Вероятностные тематические модели
Лекция 1. Введение

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • весна 2017

Содержание

1. **Мотивации и элементарная постановка задачи**
   - Задача тематического моделирования
   - Статистическая (частотная) интерпретация текста
   - Элементарный подход к решению задачи
2. **Математический инструментарий**
   - Принцип максимума правдоподобия
   - Условия Каруша–Куна–Таккера
   - Частотные оценки максимума правдоподобия
3. **Вероятностный латентный семантический анализ**
   - Тематическая модель PLSA
   - EM-алгоритм
   - Рациональный EM-алгоритм

Что такое «тема» в коллекции текстовых документов?

Неформально,

- *тема* — семантически однородное множество текстов
- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность (частота) термина w в теме t;
- *тематика* документа — условное распределение $p(t|d)$ — вероятность (частота) темы t в документе d.

Тематическая модель оценивает вероятности $p(w|t)$ и $p(t|d)$ по наблюдаемым частотам $p(w|d)$ слов w в документах d.

Цели и приложения тематического моделирования

- Выявить скрытую тематическую структуру коллекции текстов
- Выявить тематику каждого документа

Приложения:

- Семантический поиск по текстовому запросу любой длины
- Классификация, аннотирование, сегментация текстов
- Визуализация, систематизация, навигация по коллекции
- Поиск научной информации, трендов, фронта исследований
- Поиск специалистов (expert search), рецензентов, проектов
- Анализ новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендательные системы, коллаборативная фильтрация
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Коллекция текстовых документов

D — конечное множество документов
W — конечное множество терминов (слов или словосочетаний)
T — конечное множество тем, $|T| \ll |D|$, $|T| \ll |W|$

$(d_i, w_i, t_i)_{i=1}^{n} \subset D \times W \times T$ — коллекция текстовых документов

Когда автор документа d_i писал термин w_i, он думал о теме t_i, и мы хотели бы выявить, о какой именно.

Основные предположения:

- d_i, w_i — наблюдаемые, темы t_i — скрытые
- порядок терминов в документе не важен (bag of words)
- порядок документов в коллекции не важен (bag of docs)
- слова приведены к нормальным формам (лемматизированы)

Обозначения частот — счётчиков числа терминов

Ненаблюдаемые частоты, зависящие от t:

$n_{dwt} = \sum_{i=1}^{n} [d_i = d]\,[w_i = w]\,[t_i = t]$ — частота (d, w, t) в коллекции

$n_{wt} = \sum_d n_{dwt}$ — частота термина w в теме t

$n_{td} = \sum_w n_{dwt}$ — частота терминов темы t в документе d

$n_t = \sum_{d,w} n_{dwt}$ — частота терминов темы t в коллекции

Наблюдаемые частоты, не зависящие от t:

$n_{dw} = \sum_t n_{dwt}$ — частота термина w в документе d

$n_w = \sum_{d,t} n_{dwt}$ — частота термина w в коллекции

$n_d = \sum_{w,t} n_{dwt}$ — длина документа d

$n = \sum_{d,w,t} n_{dwt}$ — длина коллекции

Элементарная вероятностная формализация

- Коллекция — неслучайная последовательность $(d_i, w_i, t_i)_{i=1}^n$ равновероятных элементарных событий.

  Условные вероятности:

  $p(w|d) = \dfrac{n_{dw}}{n_d}$ — распределение терминов в документе d,

  $p(t|d) = \dfrac{n_{td}}{n_d}$ — распределение тем в документе d,

  $p(w|t) = \dfrac{n_{wt}}{n_t}$ — распределение терминов в теме t.

- **Гипотеза условной независимости:**
  «вероятность термина в теме не зависит от документа»,

$$p(w|d,t) = p(w|t)$$

$$\frac{n_{dwt}}{n_{td}} = \frac{n_{wt}}{n_t}$$

Задача тематического моделирования

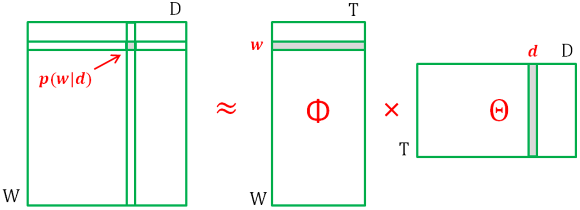
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Тематическая модель, по формуле полной вероятности:

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

Найти: параметры модели $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$

Это задача стохастического матричного разложения:



Элементарный подход к решению задачи

Выразим n_{dwt} через ϕ_{wt}, θ_{td} по формуле Байеса:

$$\frac{n_{dwt}}{n_{dw}} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

Получим систему уравнений относительно параметров модели ϕ_{wt}, θ_{td} и вспомогательных переменных n_{dwt}:

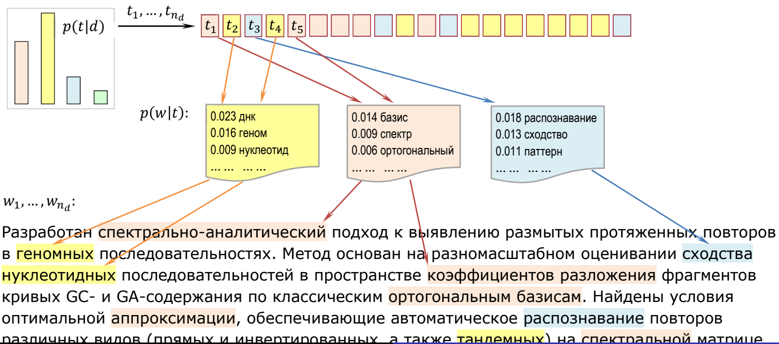
$$\begin{cases} n_{dwt} = n_{dw}\dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}, & d \in D,\ w \in W,\ t \in T; \\ \phi_{wt} \equiv \dfrac{n_{wt}}{n_t} = \dfrac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, & w \in W,\ t \in T; \\ \theta_{td} \equiv \dfrac{n_{td}}{n_d} = \dfrac{\sum_w n_{dwt}}{\sum_{t,w} n_{dwt}}, & d \in D,\ t \in T. \end{cases}$$

Численное решение — методом простых итераций

Стандартная вероятностная формализация

- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- вероятностная модель порождения текста документов:

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



Вероятностный процесс порождения текстов

- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- вероятностная модель порождения текста документов:

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

Вход: распределение $p(w|t)$ для каждой темы $t \in T$;
распределение $p(t|d)$ для каждого документа $d \in D$;
Выход: коллекция документов;

для всех документов $d \in D$
  для всех позиций слов $i = 1, \dots, n_d$ в документе d
    выбрать тему t_i из $p(t|d)$;
    выбрать слово w_i из $p(w|t_i)$;

Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки (d_i, w_i):

$$p(D) = \prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}.$$

Пусть $p(w|d, \alpha)$ — параметрическая вероятностная модель документа d, зависящая от вектора параметров $\alpha = (\Phi, \Theta)$.

Логарифм правдоподобия выборки D:

$$\ln p(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d, \alpha) p(d) \to \max_{\alpha}.$$

Избавимся от $p(d)$, не влияющего на точку максимума:

$$L(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d, \alpha) \to \max_{\alpha}.$$

Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \dots, m; \\ h_j(x) = 0, \quad j = 1, \dots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \dots, m$, λ_j, $j = 1, \dots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0;\ h_j(x) = 0;\ \text{(исходные ограничения)} \\ \mu_i \geqslant 0;\ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0;\ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Два упражнения на принцип максимума правдоподобия

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{dw} \to \max_{\xi}, \qquad \sum_{w\in W} \xi_{dw} = 1, \qquad \xi_{dw} \geqslant 0.$$

Лагранжиан: $\mathscr{L} = \sum_{d\in D}\Bigl(\sum_{w\in d} n_{dw} \ln \xi_{dw} - \lambda_d \Bigl(\sum_{w\in W} \xi_{dw} - 1\Bigr)\Bigr);$

$$\frac{\partial \mathscr{L}}{\partial \xi_{dw}} = n_{dw}\frac{1}{\xi_{dw}} - \lambda_d = 0 \quad\Rightarrow\quad \lambda_d = n_d, \quad \xi_{dw} = \frac{n_{dw}}{n_d} \equiv \hat{p}(w|d).$$

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{w} \to \max_{\xi}, \qquad \sum_{w\in W} \xi_{w} = 1, \qquad \xi_{w} \geqslant 0.$$

Лагранжиан: $\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{w} - \lambda \Bigl(\sum_{w\in W} \xi_{w} - 1\Bigr);$

$$\frac{\partial \mathscr{L}}{\partial \xi_{w}} = n_{w}\frac{1}{\xi_{w}} - \lambda = 0 \quad\Rightarrow\quad \lambda = n, \quad \xi_{w} = \frac{n_{w}}{n} \equiv \hat{p}(w).$$

Модель PLSA (Probabilistic Latent Semantic Analysis)

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

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

Ещё одна интерпретация: минимизация взвешенной суммы KL-дивергенций между тематическими $p(w|d) = \sum_t \phi_{wt}\theta_{td}$ и униграммными $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$ моделями документов:

$$\sum_{d\in D} n_d \mathrm{KL}_w\big(\hat{p}(w|d) \,\big\|\, p(w|d)\big) \to \min$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999. Pp. 50–57

Необходимые условия точки максимума правдоподобия

Теорема

Точка максимума правдоподобия Φ, Θ удовлетворяет системе уравнений со вспомогательными переменными n_{dwt}:

$$
\begin{array}{rl}
\text{E-шаг:} & \left\{\begin{array}{l}
n_{dwt} = n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_s \phi_{ws}\theta_{sd}}; \\[2ex]
\phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d \in D} n_{dwt}; \quad n_t = \sum\limits_w n_{wt} \\[2ex]
\theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w \in d} n_{dwt}; \quad n_d = \sum\limits_t n_{td}
\end{array}\right. \\
\text{M-шаг:} &
\end{array}
$$

EM-алгоритм — это чередование шагов E и M до сходимости, т. е. решение системы уравнений методом простых итераций.

ЕМ-алгоритм. Вывод формулы М-шага для ϕ_{wt}

Лагранжиан задачи максимизации правдоподобия при ограничениях нормировки но без ограничений неотрицательности:

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

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

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

$$\phi_{wt} = \frac{\sum\limits_{d\in D} n_{dw}\,p(t|d,w)}{\sum\limits_{d\in D}\sum\limits_{w'\in d} n_{dw'}\,p(t|d,w')} \equiv \frac{n_{wt}}{n_t} \quad \text{для всех } w\in W,\ \ t\in T.$$

EM-алгоритм. Вывод формулы M-шага для θ_{td}

Лагранжиан задачи максимизации правдоподобия при ограничениях нормировки но без ограничений неотрицательности:

$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} - \sum_{t\in T}\lambda_t\Bigl(\sum_{w\in W}\phi_{wt} - 1\Bigr) - \sum_{d\in D}\mu_d\Bigl(\sum_{t\in T}\theta_{td} - 1\Bigr)$$

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

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

$$\theta_{td} = \frac{\sum\limits_{w\in d} n_{dw}\,p(t|d,w)}{\sum\limits_{w\in d} n_{dw}\sum\limits_{t'\in T} p(t'|d,w)} \equiv \frac{n_{td}}{n_d} \quad \text{для всех } d\in D,\ \ t\in T.$$

Рациональный EM-алгоритм

Проблема: необходимость хранить 3D-матрицу n_{dwt}
Идея: Е-шаг встраивается внутрь М-шага

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы терминов тем Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех $i = 1, \dots, i_{\max}$ (итерация = один проход коллекции)
- $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
- **для всех** документов $d \in D$ и всех терминов $w \in d$
  - $n_{dwt} := n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $t \in T$;
  - $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dwt}$ для всех $t \in T$;
- $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
- $\theta_{td} := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Резюме

- Тематическое моделирование — это восстановление латентных тем в коллекции текстовых документов
- Цели — поиск, систематизация, классификация текстов
- Вероятностное пространство — $D \times W \times T$
- Базовая модель — PLSA
- Базовая задача — стохастическое матричное разложение
- Базовый метод оптимизации — EM-алгоритм
- Рациональный EM-алгоритм со сложностью $O(n \cdot |T|)$
- PLSA-EM примитивен, требует улучшений и расширений