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 2016

Содержание

Что такое «тема»?

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.

Более формально,

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

Когда автор писал термин w в документе d, он думал о теме t, и мы хотели бы выявить, о какой именно.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Цели и приложения тематического моделирования

- Выявить скрытую тематическую структуру коллекции текстов
- Выявить тематический профиль каждого документа

Приложения:

- Семантический поиск по текстовому запросу любой длины
- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов
- Поиск научной информации, трендов, фронта исследований
- Поиск специалистов (expert search), рецензентов, проектов
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендующие системы, коллаборативная фильтрация
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Основные предположения

1. Порядок документов в коллекции не важен
2. Порядок слов в документе не важен (bag of words)
3. Слово в разных формах — это одно и то же слово
4. Документ обычно относится к небольшому числу тем
5. Тема обычно определяется небольшим числом терминов

Предварительная обработка текстов:

- Приведение всех слов к нормальной форме (лемматизация или стемминг)
- Выделение терминов (term extraction)
- Удаление стоп-слов и слишком редких слов

Вероятностная порождающая модель

Формализация основных предположений:

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

Вероятностная модель порождения документа d:

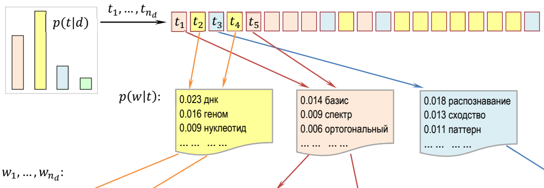
$$p(w|d) = \sum_{t \in T} p(w|d,t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

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

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t:

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



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 текстов

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

Процесс порождения коллекции документов как множества пар $(d_i, w_i)_{i=1}^n$, с точностью до произвольной их перестановки

Вход:

распределение $p(w|t)$ для каждой темы $t \in T$;
распределение $p(t|d)$ для каждого документа $d \in D$;

Выход:

коллекция документов;

1: **для всех** документов $d \in D$
2: **для всех** слов w в документе d
3: выбрать тему t из $p(t|d)$;
4: выбрать слово w из $p(w|t)$;

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дано: W — словарь терминов
D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$
n_{dw} — сколько раз термин w встретился в документе d
n_d — длина документа d

Найти: модель $p(w|d) = \sum_t \phi_{wt}\theta_{td}$ с параметрами ϕ, θ:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Критерий: максимизация логарифма правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\phi,\theta},$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geqslant 0; \quad \sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_t \theta_{td} = 1$$

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:

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

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

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

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

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

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

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

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

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

Интерпретация: стохастическое матричное разложение

$$F \approx \Phi\Theta,$$

$F = \big(\hat{p}(w|d)\big)_{W\times D}$ — 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 $\theta_{td} = p(t|d)$.

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}:

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

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

EM-алгоритм. Вывод формулы M-шага для ϕ_{wt}

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

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

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

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

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

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

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

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

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

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

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

EM-алгоритм. Элементарная интерпретация

EM-алгоритм — это чередование E и M шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t,d,w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

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

M-шаг: частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}\,p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{dwt}, \qquad n_t = \sum_{w\in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w\in d} n_{dwt}, \qquad n_d = \sum_{t\in T} n_{td}.$$

Частотные оценки условных вероятностей

Если рассматривать коллекцию как выборку троек (d, w, t), то

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

n_{dwt} — число троек (d, w, t) во всей коллекции

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

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

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

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

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

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

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

Рациональный EM-алгоритм

Проблема: необходимость хранить 3D-матрицу n_{dwt}
Идея: E-шаг встраивается внутрь M-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 итераций $i = 1, \dots, i_{\max}$

- $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
- **для всех** документов $d \in D$ и всех слов $w \in d$
  - $n_{dwt} := n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ 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 $d \in D$, $t \in T$;

Расширения тематических моделей

1. Устранение неединственности и неустойчивости решения: $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$ для невырожденных $S_{T\times T}$
2. Различность и разреженность тем
3. Выведение слов общей лексики из предметных тем
4. Модели дистрибутивной семантики
5. Тематические модели классификации и регрессии
6. Динамические модели развития тем во времени
7. Модели с автоматически определением числа тем
8. Иерархические тематические модели
9. Мультиграммные тематические модели
10. Мультиязычные тематические модели
11. Модели для сегментации и суммаризации текстов
12. Автоматическое именование тем
13. Модели гетерогенных и мультимодальных данных
14. Онлайновая, параллельная, распределённая реализация

Резюме

- Тематическое моделирование — это восстановление латентных тем в коллекции текстовых документов
- Тематическое моделирование используется для многих задач текстовой аналитики: «поиска по смыслу», классификации, сегментации, аннотирования, и др.
- Вероятностное тематическое моделирование — некорректно поставленная задача стохастического матричного разложения
- Базовая модель — PLSA
- Базовый метод оптимизации — EM-алгоритм
- Рациональный EM-алгоритм: каждая итерация — один линейный проход по коллекции
- PLSA-EM примитивен, требует улучшений и расширений