Аддитивная регуляризация вероятностных тематических моделей

Воронцов Константин Вячеславович

ВМК МГУ • 31 октября 2013

Содержание

Задача определения тематики коллекции документов

Тема — это набор терминов, неслучайно часто совместно встречающихся в относительно узком подмножестве документов.

Дано:
W — словарь, множество слов (терминов)
D — множество (коллекция, корпус) текстовых документов
n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Задача:
— найти, какими терминами определяется каждая тема
— найти, к каким темам относится каждый документ

Возможные дополнительные задачи:
— определить число статистически различимых тем
— восстановить иерархию тем
— построить динамику развития тем во времени
— найти тематику связанных с документами объектов

Цели тематического моделирования (topic modeling)

- Тематический поиск документов и объектов по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация текстовых документов

Типичные приложения:

- Поиск научной информации
- Поиск экспертов (expert search), рецензентов, проектов
- Выявление трендов и фронта исследований
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендательные сервисы (коллаборативная фильтрация)
- Аннотация генома и другие задачи биоинформатики

Вероятностная формализация постановки задачи

Базовые предположения:

- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция D — выборка троек $(d_i, w_i, t_i)_{i=1}^{n} \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

Вероятностная модель порождения документа d:

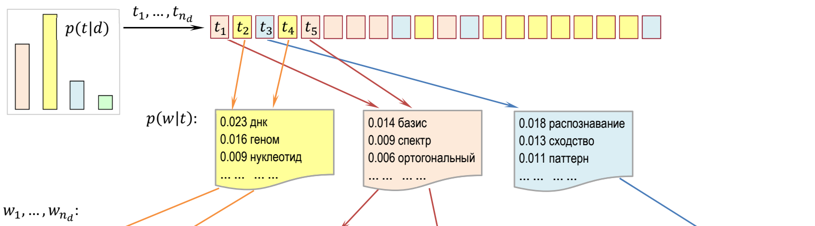
$$p(w|d) = \sum_{t \in T} p(w|d, t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

Дано $\hat{p}(w|d) \equiv n_{dw}/n_d$, **найти:**

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темеах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

Вероятностная модель порождения документа d

Вероятностная тематическая модель: $p(w|d) = \sum\limits_{t\in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Некоторые дополнительные предположения

Гипотеза разреженности:

- употребление слова связано с одной-двумя темами $\Longrightarrow$ распределение $p(t|d,w)$ разрежено
- тема характеризуется небольшой долей слов словаря $\Longrightarrow$ распределение $p(w|t) = \phi_{wt}$ разрежено
- документ относится лишь к нескольким темам $\Longrightarrow$ распределение $p(t|d) = \theta_{td}$ разрежено

Гипотеза о наличии нетематических слов:

1. некоторые слова — фоновые, $p(w|d) \approx n_w/n$
2. некоторые слова — шумовые, $p(w|d) \ll n_{dw}/n_d$
3. некоторые слова не используются, $p(w|d) \gg n_{dw}/n_d$

Задача максимизации правдоподобия

Задача: найти максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1$$

Интерпретация: найти стохастическое матричное разложение

$$\|F - \Phi\Theta\|_{\mathrm{KL}} \to \min_{\Phi,\Theta}$$

$F = \big(\hat{p}(w|d)\big)_{W\times D}$ — известная матрица исходных данных,
$\Phi = \big(\phi_{wt}\big)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$,
$\Theta = \big(\theta_{td}\big)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

EM-алгоритм, вероятностный латентный семантический анализ PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

E-шаг. Выразим $p(t|d,w)$ через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}} \propto \phi_{wt}\theta_{td}$$

$n_{dwt} = n_{dw}\,p(t|d,w)$ — оценка числа троек (d,w,t) в коллекции

M-шаг. Частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d\in D} n_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w\in d} n_{dwt}}, \qquad \theta_{td} = \frac{n_{dt}}{n_d} \equiv \frac{\sum\limits_{w\in d} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dwt}},$$

или краткая запись:

$$\phi_{wt} \propto n_{wt} \qquad \theta_{td} \propto n_{dt}$$

Недостатки классического PLSA

1. PLSA переобучается, т.к. $\dim(\Phi, \Theta) = |D| \cdot |T| + |W| \cdot |T|$
   — регуляризации: сглаживание, разреживание и др.
2. PLSA неверно оценивает вероятности новых слов: если $n_w = 0$, то $\hat{p}(w|t) = 0$ для всех $t \in T$
   — робастные модели с шумом и фоном
3. PLSA вынужден хранить 3D-матрицу $p(t|d,w)$;
   PLSA медленно сходится на больших коллекциях;
   PLSA искажает модель при добавлении документа;
   — рациональный алгоритм, онлайновый алгоритм
4. PLSA не позволяет управлять *разреженностью* Φ и Θ, т.к.
   (в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в финале $\phi_{wt} = 0$)
   (в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в финале $\theta_{td} = 0$)
   — эвристики постепенного разреживания

Рациональный EM-алгоритм

Идея: E-шаг встраивается внутрь M-шага

Вход: коллекция D, число тем $|T|$, число итераций MaxIter;
Выход: распределения Θ и Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \dots, \mathrm{MaxIter}$
- $n_{wt}, n_{dt}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
- **для всех** $d \in D$, $w \in d$
  - $p(t|d,w) \propto \phi_{wt}\theta_{td}$ для всех $t \in T$;
  - возможно, применить разреживание к $p(t|d,w)$;
  - $n_{wt}, n_{dt}, n_t, n_d \mathrel{+}= n_{dw}\,p(t|d,w)$ для всех $t \in T$;
- $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
- $\theta_{td} := n_{dt}/n_d$ для всех $d \in D$, $t \in T$;

Эвристики

Стратегии разреживания распределений $p(t|d,w)$

- пропорциональное распределение без разреживания
- сэмплирование Гиббса: $t \sim p(t|d,w)$ для каждой позиции w_i
- сэмплирование: $t \sim p(t|d,w)$ для каждого слова (d,w)
- максимизация (оптимальный байесовский классификатор): $t = \arg\max_t p(t|d,w)$ для каждого слова (d,w)

Чередование сэмплирования и максимизации приводит к лучшему локальному максимуму правдоподобия [Д. Елшин]

Стратегии частого обновления параметров ϕ_{wt}, θ_{td}:

- после каждого прохода всей коллекции
- после каждого документа
- после каждого слова (самая быстрая сходимость)

Онлайновый EM-алгоритм для модели PLSA

инициализировать ϕ_{wt} для всех $w \in W,\ t \in T$;
$n_{wt} := 0,\ n_t := 0$ для всех $w \in W,\ t \in T$;
для всех пакетов $D_j,\ j = 1, \dots, J$
- $\tilde{n}_{wt} := 0,\ \tilde{n}_t := 0$ для всех $w \in W,\ t \in T$;
- **для всех** $d \in D_j$
  - инициализировать θ_{td} для всех $t \in T$;
  - **повторять**
    - $p(t|d, w) \propto \phi_{wt}\theta_{td}$ для всех $t \in T$;
    - $\theta_{td} := \frac{1}{n_d} \sum\limits_{w \in d} n_{dw} p(t|d, w)$ для всех $t \in T$;
  - **пока** θ_d не сойдётся;
  - $\tilde{n}_{wt}, \tilde{n}_t \mathrel{+}= n_{dw} p(t|d, w)$ для всех $w \in d,\ t \in T$;
- $n_{wt} := \rho_j n_{wt} + \tilde{n}_{wt};\ \ n_t := \rho_j n_t + \tilde{n}_t$ для всех $w \in W,\ t \in T$;
- $\phi_{wt} := n_{wt} / n_t$ для всех $w \in W,\ t \in T$;

Робастная тематическая модель

Робастная тематическая модель с шумом и фоном:

$$p(w|d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1+\gamma+\varepsilon}, \qquad Z_{dw} = \sum_{t\in T} \phi_{wt}\theta_{td},$$

$\pi_{dw} \equiv p_{\text{ш}}(w|d)$ — *шумовая* компонента, γ — параметр;
$\pi_w \equiv p_{\text{ф}}(w)$ — *фоновая* компонента, ε — параметр.

Недостатки:

- неочевидно, как задавать параметры γ, ε
- приходится хранить матрицу шума $\Pi = (\pi_{dw})_{D\times W}$

Упрощённая робастная разреженная модель:

$$p(w|d) = \nu_d Z_{dw} + \big[Z_{dw} = 0\big]\pi_{dw}$$

Упрощённая робастная модель с шумом без фона

Компонента шума включается только когда тематическая компонента $Z_{dw} = 0$ в результате разреживания:

$$p(w|d) = \nu_d Z_{dw} + \bigl[Z_{dw} = 0\bigr] \pi_{dw}$$

Оптимальное значение π_{dw} находится аналитически и совпадает с частотной оценкой условной вероятности $p(w|d)$:

$$\pi_{dw} = n_{dw} / n_d,$$

Нормировочный множитель ν_d также находится аналитически и равен доле тематических терминов в документе:

$$\nu_d = \frac{1}{n_d} \sum_{w \in d} \bigl[Z_{dw} > 0\bigr] n_{dw}.$$

Латентное размещение Дирихле
LDA — Latent Dirichlet Allocation [David Blei, 2003]

Вероятностная тематическая модель: $p(w|d) = \sum\limits_{t\in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$

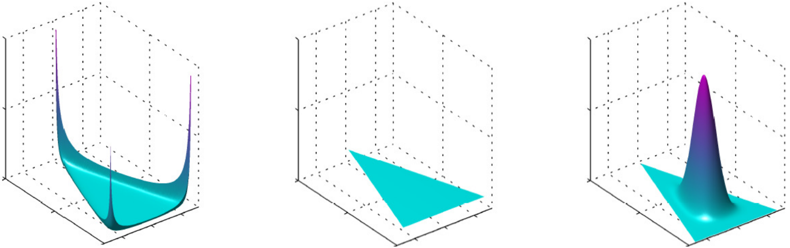
Гипотеза об априорных распределениях Дирихле:

- $\theta_d = (\theta_{td})_{t\in T} \in \mathbb{R}^{|T|}$ — случайные векторы из распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \quad \sum_t \theta_{td} = 1;$$
- $\phi_t = (\phi_{wt})_{w\in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod\limits_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \quad \sum_w \phi_{wt} = 1;$$

Почему именно распределение Дирихле?

– Является сопряжённым к мультиномиальному распределению
– Порождает как сглаженные, так и разреженные векторы
– Неплохо описывает кластерные структуры на симплексе

Пример. $\mathrm{Dir}(\theta|\alpha)$ при $|T| = 3$, $\theta, \alpha \in \mathbb{R}^3$:



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ $\quad$ $\alpha_1 = \alpha_2 = \alpha_3 = 1$ $\quad$ $\alpha_1 = \alpha_2 = \alpha_3 = 10$

Байесовская оценка параметров $\theta_{td} \equiv p(t|d)$

Пусть темы слов в документах $d \in D$ выбираются из θ_d:

$$X_d = \{t_1, \ldots, t_{n_d}\} \sim \theta_d.$$

Тогда вероятность встретить каждую из тем t ровно n_{td} раз подчиняется мультиномиальному распределению:

$$p(X_d|\theta_d) = \mathrm{Mult}(n_{1d}, \ldots, n_{Td}|\theta_d) = \frac{n_d!}{\prod_t n_{td}!} \prod_t \theta_{td}^{n_{td}}.$$

Если предположить, что $\theta_d \sim \mathrm{Dir}(\alpha)$, то по формуле Байеса апостериорное распределение также из $\mathrm{Dir}(\alpha')$, $\alpha'_t = \alpha_t + n_{td}$:

$$p(\theta_d|X_d) = \frac{p(X_d|\theta_d)\,\mathrm{Dir}(\theta_d|\alpha)}{\int p(X_d|\theta)\,\mathrm{Dir}(\theta|\alpha)\,d\theta} \propto \prod_t \theta_{td}^{n_{td}} \theta_{td}^{\alpha_t - 1} = \mathrm{Dir}(\theta_d;\alpha').$$

Распределение Дирихле — *сопряжённое* к мультиномиальному, что упрощает байесовское оценивание параметров ϕ_{wt} и θ_{td}.

Байесовская оценка параметров $\theta_{td} \equiv p(t|d)$

Оценка θ_{td} при априорном распределении:

$$\mathsf{E}p(t|d,\alpha) = \int \theta_{td} \mathrm{Dir}(\theta_d|\alpha)\, d\theta_d = \frac{\alpha_t}{\alpha_0}.$$

Пусть известна выборка тем $X_d = \{t_1, \dots, t_{n_d}\} \sim \theta_d$.
Оценка θ_{td} при апостериорном распределении:

$$\mathsf{E}p(t|d,X_d,\alpha) = \int \theta_{td} \mathrm{Dir}(\theta_d|\alpha')\, d\theta_d = \frac{n_{td} + \alpha_t}{\sum\limits_{t'} n_{t'd} + \alpha_{t'}} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0},$$

n_{td} — сколько раз слово документа d было отнесено к теме t,
n_d — длина документа в словах.

Замечание. Эта оценка переходит в МП-оценку при $\alpha_t \equiv 0$, хотя при $\alpha_t = 0$ распределение Дирихле не определено.

Байесовская оценка параметров $\phi_{wt} \equiv p(w|t)$

Оценка ϕ_{wt} при априорном распределении:

$$\mathsf{E}p(w|t,\beta) = \int \phi_{wt} \operatorname{Dir}(\phi_t|\beta)\, d\phi_t = \frac{\beta_w}{\beta_0}.$$

Коллекция порождается двумя распределениями $p(t|d)$, $p(w|t)$.
Часть коллекции, порождённая темой t:

$$X_t = \bigl\{(d,w,t)\colon\, d \in D,\ \ w \sim \phi_t\bigr\}.$$

Апостериорное распределение для ϕ_t по формуле Байеса:

$$p(\phi_t|X_t,\beta) = \frac{p(X_t|\phi_t)\operatorname{Dir}(\phi_t|\beta)}{\int p(X_t|\phi)\operatorname{Dir}(\phi|\beta)\, d\phi} = \operatorname{Dir}(\phi_t|\beta'), \quad \beta'_w = \beta_w + n_{wt}.$$

Оценка ϕ_{wt} через апостериорное распределение:

$$\mathsf{E}p(w|t,X_d,\beta) = \int \phi_{wt} \operatorname{Dir}(\phi_t|\beta')\, d\phi_t = \frac{n_{wt} + \beta_w}{n_t + \beta_0}.$$

Главное отличие LDA от PLSA

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:
$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$
- в LDA — сглаженные байесовские оценки:
- практика показывает, что на достаточно больших данных нет значимых различий между LDA и PLSA

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

Байесовский вывод для алгоритма сэмплирования Гиббса:

Yi Wang. Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.

Недостатки LDA

1. распределение Дирихле удобно математически, но имеет крайне слабые лингвистические обоснования
2. сглаживание вместо разреживания
3. байесовский вывод требует интегрирования по пространству параметров модели, которое только в базовом варианте LDA элементарно
4. построение композитных и многофункциональных моделей становится громоздкой математической задачей
5. практика показывает, что на достаточно больших данных нет значимых различий между LDA и PLSA
6. переобучение PLSA связано только с редкими словами, и это отнюдь не главный недостаток PLSA (плохо поняли суть проблемы и боремся не с тем врагом)

Эксперимент на модельных данных (а кто же враг?)

Модельные коллекции порождаются заданными матрицами Φ_0 и Θ_0 при $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

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

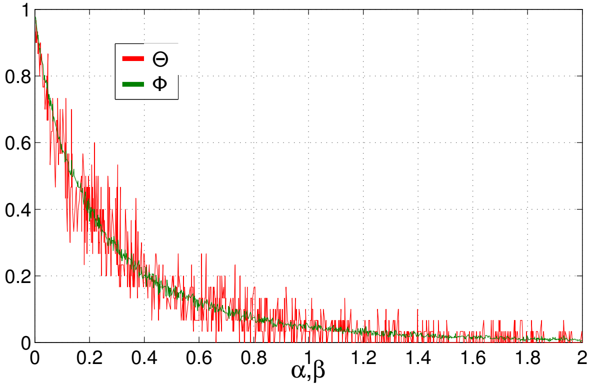
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

как для самих матриц Φ и Θ, так и для их произведения:

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

Генерация модельных данных различной степени разреженности

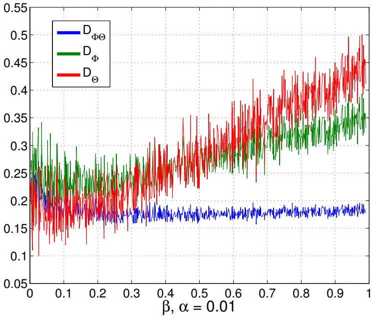
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



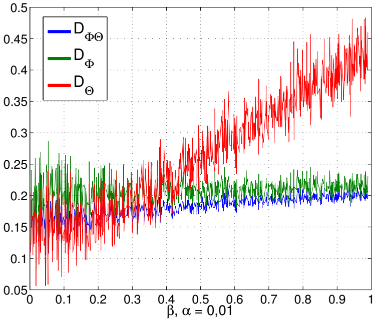
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и $\Phi\Theta$ от разреженности матрицы Φ_0

PLSA



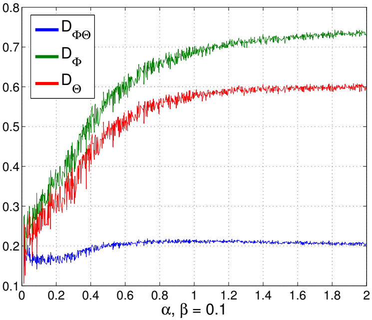
LDA



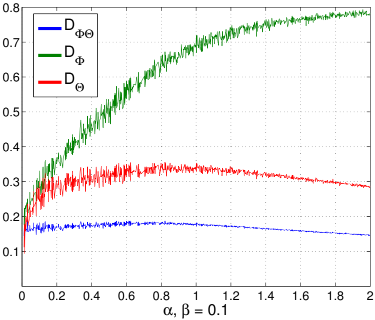
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и $\Phi\Theta$ от разреженности матрицы Θ_0

PLSA



LDA



Выводы

1. Произведение $\Phi\Theta$ восстанавливается устойчиво, точность восстановления не зависит от разреженности исходных модельных данных Φ_0, Θ_0
2. Матрицы Φ, Θ восстанавливаются неустойчиво, результат зависит от случайной инициализации
3. Методы PLSA и LDA одинаково неустойчивы (сглаживание не спасает от неединственности)
4. Устойчивое восстановление матриц Φ, Θ происходит только при сильной разреженности (более 80% нулей)

Реализация экспериментов:
Виталий Глушаченков. Магистерская диссертация. МФТИ, 2013.
Михаил Колупаев. Курсовая работа. ВШЭ, 2013.

Причина неустойчивости тематических моделей

Задача стохастического матричного разложения:

$$\hat{F} \approx F = \Phi\Theta$$

$\hat{F} = \bigl(n_{dw}/n_d\bigr)_{W\times D}$ — известная матрица исходных данных;
$F = \bigl(p(w|d)\bigr)_{W\times D}$ — матрица тематической модели;
$\Phi = \bigl(\phi_{wt}\bigr)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$;
$\Theta = \bigl(\theta_{td}\bigr)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.
Все матрицы неотрицательные, с нормированными столбцами.

Проблема неединственности матричного разложения:

$$F = \Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для любых невырожденных $S_{T\times T}$ таких, что $\Phi', \Theta' > 0$.

Регуляризация — это выбор лучшего из множества разложений

Обоснование EM-алгоритма PLSA

Теорема

Максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1$$

и фиксированных значениях $p(t|d,w)$ достигается при

$$\phi_{wt} \propto n_{wt} \equiv \sum_{d\in D} n_{dw} p(t|d,w) \qquad \theta_{td} \propto n_{dt} \equiv \sum_{w\in W} n_{dw} p(t|d,w)$$

Обоснование регуляризованного EM-алгоритма PLSA

Теорема

Максимум регуляризованного правдоподобия

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1$$

и фиксированных значениях $p(t|d,w)$ достигается, когда

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

Дивергенция Кульбака–Лейблера

Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

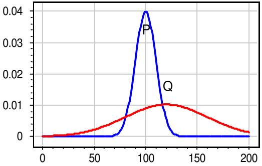
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

1. $\mathrm{KL}(P\|Q) \geqslant 0$; $\quad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q$;

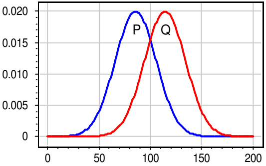
2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_\alpha \quad \Longleftrightarrow \quad \sum_{i=1}^n p_i \ln q_i(\alpha) \to \max_\alpha .$$

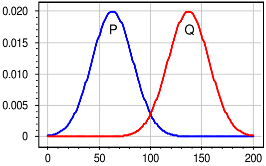
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$



$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Регуляризатор №1: Сглаживание LDA

Гипотеза сглаженности:
распределения ϕ_{wt} близки к заданным распределениям β_w
распределения θ_{td} близки к заданным распределениям α_t

$$\sum_{t\in T} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

Максимизируем сумму этих регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \qquad \theta_{td} \propto n_{dt} + \alpha_0 \alpha_t.$$

Этого вы не найдёте в *D.Blei, A.Ng, M.Jordan.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — Vol. 3. — Pp. 993–1022.

Регуляризатор №1: Сглаживание LDA

Выводы:

- Найдено альтернативное обоснование LDA: оказывается, это всего лишь притягивание столбцов Φ, Θ к заданным распределениям
- Формулы M-шага LDA получены без байесовского вывода:
  — без предположения об априорном распределении
  — без интегрирования по пространству параметров модели
  — без требования сопряжённости
- Распределение Дирихле утрачивает «особую роль», это один из многих регуляризаторов, и не самый лучший

Открытый вопрос:

- Всегда ли возможно подобрать регуляризатор (альтернативное априорное распределение для MAP), чтобы результаты байесовского вывода и MAP совпали?

Регуляризатор №2: Частичное обучение

Пусть известно, что
1) документы $d \in D_0$ относятся к темам $T_d \subset T$,
2) к темам $t \in T_0$ относятся термины $W_t \subset W$.

ϕ^0_{wt} — распределение, равномерное на W_t
θ^0_{td} — распределение, равномерное на T_d

Максимизируем сумму этих регуляризаторов:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi^0_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \ln \theta_{td} \to \max$$

Подставляем, получаем обобщение LDA:

$$\theta_{td} \propto n_{dt} + \beta_0 \theta^0_{td} \qquad \phi_{wt} \propto n_{wt} + \alpha_0 \phi^0_{wt}$$

Nigam K., McCallum A., Thrun S., Mitchell T. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, no. 2–3.

Регуляризатор №2: Частичное обучение (новое обобщение)

Гипотеза: вместо логарифма можно взять любую другую монотонно возрастающую функцию μ

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi^0_{wt} \mu(\phi_{wt}) + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \mu(\theta_{td}) \to \max$$

Подставляем, получаем ещё одно обобщение LDA:

$$\theta_{td} \propto n_{dt} + \beta_0 \theta^0_{td} \theta_{td} \mu'(\theta_{td}) \qquad \phi_{wt} \propto n_{wt} + \alpha_0 \phi^0_{wt} \phi_{wt} \mu'(\phi_{wt})$$

При $\mu(z) = z$ максимизируется сумма ковариаций $\mathrm{cov}(\theta^0_d, \theta_d)$.

Преимущество ковариационного регуляризатора:
Если θ^0_{td} равномерно на T_d, то ковариация не накладывает ограничений на распределение θ_{td} между темами из T_d.

Регуляризатор №3: Разреживание

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей.

Чем сильнее разрежено распределение, тем ниже его энтропия. Максимальной энтропией обладает равномерное распределение.

Поэтому максимизируем дивергенцию между равномерным распределением и искомыми распределениями ϕ_{wt}, θ_{td}:

$$R(\Phi, \Theta) = -\beta \sum_{t \in T} \sum_{w \in W} \ln \phi_{wt} - \alpha \sum_{d \in D} \sum_{t \in T} \ln \theta_{td} \to \max .$$

Подставляем, получаем «анти-LDA»:

$$\phi_{wt} \propto \bigl(n_{wt} - \beta\bigr)_+, \qquad \theta_{td} \propto \bigl(n_{dt} - \alpha\bigr)_+ .$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Постепенное разреживание распределений ϕ_{wt} и θ_{td}

Эвристика:
постепенно увеличивать коэффициенты регуляризации α, β

Реализация эвристики:
начиная с итерации i_0,
в конце каждой δ-й итерации
обнуляем долю r наименьших значений
в каждом распределении ϕ_t и θ_d,
так, чтобы сумма обнуляемых значений
не превышала R_θ для распределений θ_d,
не превышала R_ϕ для распределений ϕ_t

Обозначения параметров эвристики:
i_0:δ:r (если R_θ и R_ϕ не используются)
i_0:δ:r, th:R_θ, ph:R_ϕ

Альтернативное обоснование — OBD (Optimal Brain Damage)

Пусть алгоритм сошелся к локальному оптимуму.
Обнуление каких параметров меньше всего повлияет на значение правдоподобия?

Разложим правдоподобие в ряд Тейлора в окрестности точки локального максимума:

$$L(\Phi+\Delta\Phi, \Theta+\Delta\Theta) \approx L(\Phi, \Theta) + \frac{1}{2}\sum_{w,t}\sum_{u,s}\Delta\phi_{wt}\Delta\phi_{us}\frac{\partial^2 L(\Phi,\Theta)}{\partial\phi_{wt}\partial\phi_{us}} + $$

$$+\frac{1}{2}\sum_{t,d}\sum_{s,g}\Delta\theta_{td}\Delta\theta_{sg}\frac{\partial^2 L(\Phi,\Theta)}{\partial\theta_{td}\partial\theta_{sg}} + \sum_{w,t}\sum_{s,d}\Delta\phi_{wt}\Delta\theta_{sd}\frac{\partial^2 L(\Phi,\Theta)}{\partial\phi_{wt}\partial\theta_{sd}}$$

Y. LeCun, J. Denker, S. Solla R. E. Howard, L. D. Jackel.
Optimal Brain Damage // Advances in Neural Information Processing Systems II. Morgan Kauffman. — 1990.

Альтернативное обоснование — OBD (Optimal Brain Damage)

Оценим изменение функционала:

- при обнулении одного параметра ϕ_{wt}:

$$\Delta L = -\frac{1}{2}\sum_{d} n_{dw} p^2(t|d,w) \approx n_{wt}$$

- при обнулении параметров ϕ_{wt} для данного слова w:

$$\Delta L = -\frac{1}{2}\sum_{d,t,s} n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} \frac{\phi_{ws}\theta_{sd}}{p(w|d)} = -\frac{1}{2}\sum_{t} n_{wt}$$

- при обнулении параметров ϕ_{wt}, θ_{td} для всех слов, тем и документов:

$$\Delta L = -\frac{1}{2}\sum_{wt} n_{wt} - \frac{1}{2}\sum_{td} n_{td}$$

Работает ли разреживание? Эксперимент...

D — коллекция 2000 авторефератов диссертаций на русском языке суммарной длины $n \approx 8.7 \cdot 10^6$, словарь $|W| \approx 3 \cdot 10^4$.
Предобработка: лемматизация, удаление стоп-слов.
D' — коллекция 200 авторефератов, не включённых в D.
Перплексия тестовой коллекции D' (hold-out perplexity):

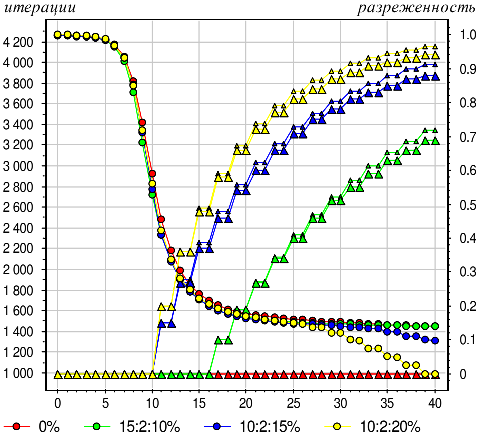
$$\mathcal{P}(D') = \exp\left(-\frac{\sum\limits_{d\in D'} \sum\limits_{w\in d''} n_{dw} \ln p(w|d)}{\sum\limits_{d\in D'} \sum\limits_{w\in d''} n_{dw}} \right)$$

Число итераций 40; число тем $|T|=100$.

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

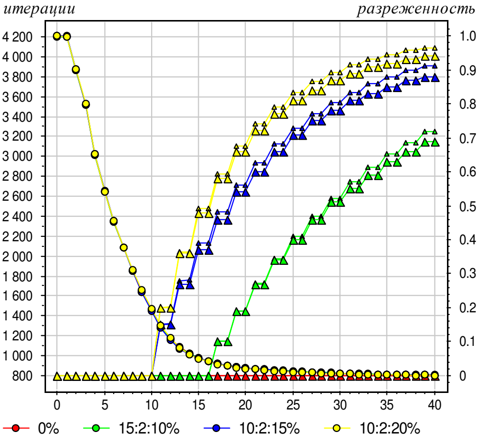
Разреживание распределений ϕ_{wt} и θ_{td}

упрощённая робастная модель без фона,
разреживание через 2 итерации



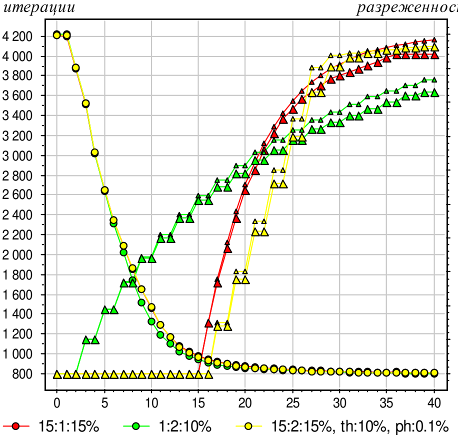
Разреживание распределений ϕ_{wt} и θ_{td}

робастная модель с фоном и шумом,
разреживание через 2 итерации



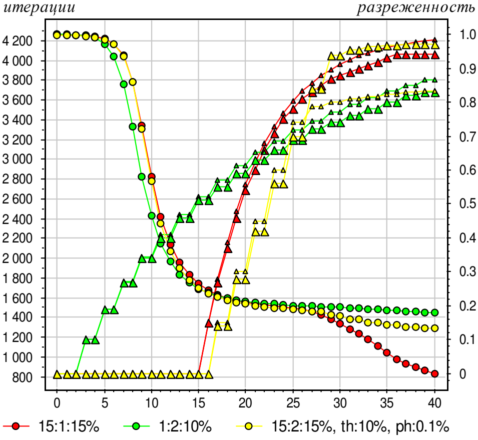
Разреживание распределений ϕ_{wt} и θ_{td}

робастная модель с фоном и шумом,
агрессивные стратегии разреживания $i_0{:}\delta{:}r$, th:R_θ, ph:R_ϕ



Разреживание распределений ϕ_{wt} и θ_{td}

упрощённая робастная модель без фона,
агрессивные стратегии разреживания i_0:δ:r, th:R_θ, ph:R_ϕ



Выводы

1. Возможно достигать разреженности 95–99% без ухудшения перплексии
2. При числе тем $|T| = 100$ это означает, что в среднем каждое слово относится к 1–5 темам
3. При этом многие строки матрицы Φ обнуляются, т.е. слово оказывается нетематическим

Регуляризатор №4: Антикорреляция

Гипотеза некоррелированности тем:
чем различнее темы, тем лучше они интерпретируются.

Минимизируем ковариации между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t\in T} \sum_{s\in T\backslash t} \sum_{w\in W} \phi_{wt}\phi_{ws} \to \max,$$

Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} \propto \Bigl(n_{wt} - \tau\phi_{wt} \sum_{s\in T\backslash t} \phi_{ws}\Bigr)_+.$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Регуляризатор №5: Максимизация когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, v \in W$.

Пусть C_{uv} — оценка когерентности, например $\hat{p}(v|u) = N_{uv}/N_u$.

$$R(\Phi, \Theta) = \tau \sum_{t \in T} \sum_{(u,v)} C_{uv} n_{ut} \ln \phi_{vt} \to \max,$$

Подставляем, получаем ещё один вариант сглаживания: векторы ϕ_{wt} притягиваются к эмпирическим оценкам распределений $p(w|t)$, вычисляемым по когерентным словам:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} n_{ut}.$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Регуляризатор №6: Выделение стоп-слов

Гипотеза: нетематические слова имеют во всех документах одинаковое распределение, близкое к $\hat{p}(w) = n_w/n$

Пусть $T_0 \subset T$ — подмножество тем для стоп-слов

Минимизируем сумму дивергенций $\mathrm{KL}_w(\hat{p}(w) \| \phi_{wt})$:

$$R(\Phi) = \beta_0 \sum_{t \in T_0} \sum_{w \in W} \frac{n_w}{n} \ln \phi_{wt} \to \max.$$

Подставляем, получаем «LDA только для стоп-слов»:

$$\phi_{wt} \propto \hat{n}_{wt} + \beta_0 n_w/n, \quad t \in T_0.$$

Совмещение с разреживанием, антикорреляцией, когерентностью должно переводить стоп-слова из других тем в темы из T_0

Регуляризатор №7: Связи между документами

Гипотеза: чем больше n_{dc} — число ссылок из d на c, тем более близки тематики документов d и c.

Минимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Phi, \Theta) = \tau \sum_{d,c \in D} n_{dc} \mathrm{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} \propto \hat{n}_{dt} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Регуляризатор №8: Классификация

Пусть C — множество классов документов (категории, пользователи, авторы, ссылки, годы, конференции,...)

Гипотеза:
классификация документа d объясняется его темами:

$$p(c|d) = \sum_{t\in T} p(c|t)p(t|d) = \sum_{t\in T} \psi_{ct}\theta_{td}$$

Минимизируем дивергенцию между моделью $p(c|d)$ и «эмпирической частотой» классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d\in D} \sum_{c\in C} m_{dc} \ln \sum_{t\in T} \psi_{ct}\theta_{td} \to \max$$

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

Регуляризатор №8: Классификация (ЕМ-алгоритм)

Е-шаг. По формуле Байеса:

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}} \qquad p(t|d,c) = \frac{\psi_{ct}\theta_{td}}{\sum\limits_{s\in T}\psi_{cs}\theta_{sd}}$$

М-шаг. Максимизация регуляризованного правдоподобия:

$$\phi_{wt} \propto n_{wt} \qquad n_{wt} = \sum_{d\in D} n_{dw}\, p(t|d,w)$$

$$\theta_{td} \propto n_{dt} + \tau m_{dt} \qquad n_{dt} = \sum_{w\in W} n_{dw}\, p(t|d,w) \qquad m_{dt} = \sum_{c\in C} m_{dc}\, p(t|d,c)$$

$$\psi_{ct} \propto m_{ct} \qquad m_{ct} = \sum_{d\in D} m_{dc}\, p(t|d,c)$$

Регуляризатор №9: Динамическая тематическая модель

Пусть классы C — это годы публикации

Гипотеза:
тематика меняется медленно, поэтому вероятности ψ_{ct} в последовательные годы $(c-1, c)$ должны быть близки:

$$R_2(\Psi) = -\frac{\tau_2}{2} \sum_{c \in C} \sum_{t \in T} \bigl(\psi_{ct} - \psi_{c-1,t}\bigr)^2 \to \max .$$

Сглаживание распределений $\psi_{ct} = p(c|t)$:
если значение ψ_{ct} меньше полусуммы соседних вероятностей $\psi_{c-1,t}$, $\psi_{c+1,t}$, то оно увеличивается, иначе — уменьшается:

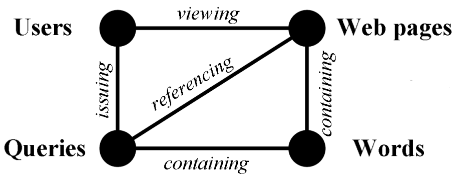
$$\psi_{ct} \propto \tau_1 m_{ct} + \tau_2 \psi_{ct} \bigl(\psi_{c-1,t} + \psi_{c+1,t} - 2\psi_{ct}\bigr).$$

Многомерные ТМ: коллаборативная фильтрация

Тематическое моделирование с классификацией документов:
документы D, термины W, классы C

Коллаборативная фильтрация:
предметы D с их описаниями, термины W, пользователи U

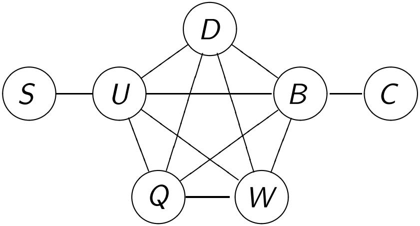
Персонализация поиска:
сайты D, термины W, пользователи U, запросы Q



Wang X., Sun J. T., Chen Z., Zhai C. X. Latent semantic analysis for multiple-type interrelated data objects // SIGIR'06, Pp. 236–243

Многомерные ТМ: рекламная сеть поисковой системы

Персонализация показов рекламы:
сайты D, термины W, пользователи U, запросы Q, баннеры B, социально-демографические классы пользователей S, рекламные кампании C



Объекты x всех типов получают тематические профили $p(t|x)$, учитывающие всевозможные взаимодействия между объектами

Подбор траекторий регуляризации

Пусть задана линейная комбинация регуляризаторов:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$$

Задача: выбрать вектор коэффициентов $\tau = (\tau_i)_{i=1}^{n}$

Ближайшие аналоги:

- Построение «Regularization Path» в задачах регрессии с двумя регуляризаторами L_1 и L_2 (Elastic Net)
- Постепенное разреживание тематической модели

Идея построения траектории в пространстве коэффициентов τ:
1) достичь сходимости нерегуляризованного PLSA,
2) усиливать регуляризаторы постепенно, в определённом порядке.

Открытые проблемы и задачи

Математические:

1. Всегда ли возможно подобрать априорные распределения так, чтобы результаты MAP и BI совпали?
2. Доказательство сходимости регуляризованного PLSA-EM
3. Разреживание $p(t|d,w)$: сэмплирование—максимизация
4. Устойчивое определение числа тем без HDP

Экспериментальные:

1. Регуляризаторы, улучшающие интерпретируемость тем
2. Многофункциональные, композитные, многомерные TM
3. Подбор траекторий регуляризации

Технические:

1. Реализация библиотеки регуляризаторов
2. Распределённая параллельная реализация (Big Data)

Воронцов Константин Вячеславович
voron@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН, 2014 (в печати).