Итеративный подбор коэффициентов регуляризации тематических моделей

Захаренков Антон

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

Научный руководитель
профессор РАН, д.ф.-м.н.
К. В. Воронцов

28 июня 2018

Решаемая задача

Цель

Разработать метод последовательного подбора коэффициентов регуляризации тематических моделей.

Проблема

Различные комбинации регуляризаторов могут значительно улучшить многие внутренние метрики модели тематического моделирования. Подбор коэффициентов перед ними на данный момент эвристика.

Предложенный подход

Применяются методы безградиентной оптимизации, байесовской оптимизации и машинного обучения для автоматического подбора коэффициентов регуляризации в модели АРТМ.

Тематическое моделирование. Модель ARTM

$$L(\Phi, \Theta) = \ln \prod_{d \in D} \prod_{w \in d} \mathsf{P}(w \,|\, d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

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

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \hat{\tau}_i R_i(\Phi, \Theta)$$

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

Тематическое моделирование. Примеры регуляризаторов

Сглаживающий регуляризатор

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \varphi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max$$

Разреживающий регуляризатор

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \varphi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max$$

Декоррелирующий регуляризатор

$$R(\Phi, \Theta) = \frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \mathrm{cov}(\varphi_t, \varphi_s) \to \max$$

$$\mathrm{cov}(\varphi_t, \varphi_s) = \sum_{w \in W} \varphi_{wt} \varphi_{ws}$$

Тематическое моделирование. Примеры метрик качества

- **Перплексия**

$$\mathscr{P}(D;p) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in W} n_{dw} \ln \mathsf{P}(w\,|\,d))\right)$$

- **Когерентность**

$$Coher(t) = \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=1}^{k} PMI(w_i; w_j)$$

- **Разреженность матриц**
- **Характеристики ядер тем**
  - $\mathrm{purity_t} = \sum_{w\in W_t} \mathsf{P}(w\,|\,t)$ — чистота темы
  - $\mathrm{contrast_t} = \frac{1}{|W_t|}\sum_{w\in W_t} \mathsf{P}(t\,|\,w)$ — контрастность темы
  - $\mathrm{kernel_t\ size} = |W_t|$ — размер ядра

Постановка задачи

Задача оптимизации

$$R(\Phi) = \sum_{i=1}^{r} \vec{\tau}_i R_i(\Phi), \quad L(\Phi) + R(\Phi) \to \max_{\Phi}$$

Для решения данной оптимизационной задачи применяется итерационный процесс, порождающий последовательность приближений $\Phi^l, \ l = 0, \ 1, \ldots,$ где используется правило

$$\Phi^{l+1} = F_l(\Phi^l, \vec{\tau}^l, D_{b(l)})$$

с вектором коэффициентов регуляризации $\vec{\tau}^l := \{\tau_i^l\}_{i=1}^n$.

Последовательность векторов $\vec{\tau}^l$, $l = 0, 1, \ldots,$ называется *траекторией регуляризации*.

Общий подход

Algorithm Адаптивное управление траекторией регуляризации

Вход: коллекция документов $D \equiv \{D_b\}_{b=1}^{B}$, Φ^0, $\vec{\tau}^0$;

Выход: оптимальный вектор коэффициентов регуляризации $\vec{\tau}^*$;

для итераций $l = 0, 1, \dots:$

$D_{b(l)}$:= *взять следующую пачку из* $\{D_b\}_{b=1}^{B}$;

$\Phi^{l+1} := F_l(\Phi^l, \vec{\tau}^l, D_{b(l)})$;

$\vec{Q}^{l+1} := \vec{Q}(x^{l+1})$;

$\vec{\tau}^{l+1}$:= *выбрать следующий* $\vec{\tau}(\vec{Q}^{l+1}, \vec{\tau}^l, \Phi^{l+1}, D_{b(l)})$;

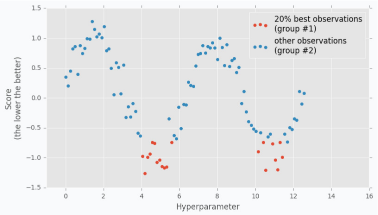
Основные проблемы

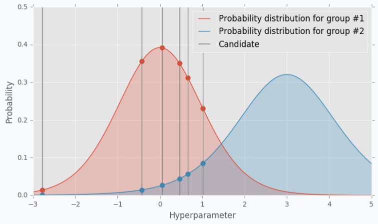
- Показатели качества Q_j, определённые значения которых требуется достичь, могут быть лишь косвенно связаны с регуляризаторами R_i
- Исходная функция $a : \vec{\tau} \to \vec{Q}$ может является трудновычислимой, недифференцируемой и даже негладкой из-за сложных регуляризаторов
- Точка в пространстве показателей качества $\vec{Q}$ определяется неоднозначно, с учётом погрешности ε, которую даёт тематическая модель APTM, производящая разложение на матрицы Φ и Θ при обработке конечной пачки документов $D_{b(l)}$
- Коэффициенты регуляризации $\vec{\tau}_i$ могут входить в противоречия и требовать изменений в разные стороны

Базовые подходы

- **Случайный поиск по сетке**
  Метод случайного поиска предполагает случайный выбор вектора коэффициентов регуляризации $\vec{\tau}$ из фиксированной сетки, которая задаётся экспертно.
- **Симплекс метод Nelder—Mead**
  Данный метод заключается в построении симплекса из $n + 1$ точки в n-мерном пространстве значений. Далее текущий симплекс предлагается итеративно деформировать и перемещать в поисках экстремума путём замены одной из точек. Деформации и перемещения обеспечиваются функциями отражения, сжатия и растяжения.

- **Tree-structured Parzen Estimators**
  TPE принимает на вход иерархическое пространство поиска с априорными вероятностями, и на каждом шаге уточняет распределения «хороших» и «плохих» параметров, основываясь на значениях целевой метрики. В начале каждой итерации выбирается точка, которая наиболее вероятно будет относиться к группе «хороших» точек, и наименее вероятно «плохих».





Source: neopy.com

Модификация базовых подходов

- **Множественная регрессия**
  В ходе итерационного процесса накапливается обучающая выборка пар векторов $(\vec{\tau}^k, \vec{Q}^k)$, где вектор $\vec{Q}^k$ — это значения показателей качества, полученные при построении АРТМ модели с коэффициентами регуляризации $\vec{\tau}^k$.
  Если зависимость $\vec{Q}(\vec{\tau})$ удалось хорошо восстановить, то в качестве следующеего вектора $\vec{\tau}$ можно брать тот, для которого предсказанная величина $\vec{Q}(\vec{\tau})$ будет наиближайшей к $\vec{Q}^*$.

Проблемы

При восстановлении регрессионной зависимости существуют следующие сложности.

Во-первых, изначально выборка $\{(\vec{\tau}^k, \vec{Q}^k)\}_{k=0}^N$ пуста, и не получится сразу хорошо предсказывать вектор показателей качества для векторов-кандидатов $\vec{\tau}^l$, $l = 0, 1, \ldots, L$.
Первые значения можно:

- брать случайными
- сэмплировать из распределения, смоделированного TPE

Проблемы

Во-вторых, достижение окрестности $\vec{Q}^*$ может оказаться трудной задачей, а регрессионная зависимость может строиться по обучающей выборке, которая находится далеко от целевых значений $(\vec{\tau}^*, \vec{Q}^*)$.

Для решения этой проблемы предлагается чередовать *исследовательские* и *максимизирующие* шаги. На этапе исследования предлагается выбирать произвольный вектор $\vec{\tau}^l$(либо сэмплировать из TPE), а на этапе максимизации делать попытку приблизиться к оптимуму.

Вычислительный эксперимент

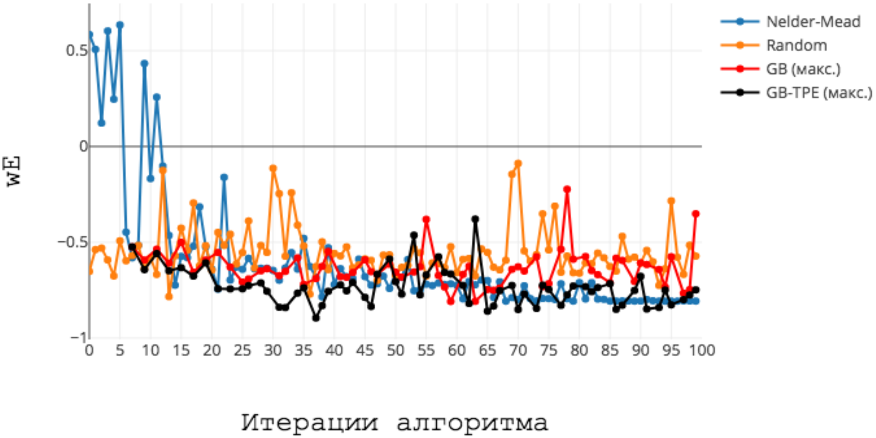
Цели:

- применить предложенные алгоритмы к реальным данным,
- сравнить полученные результаты друг с другом с точки зрения взвешенной метрики, а так же отдельных метрик

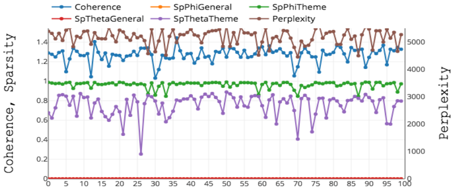
Данные:

- **postnauka** - postnauka.ru
  $|D| = 3446$ $|W| = 35531$;
- **elementy** - elementy.ru
  $|D| = 2397$ $|W| = 54349$;

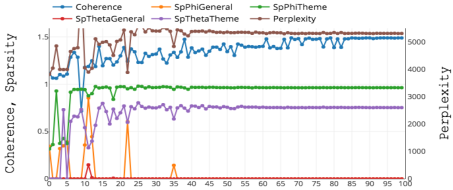
Элементы. Оффлайн эксперимент.



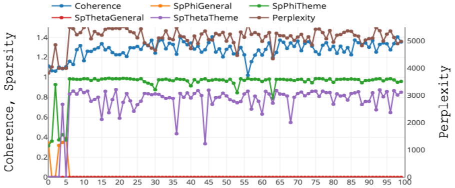
Nelder-Mead сходится к локальному оптимуму, GB-TPE продолжает исследования, переодически достигая более оптимальных значений.



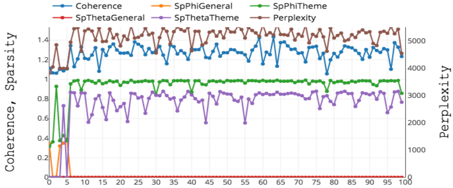
Random



Nelder-Mead



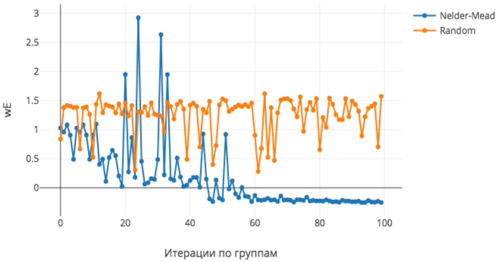
GB-Random

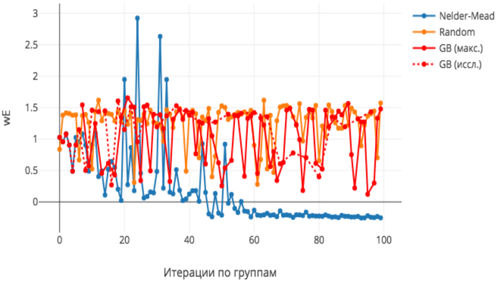


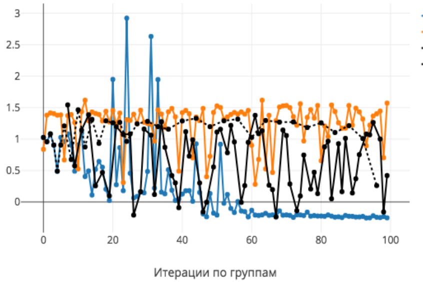
GB-TPE

Рис.: Изменение внутренних метрик качества модели

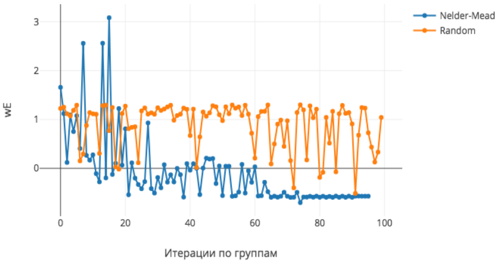
Постнаука. Онлайн эксперимент.

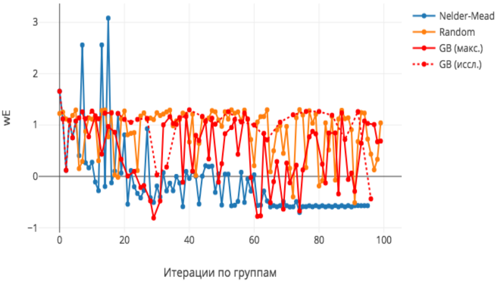


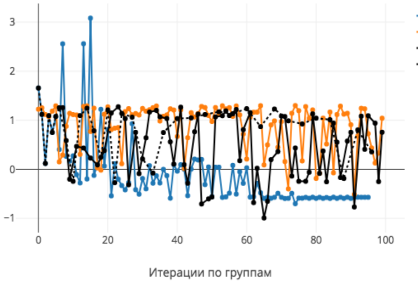




Элементы. Онлайн эксперимент.







Заключение

1. Проведены вычислительные эксперименты на реальных данных;
2. Проверены новые и существующие методы адаптивного подбора траектории регуляризации в тематическом моделировании
3. Показано, что новый алгоритм GB-TPE находит точки не хуже, чем Nelder-Mead, но ведет себя более «исследовательски».