Полувероятностная тематическая модель для задачи классификации

Смирнов Евгений Александрович

Московский физико-технический институт
Вычислительный центр РАН

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 274, весна 2015

Введение

Цель исследования:

Разработать полувероятностную тематическую модель для задачи классификации. Построить на её основе *EM*-алгоритм и сравнить с алгоритмом основанным на методе *LDA*.

Новизна:

Использование тематической модели для задачи классификации.

Метод решения:

EM-алгоритм. *ARTM.LDA*

Литература

1. David M. Blei Andrew Y. Ng, M. I. J. Latent Dirichlet Allocation Journal of Machine Learning Research, 2003
2. Herlocker J. L.; Konstan, J. A. T. L. G. R. J. T. Evaluating collaborative filtering recommender systems ACM Trans. Inf. Syst. 22 (1): 5–53. doi:10.1145/963770.963772, 2004
3. Hoffman, T. Probabilistic Latent Semantic Indexing Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, 1999
4. Vorontsov K. V., P. A. A. Additive Regularization of Topic Models Machine Learning. Special Issue “Data Analysis and Intelligent Optimization with Applications, 2014
5. K.B., B. Вероятностное тематическое моделирование http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf, 2013

Постановка задачи

Дано:

D - коллекция текстовых документов W – термины C - классы $d \in D$ - последовательность пар (w, c).

Предположения:

- для рекомендации выбирается тема t из распределения $p(t|d)$;
- из темы t выбирается слово на основании $p(w|t)$;
- документ относится к классу c на основе $p(c|t, d)$

Найти:

распределения $p(w|t)$ и $p(c, t|d)$

Теоретическая часть

Запишем тематическую модель нашей коллекции

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

Равенство можно рассматривать, как задачу стохастического разложения матрицы
$F \approx \Phi\Theta$
Нахождения параметров Φ, Θ эквивалентно максимизации по ним функционала

$$Q(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} \sum_{c \in C} n_{dwc} \log(\sum_{t \in T} \phi_{wt}\theta_{ctd}) + R(\Phi, \Theta)$$

при условиях

$$\sum_{w \in W} \phi_{wt} = 1,\ \phi_{wt} \geq 0,\ \sum_{c \in C} \sum_{t \in T} \theta_{ctd} = 1,\ \theta_{ctd} \geq 0$$

Теорема:

Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Точка локального экстремума (Φ, Θ) функции $Q(\Phi, \Theta)$ удовлетворяет системе уравнений

$$p_{twcd} \propto \phi_{wt}\theta_{ctd}$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+ \qquad n_{wt} = \sum_{d\in D}\sum_{c\in C} n_{dwc}p_{twcd}$$

$$\theta_{ctd} \propto \left(n_{ctd} + \theta_{ctd}\frac{\partial R}{\partial \theta_{ctd}}\right)_+ \qquad n_{ctd} = \sum_{w\in W} n_{dwc}p_{twcd}$$

где n_{dwc} – частота отнесения термина w пользователем d к классу c

Регуляризаторы

- Декоррелирующий $R(\Phi) = -\gamma \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt}\phi_{ws}$
- Сглаживающий
  $R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{s \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td}$

M-шаг

$$\phi_{wt} \propto \left(n_{wt} - \beta_0 \beta_w - \gamma \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \right)_+$$

$$\theta_{ctd} \propto (n_{ctd} - \alpha_0 \alpha_t)_+$$

,где α_t, β_t из равномерного распределения.

Алгоритм

EM-алгоритм

- **Инициализация:** ϕ_{wt} взять из распределения Пуассона, θ_{td} из равномерного распределения;
- *E*-**шаг:**Вычислить $p(w, c|d)t \in T, \forall w \in W, c \in C, d \in D$;
- *M*-**шаг:** Вычислить новое приближение параметров ϕ_{wt}, θ_{ctd} ;
- Повторять пока Φ и Θ не сойдутся;

Вычислительный эксперимент

Дано:

- Коллекция слов $W = (w_n)_{n=1}^{P}$, $U = (u_n)_{n=1}^{S}$ - множество пользователей.
- Выборки $(u_n, a_n)_{n=1}^{N}$ и $(u_n, o_n)_{n=1}^{M}$, где $u_n \in U$, $a_n \in W$, $o_n \in W$ id пользователя, откинутые и принятые слова.

Найти

Распределения $p(w|t), p(c, t|d)$.

Вычислительный эксперимент

$$E(err) = 4 * 10^{-4} \qquad P(err) = 0.0410$$

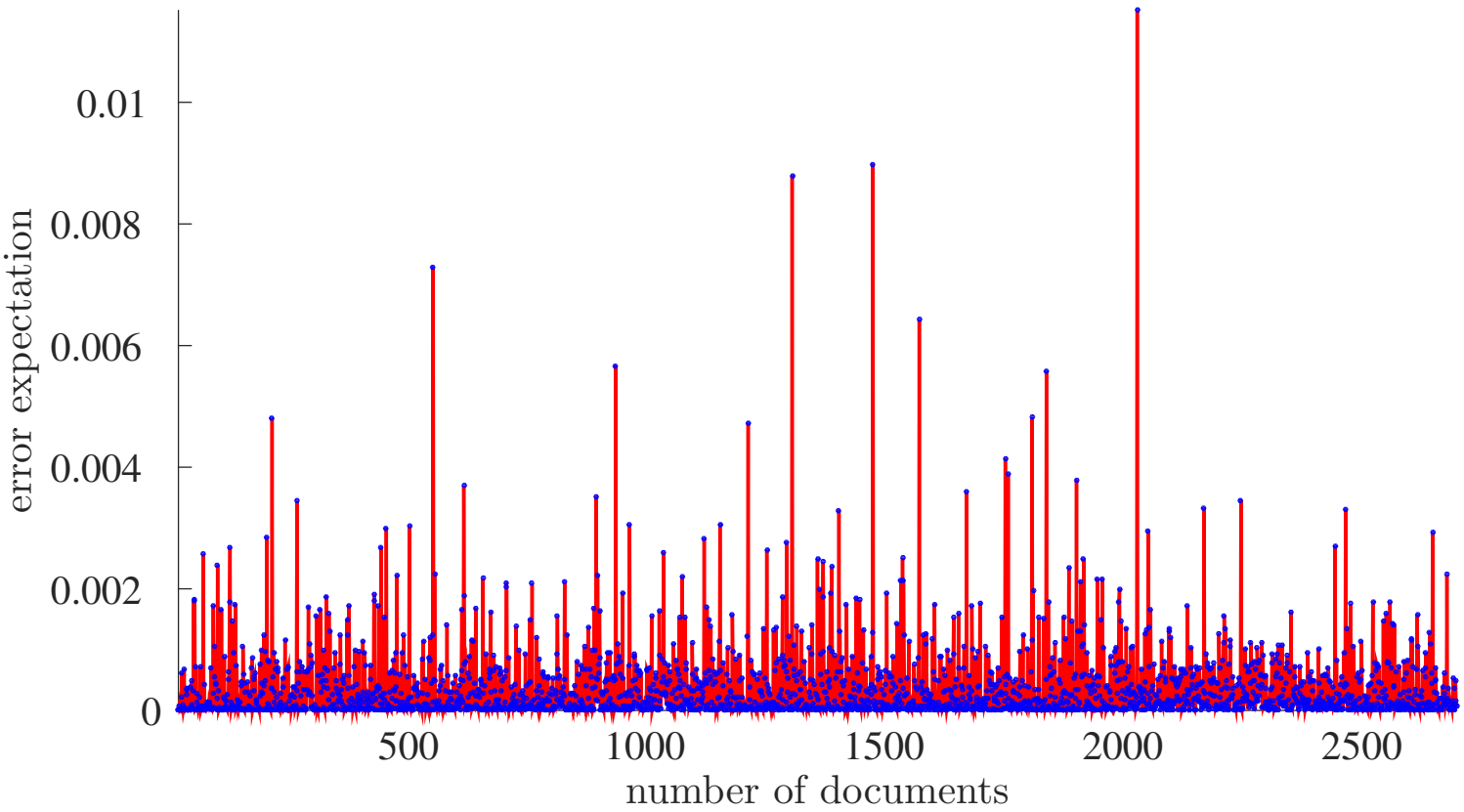

Рис.: Сэмплирование Гиббса для модели LDA

Вычислительный эксперимент

$$E(err) = 2.3 * 10^{-3} \qquad P(err) = 0.2322$$

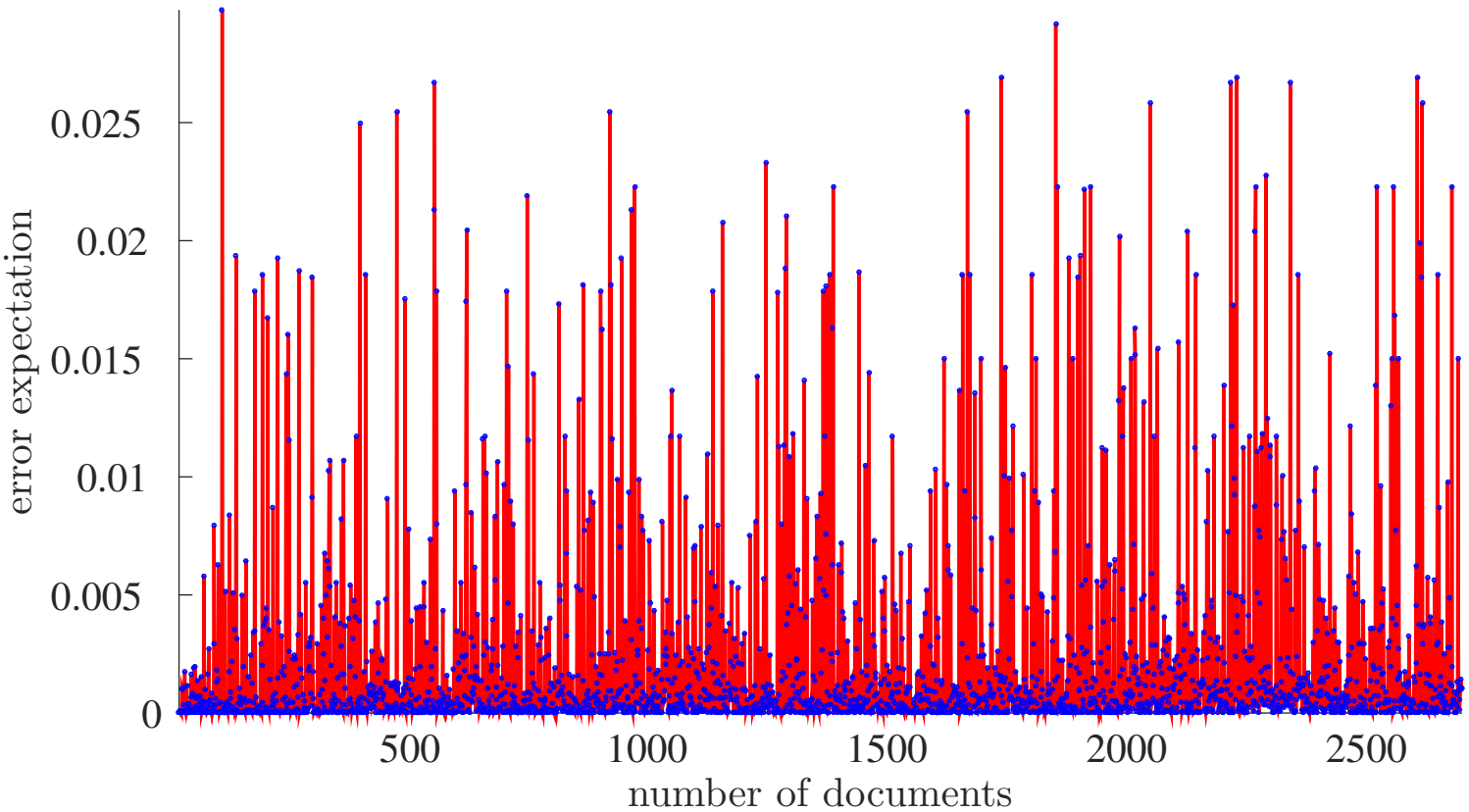

Рис.: ARTM

Заключение

- Построена тематическая модель для задачи классификации
- Проверена работоспособность алгоритма
- Проведено сравнение с алгоритмом сэмплирования для модели LDA
- Без использования регуляризаторов предложенный алгоритм работает хуже

Идеи для дальнейшей работы

- Использовать онлайн EM-алгоритм для работы с большим объёмом данных
- Придумать регуляризатор отвечающий модели генерации слов мобильного приложения