Название: Оптимизация числа тем в вероятностных тематических моделях с помощью регуляризатора строкового разреживания.

Задача: Вероятностная тематическая модель описывает вероятности появления слов $w \in W$ в документах $d \in D$ через латентные темы $t \in T$:

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

Требуется проверить гипотезу, что, накладывая ограничения на матрицу Θ с помощью регуляризатора строкового разреживания, возможно определить оптимальное число тем.

Данные: Коллекция документов D задаётся частотами слов n_{dw}. Поскольку для решения задачи необходимо знать «истинное» число тем $|T|$, эксперименты производятся на реалистичных модельных или полумодельных данных.

Решение: Для обучения модели по коллекции документов D максимизируется логарифм правдоподобия с n аддитивными регуляризаторами R_i, при ограничениях нормировки и неотрицательности:

$$L(D;\Phi,\Theta) = \sum_{d \in D}\sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

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

Предполагается задавать заведомо избыточное число тем $|T|$ и использовать регуляризатор $R(\Phi,\Theta) = \mathrm{KL}\big(\frac{1}{|T|} \,\big\|\, \sum_d \theta_{td}\frac{n_d}{n}\big)$, разреживающий строки матрицы Θ целиком, возможно, в комбинации с другими регуляризаторами. В качестве основного средства визуализации результатов предполагается использовать графики зависимости перплексии и числа тем (ненулевых строк матрицы Θ) от номера итерации.

Базовой алгоритм: Для решения оптимизационной задачи используется регуляризованный EM-алгоритм [2, 1]. Может быть использована рациональная, стохастическая или онлайновая версия EM-алгоритма.

Новизна: Для оптимизации числа тем $|T|$ обычно используется модель иерархического процесса Дирихле HDP [3]. Она определяет число тем неустойчиво, и при этом сложна как для понимания, так и для реализации. Аддитивная регуляризация тематических моделей (ARTM) — это новый подход к тематическому моделированию [2, 1], сочетающий универсальность, гибкость и простоту. Задача оптимизации числа тем ещё не рассматривалась в рамках ARTM.

Список литературы

[1] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. — Т. 455, № 3 (в печати).

[2] *Воронцов К. В.* Вероятностное тематическое моделирование. — 2014. `http://www.MachineLearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf`.

[3] *Teh Y. W., Jordan M. I., Beal M. J., Blei D. M.* Hierarchical Dirichlet processes // Journal of the American Statistical Association. — 2006. — Vol. 101, no. 476. — Pp. 1566–1581.