Latent Dirichlet allocation

Притыковская Наташа

30 октября 2019 г.

Overview

Тематическая модель

Тематическая модель (topic model) коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.
Вероятностная тематическая модель (ВТМ) описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем.

Применение тематических моделей

- выявление трендов в новостных потоках
- классификация и категоризация документов
- тегирование веб-страниц
- обнаружение текстового спама
- рекомендательные системы

Обозначения

- D — множество (коллекция) текстовых документов
- W — множество (словарь) всех употребляемых в них терминов (слов или словосочетаний).
- каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, ..., w_{n_d})$ из словаря W. Термин может повторяться в документе много раз.
- T - конечное множество тем, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна.

Вероятностное простанство и гипотеза независимости

- коллекция документов рассматривается как множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D * W * T$
- гипотеза о независимости элементов выборки эквивалентна тому, что порядок терминов в документах не важен - «bag of words»
- порядок документов в коллекции также не имеет значения - «bag of docs»

Обозначения 2.0

- D — множество (коллекция) текстовых документов
- W — множество (словарь) всех употребляемых в них терминов (слов или словосочетаний).
- каждый документ $d \in D$ представляет собой $d \subset W$, где каждому элементу w поставлено в соответствие n_{dw} - число вхождений термина w в документ d
- T - конечное множество тем, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна.

Построение тематической модели

Построить тематическую модель коллекции документов D — значит найти распределения $p(w|t)$ для всех тем $t \in T$ и распределения $p(t|d)$ для всех документов $d \in D$.

Гипотеза условной независимости

появление слов в документе d, относящихся к теме t, описывается общим для всей коллекции распределением $p(w|t)$ и не зависит от документа d, тогда:

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

Вероятностная модель порождения данных (1)

Согласно определению условной вероятности, формуле полной вероятности и гипотезе условной независимости:

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

Вероятностная модель порождения данных (2)

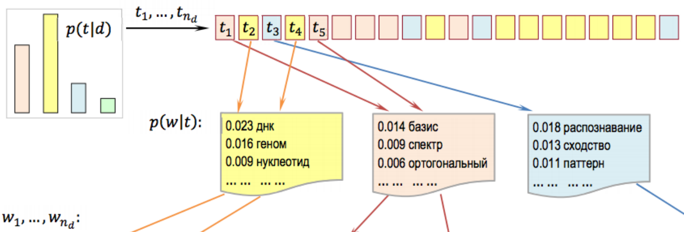
Согласно

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

коллекция D — это выборка наблюдений (d, w) сгенерированная

```
Вход: распределения p(w | t), p(t | d);
Выход: выборка пар (d_i, w_i),  i = 1, ..., n;
для всех d ∈ D
    задать длину n_d документа d;
    для всех i = 1, ..., n_d
        выбрать случайную тему t из распределения p(t | d);
        выбрать случайный термин w из распределения p(w | t);
        добавить в выборку пару (d, w), при этом тема t «забывается»;
```

Вероятностная модель порождения данных (3)



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).

Частотные оценки (1)

Вероятности, связанные с наблюдаемыми переменными d и w, можно оценивать по выборке как частоты $\hat{p}(d,w) = \frac{n_{dw}}{n}$ $\hat{p}(d) = \frac{n_d}{n}$ $\hat{p}(w) = \frac{n_w}{n}$ $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

- 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 термина w во все документы коллекции
- $n = \sum_{d \in D} \sum_{w \in d} n_{dw}$ — длина коллекции в терминах

Частотные оценки (2)

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

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

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

Запишем задачу в матричной форме

Равенство

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

можно понимать, как задачу представления заданной матрицы частот: $F = (\hat{p_{wd}})_{W*D}$, $\hat{p_{wd}} = \hat{p}(w|d) = \frac{n_{dw}}{n_d}$
в виде произведения $F \approx \Phi\Theta$, где

- Φ — матрица терминов тем, $\Phi = (\phi_{wt})_{W*T}$, $\phi_{wt} = p(w|t)$
- Θ — матрица тем документов, $\Theta = (\theta_{td})_{T*D}$, $\theta_{td} = p(t|d)$

Метод максимума правдоподобия

Для оценки параметров Φ, Θ тематической модели по коллекции документов D будем максимизировать правдоподобие (плотность распределния выборки):

$$p(D; \Phi, \Theta) = C \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}} = \prod_{d \in D} \prod_{w \in d} p(w|d)^{n_{dw}} C p(d)^{n_{dw}} \to \max_{\Phi, \Theta}$$

$$L(D; \Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \qquad (1)$$

При ограничениях неотрицательности $\theta_{td} \geq 0$, $\phi_{wt} \geq 0$ и нормировки:

$$\sum_{w \in W} \phi_{wt} = 1$$

,

$$\sum_{t \in T} \theta_{td} = 1$$

Probalistic Latent Semantic Analysis

Задача максимизации правдоподобия для вероятностной модели не имеет простого аналитического решения и решается численно с помощью EM-алгоритма.

EM-алгоритм(1)

EM-алгоритм - итерационный процесс, в котором каждая итерация состоит из двух шагов - E (expectation) и M (maximization).

- Перед первой итерацией выбирается начальное приближение параметров ϕ_{wt}, θ_{td}.
- На E-шаге по текущим значениям параметров ϕ_{wt}, θ_{td} с помощью формулы Байеса вычисляются условные вероятности $p(t|d,w)$ для всех тем $t \in T$ для каждого термина $w \in d$ в каждом документе d:

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

EM-алгоритм(2)

- Е-шаг:

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

- На М-шаге, наоборот, по условным вероятностям тем H_{dwt} вычисляется новое приближение параметров ϕ_{wt}, θ_{dt}:

$$\hat{n}_{dwt} = n_{dw}p(t|d,w) = n_{dw}H_{dwt}$$

оценивает число n_{dwt} вхождений термина w в документ d, связанных с темой t. Просуммировав $\hat{n}_{dwt}$ по документам d и по терминам w, получим оценки $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ и далее

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \hat{n}_t = \sum_{w\in W}\hat{n}_{wt}, \hat{n}_{wt} = \sum_{d\in D} n_{dw}H_{dwt} \tag{2}$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \hat{n}_d = \sum_{t\in T}\hat{n}_{dt}, \hat{n}_{dt} = \sum_{w\in D} n_{dw}H_{dwt} \tag{3}$$

EM-алгоритм(3)

Покажем теперь, что оценки 2 и 3:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \hat{n}_t = \sum_{w \in W} \hat{n}_{wt}, \hat{n}_{wt} = \sum_{d \in D} n_{dw} H_{dwt}$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}, \hat{n}_{dt} = \sum_{w \in D} n_{dw} H_{dwt}$$

действительно являются решением задачи максимизации правдоподобия 1:

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

Метод множителей Лагранжа

Метод нахождения условного экстремума функции $f(x)$, где $x \in R^n$, относительно m ограничений $\phi_i(x) = 0$, где i меняется от 1 до m.

- составим функцию Лагранжа:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i \phi_i(x)$$

  , где

$$\lambda = (\lambda_1, \dots, \lambda_m)$$

- составим систему из $n + m$ уравнений, приравняв к нулю частные производные функции Лагранжа $L(x, \lambda)$ по x_i и λ_j
- если полученная система имеет решение отномительно x_i и λ_j, то для точки x выполняются необходимые условия минимума.

Доказательство сходимости EM-алгоритма (1)

Запишем Ланграножиан задачи 1 при ограничениях нормировки, проигнорировав ограничения неотрицательности:

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

Продифференцировав ланграножиан по ϕ_{wt} и приравняв нулю производную, получим

$$\lambda_t = \sum_{d \in D} n_{dw} \frac{\theta_{td}}{p(w|d)} \tag{4}$$

Домножим обе части этого равества на ϕ_{wt}, просуммируем по всем терминам $w \in W$, применим условие нормировки вероятностей ϕ_{wt} в левой части и выделим переменную H_{dwt} в правой части. Получим

$$\lambda_t = \sum_{d \in D} \sum_{w \in W} n_{dw} H_{dwt}$$

Доказательство сходимости EM-алгоритма (2)

Снова домножим обе части 4 на ϕ_{wt}, выделим переменную H_{dwt} в правой части и выразим ϕ_{wt} из левой части, подставив выражение для λ_t, получим:

$$\phi_{wt} = \frac{\sum_{d \in D} n_{dw} H_{dwt}}{\sum_{w' \in W} \sum_{d \in D} n_{dw'} H_{dw't}}$$

Обозначив числитель через $\hat{n}_{wt}$ получим 2.
Проделав аналогичные действия с производной лангранажиана по θ_{td} получим 3 ДЗ.

Латентное размещение Дирихле (1)

Основным недостатком PLSA считается высокая размерность пространства параметров. Для сокращеня размерности ипользуется либо:

- отбор признаков
- регулиризация - наложение дополнительных ограничений на параметры

Байесовская регуляризация - введение априорного распределения вероятности в пространстве параметров.

Латентное размещение Дирихле (2)

Тематическая модель латентного размещения Дирихле (LDA) основана на дополнительном предположении, что векторы документов $\theta_d = (\theta_{td}) \in R^{|T|}$ и векторы тем $\phi_t = (\phi_{wt}) \in R^{|W|}$ порождаются распределением Дирихле с параметрами $\alpha \in R^{|T|}$ и $\beta \in R^{|W|}$:

$$Dir(\theta_d, \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \alpha_t > 0, \alpha_0 = \sum_t \alpha_t, \theta_{td} > 0, \sum_t \theta_{td} = 1$$

$$Dir(\phi_t, \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \beta_w > 0, \beta_0 = \sum_w \beta_w, \phi_{wt} > 0, \sum_w \phi_{wt} = 1$$

Латентное размещение Дирихле (3)

Если вектор θ_d порождается распределим Дирихле, а документ d представляется в виде выборки n_d пар тема-темин $X_d = (t_1, w_1), \dots, (t_{n_d}, w_{n_d})$, где в каждой паре (t_i, w_i) тема t_i выбирается из дискретного распределения $p(t|d) = \theta_{td}$, а потом выбирается w_i из дискретного распределения $p(w|t) = \phi_{wt}$, то апостериорное распределение

$$p(\theta_d|X_d, \alpha) = Dir(\theta_d; \alpha'), \text{где } \alpha'_t = \alpha_t + n_{td}$$

Оценим случайную величину θ_{td} ее математическим ожиданием по апосртериорному распределению

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

Латентное размещение Дирихле (4)

Оценим случайную величину θ_{td} ее математическим ожиданием по апостериорному распределению

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

Заменив величину n_{td} ее оценкой $\hat{n}_{td}$ получим сглаженную байсевскую оценку параметра θ_{td} для EM-алгоритма, обощающую 2

$$\theta_{td} = \frac{\hat{n}_{td} + \alpha_t}{\hat{n}_d + \alpha_0}$$

Аналогично выводится сглаженная байсевская оценка для ϕ_{wt}:

$$\phi_{wt} = \frac{\hat{n}_{wt} + \beta_t}{\hat{n}_t + \beta_0}$$