Курс: Байесовские методы машинного обучения, 2011

Тематические модели для коллекций текстов

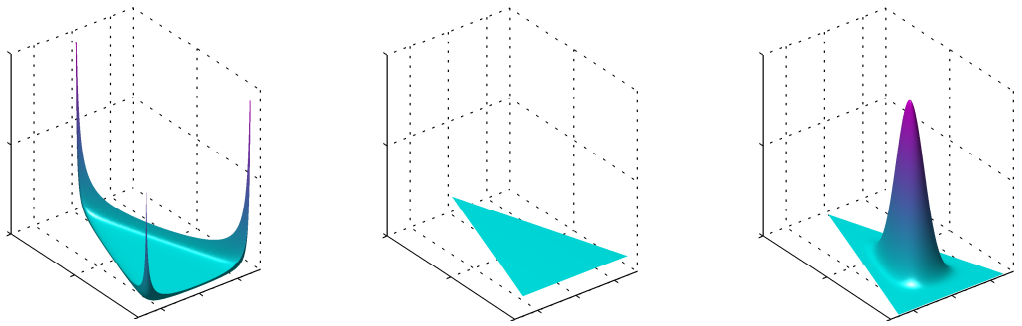
Дата: 14 декабря 2011

Распределение Дирихле

Случайная величина $\boldsymbol{\theta} \in \mathbb{R}^K$, определенная на симплексе $(\theta_k \geq 0, \sum_{k=1}^K \theta_k = 1)$, имеет распределение Дирихле, если ее плотность определяется как

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}, \;\; \alpha_k > 0.$$

Здесь $\Gamma(\cdot)$ – гамма-функция, $\boldsymbol{\alpha}$ – набор параметров распределения. Различные виды распределения Дирихле для случая $K = 3$ показаны на рис. 1. Заметим, что в случае $\alpha_1 = \cdots = \alpha_K = 1$ распределение Дирихле переходит в равномерное распределение на симплексе.



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

Рис. 1: Различные виды распределения Дирихле

Математическое ожидание и ковариация для распределения Дирихле определяются как

$$\mathbb{E}_p \theta_i = \frac{\alpha_i}{\alpha_0},$$
$$\mathrm{Cov}(\theta_i, \theta_j) = \frac{\alpha_i \alpha_0 [i = j] - \alpha_i \alpha_j}{\alpha_0^2 (\alpha_0 + 1)},$$
$$\alpha_0 = \sum_k \alpha_k.$$

Распределение Дирихле является представителем экспоненциального семейства распределений с набором параметров $[\alpha_1 - 1, \ldots, \alpha_K - 1]$ и набором достаточных статистик $\boldsymbol{u}(\boldsymbol{\theta}) = [\log \theta_1, \ldots, \log \theta_K]$. Известно, что для распределения из экспоненциального семейства все моменты достаточных статик определяются соответствующими производными нормировочной константы. Отсюда получаем, что

$$\mathbb{E}_p \log \theta_i = \frac{\partial}{\partial(\alpha_i - 1)} \log \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} = \Psi(\alpha_i) - \Psi(\sum_k \alpha_k). \tag{1}$$

Здесь $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$ – дигамма функция.

Распределение Дирихле часто используется в качестве априорного распределения для набора дискретных вероятностей. Рассмотрим дискретную случайную величину, принимающую K значений:

$$\begin{array}{cccc} 1 & 2 & \dots & K \\ \theta_1 & \theta_2 & \dots & \theta_K \end{array}$$

Рассмотрим задачу оценки параметров $\boldsymbol{\theta}$ этой случайной величины по выборке из нее объема N с помощью метода максимального правдоподобия:

$$p(X|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(x_n|\boldsymbol{\theta}) = \prod_{n=1}^{N} \theta_{x_n} \to \max_{\boldsymbol{\theta}:\theta_k\geq 0, \sum_k \theta_k = 1}$$

Данная задача условной оптимизации может быть решена аналитически с помощью функции Лагранжа L:

$$L(\boldsymbol{\theta}, \lambda) = \log p(X|\boldsymbol{\theta}) + \lambda\left(\sum_k \theta_k - 1\right) = \sum_{n=1}^{N} \log\theta_{x_n} + \lambda\left(\sum_k \theta_k - 1\right) =$$
$$= \sum_{k=1}^{K} \log\theta_k \left(\sum_{n=1}^{N} [x_n = k]\right) + \lambda\left(\sum_{k=1}^{K}\theta_k - 1\right).$$

Приравнивая производные функции Лагранжа к нулю и суммируя по k, получаем:

$$\frac{\partial}{\partial\theta_k} L(\boldsymbol{\theta}, \lambda) = \frac{\sum_{n=1}^{N}[x_n = k]}{\theta_k} + \lambda = 0 \;\Rightarrow\; \theta_k = -\frac{1}{\lambda}\sum_{n=1}^{N}[x_n = k], \;\Rightarrow\; \lambda = -\sum_{k=1}^{K}\sum_{n=1}^{N}[x_n = k] = -N.$$

Таким образом, оценка максимального правдоподобия для параметров $\boldsymbol{\theta}$ определяется частотами:

$$\theta_k = \frac{\sum_{n=1}^{N}[x_n = k]}{N}. \tag{2}$$

Введем распределение Дирихле $\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ в качестве априорного распределения для параметров $\boldsymbol{\theta}$ и рассмотрим оценку максимума апостериорного распределения:

$$p(\boldsymbol{\theta}|X, \boldsymbol{\alpha}) \to \max_{\boldsymbol{\theta}} \;\Leftrightarrow\; p(X|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \to \max_{\boldsymbol{\theta}}. \tag{3}$$

Действуя аналогично случаю максимума правдоподобия, получаем следующее решение данной задачи оптимизации:

$$\theta_k = \frac{\alpha_k - 1 + \sum_{n=1}^{N}[x_n = k]}{\sum_{j=1}^{K}\alpha_j - K + N}. \tag{4}$$

Заметим, что в случае равномерного априорного распределения ($\alpha_1 = \dots \alpha_K = 1$) данное решение переходит в оценку максимального правдоподобия (2). При всех $\alpha_k > 1$ решение (4) является менее контрастным, чем решение (2), и, в частности, задает ненулевую вероятность для исходов, ни разу не наблюдавшихся в обучающей выборке. В этом случае происходит сглаживание вероятностей. Напротив, при $\alpha_k < 1$ решение (4) является более контрастным по сравнению с (2), т.к. в этом случае априорное распределение имеет большой вес у границ симплекса. Например, в случае $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ и выборки из одной единицы, двух двоек и трех

троек оценки максимального правдоподобия и максимального апостериорного распределения соответственно равны:

$$\theta_{ML,1} = \frac{1}{6},\ \theta_{ML,2} = \frac{1}{3},\ \theta_{ML,3} = \frac{1}{2},$$
$$\theta_{MP,1} = \frac{1}{33},\ \theta_{MP,2} = \frac{11}{33},\ \theta_{MP,3} = \frac{21}{33}.$$

Пусть для некоторых $k \in \{1, \ldots, K\}$ значение $\alpha_k - 1 + \sum_n [x_n = k] \leq 0$. Обозначим множество таких индексов через $K_{\leq 0}$, а множество оставшихся индексов — через $K_{>0}$. Тогда можно показать, что решение задачи (3) вместо (4) становится следующим:

$$\theta_{MP,k} = \begin{cases} 0, & \text{если } k \in K_{\leq 0}, \\ \frac{\alpha_k - 1 + \sum_n [x_n = k]}{\sum_{j \in K_{>0}} (\alpha_j - 1 + \sum_n [x_n = j])}, & \text{иначе.} \end{cases}$$

Модель LDA

Тематические модели (чтение статей по данной теме рекомендуется начать с вводных обзоров [1, 2]) предназначены для описания текстов с точки зрения их тематик (рубрик). Поэтому в тематических моделях, как правило, используется модель мешка слов, т.е. каждый документ рассматривается как набор терминов (слов, словосочетаний), которые в нем используются. При этом порядок употребления терминов в документе игнорируется. Соответственно, для применения тематических моделей набор текстов (корпус) подвергается предобработке, которая выделяет термины (значимые слова) в каждом документе. В частности, на этом этапе отбрасываются слова, которые встречаются практически в каждом тексте, — союзы, предлоги, вводные слова и проч.

Основное предположение тематической модели Latent Dirichlet Allocation (сокращенно LDA) [3] состоит в том, что каждый документ имеет несколько тематик, смешанных в некоторой пропорции. Например, в тексте на рис. 2 можно выделить тематику, связанную с анализом данных (выделено синим) со словами «computer», «prediction» и т.д., тематику, связанную с процессом эволюции (выделено розовым, слова-представители «life», «organism»), тематику, связанную с генетикой (выделено желтым, слова-представители «genes», «sequenced») и проч.

LDA — это вероятностная модель порождения текста. Здесь тематика рассматривается как некоторое распределение вероятностей в пространстве слов из общего словаря. Например, генетическая тематика задает высокие вероятности для слов «ген», «секвенирование» и т.д., а компьютерная тематика задает высокие вероятности для слов «вычисления», «компьютер», «память», «алгоритм» и т.д. Процесс порождения текста в модели LDA состоит из двух этапов. На первом этапе для данного текста выбирается некоторое распределение вероятностей в пространстве тематик. На втором этапе сначала для каждого слова (термина) в тексте генерируется тематика из распределения вероятностей на тематиках, а затем само слово генерируется из соответствующего распределения для выбранной тематики.

Введем формальные обозначения для всех упомянутых понятий, таких как документы, тематики, слова в документе, тематики для каждого слова в документе и т.д. (см. табл. 1). Тогда

Таблица 1: Основные обозначения

$w \in \{1, \dots, W\}$	— номер слова в словаре
$t \in \{1, \dots, T\}$	— номер тематики
$d \in \{1, \dots, D\}$	— номер документа в корпусе
N_d	— число слов в документе d
$\boldsymbol{w}_d = [w_{d,1}, \dots, w_{d,N_d}]$	— слова в документе d, $w_{d,n} \in \{1, \dots, W\}$
$\boldsymbol{z}_d = [z_{d,1}, \dots, z_{d,N_d}]$	— тематики слов в документе d, $z_{d,n} \in \{1, \dots, T\}$
$\boldsymbol{\theta}_d = [\theta_{d,1}, \dots, \theta_{d,T}]$	— вероятности тематик в документе d
$\boldsymbol{\phi}_t = [\phi_{t,1}, \dots, \phi_{t,W}]$	— вероятности слов в тематике t
$\Theta = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D]^T \in \mathbb{R}^{D \times T}$	— вероятности тематик во всех документах
$\Phi = [\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_T]^T \in \mathbb{R}^{T \times W}$	— вероятности слов во всех тематиках
$\mathcal{W} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_D\}$	— набор всех слов в корпусе
$\mathcal{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_D\}$	— разбиение всех слов по тематикам

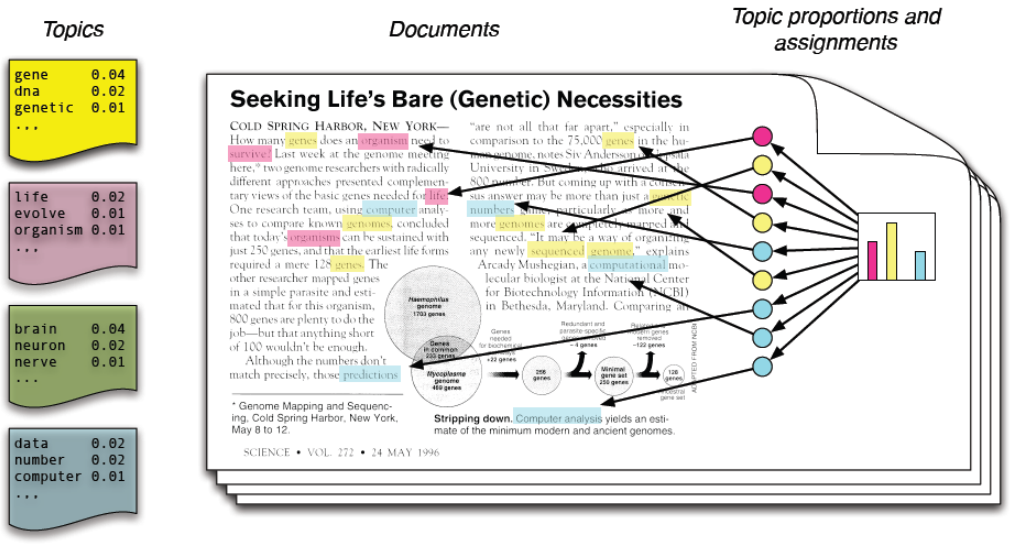


Рис. 2: Иллюстрация порождающей модели текста с помощью набора тематик

вероятностная модель LDA задается как

$$
\begin{aligned}
&p(\mathcal{W}, \mathcal{Z}, \Theta | \Phi, \alpha) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d | \alpha) \prod_{n=1}^{N} p(w_{d,n} | z_{d,n}, \Phi) p(z_{d,n} | \boldsymbol{\theta}_d), \\
&p(\boldsymbol{\theta}_d | \alpha) = \mathrm{Dir}(\boldsymbol{\theta}_d | \alpha), \\
&p(w_{d,n} | z_{d,n}, \Phi) = \Phi_{z_{d,n}, w_{d,n}}, \;\; p(z_{d,n} | \boldsymbol{\theta}_d) = \Theta_{d, z_{d,n}}.
\end{aligned}
\tag{5}
$$

Рассмотрим подробнее модель (5). Если мысленно отбросить априорные распределения на параметры Θ, а также просуммировать по всем $\mathcal{Z}$, то получится, что модель (5) представляет каждый документ $\boldsymbol{w}_d$ как выборку из смеси дискретных распределений, соответствующих тематикам. Поэтому обучение модели LDA в данном случае эквивалентно одновременной кластеризации нескольких выборок (одна выборка соответствует одному документу) с общими компонентами. При добавлении априорных распределений на параметры Θ к общим компо-

нентам добавляются общие сглаживающие факторы (или, наоборот, факторы, способствующие разреженности в пространстве тематик для документов). Заметим также, что модель (5) игнорирует порядок документов в корпусе.

Наряду с моделью (5) рассмотрим расширение этой модели путем добавления априорных распределений Дирихле на параметры Φ:

$$\begin{aligned} &p(\mathcal{W}, \mathcal{Z}, \Theta, \Phi|\alpha, \beta) = \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\beta) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N} p(w_{d,n}|z_{d,n}, \Phi) p(z_{d,n}|\boldsymbol{\theta}_d), \\ &p(\boldsymbol{\phi}_t|\beta) = \mathrm{Dir}(\boldsymbol{\phi}_t|\beta),\ p(\boldsymbol{\theta}_d|\alpha) = \mathrm{Dir}(\boldsymbol{\theta}_d|\alpha), \\ &p(w_{d,n}|z_{d,n}, \Phi) = \Phi_{z_{d,n},w_{d,n}},\ p(z_{d,n}|\boldsymbol{\theta}_d) = \Theta_{d,z_{d,n}}. \end{aligned} \tag{6}$$

Введение априорных распределений на параметры Φ позволяет, в зависимости от желаний пользователя, сглаживать данные величины или способствовать разреженности в пространстве слов для тематик. Заметим, что модель (6) переходит в модель (5) при $\beta = 1$. Заметим также, что модель (6) по своей структуре полностью повторяет байесовскую модель смеси нормальных распределений. В обоих случаях предлагается вводить априорное распределение Дирихле на веса компонент смеси, а на сами компоненты вводить сопряженное распределение (Гаусса-Уишарта для нормального распределения и Дирихле для дискретного распределения).

Вариационный подход для вывода в модели LDA

Рассмотрим решение задачи обучения модели LDA (5) с помощью метода максимального правдоподобия:

$$p(\mathcal{W}|\Phi, \alpha) \to \max_{\Phi,\alpha}. \tag{7}$$

Заметим, что здесь пользователь заранее задает количество тематик T. Величина правдоподобия $p(\mathcal{W}|\Phi, \alpha)$ не может быть вычислена аналитически даже для небольших T и объемов документов в корпусе, т.к. требует, в частности, суммирования по всем $\mathcal{Z}$, что соответствует суммированию по $T^{\sum_d N_d}$ слагаемым. Для решения задачи (7) воспользуемся вариационным EM-алгоритмом:

$$\text{E-шаг: } q(\Theta, \mathcal{Z}) \simeq p(\Theta, \mathcal{Z}|\mathcal{W}, \Phi, \alpha), \tag{8}$$

$$\text{M-шаг: } \mathbb{E}_q \log p(\mathcal{W}, \mathcal{Z}, \Theta|\Phi, \alpha) \to \max_{\Phi,\alpha}. \tag{9}$$

Апостериорное распределение на E-шаге не может быть вычислено аналитически. Поэтому воспользуемся вариационным подходом и будем искать аппроксимирующее распределение q в семействе факторизованных распределений:

$$q(\Theta, \mathcal{Z}) = \prod_d q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \prod_{n=1}^{N_d} q(z_{d,n}|\boldsymbol{\mu}_{d,n}). \tag{10}$$

Здесь $\boldsymbol{\gamma}_d = [\gamma_{d,1}, \ldots, \gamma_{d,T}]$ и $\boldsymbol{\mu}_{d,n} = [\mu_{d,n,1}, \ldots, \mu_{d,n,T}]$ — параметры распределения q. Вариационный подход состоит в покоординатной оптимизации факторов в распределении q (оптимизации одного фактора при фиксировании всех остальных). При этом оптимальный фактор находится путем усреднения $\log p(\mathcal{W}, \mathcal{Z}, \Theta|\Phi, \alpha)$ по всем фиксированным факторам. Рассмотрим решение этой задачи отдельно для всех факторов распределения q.

Компонента $q(\boldsymbol{\theta}_c|\boldsymbol{\gamma}_c)$.

$$\log q(\boldsymbol{\theta}_c|\boldsymbol{\gamma}_c) = \int \log p(\mathcal{W},\mathcal{Z},\Theta|\Phi,\alpha)\prod_{d\neq c} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)d\boldsymbol{\theta}_d \prod_{d=1}^{D}\prod_{n=1}^{N_d} q(z_{d,n}|\boldsymbol{\mu}_{d,n})dz_{d,n} + \text{Const} =$$

$$= \log p(\boldsymbol{\theta}_c|\alpha) + \sum_{n=1}^{N_c} \mathbb{E}_{q(z_{c,n}|\boldsymbol{\mu}_{c,n})} \log \underbrace{p(z_{c,n}|\boldsymbol{\theta}_c)}_{\theta_{c,z_{c,n}}} + \text{Const} = \sum_{t=1}^{T}(\alpha-1)\log\theta_{c,t} + \sum_{n=1}^{N_c}\sum_{t=1}^{T}\mu_{c,n,t}\log\theta_{c,t} + \text{Const}.$$

Это выражение как функция от $\boldsymbol{\theta}_c$ представляет собой логарифм распределения Дирихле. Следовательно,

$$q(\boldsymbol{\theta}_c|\boldsymbol{\gamma}_c) = \text{Dir}(\boldsymbol{\theta}_c|\boldsymbol{\gamma}_c),\ \ \gamma_{c,t} = \alpha + \sum_{n=1}^{N_c}\mu_{c,n,t}. \tag{11}$$

Компонента $q(z_{d,n}|\boldsymbol{\mu}_{d,n})$.

$$\log q(z_{c,m}|\boldsymbol{\mu}_{c,m}) = \int \log p(\mathcal{W},\mathcal{Z},\Theta|\Phi,\alpha)\prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)d\boldsymbol{\theta}_d \prod_{(d,n)\neq(c,m)} q(z_{d,n}|\boldsymbol{\mu}_{d,n})dz_{d,n} + \text{Const} =$$

$$= \mathbb{E}_{q(\boldsymbol{\theta}_c|\boldsymbol{\gamma}_c)} \log \underbrace{p(z_{c,m}|\boldsymbol{\theta}_c)}_{\theta_{c,z_{c,m}}} + \log p(w_{c,m}|z_{c,m},\Phi) + \text{Const} =$$

$$= \Psi(\gamma_{c,z_{c,m}}) - \Psi\left(\sum_{t=1}^{T}\gamma_{c,t}\right) + \log \underbrace{p(w_{c,m}|z_{c,m},\Phi)}_{\Phi_{z_{c,m},w_{c,m}}} + \text{Const}.$$

Здесь при вычислении $\mathbb{E}_{q(\boldsymbol{\theta}_c|\boldsymbol{\gamma}_c)}\log\theta_{c,z_{c,m}}$ мы воспользовались результатом (1). Таким образом, $q(z_{c,m}|\boldsymbol{\mu}_{c,m})$ представляет собой дискретное распределение, принимающее T значений с вероятностями

$$\mu_{c,m,s} = \frac{\Phi_{s,w_{c,m}}\exp(\Psi(\gamma_{c,s}))}{\sum_{k=1}^{T}\Phi_{k,w_{c,m}}\exp(\Psi(\gamma_{c,k}))}, \quad s = 1,\ldots,T. \tag{12}$$

Итеративный пересчет по формулам (11),(12) производится до сходимости по значению функционала

$$\mathcal{L} = \mathbb{E}_q \log p(\mathcal{W},\mathcal{Z},\Theta|\Phi,\alpha) - \mathbb{E}_q \log q(\Theta|\mathcal{Z}).$$

Легко показать, что значение $\mathcal{L}$ может быть вычислено аналитически:

$$\mathbb{E}_q \log p(\mathcal{W},\mathcal{Z},\Theta|\Phi,\alpha) = D\log\Gamma(\alpha T) - DT\log\Gamma(\alpha) + \sum_{d=1}^{D}\Bigg[(\alpha-1)\sum_{t=1}^{T}\Psi(\gamma_{d,t}) -$$

$$- (N_d - T(\alpha-1))\Psi(\sum_{k=1}^{T}\gamma_{d,k}) + \sum_{n=1}^{N_d}\sum_{t=1}^{T}(\log\Phi_{t,w_{d,n}} + \Psi(\gamma_{d,t}))\mu_{d,n,t}\Bigg], \tag{13}$$

$$\mathbb{E}_q \log q(\Theta|\mathcal{Z}) = \sum_{d=1}^{D}\Bigg[\log\Gamma(\sum_{t=1}^{T}\gamma_{d,t}) - \sum_{t=1}^{T}\log\Gamma(\gamma_{d,t}) + \sum_{t=1}^{T}(\gamma_{d,t}-1)(\Psi(\gamma_{d,t}) - \Psi(\sum_{k=1}^{T}\gamma_{d,k}))\Bigg] +$$

$$+ \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}\mu_{d,n,t}\log\mu_{d,n,t}.$$

Заметим, что выражение (13) — это функция, которая оптимизируется на М-шаге (9) по параметрам Φ, α. Легко показать, что оптимальное значение Φ вычисляется как

$$\Phi_{t,w} = \frac{\sum_{(d,n):[w_{d,n}=w]} \mu_{d,n,t}}{\sum_{d,n} \mu_{d,n,t}}.$$

Оптимальное значение α может быть найдено с помощью численного метода одномерной оптимизации. Возникающая здесь задача оптимизации по α является выпуклой и равносильна поиску максимума плотности некоторого распределения Дирихле $\mathrm{Dir}(\boldsymbol{x}|\alpha)$.

Заметим также, что функционал $\mathcal{L}$ является нижней оценкой для значения $\log p(\mathcal{W}|\Phi, \alpha)$. Таким образом, в результате вариационной оптимизации (11),(12),(13) находится не только аппроксимант для апостериорного распределения $p(\mathcal{Z}, \Theta|\mathcal{W}, \Phi, \alpha)$, но и оценка обоснованности $p(\mathcal{W}|\Phi, \alpha)$.

Найденные в результате вариационной оптимизации параметры $\boldsymbol{\gamma}_d$ и $\boldsymbol{\mu}_{d,n}$ для оптимальных значений параметров Φ, α используются для оценки искомых распределений по тематикам для каждого документа и разбиения слов документов по тематикам путем взятия статистик распределений $q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)$ и $q(z_{d,n}|\boldsymbol{\mu}_{d,n})$:

$$\hat{\theta}_{d,t} = \frac{\gamma_{d,t}}{\sum_{k=1}^{T} \gamma_{d,k}}, \quad \hat{z}_{d,n} = \arg\max_t \mu_{d,n,t}. \tag{14}$$

При добавлении в коллекцию нового документа $\boldsymbol{w}_{new}$ его вероятность $p(\boldsymbol{w}_{new}|\Phi, \alpha)$, а также распределение по тематикам и разбиение слов по тематикам может быть найдено с помощью вариационного приближения для апостериорного распределения $p(\boldsymbol{z}_{new}, \boldsymbol{\theta}_{new}|\boldsymbol{w}_{new}, \Phi, \alpha)$.

Возможность вычисления правдоподобия тестовой выборки $p(\mathcal{W}_{test}|\Phi, \alpha)$ открывает путь к автоматическому определению количества тематик T с помощью скользящего контроля. Параметр T является типичным структурным параметром, который обеспечивает компромисс между точностью восстановления обучающих данных $p(\mathcal{W}|\Phi, \alpha)$ и интерпретабельностью восстанавливаемых тематик. В скользящем контроле для поиска T обучающая выборка разбивается несколько раз на обучающую $\mathcal{W}_{train}$ и тестовую $\mathcal{W}_{test}$. По обучающей выборке находятся оптимальные значения $(\Phi_*, \alpha_*) = \arg\max p(\mathcal{W}_{train}|\Phi, \alpha)$ и оценивается правдоподобие тестовой выборки $p(\mathcal{W}_{test}|\Phi_*, \alpha_*)$. Величина правдоподобия тестовой выборки усредняется по всем разбиениям исходной выборки на обучение и тест. Значение T, максимизирующее среднее правдоподобие тестовой выборки, признается наилучшим. Альтернативный способ автоматического подбора T, основанный на непараметрическом байесовском подходе, описан в работе [4].

В заключении данного раздела отметим, что представленный здесь вариационный подход может быть напрямую обобщен для расширенной тематической модели LDA (6).

Схема Гиббса для вывода в модели LDA

Вариационный подход, представленный в предыдущем разделе, ориентирован, в первую очередь, на оптимизацию правдоподобия $p(\mathcal{W}|\Phi, \alpha)$. При этом происходит одновременный поиск аппроксиманта для апостериорного распределения $p(\mathcal{Z}, \Theta|\mathcal{W}, \Phi, \alpha)$ в семействе факторизованных распределений (10). Однако, поиск подобного аппроксиманта не соответствует, вообще говоря, индивидуальным приближениям $p(\boldsymbol{z}_d, \boldsymbol{\theta}_d|\mathcal{W}, \Phi, \alpha)$ с помощью факторов $q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)$ и $q(z_{d,n}|\boldsymbol{\mu}_{d,n})$. В результате, величины $\hat{\boldsymbol{\theta}}_d$ и $\hat{z}_{d,n}$, вычисляемые по формулам (14), могут оказаться плохими приближениями для истинных статистик индивидуальных апостериорных распределений $p(\boldsymbol{z}_d, \boldsymbol{\theta}_d|\mathcal{W}, \Phi, \alpha)$. Возможным решением обозначенной проблемы является дополнительный

поиск вариационного приближения для всех индивидуальных апостериорных распределений $p(\boldsymbol{z}_d, \boldsymbol{\theta}_d|\mathcal{W}, \Phi, \alpha)$ после определения наилучших параметров (Φ, α). Однако, такой многократный запуск может требовать значительных временных ресурсов. Альтернативной процедурой здесь является схема Гиббса, которая подробно изложена в работе [5].

Рассмотрим расширенную модель LDA (6). В этой модели маргинальное распределение $p(\mathcal{W}, \mathcal{Z}|\alpha, \beta)$ может быть вычислено аналитически:

$$p(\mathcal{W}, \mathcal{Z}|\alpha, \beta) = \int p(\mathcal{W}, \mathcal{Z}, \Theta, \Phi|\alpha, \beta) d\Theta d\Phi =$$

$$= \left[\int \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\beta) \prod_{d=1}^{D}\prod_{n=1}^{N_d} p(w_{d,n}|z_{d,n}, \Phi) d\Phi\right]\left[\int \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N_d} p(z_{d,n}|\boldsymbol{\theta}_d) d\Theta\right] =$$

$$\left(\frac{\Gamma(\beta W)}{(\Gamma(\beta))^W}\right)^T \left[\int \prod_{t=1}^{T}\prod_{w=1}^{W} \phi_{t,w}^{\beta-1+\sum_{d,n=1}^{D,N_d}[z_{d,n}=t, w_{d,n}=w]} d\boldsymbol{\phi}_t\right] \left(\frac{\Gamma(\alpha T)}{(\Gamma(\alpha)^T)}\right)^D \left[\int \prod_{d=1}^{D}\prod_{t=1}^{T} \theta_{d,t}^{\alpha-1+\sum_{n=1}^{N_d}[z_{d,n}=t]} d\boldsymbol{\theta}_d\right] =$$

$$\left(\frac{\Gamma(\beta W)}{(\Gamma(\beta))^W}\right)^T \prod_{t=1}^{T} \frac{\prod_{w=1}^{W}\Gamma(\beta + \sum_{d,n=1}^{D,N_d}[z_{d,n}=t, w_{d,n}=w])}{\Gamma(\beta W + \sum_{d,n=1}^{D,N_d}[z_{d,n}=t])} \left(\frac{\Gamma(\alpha T)}{(\Gamma(\alpha)^T)}\right)^D \prod_{d=1}^{D} \frac{\prod_{t=1}^{T}\Gamma(\alpha + \sum_{n=1}^{N_d}[z_{d,n}=t])}{\Gamma(\alpha T + N_d)}.$$

Применим схему Гиббса для генерации выборки из распределения $p(\mathcal{Z}|\mathcal{W}, \alpha, \beta)$. Для этого найдем одномерное условное распределение $p(z_{c,m}|\mathcal{Z}_{\setminus(c,m)}, \mathcal{W}, \alpha, \beta)$ для всех пар (c, m). Нетрудно показать, что данное распределение равно

$$p(z_{c,m} = j|\mathcal{Z}_{\setminus(c,m)}, \mathcal{W}, \alpha, \beta) \propto \frac{(\alpha + \sum_{n\neq m}[z_{c,n}=j])(\beta + \sum_{(d,n)\neq(c,m)}[z_{d,n}=j, w_{d,n}=w_{c,m}])}{\beta W + \sum_{(d,n)\neq(c,m)}[z_{d,n}=j]}.$$

Таким образом, данное распределение зависит от трех показателей: количества слов, за исключением текущего $w_{c,m}$, отнесенных к тематике j в текущем документе c, общего количества слов во всех документах, за исключением текущего $w_{c,m}$, отнесенных к тематике j, и общего количества слов данного типа $w_{c,m}$, отнесенных к тематике j во всех документах (не считая текущего слова $w_{c,m}$). Все эти показатели легко пересчитываются при переходе между парами (c, m) в одномерных условных распределениях.

В результате применения схемы Гиббса мы получаем выборку

$$\mathcal{Z}_1, \ldots, \mathcal{Z}_P \sim p(\mathcal{Z}|\mathcal{W}, \alpha, \beta).$$

Эта выборка может быть использована для оценки наиболее вероятного разбиения слов $\mathcal{W}$ по тематикам, распределения тематик для каждого документа и вероятностей слов в каждой тематике:

$$\hat{\mathcal{Z}} = \arg\max_{\mathcal{Z}_p} p(\mathcal{Z}, \mathcal{W}|\alpha, \beta),$$

$$\hat{\theta}_{d,t} = \frac{1}{P}\sum_{p=1}^{P} \frac{\alpha + \sum_{n=1}^{N_d}[z_{d,n}^p = t]}{\alpha T + N_d},$$

$$\hat{\Phi}_{t,w} = \frac{1}{P}\sum_{p=1}^{P} \frac{\beta + \sum_{d=1}^{D}\sum_{n=1}^{N_d}[z_{d,n}^p = t, w_{d,n} = w]}{\beta W + \sum_{d=1}^{D}\sum_{n=1}^{N_d}[z_{d,n}^p = t]}.$$

Наряду с вариационным подходом и методами Монте Карло для приближенного вывода в модели LDA можно использовать также подход распространения ожидания [6].

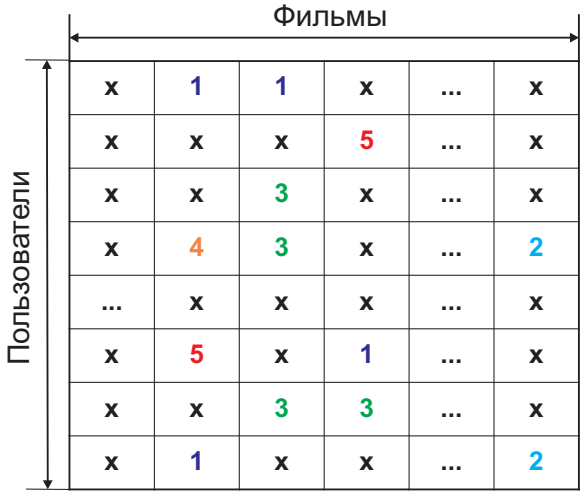


Рис. 3: Иллюстрация к задаче коллаборативной фильтрации.

Способы применения тематической модели LDA

Тематическая модель LDA является вероятностной моделью и поэтому может быть использована как метод восстановления плотности в пространстве текстовых коллекций. При этом высокие значения плотности будут соответствовать как документам из обучающей коллекции, так и документам со схожим тематическим наполнением. Одной из возможных областей применения такого метода восстановления плотности является решение задачи идентификации принадлежности документа определенному дискурсу. В том случае, если значение восстановленной плотности для тестового документа меньше определенного порога, то такой документ признается не соответствующим текстовой коллекции.

Модель LDA может быть использована как метод уменьшения размерности для документов. Исходным вектором признаков для документа является разреженный вектор длины W, где в позиции w стоит количество употреблений слова w в документе. С помощью модели LDA можно получить представление для документа длины T, состоящего из вероятностей тематик $\boldsymbol{\theta}_{d,t}$. Заметим, что количество слов W в словаре, как правило, составляет десятки или сотни тысяч, а количество тематик T обычно выбирается равным нескольким десяткам. Таким образом, с помощью LDA можно произвести радикальное сокращение размерности для представления документов. Пример применения такого метода уменьшения размерности для классификации документов из базы данных Reuters-21578 по тематикам можно найти в [3].

Модель LDA (5) используется также для решения задачи коллаборативной фильтрации. Здесь входной информацией является разреженная матрица действительных чисел, и задача состоит в восстановлении значений этой матрицы в пропущенных позициях. Например, данная матрица может описывать предпочтения пользователей для коллекции фильмов (см. рис. 3), и задача состоит в выработке рекомендации для каждого пользователя по просмотру новых для него фильмов. При применении тематической модели LDA в этой задаче каждый пользователь интерпретируется как документ, каждый фильм — как слово из словаря, а рейтинг пользователя за данный фильм — как количество употреблений слова-фильма в документе-пользователе. На этапе обучения находятся параметры LDA (Φ, α) по данным пользователей, отрейтинговавших большое количество фильмов. Затем для очередного пользователя d, отрейтинговавшего ряд фильмов (для него известен набор слов $\boldsymbol{w}_d$ длины N), рейтинг за новый

фильм w вычисляется как

$$\hat{k} = \arg\max_k p(w_{d,N+1} = w, \ldots, w_{d,N+k} = w | \boldsymbol{w}_d, \Phi, \alpha).$$

Здесь соответствующая вероятность находится как

$$p(w_{d,N+1} = w, \ldots, w_{d,N+k} = w | \boldsymbol{w}_d, \Phi, \alpha) =$$
$$= \int \sum_{z_{d,N+1},\ldots,z_{d,N+k}} \left(\prod_{i=1}^{k} p(w_{d,N+i} | z_{d,N+i}, \Phi) \right) p(z_{d,N+1}, \ldots, z_{d,N_k}, \boldsymbol{\theta}_d | \boldsymbol{w}_d, \Phi, \alpha) d\boldsymbol{\theta}_d =$$
$$\{p(z_{d,N+1}, \ldots, z_{d,N+k}, \boldsymbol{\theta}_d | \boldsymbol{w}_d, \Phi, \alpha) \approx q(\boldsymbol{\theta}_d | \boldsymbol{\gamma}_d) \prod_{i=1}^{k} q(z_{d,N+i} | \boldsymbol{\mu}_{d,N+i})\} \approx \prod_{i=1}^{k} \sum_{t=1}^{T} \Phi_{t,w} \mu_{d,N+i,t}.$$

В данном вычислении мы воспользовались вариационным приближением для апостериорного распределения $p(z_{d,N+1}, \ldots, z_{d,N+k}, \boldsymbol{\theta}_d | \boldsymbol{w}_d, \Phi, \alpha)$.

Наконец, модель LDA может быть использована для решения задачи информационного поиска документов из коллекции, наиболее близких к заданному. Для этого вводится функция расстояния между двумя документами c и d:

$$\rho(c, d) = \frac{1}{2}(\mathrm{KL}(\boldsymbol{\theta}_c \| \boldsymbol{\theta}_d) + \mathrm{KL}(\boldsymbol{\theta}_d \| \boldsymbol{\theta}_c)).$$

Здесь через KL обозначена дивергенция Кульбака-Лейблера между двумя вероятностными распределениями, а через $\boldsymbol{\theta}_d$ – распределение тематик в документе d, получаемое в помощью LDA.

Список литературы

[1] D.M. Blei. Introduction to Probabilistic Topic Models.
http://www.cs.princeton.edu/ blei/papers/Blei2011.pdf

[2] A. Daud, J. Li, L. Zhou, F. Muhammad. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China, V. 4, No. 2., 2010, pp. 280–301.

[3] D.M. Blei, A.Y. Ng, M.I. Jordan. Latent Dirichlet Allocation // Journal of Machine Learning Research, V. 3, 2003, pp. 993–1022.

[4] Y. Teh, M. Jordan, M. Beal, D. Blei. Hierarchical Dirichlet processes // Journal of the American Statistical Association, V. 101, 2006, pp. 1566–1581.

[5] M. Steyvers, T. Griffiths. Probabilistic topic models // T. Landauer, D.S. McNamara, S.Dennis, W. Kintsch (Eds.), Handbook of Latent Semantic Analysis. Hillsdale, NJ: Erlbaum. 2007.

[6] T. Minka, J. Lafferty. Expectation-propagation for the generative aspect model // UAI, 2002.