $\phi_{wt} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Постановка задачи максимизации 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

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

приводит к задаче математического программирования:

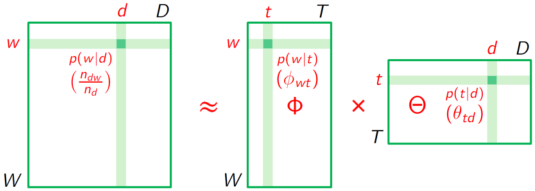
$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

при ограничениях неотрицательности и нормировки

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

Три интерпретации задачи тематического моделирования

- **Мягкая кластеризация** документов по кластерам-темам
- **Стохастическое матричное разложение:**



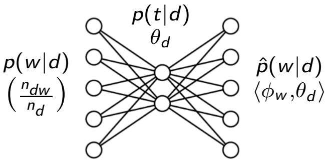
- **Автокодировщик** документов в тематические эмбединги:

кодировщик $f_\Phi\colon \frac{n_{dw}}{n_d} \to \theta_d$

декодировщик $g_\Phi\colon \theta_d \to \Phi\theta_d$

задача реконструкции:

$$\sum_d \mathrm{KL}\Big(\frac{n_{dw}}{n_d} \,\Big\|\, \langle\phi_w, \theta_d\rangle\Big) \;\to\; \min_{\Phi,\Theta}$$

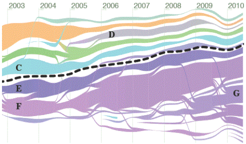


Некоторые приложения тематического моделирования

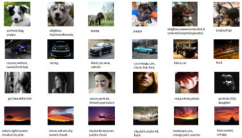
разведочный поиск в электронных библиотеках

поиск тематического контента в соцсетях

детектирование и трекинг новостных сюжетов

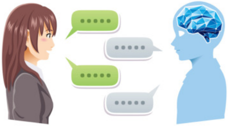


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

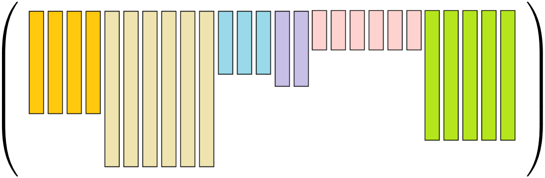
управлением диалогом в разговорном интеллекте



Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$\Omega =$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \operatorname*{norm}_{i \in I}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Теорема. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \operatorname*{norm}_{i \in I_j} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Решения $\omega_j \equiv 0$ будем считать вырожденным и отбрасывать
- Итерации похожи на градиентную оптимизацию, но учитывают ограничения и не требуют подбора шага η:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}$$

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Доказательство Леммы о максимизации на симплексах

Запишем условия Каруша–Куна–Таккера для ω_{ij}:

$$\frac{\partial f}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}; \quad \mu_{ij}\omega_{ij} = 0.$$

Предполагая $\omega_{ij} > 0$, умножим обе части равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = \omega_{ij}\lambda_j.$$

Возможны три случая:

1. Если $\lambda_j > 0$, то либо $A_{ij} > 0$, либо $\omega_{ij} = 0$. Тогда $\omega_{ij}\lambda_j = (A_{ij})_+;\ \lambda_j = \sum\limits_i (A_{ij})_+ \ \Rightarrow\ \omega_{ij} = \operatorname*{norm}\limits_i(A_{ij})$.
2. Если $\lambda_j < 0$ и $(\exists i)\ A_{ij} < 0$, то $(\forall i)\ A_{ij} \leqslant 0$. Тогда $\omega_{ij}\lambda_j = -(-A_{ij})_+;\ \lambda_j = -\sum\limits_i (-A_{ij})_+ \ \Rightarrow\ \omega_{ij} = \operatorname*{norm}\limits_i(-A_{ij})$.
3. Иначе $\lambda_j = 0$ и ω_j находится из уравнений $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = 0$.

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена по Адамару*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Наша задача матричного разложения *некорректно поставлена*: если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta), \;\; \operatorname{rank} S = |T|$
- $L(\Phi', \Theta') = L(\Phi, \Theta)$
- $L(\Phi', \Theta') \leqslant L(\Phi, \Theta) + \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned}
&\text{E-шаг:} \quad && \begin{cases} p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\mathrm{norm}}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \underset{w\in W}{\mathrm{norm}}\Bigl(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\Bigr), \quad n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\mathrm{norm}}\Bigl(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Bigr), \quad n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw} \end{cases}\\
&\text{M-шаг:} &&
\end{aligned}$$

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

$$\phi_{wt} = \mathop{\mathrm{norm}}_{w\in W}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \mathop{\mathrm{norm}}_{w\in W}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) =$$

$$= \mathop{\mathrm{norm}}_{w\in W}\left(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right);$$

$$\theta_{td} = \mathop{\mathrm{norm}}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \mathop{\mathrm{norm}}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\phi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \mathop{\mathrm{norm}}_{t\in T}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ).

Тема t вырождена, если для всех термов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leqslant 0.$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$; это означает, что тема исключается из модели (происходит отбор тем).

Документ d вырожден, если для всех тем $t \in T$

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leqslant 0.$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$; это означает, что модель не в состоянии описать данный документ.

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.

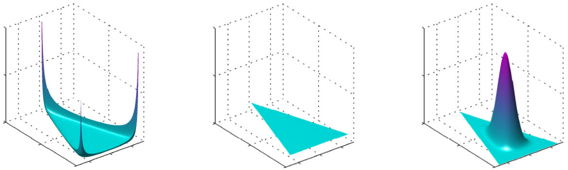
Распределение Дирихле

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

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

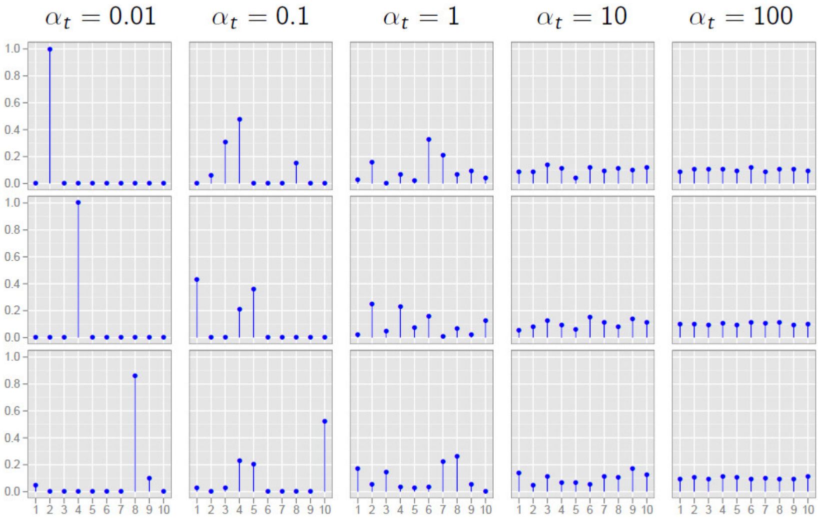
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

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



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

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



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:

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

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

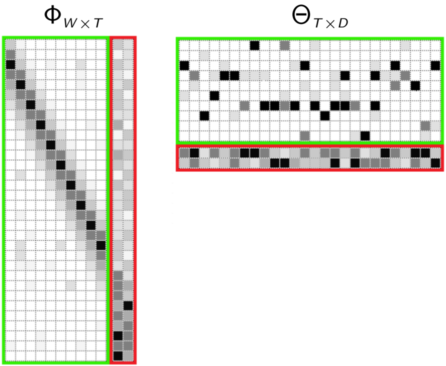
Возможные применения сглаживания и разреживания:

- 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



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 M-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

Основы байесовской регуляризации

Введём более общие обозначения:

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*
$\Omega = (\Phi, \Theta)$ — параметры порождающей модели $p(X|\Omega)$
$\gamma = (\beta, \alpha)$ — гиперпараметры *априорного распределения* $p(\Omega|\gamma)$

Задача: по X найти Ω.

Формула Байеса даёт *апостериорное распределение* $p(\Omega|X,\gamma)$, где символ $\propto$ означает «равно с точностью до нормировки»:

$$p(\Omega|X,\gamma) = \frac{p(\Omega,X|\gamma)}{p(X|\gamma)} \propto p(\Omega,X|\gamma) \propto p(X|\Omega)\, p(\Omega|\gamma)$$

Далее есть два пути:

- Максимизация правдоподобия: $\Omega = \arg\max_\Omega \ln p(\Omega|X,\gamma)$
- Байесовский вывод: вычисление распределения $p(\Omega|X,\gamma)$

Вероятностная модель со скрытыми переменными

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*
$Z = (t_i)_{i=1}^n$ — *скрытые переменные*
$\Omega = (\Phi, \Theta)$ — параметры порождающей модели $p(X|\Omega)$
$\gamma = (\beta, \alpha)$ — гиперпараметры *априорного распределения* $p(\Omega|\gamma)$

Задача: по X найти Ω.

Апостериорное распределение:

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

Принцип максимума апостериорной вероятности:

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

$R(\Omega)$ 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)

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

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

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

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

Это общий вид EM-алгоритма, используемый не только в тематическом моделировании.

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:

$$\frac{\partial}{\partial\Omega}\biggl(\ln\sum_Z p(X,Z|\Omega) + R(\Omega)\biggr) = \frac{1}{p(X|\Omega)}\sum_Z \frac{\partial p(X,Z|\Omega)}{\partial\Omega} + \frac{\partial R(\Omega)}{\partial\Omega} = 0$$

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

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

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

Это необходимые условия локального экстремума задачи M-шага, если $q(Z)$ рассматривать как константу, а не как функцию от Ω.

■

Ещё более общий EM-алгоритм и его сходимость

Теорема. Значение маргинализованного правдоподобия

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

не убывает на каждом шаге итерационного процесса

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

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

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

Минимизация KL на E-шаге используется в тех случаях, когда $p(Z|X,\Omega)$ не удаётся вычислить 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 $q(Z)$

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

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

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

E-шаг: $L(q,\Omega) + \cancel{R(\Omega)} \to \max\limits_q \;\Leftrightarrow\; \mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X,\Omega)\big) \to \min\limits_q$

M-шаг: $L(q,\Omega) + R(\Omega) \to \max\limits_\Omega \;\Leftrightarrow\; \sum\limits_Z q(Z) \ln p(X,Z|\Omega) + R(\Omega) \to \max\limits_\Omega$

На каждом шаге значение функционала может только возрастать.

■

Регуляризованный EM-алгоритм для тематической модели

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

Лемма. Точка (Φ,Θ) 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:

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

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

Доказательство леммы

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

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

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

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

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

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

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

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

■

Вывод M-шага ARTM из общего EM-алгоритма

Оптимизационная задача M-шага:

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

Применим лемму о максимизации на единичных симплексах:

$$\phi_{wt} = \mathop{\mathrm{norm}}_{w\in W}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \mathop{\mathrm{norm}}_{w\in W}\left(\underbrace{\sum_{d\in D} n_{dw} p(t|d, w)}_{n_{wt}} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$$

$$\theta_{td} = \mathop{\mathrm{norm}}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \mathop{\mathrm{norm}}_{t\in T}\left(\underbrace{\sum_{w\in d} n_{dw} p(t|d, w)}_{n_{td}} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$$

Таким образом, снова получили формулы ARTM

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, а их точечные оценки
- В ARTM те же модели выводятся намного проще — с помощью Теоремы о максимизации на симплексах,
- которая применима для оптимизации любых моделей с векторами дискретных вероятностных распределений

Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. 2017–2022. http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf