Мультиграммные аддитивно регуляризованные тематические модели

Стенин Сергей

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. К. В. Воронцов

Москва,
2015 г.

Задача тематического моделирования

D — множество документов
W — множество терминов, словарь
n_{dw} — частоты слов в документах

Основное предположение

Появление слов в документах связано с набором скрытых тем $t \in T$.

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

Задача

Полагая, что

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

по имеющимся частотам слов в документах n_{wd} найти
$p(w|t) = \varphi_{wt}$ — как распределены слова в темах, матрица Φ
$p(t|d) = \theta_{td}$ — как распределены темы в документах, матрица Θ

Задача тематического моделирования

Задача стохастического матричного разложения

$$P = (p(w|d))_{W \times D} \approx \Phi\Theta$$

Недостатки

- Модель, построенная на словах, плохо интерпретируется. Нужно добавить в словарь словосочетания.
- Задача некорректно поставлена

$$P \approx \Phi\Theta = \Phi S S^{-1}\Theta = \widetilde{\Phi}\widetilde{\Theta}$$

Нужна регуляризация

Обозначения

Мультиграммные модели

- v — слово, идущее перед w в коллекции документов.
- n_{dvw} — частота пары слов v, w в документе d
- $p(w|v,t) = \varphi_{vwt}$ — вектор вероятностей слов в теме t, матрица Φ_v. Всего таких матриц W. Их объединение обозначим $\overline{\Phi}$.
- $p(t|d) = \theta_{td}$ — образуют матрицу Θ
- σ_{vw} — языковая модель .
- $x_{vw} \in [0,1]$ — метка биграммности для пары слов v, w.

Существующие модели

- PLSA [Hofmann 1999]

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

Модель на униграммах плохо интерпретируется.

- Bigram TM [Wallach 2006]

$$\sum_{d\in D}\sum_{(v,w)\in d} n_{vwd} \ln \sum_{t\in T} \varphi_{vwt}\theta_{td} \to \max_{\Phi,\Theta}$$

Нет возможности выделять наиболее вероятные биграммы в темах, используя только параметры модели.

Существующие модели

- LDACOL [Steyvers 2006]

$$\sum_{d\in D}\sum_{(v,w)\in d} n_{vwd}\ln\Big(x_{vw}\sigma_{vw} + (1-x_{vw})\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) \to \max_{\Phi,\Theta,\Sigma,x}$$

Построение плотной матрицы σ_{vw} в биграммных моделях вычислительно неэффективно.

- Topical n-grams (TNG) [Wang 2007]

$$\sum_{d\in D}\sum_{(v,w)\in d} n_{vwd}\ln\Big(x_{vw}\sum_{t\in T}\varphi_{vwt}\theta_{td} + (1-x_{vw})\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) \to \max_{\Phi,\Theta,\overline{\Phi},x}$$

Требует дополнительного хранения меток биграммности, что уменьшает вычислительную эффективность.

Постановка задачи

Задача

Построить биграммную вероятностную тематическую модель (BTM) с аддитивной регуляризацией и вычислительно эффективным алгоритмом оптимизации параметров.

Модель

Предлагается модель Bigram Additively Regularized Topic Model (Bigram-ARTM).

Целевая функция

$$\sum_{d\in D}\sum_{w\in d} n_{wd}\ln\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) + \lambda\sum_{d\in D}\sum_{v,w\in d} n_{vwd}\ln\Big(\sum_{t\in T}\varphi_{wtv}\theta_{td}\Big) +$$

$$+R(\Phi,\overline{\Phi},\Theta)\to\max_{\Phi,\overline{\Phi},\Theta}$$

$$\sum_{w\in W}\varphi_{wtv}=1,\quad \varphi_{wtv}\geq 0$$

$$\sum_{t\in T}\theta_{td}=1,\quad \theta_{td}\geq 0$$

$$\sum_{w\in W}\varphi_{wt}=1,\quad \varphi_{wt}\geq 0$$

Связь с моделью TNG

Оценка снизу целевой функции модели TNG, основываясь на вогнутости логарфима

$$\sum_{d\in D}\sum_{v,w\in d} n_{vwd}\ln\Big(x_{vw}\sum_{t\in T}\varphi_{wtv}\theta_{td} + (1-x_{vw})\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) \geq$$

$$\geq \sum_{d\in D}\sum_{v,w\in d} n_{vwd}x_{vw}\ln\Big(\sum_{t\in T}\varphi_{wtv}\theta_{td}\Big) + \sum_{d\in D}\sum_{v,w\in d} n_{vwd}(1-x_{vw})\ln\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) \to \max_{\Phi,\Theta,\overline{\Phi},x}$$

Положим все метки равными: $x_{vw} = \frac{\lambda}{1+\lambda}$

$$\frac{\lambda}{1+\lambda}\sum_{d\in D}\sum_{v,w\in d} n_{vwd}\ln\Big(\sum_{t\in T}\varphi_{wtv}\theta_{td}\Big) + \frac{1}{1+\lambda}\sum_{d\in D}\sum_{w\in d} n_{wd}\ln\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) \to \max_{\Phi,\Theta,\overline{\Phi}}$$

N-граммы — это модальности

Документ — набор токенов (слова словаря, авторы и т. д.).
Токены w одного типа — модальность $m(w) \in M$ со словарем W^m, $\bigcup\limits_{m \in M} W^m = W$.
Примем гипотезу условной независимости

$$p(w|d) = \sum_{i \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \quad w \in W^m, \quad d \in D$$

Параметры модели $\Phi^m = (\varphi_{wt})_{W^m \times T}, \Theta = (\theta_{td})_{T \times D}, \bigcup \Phi^m = \Phi$

Введем оператор нормировки

$$\underset{i \in I}{norm}(x_i) = \frac{\max\{x_i, 0\}}{\sum\limits_{j \in J} \max\{x_j, 0\}}.$$

Локальный максимум в мультимодальной модели

Логарифм правдоподобия мультимодальной модели с регуляризацией

$$\sum_{m\in M}\tau_m\sum_{d\in D}\sum_{w\in W^m} n_{wd}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta)\to\max_{\Phi,\Theta}$$

$$\sum_{w\in W^m}\varphi_{wt}=1,\varphi_{wt}\geq 0; \qquad \sum_{t\in T}\theta_{td}=1,\theta_{td}\geq 0$$

Теорема (Воронцов, Фрей, Апишев, Ромов, Дударенко, 2015)

Для непрерывно дифференцируемой $R(\Phi,\Theta)$ в точке локального экстремума описанной выше задачи выполнены соотношения

$$p_{tdw} = \underset{t\in T}{norm}\,\varphi_{wt}\theta_{td}$$

$$\varphi_{wt} = \underset{w\in W^m}{norm}\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \qquad n_{wt}=\sum_{d\in D}\tau_m n_{dw}p_{tdw}$$

$$\theta_{td} = \underset{t\in T}{norm}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td}=\sum_{m\in M}\tau_m\sum_{w\in W^m} n_{dw}p_{tdw}$$

Частный случай

Предлагаемая в работе модель оказывается частным случаем мультимодальной модели с регуляризацией.

$$W = W^0 \bigcup W^{v_1} \bigcup \cdots \bigcup W^{v_W},$$

W^0 — униграммный словарь коллекции
W^{v_i} — множество пар слов, идущих подряд в коллекции, начинающихся на слово $v_i \in W^0$.

Условия оптимальности

Теорема: оптимум Bigram-ARTM (Стенин, 2015)

В точке максимума целевой функции для непрерывно дифференцируемой функции R выполнены соотношения

$$p_{twd} = \underset{t \in T}{norm}\left(\varphi_{wt}\theta_{td}\right) \qquad p_{tvwd} = \underset{t \in T}{norm}\left(\varphi_{vwt}\theta_{td}\right)$$

$$\varphi_{wt} = \underset{w \in W}{norm}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\varphi_{wt}}\right) \qquad n_{wt} = \sum_{d} n_{wd}\, p_{twd}$$

$$\varphi_{vwt} = \underset{w \in W}{norm}\left(n_{vwt} + \varphi_{vwt}\frac{\partial R}{\partial \varphi_{vwt}}\right) \qquad n_{vwt} = \sum_{d} \lambda n_{vwd}\, p_{tvwd}$$

$$\theta_{td} = \underset{t \in T}{norm}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \qquad n_{td} = \sum_{vw} \lambda n_{vwd}\, p_{tvwd} + \sum_{w} n_{wd}\, p_{twd}$$

Данные

Документы тезисов конференций ИОИ и ММРО за 2007-2013 года. Всего

- $D = 1009$ документов
- $W = 13678$ слов
- $\left|\bigcup_{v \in W} W_v\right| = 327694$ пар слов
- $T = 70$ (согласно работе [6], оптимальное число тем для этой коллекции - 60-80)
- Для ранжирования биграмм в темах используются параметры модели

$$p(w, v|t) = p(w|v, t) \cdot p(v|t)$$

Сравнение интерпретируемости, $\lambda = 0.01$

Тема 1	Тема 1
СКЕЛЕТ	МНОГОУГОЛЬНЫЙ ФИГУРА
ТОЧКА	БАЗОВЫЙ СКЕЛЕТ
ФИГУРА	РЕБРО СКЕЛЕТ
РЕБРО	ПУСТОЙ КРУГ
ВЕРШИНА	КРУГ ЦЕНТР
ГРАНИЦА	СКЕЛЕТ ФИГУРА
КРУГ	ТЕРМИНАЛЬНЫЙ ВЕРШИНА
РИС	МАКСИМАЛЬНЫЙ ПУСТОЙ
ЦЕНТР	СКЕЛЕТНЫЙ ГРАФ
МНОГОУГОЛЬНЫЙ	ТЕРМИНАЛЬНЫЙ РЕБРО
БАЗОВЫЙ	МНОЖЕСТВО ТОЧКА
МНОЖЕСТВО	СКЕЛЕТ МНОГОУГОЛЬНЫЙ
ФОРМА	ТОЧКА СКЕЛЕТ
СКЕЛЕТНЫЙ	ФРАГМЕНТ ГРАНИЦА
МОДЕЛЬ	БАЗОВЫЙ КРУГ
РАДИУС	ВЕТВЬ СКЕЛЕТ
ЦЕПОЧКА	ЦЕНТР ТОЧКА
МАКСИМАЛЬНЫЙ	ФИГУРА СКЕЛЕТ
СРАВНЕНИЕ	ТОЧНОСТЬ АППРОКСИМАЦИЯ
ПРОЕКЦИЯ	СЕРЕДИННЫЙ ОСЬ

Сравнение интерпретируемости, $\lambda = 0.01$

Тема 2	Тема 2
ЗАДАЧА	ЗАДАЧА РАСПОЗНАВАНИЕ
МЕТРИКА	АЛГЕБРАИЧЕСКИЙ ЗАМЫКАНИЕ
ОБЪЕКТ	МОДЕЛЬ АВО
АЛГЕБРАИЧЕСКИЙ	ВЫБОР МЕТРИКА
МНОЖЕСТВО	АЛГЕБРАИЧЕСКИЙ ПОДХОД
МАТРИЦА	ЛИНЕЙНЫЙ ЗАМЫКАНИЕ
МОДЕЛЬ	КРИТЕРИЙ КОРРЕКТНОСТЬ
АЛГОРИТМ	РЕШЕНИЕ ЗАДАЧА
РАСПОЗНАВАНИЕ	СИСТЕМА ТОЧКА
ЗАМЫКАНИЕ	КОРРЕКТНОСТЬ АЛГЕБРАИЧЕСКИЙ
ОЦЕНКА	РАСПОЗНАВАТЬ ОПЕРАТОР
АВО	МАТРИЦА ОЦЕНКА
КОРРЕКТНЫЙ	ПОПАРНЫЙ РАССТОЯНИЕ
УСЛОВИЕ	УСЛОВИЕ РЕГУЛЯРНОСТЬ
КОРРЕКТНОСТЬ	ВЫЧИСЛЕНИЕ ОЦЕНКА
РАССТОЯНИЕ	МАТРИЦА ПОПАРНЫЙ
ОПЕРАТОР	ЗАМЫКАНИЕ МОДЕЛЬ
СИСТЕМА	КОНТРОЛЬНЫЙ ОБЪЕКТ
ПРИЗНАК	ЗАМЫКАНИЕ АВО
ЛИНЕЙНЫЙ	АЛГОРИТМ ВЫЧИСЛЕНИЕ

Сравнение интерпретируемости, $\lambda = 0.01$

Тема 3	Тема 3
ОЦЕНКА	ВЕРОЯТНОСТЬ ПЕРЕОБУЧЕНИЕ
АЛГОРИТМ	ОЦЕНКА ВЕРОЯТНОСТЬ
ВЫБОРКА	ЭМПИРИЧЕСКИЙ РИСКА
ВЕРОЯТНОСТЬ	МИНИМИЗАЦИЯ ЭМПИРИЧЕСКИЙ
МЕТОД	МЕТОД ОБУЧЕНИЕ
ПЕРЕОБУЧЕНИЕ	ОБУЧАТЬ ВЫБОРКА
ОШИБКА	ТОЧНЫЙ ОЦЕНКА
МОНОТОННЫЙ	ВЕРХНИЙ ОЦЕНКА
СЕМЕЙСТВО	ГЕНЕРАЛЬНЫЙ ВЫБОРКА
ОБЪЕКТ	ЧАСТОТА ОШИБКА
МНОЖЕСТВО	СЕМЕЙСТВО АЛГОРИТМ
ЭМПИРИЧЕСКИЙ	ЧИСЛО ОШИБКА
ОБУЧЕНИЕ	ВЕКТОР ОШИБКА
РИСКА	ОБОБЩАТЬ СПОСОБНОСТЬ
ЧИСЛО	ЭМПИРИЧЕСКИЙ РИСК
ТОЧНЫЙ	КОМБИНАТОРНЫЙ ОЦЕНКА
ОБУЧАТЬ	ОЦЕНКА CCV
КОМБИНАТОРНЫЙ	МНОЖЕСТВО АЛГОРИТМ
СЛУЧАЙ	ОШИБКА АЛГОРИТМ
КЛАССИФИКАЦИЯ	МЕТОД МИНИМИЗАЦИЯ

Сглаживающий регулизатор

Уменьшает расстояние Кульбака-Лейблера от заданных столбцов матриц $\Phi, \overline{\Phi}$ и заданных строк матрицы Θ до равномерного распределения.
Используется на 2 последний темах (из 70 тем), которые будут содержать слова общей лексики.

$$R(\Phi, \overline{\Phi}, \Theta) = \sum_{t \in \{T-1, T\}} \sum_{w \in W} \alpha \cdot \ln(\varphi_{wt}) + \sum_{d \in D} \sum_{t \in \{T-1, T\}} \alpha \ln(\theta_{td})$$

Сглаживающий регулизатор, $\alpha = 3.0$

Тема 69	Тема 69
ВЛАДИМИР	РАБОТА ПОДДЕРЖАТЬ
АЛЕКСАНДР	ГРАНТ РФФИ
ЮРИЙ	ПОДДЕРЖАТЬ ГРАНТ
СЕРГЕЙ	ВЛАДИМИР ВЛАДИМИР
МИХАИЛ	КРАСОТКИН ОЛЬГА
НИКОЛАЙ	ОЛЬГА ВЯЧЕСЛАВ
ВЯЧЕСЛАВ	НАГОРНЫЙ ЮРИЙ
ПОДДЕРЖАТЬ	ОЛЕГ СЕРГЕЙ
АЛЕКСЕЙ	ВАДИМ ВЯЧЕСЛАВ
ГРАНТ	СЕРГЕЙ ВЛАДИМИР
РФФИ	ВИШНЯК БОРИС
КОНСТАНТИН	СЕРЕДИНА ОЛЕГ
ОЛЕГ	МИХАИЛ СЕРГЕЙ
ИВАН	РАСТЯЖИМОСТЬ СОПРОТИВЛЕНИЕ
ИГОРЬ	ФИЗИК РАДИОЭЛЕКТРОНИКА
ДМИТРИЙ	РАЗИН НИКОЛАЙ
БОРИС	СУЛИМОВ ВАЛЕНТИН
ОЛЬГА	ОБЩИЙ ПРЕДОК
ВАЛЕРИЙ	ИВАН КОНСТАНТИН

Сглаживающий регулизатор, $\alpha = 30.0$

Тема 70	Тема 70
РАБОТА	ПОДДЕРЖКА РФФИ
МЕТОД	ГРАНТ РФФИ
ОСНОВА	РФФИ ПРОЕКТ
ПОЗВОЛЯТЬ	РАБОТА ПОДДЕРЖАТЬ
АНАЛИЗ	ПОДДЕРЖАТЬ ГРАНТ
ЗАДАЧА	ВЫПОЛНИТЬ ПОДДЕРЖКА
ИССЛЕДОВАНИЕ	ТОЧКА ЗРЕНИЕ
ПОДХОД	МАТЕМАТИЧЕСКИЙ ОЖИДАНИЕ
РЕЗУЛЬТАТ	РАБОТА ВЫПОЛНИТЬ
ИСПОЛЬЗОВАНИЕ	РАБОТА ПОСВЯТИТЬ
ВЫПОЛНИТЬ	СКАЛЯРНЫЙ ПРОИЗВЕДЕНИЕ
СЛЕДОВАТЬ	PATTERN RECOGNITION
ПОКАЗАТЬ	ОБОБЩАТЬ СПОСОБНОСТЬ
ВИД	ДРУГ ДРУГ
РЕШЕНИЕ	DATUM MINE
НЕОБХОДИМЫЙ	ДИНАМИЧЕСКИЙ ПРОГРАММИРОВАНИЕ
РАЗЛИЧНЫЙ	КОСТНЫЙ ТКАНЬ
ОСНОВАТЬ	ЛИНЕЙНЫЙ КОМБИНАЦИЯ
ОБРАЗ	РЕШЕНИЕ ПОСТАВИТЬ

Выводы

1. Предложена мультиграммная тематическая модель с аддитивной регуляризацией.
2. Показано, что мультиграммы — частный случай модальностей.
3. Построена мультиграммная тематическая модель ММРО.
4. Показано, что модель выделяет интерпертируемые темы, и использование регуляризации улучшает качество модели.

Hofmann T. Probabilistic latent semantic indexing //Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 1999. – С. 50-57.

Wallach H. M. Topic modeling: beyond bag-of-words //Proceedings of the 23rd international conference on Machine learning. – ACM, 2006. – С. 977-984.

Griffiths T. L., Steyvers M., Tenenbaum J. B. Topics in semantic representation //Psychological review. – 2007. – Т. 114. – №. 2. – С. 211.

Wang X., McCallum A., Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval //Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. – IEEE, 2007. – С. 697-702.

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov and Marina Dudarenko BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections//Proceedings of the 4th international conference on analisys of images, social networks, and texts

Vorontsov K., Potapenko A., Plavin A. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization.