

# Устойчивость алгоритмов обучения вероятностных тематических моделей

автор задачи: *Воронцов Константин Вячеславович*, voron@forecsys.ru  
консультант: *Потапенко Анна Александровна*, anya\_potapenko@mail.ru

26 февраля 2014 г.

## 1 Предварительные сведения

### 1.1 Вероятностная тематическая модель коллекции текстов

Пусть  $D$  — множество (коллекция) текстовых документов,  $W$  — множество (словарь) терминов. Известно число  $n_{dw}$  вхождений каждого из терминов  $w$  в каждый документ  $d \in D$ . Предполагается, что каждое слово в каждом документе связано с некоторой неизвестной (латентной) темой  $t$  из множества тем  $T$ . Обычно число тем  $|T|$  много меньше  $|D|$  и  $|W|$ . Задача заключается в том, чтобы выявить латентные темы по наблюдаемой коллекции  $D$ .

*Вероятностная тематическая модель* описывает вероятность появления терминов в документе:

$$p(w | d) = \sum_{t \in T} \phi_{wt} \theta_{td}, \quad (1)$$

где  $\phi_{wt} = p(w | t)$  — неизвестное распределение на множестве терминов, определяющее тему  $t$ ;  $\theta_{td} = p(t | d)$  — неизвестное распределение на множестве тем, определяющее тематическую структуру документа  $d$ .

В *вероятностном латентном семантическом анализе* PLSA [4] для обучения модели (1) по коллекции документов  $D$  максимизируется логарифм правдоподобия при ограничениях нормировки и неотрицательности:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}; \quad (2)$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \quad (3)$$

Задача (2), (3) является некорректно поставленной, поскольку имеет в общем случае бесконечно много решений. Действительно, в правой части (1) записано произведение *матрицы терминов тем*  $\Phi = (\phi_{wt})_{W \times T}$  и *матрицы тем документов*  $\Theta = (\theta_{td})_{T \times D}$ . Матричное произведение определено с точностью до невырожденного линейного преобразования:  $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$ . Следствием неединственности является неустойчивость любых численных методов оптимизации, используемых для решения данной задачи, в том числе EM-алгоритма.

Известны различные условия единственности неотрицательных матричных разложений [2, 5, 6, 3]. В частности, разложение может быть единственным при сильной разреженности матриц  $\Phi$  и  $\Theta$ . Однако остаются не ясны вопросы: какой должна быть структура разреженности, т.е. как должны быть расположены нулевые элементы в матрицах  $\Phi$  и  $\Theta$ ; какие дополнительные ограничения необходимо на них наложить, чтобы решение стало устойчивым; возможно ли связать эти ограничения с требованиями интерпретируемости (понятности) тем.

Исследовательские задания направлены на поиск ответов на эти вопросы.

## 1.2 EM-алгоритм

Для решения оптимизационной задачи (2), (3) используется итерационный процесс, называемый EM-алгоритмом [1]. Столбцы матриц  $\Phi$  и  $\Theta$  инициализируются случайными распределениями. На каждой итерации выполняются два шага.

На E-шаге для всех терминов в документах  $(d, w)$  по формуле Байеса вычисляются условные распределения  $p(t | d, w)$ . Вводятся вспомогательные переменные  $n_{dwt}$ , интерпретируемые как число вхождений термина  $w$  в документ  $d$  по теме  $t$ :

$$n_{dwt} = n_{dw}p(t | d, w), \quad p(t | d, w) = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}. \quad (4)$$

На M-шаге суммирование переменных  $n_{dwt}$  по  $d, w, t$  даёт частотные оценки искомых условных вероятностей:

$$\begin{aligned} \phi_{wt} &= \frac{n_{wt}}{n_t}, & n_{wt} &= \sum_{d \in D} n_{dwt}, & n_t &= \sum_{w \in W} n_{wt}, \\ \theta_{td} &= \frac{n_{dt}}{n_d}, & n_{dt} &= \sum_{w \in d} n_{dwt}, & n_d &= \sum_{t \in T} n_{dt}, \end{aligned}$$

или, в краткой записи через знак пропорциональности  $\propto$ :

$$\phi_{wt} \propto n_{wt}, \quad \theta_{td} \propto n_{dt}. \quad (5)$$

Система уравнений (4), (5) определяет стационарную точку оптимизационной задачи (2), (3). Существуют различные способы построения итерационного процесса по формулам (4), (5): рациональный, стохастический, онлайнный [1].

## 1.3 Регуляризованный EM-алгоритм

Допустим, что наряду с правдоподобием (2) требуется максимизировать ещё  $k$  критериев  $R_i(\Phi, \Theta)$ ,  $i = 1, \dots, k$ , называемых *регуляризаторами*. Для решения задачи многокритериальной оптимизации будем максимизировать линейную комбинацию критериев с неотрицательными *коэффициентами регуляризации*  $\tau_i$ :

$$R(\Phi, \Theta) = \sum_{i=1}^k \tau_i R_i(\Phi, \Theta), \quad L(\Phi, \Theta) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}. \quad (6)$$

Данная задача по-прежнему решается с помощью EM-алгоритма, но вместо (5) используется модифицированная формула M-шага [1]:

$$\phi_{wt} \propto \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+, \quad \theta_{td} \propto \left( n_{dt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+, \quad (7)$$

где  $(z)_+ = \max\{z, 0\}$  — операция положительной срезки.

## 2 Постановка задачи

Задания предполагают выполнение вычислительных экспериментов на модельных данных. Каждый эксперимент заключается в многократном восстановлении модельных матриц  $\Phi_0, \Theta_0$  при различных значениях некоторой выбранной характеристики задачи.

Отклонение восстановленных распределений  $p(i | j)$  от модельных  $p_0(i | j)$  измеряется средним расстоянием Хеллингера

$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^m \sqrt{\frac{1}{2} \sum_{i=1}^n \left( \sqrt{p(i | j)} - \sqrt{p_0(i | j)} \right)^2},$$

как для самих матриц  $\Phi, \Theta$ , так и для их произведения  $\Phi\Theta$ :

$$\begin{aligned} D_\Phi(\Phi, \Phi_0) &= H(\Phi, \Phi_0); \\ D_\Theta(\Theta, \Theta_0) &= H(\Theta, \Theta_0); \\ D_{\Phi\Theta}(\Phi\Theta, \Phi_0\Theta_0) &= H(\Phi\Theta, \Phi_0\Theta_0). \end{aligned}$$

---

**Алгоритм 2.1.** Восстановление модельных матриц  $\Phi_0$  и  $\Theta_0$ .

---

- 1 сгенерировать модельные матрицы  $\Phi_0$  и  $\Theta_0$ ;
  - 2 сгенерировать коллекцию  $D$ ;
  - 3 EM-алгоритм: восстановить по коллекции  $D$  матрицы  $\Phi$  и  $\Theta$ ;
  - 4 венгерский алгоритм: найти соответствие между темами в  $(\Phi_0, \Theta_0)$  и  $(\Phi, \Theta)$ ;
  - 5 вычислить расстояния  $D_\Phi, D_\Theta, D_{\Phi\Theta}$ ;
- 

Генерация реалистичных модельных данных должна учитывать гипотезы разреженности, слабой корреляции тем, наличия фоновых тем. Возможно взять в качестве  $\Phi_0, \Theta_0$  результат тематического моделирования реальной коллекции.

Рекомендуемый способ сгенерировать коллекцию с минимальным уровнем шума:

$$n_{dw} = \text{pround} \left( n_d \sum_{t \in T} \phi_{wt} \theta_{td} \right),$$

где  $\text{pround}$  — функция вероятностного округления,  $n_d$  — заданные длины документов.

Задания предполагают построение графиков зависимости всех трёх расстояний от выбранной характеристики задачи.

На этапе тестирования модели строится зависимость перплексии и расстояний  $D_\Phi, D_\Theta, D_{\Phi\Theta}$  от номера итерации в EM-алгоритме. Это необходимо для выяснения необходимого числа итераций при заданных параметрах задачи.

### 2.1 Общая часть задания

**Задание 1.** Проверяется гипотеза, что чем выше разреженность матриц  $\Phi$  и  $\Theta$ , тем выше устойчивость решения. Вторая гипотеза: регуляризатор разреживания повышает устойчивость решения.

Исследуется зависимость расстояний  $D_\Phi, D_\Theta, D_{\Phi\Theta}$  от разреженности (доли ненулевых элементов) исходных матриц.

Это задание обязательное. Оно повторяет эксперимент В. Глушаченкова, описанный в [1]. Выбор одного из следующих заданий должен быть согласован с авторами задания.

## 2.2 Вариативная часть задания

**Задание 2.** Проверяется гипотеза, что чем выше различность тем (столбцов  $\Phi$ ), тем выше устойчивость решения. Вторая гипотеза: регуляризатор декоррелирования повышает устойчивость решения. Третья гипотеза: регуляризаторы разреживания и декоррелирования вместе повышают устойчивость решения.

Исследуется зависимость расстояний  $D_\Phi$ ,  $D_\Theta$ ,  $D_{\Phi\Theta}$  от средней корреляции между темой и ближайшей к ней темой в исходной матрице  $\Phi$ .

**Задание 3.** Проверяется гипотеза, что регуляризатор разреживания строк матрицы  $\Theta$  способен устойчиво определять истинное число тем  $|T|$ , если начинать EM-алгоритм с заведомо избыточного числа тем.

Исследуется зависимость расстояний  $D_\Phi$ ,  $D_\Theta$ ,  $D_{\Phi\Theta}$ , а также перплексии и числа ненулевых тем ( $n_t > 0$ ) от номера итерации в EM-алгоритме.

**Задание 4.** Проверяется гипотеза, что наличие фоновых тем не сильно мешает восстановить основные предметные темы. Вторая гипотеза: при наличии фоновых тем совместное применение регуляризаторов разреживания, сглаживания и декорреляции повышает устойчивость решения. Третья гипотеза: ошибка при назначении числа фоновых тем не сильно влияет на результат.

Исследуется зависимость расстояний  $D_\Phi$ ,  $D_\Theta$ ,  $D_{\Phi\Theta}$ , (они вычисляются только по предметным темам), от доли фоновых тем в документах коллекции. Затем, при фиксированной реалистичной доле фоновых тем (30%–70%) исследуется зависимость расстояний  $D_\Phi$ ,  $D_\Theta$ ,  $D_{\Phi\Theta}$  от средней корреляции между темой и ближайшей к ней темой в исходной матрице  $\Phi$ .

**Задание 5.** Проверяется гипотеза, что добавление в коллекцию «виртуальных документов», каждый из которых содержит список топовых слов одной темы, повышает устойчивость решения (применяется регуляризатор для частичного обучения [1]). Вторая гипотеза: достаточно лишь небольшого числа топовых слов. Третья гипотеза: достаточно задать топовые слова лишь для небольшой части тем.

Исследуется зависимость расстояний  $D_\Phi$ ,  $D_\Theta$ ,  $D_{\Phi\Theta}$  от числа виртуальных документов и суммарного числа слов в виртуальных документах.

**Задание 6.** Проверяется гипотеза, что если темы сгруппированы по классам, и для каждого документа известны его классы, то это повышает устойчивость решения (число тем порядка 100, число классов порядка 20; применяется ковариационный регуляризатор для категоризации текстов [1]).

## Список литературы

- [1] *Воронцов К. В.* Вероятностное тематическое моделирование. — 2014. <http://www.MachineLearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>.
- [2] *Donoho D., Stodden V.* When does non-negative matrix factorization give a correct decomposition into parts? // *Advances in Neural Information Processing Systems 2003* / Ed. by S. Thrun, L. Saul, B. Schölkopf. — Cambridge, MA: MIT Press, 2004. <http://www-stat.stanford.edu/~donoho/Reports/2003/NMFCDP.pdf>.
- [3] *Gillis N.* Sparse and unique nonnegative matrix factorization through data preprocessing // *Journal of Machine Learning Research.* — 2012. — Vol. 13, no. 1. — Pp. 3349–3386.
- [4] *Hofmann T.* Probabilistic latent semantic indexing // *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* — New York, NY, USA: ACM, 1999. — Pp. 50–57.
- [5] *Lauerberg H., Christensen M., Pumbley M., Hansen L., Jensen S.* Theorems on positive data: On the uniqueness of nmf // *Computational Intelligence and Neuroscience.* — 2008. — Vol. 2008. — P. 10.
- [6] *Schachtner R., Pöppel G., Lang E. W.* Towards unique solutions of non-negative matrix factorization problems by a determinant criterion // *Digital Signal Processing.* — 2011. — Vol. 21, no. 4. — Pp. 528–534.