Вероятностное тематическое моделирование

Мурат Апишев
great-mel@yandex.ru

МФТИ (ГУ)

25 сентября 2015

Материал лекции полностью основан на работах К. В. Воронцова.

Понятие «латентной темы»

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или слововсочетаний), совместно часто встречающихся в документах.
- *Тема* — вероятностное распределение на терминах.

Документ имеет ненаблюдаемый *тематический профиль*:

$p(t|d)$ — неизвестная частота темы t в документе d.

Когда автор писал термин w в документе d, он думал о теме t.

Документ d состоит из наблюдаемых терминов $w_1, \ldots, w_{n_d}$,

$p(w|d)$ — известная частота термина w в документе d.

Цели и предположения

1. Выявить скрытую тематическую структуру коллекции текстов
2. Выявить тематический профиль каждого документа

Основные предположения:

- Порядок документов в коллекции не важен
- Порядок слов в документе не важен (Bag-of-Words)
- Слова, встречающиеся почти во всех документах, не важны
- Слово в разных формах — это одно и то же слово
- Документ обычно относится к небольшому числу тем
- Тема обычно определяется небольшим числом терминов

Последние два пункта — *гипотеза разреженности*.

Постановка

Формализация основных предположений:

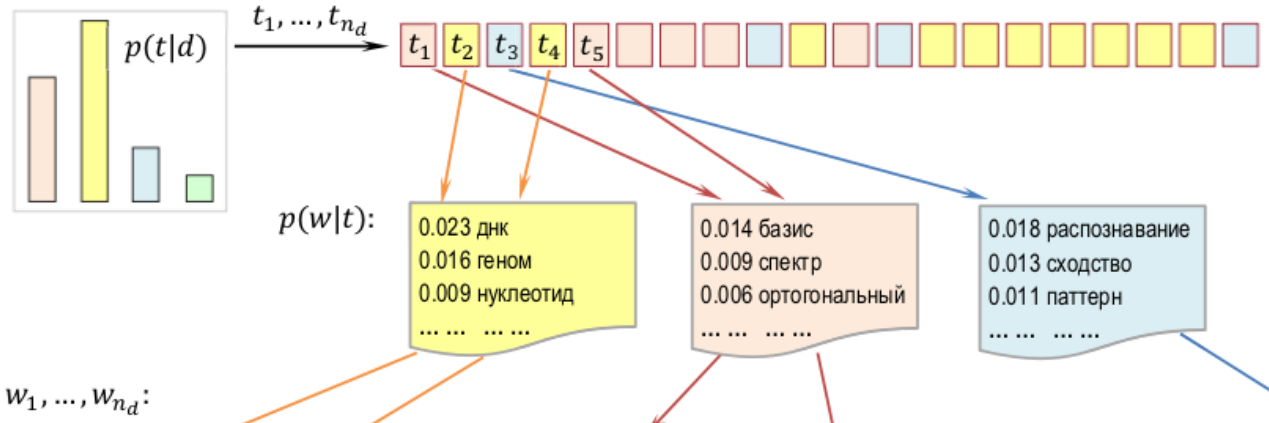
- каждое слово в документе связано с некоторой темой $t \in T$
- $W \times D \times T$ — дискретное вероятностное пространство
- коллекция D — это i.i.d выборка троек $(w_i, d_i, t_i)_{i=1}^n$ $p(w, d, t)$
- d_i, w_i — наблюдаемые, t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

Вероятностная модель порождения документа d:

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

Вероятностная модель порождения документа

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



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).

Вероятностная модель порождения документа

Алгоритм генерации текстовой коллекции

Входные данные: распределения $p(w|t), p(t|d)$;
Выходные данные: выборка пар $(d_i, w_i), i = \overline{1,n}$;

для каждого $d \in D$ **выполнять**
  задать длину n_d документа d;
  для каждого $i = \overline{1, n_d}$ **выполнять**
    $d_i := d$;
    выбрать случайную тему t_i из распределения $p(t|d_i)$;
    выбрать случайный термин w_i из распределения $p(w|t_i)$;

Задача матричного разложения

Задача тематического моделирования — обратная: по известной коллекции D оценить параметры модели $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$.

Число тем обычно гораздо меньше числа документов и объёма словаря.

Раскладываем матрицу частот $F = (f_{wd})_{W \times D}, f_{wd} = \dfrac{n_{dw}}{n_d}$ в виде произведения $F \approx \Phi\Theta$ двух неизвестных матриц меньшего размера — *матрицы терминов тем* $\Phi = (\phi_{wt})_{W \times T}$ и *матрицы тем документов* $\Theta = (\theta_{td})_{T \times D}$.

Все три матрицы — стохастические.

Частотные оценки вероятности

Вероятности, связанные с наблюдаемыми переменными d, w можно оценивать по выборке как частоты.

Такие частотные оценки являются несмещёнными оценками максимального правдоподобия.

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w|d) = \frac{n_{dw}}{n_d}$$

n_{dw} — число вхождений термина w в документ d;

$n_d = \sum_{w \in W} n_{dw}$ — длина документа d в терминах;

$n_w = \sum_{d \in D} n_{dw}$ — число вхождений документа w во все документы коллекции;

$n = \sum_{d \in D} \sum_{w \in d} n_{dw}$ — длина коллекции D в терминах.

Частотные оценки вероятности

Вероятности, связанные со скрытой переменной t, также можно оценивать как частоты, если рассматривать коллекцию документов как выборку троек (d, w, t)

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \quad \hat{p}(t|d) = \frac{n_{dt}}{n_d}, \quad \hat{p}(t|d,w) = \frac{n_{dwt}}{n_{dw}}$$

n_{dwt} — число троек, в которых термин w встретился в документе d и связан с темой t;

$n_{dt} = \sum_{w \in d} n_{dwt}$ — число троек, в которых термин из документа d связан с темой t;

$n_{wt} = \sum_{d \in D} n_{dwt}$ — число троек, в которых термин w связан с темой t;

$n_t = \sum_{d \in D} \sum_{w \in d} n_{dwt}$ — число троек, связанных с темой t.

ПМП

Для оценивания параметров Φ и Θ тематической модели по коллекции документов D будем максимизировать правдоподобие выборки:

$$p(D;\Phi,\Theta)=\prod_{i=1}^{n}p(d_i,w_i)=\prod_{d\in D}\prod_{w\in d}p(w|d)^{n_{dw}}p(d)^{n_{dw}}\to\max_{\Phi,\Theta}.$$

Прологарифмируем правдоподобие, чтобы превратить произведения в суммы, и отбросим константные слагаемые, не зависящие от параметров модели:

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

PLSA

Добавим ограничения нормировки и неотрицательности столбцов Φ и Θ:

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

$$\sum_{w\in W}\phi_{wt}=1,\ \phi_{wt}\geq 0;$$

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

Эта задача — основа модели *латентного семантического вероятностного анализа* (PLSA).

Решение — искомые параметры матричного разложения Φ и Θ.

EM-алгоритм

$\phi_{wt} = \frac{n_{wt}}{n_t}$ и $\theta_{td} = \frac{n_{dt}}{n_d}$. Их можно оценить, зная $n_{tdw} = n_{dw} p(t|d,w)$.
Формула Байеса:

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

Получаем систему уравнений относительно параметров модели ϕ_{wt} и θ_{td} и вспомогательных переменных p_{tdw}, n_{wt}, n_{dt}:

$$\mathrm{E-step}: p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}};$$

$$\mathrm{M-step}: \phi_{wt} = \frac{n_{wt}}{\sum_{v\in W} n_{vt}}; \qquad n_{wt} = \sum_{d\in D} n_{dw} p_{tdw};$$

$$\theta_{td} = \frac{n_{dt}}{\sum_{s\in T} n_{sd}}; \qquad n_{dt} = \sum_{w\in d} n_{dw} p_{tdw}.$$

Вывод формул M-шага для ϕ_{wt}

Лагранжиан задачи при ограничениях нормировки, но без ограничений неотрицательности:

$$\mathcal{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T}\lambda_t\left(\sum_{w\in W}\phi_{wt} - 1\right) - \sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right).$$

$$\frac{\partial \mathcal{L}}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} - \lambda_t = 0;$$

$$\sum_{d\in D} n_{dw}\frac{\theta_{td}\phi_{wt}}{p(w|d)} - \lambda_t\phi_{wt} \quad \Rightarrow \quad \lambda_t = \sum_{d\in D}\sum_{w\in W} n_{dw}p_{tdw};$$

$$\phi_{wt} = \frac{\sum_{d\in D} n_{dw}p_{tdw}}{\sum_{v\in W}\sum_{d\in D} n_{dv}p_{tdv}} \equiv \frac{n_{wt}}{n_t}, \forall w \in W, t \in T.$$

Вывод формул M-шага для θ_{td}

Лагранжиан задачи при ограничениях нормировки, но без ограничений неотрицательности:

$$\mathcal{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T}\lambda_t\left(\sum_{w\in W}\phi_{wt} - 1\right) - \sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right).$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{td}} = \sum_{w\in d} n_{dw}\frac{\phi_{wt}}{p(w|d)} - \mu_d = 0;$$

$$\sum_{w\in d} n_{dw}\frac{\phi_{wt}\theta_{td}}{p(w|d)} - \mu_d\theta_{td} \quad \Rightarrow \quad \mu_d = \sum_{t\in T}\sum_{w\in d} n_{dw}p_{tdw};$$

$$\theta_{td} = \frac{\sum_{w\in d} n_{dw}p_{tdw}}{\sum_{w\in d}\sum_{s\in T} n_{dw}p_{sdv}} \equiv \frac{n_{td}}{n_d}, \forall d \in D, t \in T.$$

Рациональный EM-алгоритм

Проблема: необходимость хранить 3D-матрицу p_{tdw}

Идея: E-шаг встраивается внутрь M-шага

PLSA-EM: рациональный EM-алгоритм для модели PLSA

Входные данные: коллекция D, число тем $|T|$, нач. приближения Φ, Θ;
Выходные данные: распределения Φ, Θ;

повторять
  обнулить n_{wt}, n_{dt}, n_t для всех $d \in D$, $w \in W$, $t \in T$;
  для каждого $d \in D$, $w \in d$ **выполнять**
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$; увеличить n_{wt}, n_{dt}, n_t на $\frac{n_{dw}\varphi_{wt}\theta_{td}}{Z}$ для всех тем $t \in T$;
  $\varphi_{wt} := \frac{n_{wt}}{n_t}$, для всех $w \in W, t \in T$;
  $\theta_{td} := \frac{n_{dt}}{n_t}$, для всех $d \in D, t \in T$;
до тех пор, пока Φ, Θ *не сойдутся*;

APTM

Поставленная задача стохастического матричного разложения — некорректно поставленная.

Если $F = \Phi\Theta$ — решение, то $F = (\Phi S)(S^{-1}\Theta)$ тоже является решением для всех невырожденных S, при которых матрицы $\Phi' = \Phi S$ и $\Theta' = S'\Theta$ являются стохастическими.

Для решения некорректно поставленных задач существует общий подход, называемый *регуляризацией*.

Аддитивная регуляризация тематических моделей (APTM) основана на введении дополнительных критериев-регуляризаторов $R_i(\Phi, \Theta), i = \overline{1, r}$,и максимизации их линейной комбинации с логарифмом правдоподобия $L(\Phi, \Theta)$.

APTM

$$R(\Phi,\Theta)=\sum_{i=1}^{k}\tau_i R_i(\Phi,\Theta), \qquad \mathcal{L}(\Phi,\Theta)+R(\Phi,\Theta)\to\max_{\Phi,\Theta}$$

$$\sum_{w\in W}\phi_{wt}\in\{0,1\},\ \phi_{wt}\geq 0;$$

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

τ_i — неотрицательные *коэффициенты регуляризации*.

Модель PLSA соответствует частному случаю отсутствия регуляризатора ($R(\Phi,\Theta)=0$).

Введём оператор:

$$p_i=\operatorname*{norm}_{i\in I} x_i=\frac{\max\{x_i,0\}}{\sum_{j\in I}\max\{x_j,0\}},\forall i\in I.$$

Регуляризованный EM-алгоритм

Теорема

Пусть функция $R(\Phi, \Theta)$ *непрерывно дифференцируема. Точка* (Φ, Θ) *локального экстремума регуляризованной задачи удовлетворяет системе уравнений со вспомогательными переменными* $p_{tdw} = p(t|d, w)$*:*

$$p_{tdw} = \underset{t \in T}{\text{norm}} \left(\phi_{wt} \theta_{td} \right);$$

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

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}.$$

Дивергенция Кульбака-Лейблера

KL-дивергенция — несимметричная функция расстояния между дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

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

Основные свойства KL-дивергенции:

1. KL-дивергенция неотрицательна и равна нулю тогда, и только тогда, когда распределения совпадают.
2. KL-дивергенция является мерой вложенности двух распределений. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то распределение P сильнее вложено в Q, чем Q в P.
3. Минимизация KL-дивергенции эквивалентна максимизации правдоподобия:

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

Сглаживание

Гипотеза сглаженности: ϕ_t близки к заданным распределениям $\beta_t = (\beta_{wt})_{w \in W}$

θ_d близки к заданным распределениям $\alpha_d = (\alpha_{td})_{t \in T}$

$$\sum_{t \in T} \mathrm{KL}_w(\beta_{wt} \| \phi_{wt}) \to \min_{\Phi}, \quad \sum_{d \in D} \mathrm{KL}_t(\alpha_{td} \| \theta_{td}) \to \min_{\Theta}.$$

Максимизируем сумму регуляризаторов в коэффициентами β_0, α_0:

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

Применение общих формул даёт выражение для M-шага:

$$\phi_{wt} = \underset{w \in W}{\mathrm{norm}}(n_{wt} + \beta_0 \beta_{wt}); \qquad \theta_{td} = \underset{t \in T}{\mathrm{norm}}(n_{td} + \alpha_0 \alpha_{td}).$$

Этот регуляризатор с положительными гиперпараметрами эквивалентен модели LDA. Гораздо проще!

Разреживание

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

Чем сильнее разреженно распределение, тем ниже его энтропия.

Максимальной энтропия обладает равномерное распределение.

Максимизируем дивергенцию между распределениями $\beta_t = (\beta_{wt})_{w \in W}$ и $\alpha_d = (\alpha_{td})_{t \in T}$:

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

Подставляем в формулы, получаем регуляризатор разреживания:

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}(n_{wt} - \beta_0 \beta_{wt}); \qquad \theta_{td} = \underset{t \in T}{\operatorname{norm}}(n_{td} - \alpha_0 \alpha_{td}).$$

Декорреляция тем

Тематическая модель тем полезнее, чем более различные темы она находит.

Минимизируем ковариации между столбцами ϕ_t:

$$R(\Phi,\Theta) = -\frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\operatorname{cov}(\phi_t,\phi_s) \to \max,\ \operatorname{cov}(\phi_t,\phi_s) = \sum_{w\in W}\phi_{wt}\phi_{ws}.$$

Формула для ϕ_{wt} примет вид

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} - \tau\phi_{wt}\sum_{s\in T\setminus t}\phi_{ws}\right).$$

Условные вероятности $\phi_{wt} = p(w|t)$ постепенно уменьшаются для тех терминов w, которые имеют большие значения вероятности ϕ_{ws} в других темах.

Приложения тематического моделирования

- Семантический поиск по документу любой длины
- Категоризация и классификация документов
- Поиск научной информации, трендов
- Анализ и агрегирование новостных потоков
- Рекомендательные системы коллаборативная фильтрация
- Анализ дискретизированных биомедицинских сигналов