Байесовские методы в тематическом моделировании

Анна Потапенко (anya_potapenko@mail.ru)

Конспект ориентирован на студентов второго курса ВМК. Излагаются методы математической статистики и байесовские методы машинного обучения, необходимые для решения задач тематического моделирования.

Тематическое моделирование – это метод автоматической обработки текстов, который позволяет выделить и описать присущие им темы.

Исходные данные – коллекция (набор) документов, будем индексировать их $d = 1, 2 \dots D$. Каждый документ – это последовательность слов, будем индексировать их $i = 1, 2, \dots N_d$, где N_d – длина документа d. Обозначим за x_{di} слово в документе d, стоящее на позиции i, а за X – всю совокупность слов. Это *наблюдаемые* переменные, они нам даны.

Построить модель – это значит как-то описать реальное явление, возможно, упрощенно. Мы будем строить вероятностную модель коллекции документов. Но перед тем, как это сделать, отвлечемся на простой поясняющий пример.

Простой пример: вероятностная модель волшебной монетки

Кто-то нашел на улице волшебную монетку и подбросил ее 1000 раз. Он записал, что выпадало каждый раз: $b = (b_1, \dots b_{1000})$, и рассказал об этом нам. Теперь у нас есть наблюдаемые переменные: вектор b, в котором встретилось H орлов и T решек. Мы не знаем, как устроена эта волшебная монетка. Например, вдруг в руках хороших людей она падает орлом, а в руках плохих – решкой. Но мы пробуем описать ее вероятностной моделью и делаем ряд простых предположений:

- каждый бросок не зависит от результатов всех других бросков;
- есть некоторая вероятность выпадения орла – θ.

Тогда мы можем записать следующую модель:

$$p(b) = \prod_{i=1}^{1000} p(b_i|\theta) = \theta^H (1-\theta)^T$$

Это *порождающая, генеративная* модель, а θ - ее параметр. Теперь выбрав какое-то значение θ мы можем сгенерировать новые данные, моделирующие броски волшебной монетки. Чем лучше наша модель, тем больше такие данные будут похожи на реальные. А как выбрать подходящее значение θ, соответствующее наблюдаемым данным? Это важный вопрос, которым мы еще займемся.

PLSA – вероятностная тематическая модель коллекции документов

Вернемся теперь к текстам. Как и в случае с волшебной монеткой мы предполагаем, что есть некоторая вероятностная модель, согласно которой порождаются все документы. Конечно же, для такого сложного объекта как текст, можно придумать много разных моделей, которые будут основанны на разных предположениях. Ввести модель, которая будет хорошо описывать реальные данные и при этом оставаться достаточно простой для дальнейших математических выкладок – это большое искусство. Сейчас опишем классическую *модель вероятностного латентного семантического анализа* (Probabilistic Latent Semantic Analysis, PLSA [1]).

В рамках этой модели предполагается, что есть некоторый набор тем, их число заранее фиксируется (пусть это будет T), и с каждым словом в документе связывается одна определенная тема $z_{di} \in 1, \dots T$. Это та тема, о которой думал автор, когда употребил конкретное слово x_{di}. Обозначим за Z темы всех словпозиций (d, i) в коллекции. В отличие от слов, темы мы не видим, поэтому это *скрытые* переменные, значения которых нам хотелось бы узнать.

Далее, предполагаем, что все словопозиции (d, i) независимы друг от друга. Согласно формуле условной вероятности для одной словопозиции можем записать вероятность того, что она содержит слово w и связана с темой t:

$$p(x_{di} = w, z_{di} = t) = p(x_{di} = w | z_{di} = t) p(z_{di} = t) \tag{1}$$

В этой формуле каждая вероятность формально зависит от документа d и позиции внутри него i. Однако для упрощения делаются два предположения:

1. Гипотеза «мешка слов»: порядок слов в документе не важен для тематического анализа, т.е. зависимости от i нет. Действительно, если перемешать слова в учебнике по биологии и в учебнике по математике, мы все равно сможем различить их тематику.

2. Гипотеза условной независимости: вероятность слова при условии темы не зависит от документа: $p(w|t, d) = p(w|t)$. Т.е. если мы зафиксировали тему, то в каком бы документе мы ни находились, вероятность увидеть определенное слово подчиняется распределению $p(w|t)$.

Тогда в правой части формулы (1) стоят условные вероятности p(w|t) и p(t|d). Это важные объекты, для которых мы введем отдельные обозначения, и поймем их смысл.

Матрица $\Phi^{W \times T}$ размерности число слов в словаре на число тем содержит дискретные вероятностные распределения на множестве слов для каждой темы: $\phi_{wt} = p(w|t)$. Таким образом, в нашей моделе тема характеризуется частотой слов, используемых для ее описания. Например, в теме, посвященной театру, большие вероятности будут иметь слова «актер» или «номер» и маленькую вероятность слова «генетика» или «число». Матрица $\Theta^{T \times D}$ размерности число документов на число тем содержит вероятностные распределения на множестве тем для каждого документа: $\theta_{td} = p(t|d)$. Таким образом, каждый документ характеризуется темами, которым он посвящен.

Наконец, мы готовы к тому, чтобы полностью выписать вероятностную модель:

$$p(X, Z | \Phi, \Theta) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(x_{di}, z_{di} | \Phi, \Theta) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} \phi_{x_{di} z_{di}} \theta_{z_{di} d} \tag{2}$$

Здесь Φ и Θ – это параметры модели. Именно их мы будем искать в течение оставшейся части лекции.

Общий подход: метод максимума правдоподобия

Метод максимума правдоподобия – один из наиболее часто используемых инструментов математической статистики. Это способ оценить неизвестный параметр θ, если есть выборка $X = (X_1, \dots X_n)$ из распределения $p(X|\theta)$, известного с точностью до θ. Для этого нужно подставить в плотность распределения наблюдаемые данные, т.е. записать функцию правдоподобия $L(\theta) = p(X|\theta)$, максимизировать ее по θ, и точка максимума будет является искомой оценкой. Оценки максимума правдоподобия обладают рядом хороших свойств, про которые будет рассказано в других курсах. Здесь же

отметим, что оценки максимума правдоподобия обычно имеют простую интерпретацию частотных оценок. Так, если применить метод для волшебной монетки из первого примера, то оценка выпадения орла окажется равной: $\theta = \frac{H}{H+T}$. Это отношение числа «успешных» бросков, закончившихся орлом, к общему числу бросков. Очень разумно.

Метод максимума правдоподобия для модели PLSA: гипотетический случай, когда темы Z известны

Мы задали модель PLSA вероятностным распределением $p(X, Z|\Phi, \Theta)$. Если мы хотим сказать, что это то самое распределение из метода максимального правдоподобия, которое известно с точностью до парметров Φ и Θ, то нам нужна выборка слов и тем X, Z. К сожалению, темы мы не наблюдаем, но для начала предположим, что они у нас есть. Как бы мы тогда оценивали параметры?

$$\log L(\Phi, \Theta) = \log p(X, Z|\Phi, \Theta) = \sum_{d=1}^{D} \sum_{i=1}^{N_d} (\log \phi_{x_{di} z_{di}} + \log \theta_{z_{di} d}) \to \max_{\Phi, \Theta} \tag{3}$$

Это задача максимизации правдоподобия. Мы подставили совместное распределение из (2) и для удобства взяли логарифм. Далее вспомним, что есть дополнительные ограничения на Φ и Θ, т.к. их столбцы образуют дискретные распределения:

$$\sum_{w \in W} \phi_{wt} = 1, \sum_{t \in T} \theta_{td} = 1, \phi_{wt} \geq 0, \theta_{td} \geq 0, \quad \forall w = 1, \dots W, t = 1, \dots T, d = 1 \dots D \tag{4}$$

Ограничения неотрицательности выполнятся автоматически. Для того, чтобы учесть ограничения нормировки, воспользуемся методом множителей Лагранжа:

$$\sum_{d=1}^{D} \sum_{i=1}^{N_d} (\log \phi_{x_{di} z_{di}} + \log \theta_{z_{di} d}) - \sum_{t=1}^{T} \lambda_t \left(\sum_{w=1}^{W} \phi_{wt} - 1 \right) - \sum_{d=1}^{D} \mu_d \left(\sum_{t=1}^{T} \theta_{td} - 1 \right) \to \max_{\Phi, \Theta}$$

Возьмем производную по одному элементу ϕ_{wt} матрицы Φ и приравняем ее к нулю:

$$\frac{1}{\phi_{wt}} \sum_{d=1}^{D} \sum_{i=1}^{N_d} [x_{di} = w][z_{di} = t] - \lambda_t = 0 \tag{5}$$

Здесь квадратные скобки обозначают индикатор: 1, если выражение внутри скобок истинно, и 0 иначе. Таким образом, $\sum_{d=1}^{D} \sum_{i=1}^{N_d} [x_{di} = w][z_{di} = t]$ – это число раз, когда в коллекции встретилось слово w, отнесенное к теме t. Обозначим эту величину за n_{wt}.

$$n_{wt} = \lambda_t \phi_{wt} \; \forall w = 1, \dots W \quad \Rightarrow \quad \sum_{w=1}^{W} n_{wt} = \sum_{w=1}^{W} \lambda_t \phi_{wt}, \quad \Rightarrow \quad \sum_{w=1}^{W} n_{wt} = \lambda_t$$

Тогда искомая оценка:

$$\phi_{wt} = \frac{n_{wt}}{\sum_{w=1}^{W} n_{wt}}$$

Получился хорошо интерпретируемый результат: вероятность слова w в теме t – это отношение числа раз, когда тема t связывалась со словом w, к общему числу появлений темы t в коллекции.

Совершенно аналогичный результат можно получить для параметров θ_{td}:

$$\theta_{td} = \frac{n_{td}}{\sum_{t=1}^{T} n_{td}},$$

где $n_{td} = \sum_{i=1}^{N_d} [z_{di} = t]$.

Вернемся к реальности: переменные Z – это скрытые переменные, они нам не известны. Как тогда оценить параметры модели? Это очень распространенная постановка задачи:

$$p(X|\Phi,\Theta) = \sum_Z p(X,Z|\Phi,\Theta) \to \max_{\Phi\Theta} \tag{6}$$

Такую задачу называют максимизацией *неполного* правдоподобия. Неполного – потому что из функции правдоподобия выведены скрытые переменные Z. Чтобы их вывести, происходит суммирование по всем возможным значениям набора Z, которых может быть очень много. Поэтому просто взять производную от функции правдоподобия с учетом ограничений, как мы делали раньше, не получится.

Здесь на помощь приходит мощный и очень полезный алгоритм машинного обучения – *EM-алгоритм.* В нашем случае предполагается независимость каждой словопозиции (d,i), поэтому от суммирования по всевозможным наборам можно перейти к суммам по темам каждой отдельной словопозиции z_{di} и довольно просто вывести EM-алгоритм для конкретно этой задачи. Однако мы сейчас рассмотрим EM-алгоритм в общем случае, так как его более сложные варианты нам понадобятся для других задач в тематическом моделировании.

Общий подход: EM-алгоритм

Запишем задачу максимизации неполного правдоподобия для некой вероятностной модели, в которой есть наблюдаемые переменные X, скрытые переменные Z и параметры Ω:

$$\log p(X|\Omega) \to \max_{\Omega}$$

Справедлива следующая цепочка равенств:

$$\log p(X|\Omega) = \{q(Z) \text{ – произвольное распределение}\} = \int q(Z)\log p(X|\Omega)dZ =$$

$$\int q(Z)\log\frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}dZ = \int q(Z)\log\frac{p(X,Z|\Omega)}{q(Z)}\frac{q(Z)}{p(Z|X,\Omega)}dZ =$$

$$\underbrace{\int q(Z)\log p(X,Z|\Omega)dZ - \int q(Z)\log q(Z)dZ}_{L(q,\Omega)} + \underbrace{\int q(Z)\log\frac{q(Z)}{p(Z|X,\Omega)}dZ}_{KL(q(Z)||p(Z|X,\Omega))} \tag{7}$$

Дивергенция Кульбака-Лейблера $KL(q(Z)||p(Z|X,\Omega))$ оценивает расстояние между двумя распределениями. Дивергенция Кульбака-Лейблера

- неотрицательна;
- равна нулю тогда и только тогда, когда распределения совпадают;
- несимметрична.

В силу неотрицательности $KL(q(Z)||p(Z|X,\Omega))$ слагаемое $L(q,\Omega)$ является нижней оценкой на величину $\log p(X|\Omega)$. От максимизации $\log p(X|\Omega)$ по Ω предлагается перейти к максимизации нижней границы $L(q,\Omega)$ по q и Ω. Такая постановка в общем случае может давать приближеный ответ, однако оказывается существенно более простой. Основная идея EM-алгоритма заключается в том, чтобы итеративно повторять два шага:

1. $L(q,\Omega) \to \max\limits_{q}$
2. $L(q,\Omega) \to \max\limits_{\Omega}$.

Максимизация $L(q, \Omega)$ по q эквивалентна минимизации $KL(q(Z)||p(Z|X, \Omega))$, т.к. их сумма $\log p(X|\Omega)$ от q не зависит. Из свойств дивергенции Кульбака-Лейблера следует, что минимум, то есть 0, достигается при $q(Z) = p(Z|X, \Omega)$. Поэтому если нам удается выписать аналитически распределение $p(Z|X, \Omega)$, то именно его и нужно взять в качестве q, при этом нижняя оценка $L(q, \Omega)$ будет являться точной нижней оценкой.

Рассматрим теперь второй шаг:

$$\int q(Z) \log p(X, Z|\Omega) dZ - \int q(Z) \log q(Z) dZ \to \max_{\Omega} \Leftrightarrow \int q(Z) \log p(X, Z|\Omega) dZ \to \max_{\Omega},$$

т.к.второе слагаемое не зависит от Ω. В первом слагаемом узнаем формулу мат. ожидания:

$$\int q(Z) \log p(X, Z|\Omega) dZ = \mathbb{E}_{q(Z)} \log p(X, Z|\Omega)$$

Таким образом, EM-алгоритм заключается в чередовании двух шагов. E (Expectation) соответсвует подготовке к вычислению мат. ожидания; M (Maximization) – максимизации мат. ожидания логарифма правдоподобия по параметрам.

- **E-step:** $KL(q(Z)||p(Z|X, \Omega)) \to \min_{q(Z)} \Leftrightarrow q(Z) = p(Z|X, \Omega)$
- **M-step:** $\mathbb{E}_{q(Z)} \log p(X, Z|\Omega) \to \max_{\Omega}$

EM-алгоритм максимизации неполного правдоподобия в модели PLSA

Решим задачу (6), действуя согласно общей схеме. На E-шаге необходимо оценить распределение на скрытые переменные при условии наблюдаемых и параметров: $p(Z|X, \Phi, \Theta)$. Т.к. словопозиции независимы, то сразу перейдем к отдельным вероятностям:

$$p(Z|X, \Phi, \Theta) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(z_{di}|x_{di}, \Phi, \Theta)$$

Чтобы найти эти вероятности, воспользуемся формулой Байеса:

$$p(z_{di}|x_{di}, \Phi, \Theta) = \frac{p(x_{di}|z_{di}, \Phi, \Theta) p(z_{di}|\Phi, \Theta)}{\sum_{t=1}^{T} p(x_{di}|t, \Phi, \Theta) p(t|\Phi, \Theta)} = \frac{\phi_{x_{di} z_{di}} \theta_{z_{di} d}}{\sum_{t=1}^{T} \phi_{wt} \theta_{td}} \tag{8}$$

Теперь запишем выражение, которое нужно максимизировать на M-шаге:

$$\mathbb{E}_{p(Z|X,\Phi,\Theta)} p(X, Z|\Phi, \Theta) = \sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{E}_{p(z_{di}|x_{di},\Phi,\Theta)} \left(\log \phi_{x_{di} z_{di}} + \log \theta_{z_{di} d} \right) \to \max_{\Phi,\Theta}$$

Мы пронесли мат. ожидание внутрь суммы в силу независимости словопозиций. Теперь распишем мат. ожидание по определнию:

$$\sum_{d=1}^{D} \sum_{i=1}^{N_d} \sum_{t=1}^{T} p(z_{di} = t|x_{di}, \Phi, \Theta) \left(\log \phi_{x_{di} t} + \log \theta_{td} \right) \to \max_{\Phi,\Theta} \tag{9}$$

Эта задача очень похожа на задачу (3), которую мы уже решали, когда считали темы Z известными. Если записать функцию Лагранжа для учета ограничений нормировки и взять производную по одному элементу ϕ_{wt}, то получим:

$$\frac{1}{\phi_{wt}} \sum_{d=1}^{D} \sum_{i=1}^{N_d} [x_{di} = w] p(z_{di} = t) - \lambda_t = 0$$

Здесь $\sum_{d=1}^{D} \sum_{i=1}^{N_d} [x_{di} = w] p(z_{di} = t)$ аналогично варыжению в (5) интерпретируется как число раз, когда слово w было отнесено к теме t. Однако если раньше мы это число знали точно, то теперь вместо индикаторов тем появляются вероятности, посчитанные на E-шаге. Таким образом, это наилучшая оценка интересующей величины, которой мы располагаем на текущий момент. Обозначим ее как и прежде за n_{wt}, тогда итоговые формулы M-шага будут иметь знакомый вид:

$$\phi_{wt} = \frac{n_{wt}}{\sum_{w=1}^{W} n_{wt}}; \quad \theta_{td} = \frac{n_{td}}{\sum_{t=1}^{T} n_{td}} \tag{10}$$

Итак, итерационно повторяя формулы (8) и (10), мы оценим параметры Φ и Θ, т.е. обучим модель PLSA с помощью EM-алгоритма.

Общий подход: введение априорных распределения

Вернемся к примеру с волшебной монеткой. Согласно вероятностной модели, которую мы рассматривали, вероятность орла на любом броске была равна параметру θ. Этот параметр мог быть любым. До наблюдения результатов бросков у нас не было никаких предпочтений. Но предположим теперь, что мы все же что-то знаем о монетке заранее. Например, она на вид кривая, и наверняка орлом выпадает гораздо чаще, чем решкой.

Способ учесть такого рода знания в вероятностной модели – это ввести *априорное* распределение на параметр. То есть сказать, что параметр θ сгенерирован из определенного вероятностного распределения $p(\theta|\alpha)$. Здесь α – это параметр только что введенного распределения, его также называют *гиперпараметром* модели.

В новой модели можно построить следующую логику рассуждений, очень часто используемую в байесовском подходе. Априорное распределени $p(\theta|\alpha)$ описывает наши исходные знания о монетке. Потом к нам приходят какие-то данные (наблюдаемые переменные), в нашем случае, результаты бросков b. Это новые знания, которые, безусловно, могут уточнить наши представления о значении параметра θ. И уточняют:

$$\underbrace{p(\theta|X,\alpha)}_{\text{posterior}} = \frac{p(\theta, X|\alpha)}{p(X|\alpha)} \propto p(\theta, X|\alpha) = \underbrace{p(X|\theta,\alpha)}_{\text{likelihood}} \underbrace{p(\theta|\alpha)}_{\text{prior}} \tag{11}$$

Таким образом, определенное значение параметра θ имеет большую *апостериорную* вероятность, если оно одновременно хорошо вписывается в наши представления о полете кривой монетки (prior) и хорошо описывает результаты реальных бросков (likelihood, правдоподобие).

Максимизация апостериорной вероятности – один из возможных способов оценивания параметров модели, на которые заданы априорные распределения.

Далее нам потребуется распределение Дирихле

Рассмотрим вектор $\theta = (\theta_1, \dots \theta_K)$, такой что $\theta_k \geq 0$, $\forall k = 1, \dots K$ и $\sum_{k=1}^{K} \theta_k = 1$. То есть это вектор, задающий вероятности K возможных исходов некоторой дискретной случайной величины. Распределение Дирихле – это непрерывное вероятностное распределение на пространстве таких векторов θ:

$$Dir(\theta|\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \quad \alpha_k > 0, \forall k = 1, \dots K,$$

где $\alpha = (\alpha_1, \dots \alpha_K)$ – параметры. Важное свойство распределения Дирихле заключается в том, что если $\alpha_k < 1$, $\forall k = 1, \dots K$, то наиболее вероятными будут *разреженные* вектора θ, в которых лишь несколько значений существенно отличны от нуля. При этом заметим, что остальные значения не будут нулевым, а будут положительными, близкими у нулю величинами. Также нам понадобятся еще несколько свойств:

1. Математическое ожидание: $\mathbb{E}\theta_k = \frac{\alpha_k}{\sum_{i=1}^{K} \alpha_i}$
2. Мода (точка максимума вероятности): $\hat{\theta}_k^{MP} = \frac{\alpha_k - 1}{\sum_{i=1}^{K} \alpha_i - K}$
3. Математическое ожидание логарифма: $\mathbb{E} \ln \theta_k = \psi(\alpha_k) - \psi(\sum_{i=1}^{K} \alpha_i)$,
   где $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ – дигамма-функция.
   При $x > 1$ справедливо приближение: $\exp(\psi(x)) \approx x - \frac{1}{2}$.

Вероятностная тематическая модель LDA

Вероятностная тематическая модель LDA (Latent Dirichlet Allocation [2]) отличается от PLSA введением *априорных распределений Дирихле* на параметры Φ и Θ. Предполагается, что распределение на словах для каждой темы ϕ_t имеет априорное распределение Дирихле с вектор-праметром β. И аналогично, распределение на темах для каждого документа θ_d имеет априорное распределение Дирихле с вектор-параметром α:

$$\phi_t \sim Dir(\phi_t|\beta), \forall t = 1 \dots T; \quad \theta_d \sim Dir(\theta_d|\alpha), \forall d = 1, \dots D$$

Одна из основных мотиваций такой модели – разреживающее свойство распределения Дирихле. Таким образом, это попытка учесть то, что в реальности каждый документ содержит лишь небольшое число тем, а каждая тема описывается лишь небольшим числом слов. Запишем совместную вероятность, задающую модель LDA:

$$p(X, Z, \Phi, \Theta|\alpha, \beta) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} \phi_{x_{di} z_{di}} \theta_{z_{di} d} \prod_{d=1}^{D} Dir(\theta_d|\alpha) \prod_{t=1}^{T} Dir(\phi_t|\beta) \tag{12}$$

Если сравнить это выражение с аналогичным для PLSA (2), то заметим, что наборы переменных Φ и Θ переместились налево от черты, т.е. теперь мы оцениваем их совместную вероятность с наблюдаемыми переменными X и скрытыми Z. Это означает, например, то, что теперь мы можем зафиксировать некоторые α и β и сгенерировать согласно (12) X, Z, Φ и Θ, раньше Φ и Θ мы сгенерировать никак не могли. Еще это означает, что мы можем относиться теперь к Φ и Θ как к скрытым переменным (наравне с Z), а α и β трактовать как параметры модели.

За счет усложения модели, в LDA появляется несколько различных сценариев оценивания интересующих нас матриц Φ и Θ. В PLSA практически единственным разумным сценарием была максимизация неполного правдоподобия. Далее мы рассмотрим три таких сценария:

- Максимизация апостериорной вероятности (MAP: maximum a posteriori probability)
- Вариационный байесовский вывод (VB: Variational Bayes)
- Сэмплирование Гиббса (CGS: Collapsed Gibbs Sampling)

Они строятся на существенно разной логике, но приводят к близким оценкам. В литературе также используются их различные гибриды и модификации.

Метод максимума апостериорной вероятности для модели LDA

Этот метод является наиболее близким аналогом максимизации неполного правдоподобия в модели PLSA. Согласно нему, нужно взять апостериорную вероятность параметров Φ и Θ и найти значения, в которых достигается максимум. Апостериорную вероятность мы уже умеем расписывать через правдоподобие модели и априорную вероятность. Переписывая (11) для модели LDA, получаем задачу:

$$p(\Phi, \Theta|X, \alpha, \beta) \propto p(X|\Phi, \Theta, \alpha, \beta)p(\Theta|\alpha)p(\Phi|\beta) \to \max_{\Phi,\Theta} \tag{13}$$

Эта задача как две капли воды похожа на задачу максимизации неполного правдоподобия $p(X|\Phi, \Theta) \to \max\limits_{\Phi,\Theta}$, которую мы уже научились решать с помощью EM-алгоритма. Все отличие заключается в дополнительных сомножителях $p(\Theta|\alpha)$ и $p(\Phi|\beta)$.

Вернемся к общей схеме EM-алгоритма и посмотрим, на что повлияет это отличие:

$$\log p(X|\Omega) + \log p(\Omega) = L(q, \Omega) + KL(q(Z)||p(Z|X, \Omega)) + \log p(\Omega),$$

где $p(\Omega)$ – априорное распределение. При максимизации $L(q, \Omega) + \log p(\Omega)$ по q и Ω:

- **E-шаг** остается без изменения, т.к. добавочный член не зависит от q:
  $KL(q(Z)||p(Z|X, \Omega)) \to \min\limits_{q(Z)} \Leftrightarrow q(Z) = p(Z|X, \Omega)$
- **M-шаг** меняется естественным образом: $\mathbb{E}_{q(Z)} \log p(X, Z|\Omega) + \log p(\Omega) \to \max\limits_{\Omega}$

Таким образом, чтобы записать алгоритм LDA-MAP, нам необходимо максимизировать модифицированное выражение M-шага:

$$\sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{t=1}^{T} p(z_{di} = t|x_{di}, \Phi, \Theta)\left(\log \phi_{x_{di}t} + \log \theta_{td}\right) + \\ + \sum_{d=1}^{D}\sum_{t=1}^{T}(\alpha_t - 1)\log \theta_{td} + \sum_{t=1}^{T}\sum_{w=1}^{W}(\beta_w - 1)\log \phi_{wt} \to \max_{\Phi,\Theta} \tag{14}$$

при условиях неотрицательности и нормировки (4). Если записать функцию Лагранжа и взять производную по одному элементу ϕ_{wt}, то получим:

$$\frac{1}{\phi_{wt}}\left(\sum_{d=1}^{D}\sum_{i=1}^{N_d}[x_{di} = w]p(z_{di} = t) + \beta_w - 1\right) - \lambda_t = 0$$

Отсюда, аналогично уже разобранным случаям:

$$n_{wt} + \beta_w - 1 = \lambda_t \phi_{wt} \quad \Rightarrow \quad \sum_{w=1}^{W}(n_{wt} + \beta_w - 1) = \sum_{w=1}^{W} \lambda_t \phi_{wt}, \quad \Rightarrow \quad \sum_{w=1}^{W}(n_{wt} + \beta_w - 1) = \lambda_t$$

И искомая оценка:

$$\phi_{wt} = \frac{n_{wt} + \beta_w - 1}{\sum_{w=1}^{W}(n_{wt} + \beta_w - 1)} \tag{15}$$

Заметим, что в этот раз условия неотрицательности автоматически не выполняются, и возможна ситуация, когда $n_{wt} + \beta_w - 1 < 0$. Таким образом, формула (15) корректна для значений $\beta_w > 1$. Можно показать, что в случае произвольных значений параметров необходимо заменить отрицательные значения на нули, и итоговые формулы М-шага будут иметь вид:

$$\phi_{wt} = \frac{(n_{wt} + \beta_w - 1)_+}{\sum_{w=1}^{W}(n_{wt} + \beta_w - 1)_+}; \quad \theta_{td} = \frac{(n_{td} + \alpha_t - 1)_+}{\sum_{t=1}^{T}(n_{td} + \alpha_t - 1)_+} \tag{16}$$

В результате, при маленьких значениях счетчиков n_{wt} или n_{td} соответсвующие вероятности обнулятся, таким образом, сработает разреживающее свойство априорного распределения Дирихле. Тем не менее, на практике этих обнулений может оказаться недостаточно.

Вариационный байесовский вывод для модели LDA

Приведем другой подход к оцениванию матриц Φ и Θ в модели LDA, заданной совместным вероятностным распределением $p(X, Z, \Phi, \Theta|\alpha, \beta)$. Поставим задачу максимизации неполного правдоподобия, т.е. максимизации вероятности наблюдаемых данных при условии параметров модели. Напомним, что в PLSA это было $p(X|\Phi, \Theta) \to \max\limits_{\Phi,\Theta}$. Здесь же это будет $p(X|\alpha, \beta) \to \max\limits_{\alpha,\beta}$. Казалось бы, постановка странная, т.к. интересующие нас матрицы Φ и Θ пропали, а максимизация осуществляется по гиперпараметрам. Но попробуем расписать ЕМ-алгоритм для этой задачи:

- Е-шаг: $KL(q(Z, \Phi, \Theta)||p(Z, \Phi, \Theta|X, \alpha, \beta)) \to \min\limits_{q}$
- М-шаг: $\mathbb{E}_q \, p(X, Z, \Phi, \Theta|\alpha, \beta) \to \max\limits_{\alpha,\beta}$

Чтобы получить нулевую дивергенцию Кульбака-Лейблера на Е-шаге, необходимо вычислить совместное распределение ни три группы скрытых переменных: $p(Z, \Phi, \Theta|X, \alpha, \beta)$. Это сложнее, чем вычислить распределение на темах $p(Z|X, \Phi, \Theta)$, которое нам требовалось на Е-шаге обучения PLSA. К сожалению, аналитически выписать $p(Z, \Phi, \Theta|X, \alpha, \beta)$ не удается. Что в таком случае мы можем сделать?

Мы можем ввести дополнительные упрощающие предположения и в этих предположениях приближенно оценить искомое распределение. А именно, предположим, что группы переменных взаимно независимы (в реальности это конечно же не так) и будем искать распределение q в виде:

$$q(Z, \Phi, \Theta) \approx \prod_{d=1}^{D}\prod_{i=1}^{N_d} q(z_{di}) \prod_{t=1}^{T} q(\phi_t) \prod_{d=1}^{D} q(\theta_d) \tag{17}$$

Это можно интерпретировать так: при минимизации KL-дивергенции по q мы сужаем множество допустимых q и ищем не среди всех возможных распределений, а только среди тех, которые мы умеем аналитически находить, в данном случае, только среди распределений вида (17). При этом мы не сможем найти такое q, что KL-дивергенция обнулится, но все же сможем как-то ее минимизировать. При этом нижняя оценка $L(q, \alpha, \beta)$ будет неточной оценкой на $\log p(X|\alpha, \beta)$, а итоговые оценки алгоритма – приближенными. Этот прием называется *приближенным байесовским выводом.*

Итак, теперь задача формулируется так:

$$KL\left(\prod_{d=1}^{D}\prod_{i=1}^{N_d} q(z_{di})\prod_{t=1}^{T} q(\phi_t)\prod_{d=1}^{D} q(\theta_d) \,||\, p(Z, \Phi, \Theta|X, \alpha, \beta)\right) \to \min_{q(z_{di}),q(\phi_t),q(\theta_d)} \tag{18}$$

Она решается итерационным процессом, который на каждом шаге оценивает очередной фактор q_j по всем остальным факторам $q_{\setminus j}$:

$$\log q_j \propto \mathbb{E}_{q_{\setminus j}} \log p(X, Z, \Phi, \Theta|\alpha, \beta) \tag{19}$$

Доказательство этого утверждения можно найти, например, в [5].

Распишем формулы байесовского вывода для оценки распределений на Z, Φ и Θ. Начнем с $q(\theta_d)$ для некоторого документа d. В последующих выкладках нас будет интересовать только зависимость от θ_d, все остальные члены смело опускаются. Они повлияют только на нормировочную константу, которую мы найдем из других соображений.

$$\log q(\theta_d) \propto \mathbb{E}_{\setminus q(\theta_d)} \log p(X, Z, \Phi, \Theta|\alpha, \beta) \propto \mathbb{E}_{\setminus q(\theta_d)} \sum_{d=1}^{D}\sum_{i=1}^{N_d} (\log \phi_{x_{di} z_{di}} + \log \theta_{z_{di} d}) +$$
$$+ \sum_{d=1}^{D} \log Dir(\theta_d|\alpha) + \sum_{t=1}^{T} \log Dir(\phi_t|\beta) \propto \mathbb{E}_{\setminus q(\theta_d)} \sum_{i=1}^{N_d} \log \theta_{z_{di} d} + \log Dir(\theta_d|\alpha)$$

Теперь раскрываем мат. ожидание, причем если выражение не зависит от каких-то величин, то и брать мат. ожидание по распределениям над ними нет смысла (например, по $q(\phi_t)$). И расписываем плотность распределения Дирихле, опуская его нормировочную константу.

$$\log q(\theta_d) \propto \sum_{i=1}^{N_d} \mathbb{E}_{q(z_{di})} \log \theta_{z_{di} d} + \log Dir(\theta_d|\alpha) \propto$$
$$\sum_{i=1}^{N_d}\sum_{t=1}^{T} q(z_{di} = t) \log \theta_{td} + \sum_{t=1}^{T} (\alpha_t - 1) \log \theta_{td} \propto \sum_{t=1}^{T} \left(\sum_{i=1}^{N_d} q(z_{di} = t) + \alpha_t - 1\right) \log \theta_{td}$$

Посмотрим внимательно на это выражение, как функцию от θ_{td}. Можно заметить, что с точностью до константы оно совпадает с распределением Дирихле с параметрами $\sum_{i=1}^{N_d} q(z_{di} = t) + \alpha_t$. Значит, это оно и есть!

$$\theta_d \sim Dir(\theta_d|\gamma); \quad \gamma_t = \sum_{i=1}^{N_d} q(z_{di} = t) + \alpha_t \tag{20}$$

То, что мы так удачно получили зависимость от θ_{td}, соответствующую известному распределению, изабвило нас от необходимости аналитически вычислять нормировочную константу. Более того, подсчитать ее честно и взять все необходимые для этого интегралы у нас, скорее всего, и не получилось бы. Такая удача, конечно, не случайна, автор модели LDA догадывался о ней :) Отчасти именно этим мотивирован выбор Дирихле в качестве априорного распределения.

Если апостериорное распределение лежит в том же семействе, что и априорное распределение, то *говорят, что априорное распределение является сопряженным* к функции правдоподобия. Вспоминая представление апостериорного распределения через функцию правдоподобия и априорное распределение $p(\Theta|X,\alpha,\beta) \propto p(X|\alpha,\beta)p(\Theta|\alpha)$ можно перефразировать определение так: априорное распределение образует сопряженную пару с функцией правдоподобия, если при их перемножении получается распределение из того же семейства. В нашем случае правдоподобие является мультиномиальным распределением. Распределение Дирихле является сопряженным к мультиномиальному распределению.

Остановимся теперь подробнее на результате (20), который мы получили. Это оценка апостериорного распределения на θ_d. В отличие от предыдущего метода (максимума апостериорной вероятности) мы получили распределение целиком, а не одну точку, соответсвующую максимальной плотности вероятности. С одной стороны, это хорошо, у нас появилось больше информации, и мы вольны распоряжаться ею согласно своим целям. С другой стороны, в большинстве случаев нас по-прежнему интересует одно конкретное значение θ_d, т.е. точечная оценка. Чтобы получить ее, нужно взять какую-либо статистику распределения (20), например, мат. ожидание. Вспоминая формулу мат. ожидания для распределения Дирихле, получим:

$$\mathbb{E}\,\theta_{td} = \frac{\gamma_t}{\sum_{t=1}^{T}\gamma_t}$$

И следуя введенным ранее обозначениям,

$$\mathbb{E}\,\theta_{td} = \frac{\sum_{i=1}^{N_d} q(z_{di}=t) + \alpha_t}{\sum_t \sum_{i=1}^{N_d} q(z_{di}=t) + \alpha_t} = \frac{n_{td}+\alpha_t}{\sum_{t=1}^{T}(n_{td}+\alpha_t)}$$

Вместо мат. ожидания можно взять моду, т.е. точку с максимальной плотностью вероятности. Тогда

$$\hat{\theta}_{td}^{MP} = \frac{\gamma_t - 1}{\sum_{t=1}^{T}(\gamma_t - 1)} = \frac{n_{td}+\alpha_t - 1}{\sum_{t=1}^{T}(n_{td}+\alpha_t - 1)}$$

Такая оценка в точности соответствует оценкам (16), полученными другим методом, но также из соображений максимизации апостериорного распределения на θ_d.

Однако взятие точечных оценок остается за рамками байесовского вывода, который заключается в итерационном пересчете всех q_j с помощью текущих значений остальных $q_{\setminus j}$ согласно (19). Нетрудно показать, что $q(\phi_t)$ вычисляются аналогичным образом:

$$\phi_t \sim Dir(\phi_t|\lambda); \quad \lambda_w = \sum_{d=1}^{D}\sum_{i=1}^{N_d}[x_{di}=w]q(z_{di}=t) + \beta_w \tag{21}$$

Остается вывести формулы для $q(z_{di})$.

$$\log q(z_{di}=t|x_{di}=w) \propto \mathbb{E}_{\setminus q(z_{di})} \log p(X,Z,\Phi,\Theta|\alpha,\beta) \propto \mathbb{E}_{q(\phi_t)} \log \phi_{wt} + \mathbb{E}_{q(\theta_d)} \log \theta_{td} \propto$$
$$\propto \psi(n_{wt}+\beta_w) - \psi\left(\sum_{w=1}^{W}(n_{wt}+\beta_w)\right) + \psi(n_{td}+\alpha_t) - \psi\left(\sum_{t=1}^{T}(n_{td}+\alpha_t)\right) \quad (22)$$

Чтобы проинтерпретировать полученный результат, перейдем от логарифма к вероятности и воспользуемся приближением для экспонент дигамма-функций:

$$q(z_{di}=t|x_{di}=w) \propto \frac{n_{wt}+\beta_w-0.5}{\sum_{w=1}^{W}(n_{wt}+\beta_w)-0.5} \frac{n_{td}+\alpha_t-0.5}{\sum_{t=1}^{T}(n_{td}+\alpha_t)-0.5} \approx \phi_{wt}\,\theta_{td}$$

Таким образом, приближенно выражение (22) соответствует формулам пересчета распределения на темы в PLSA-ML и LDA-MAP.

Итак, мы закончили с байесовским выводом, т.е. с приближением распределения на скрытые переменные на E-шаге EM-алгоритма. M-шаг заключается в нахождени оценок максимума правдоподобия для гиперпараметров α и β. На практике он часто опускается, а α и β фиксируются заранее. Рекомендации по оптимизации гиперпараметров можно найти в [4].

Сэмплирование Гиббса для модели LDA

Collapsed Gibbs Sampling – еще один часто используемый алгоритм обучения модели LDA. Он приближенно оценивает распределение на темах $p(Z|X,\alpha,\beta)$, забывая про матрицы Φ и Θ, а потом вычисляет их одним действием, снова из соображений максимизации правдоподобия.

Первый шаг (collapsing) заключается в том, чтобы выинтегрировать из совместного распределения Φ и Θ:

$$p(Z|X,\alpha,\beta) = \int p(X,Z,\Phi,\Theta|\alpha,\beta)d\Phi d\Theta \quad (23)$$

Это удается сделать аналитически за счет того, что интеграл распадается на произведение двух интегралов – один по Φ, другой по Θ:

$$p(X,Z|\alpha,\beta) = \underbrace{\int \prod_{d=1}^{D}\prod_{i=1}^{N_d}\theta_{z_{di}d}\prod_{d=1}^{D}Dir(\theta_d|\alpha)d\Theta}_{I_1}\underbrace{\int \prod_{d=1}^{D}\prod_{i=1}^{N_d}\phi_{x_{di}z_{di}}\prod_{t=1}^{T}Dir(\phi_t|\beta)d\Phi}_{I_2} \quad (24)$$

Распишем интеграл по Θ, по Φ все будет аналогично.

$$I_1 = \prod_{d=1}^{D}\int\prod_{i=1}^{N_d}\prod_{t=1}^{T}\theta_{td}^{[z_{di}=t]}Dir(\theta_d|\alpha)d\theta_d = \left(\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\right)^D\prod_{d=1}^{D}\int\prod_{t=1}^{T}\theta_{td}^{\sum_{i=1}^{N_d}[z_{di}=t]+\alpha_t-1}d\theta_d$$

Под знаком каждого интеграла стоит распределение Дирихле с параметрами $\tilde{\alpha}_t = \sum_{i=1}^{N_d}[z_{di}=t]+\alpha_t$, $t=1,\dots T$ без учета нормировочной константы. Следовательно, окончательно получаем:

$$I_1 = \left(\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\right)^D\prod_{d=1}^{D}\frac{\prod_{t=1}^{T}\Gamma(\sum_{i=1}^{N_d}[z_{di}=t]+\alpha_t)}{\Gamma(\sum_{t=1}^{T}\sum_{i=1}^{N_d}[z_{di}=t]+\alpha_t)} \quad (25)$$

Теперь у нас есть распределение $p(X, Z|\alpha, \beta)$. Чтобы получить из него распределение на темах, нужно воспользоватся формулой Байеса:

$$p(Z|X, \alpha, \beta) = \frac{p(X, Z|\alpha, \beta)}{\sum\limits_{Z} p(X, Z|\alpha, \beta)}$$

В знаменателе стоит сумма по всем наборам Z, которую невозможно взять аналитически. Поэтому возникает задача приближенного оценивания этого распределения. Например, если бы у нас была выборка, сгенерированная из этого распределения, мы могли бы почитать по ней эмпирические оценки. Но как сгенерировать такую выборку?

Сэмплирование Гиббса принадлежит семейству методов Монте-Карло на марковских цепях (MCMC – Markov Chain Monte Carlo) и предлагает эффективный способ генерации выборки из многомерного распределения, известного с точностью до нормировочной константы. Согласно нему, нужно запустить итерационный процесс, который на каждом шаге генерирует точку z_{di} из одномерного распределения $p(z_{di}|X, Z_{\setminus(d,i)}, \alpha, \beta)$, где $Z_{\setminus(d,i)}$ – текущие темы всех словопозиций кроме (d, i). Тогда через некоторое время процесс сойдется к генерации сэмплов из искомого распределения $p(Z|X, \alpha, \beta)$. Доказательство этого факта можно найти, например, в [6].

Чтобы выписать распределение $p(z_{cj}|X, Z_{\setminus(c,j)}, \alpha, \beta)$ для фиксированного документа c и позиции j, нужно взять уже известное нам совместное распределение $p(X, Z|\alpha, \beta) = I_1 I_2$, оставить только члены, зависящие от z_{cj}, и отнормировать получившуюся зависимость. Т.к. это одномерное дискретное распределение, то в данном случае нормировка не является проблемой. Снова сократим выкладки вдвое и будем расписывать только сомножитель I_1. Выделим из всех сумм отдельно слагаемые, содержащие z_{cj}:

$$I_1 = Const \frac{\prod\limits_{t=1}^{T} \Gamma\left(\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = t] + [z_{cj} = t] + \alpha_t\right)}{\Gamma\left(\sum\limits_{t=1}^{T}\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = t] + \alpha_t + 1\right)}$$

В знаменателе мы воспользовались очевидным тождеством $\sum_{t=1}^{T} [z_{cj} = t] = 1$. Посмотрим на числитель. Условие $[z_{cj} = t]$ при всех темах $t \neq z_{cj}$ равно 0, и его можно не писать. При $t = z_{cj}$ воспользуемся свойством гамма-функции $\Gamma(x + 1) = x\Gamma(x)$.

$$I_1 = Const \frac{\left(\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = z_{cj}] + \alpha_t\right) \prod\limits_{t=1}^{T} \Gamma\left(\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = t] + \alpha_t\right)}{\left(\sum\limits_{t=1}^{T}\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = t] + \alpha_t\right) \Gamma\left(\sum\limits_{t=1}^{T}\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = z_{cj}] + \alpha_t\right)}$$

Здесь все оставшиеся гамма-функции не зависят от z_{cj}, поэтому их можно опустить в счет нормировочной константы. Проделывая то же самое для I_2 в итоге получаем:

$$p(z_{cj}|X, Z_{\setminus(c,j)}, \alpha, \beta) \propto \frac{\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = z_{cj}] + \alpha_t}{\sum\limits_{t=1}^{T}\left(\sum\limits_{i=1, i\neq j}^{N_c} [z_{ci} = t] + \alpha_t\right)} \frac{\sum\limits_{(d,i)\neq(c,j)} [x_{di} = x_{cj}][z_{di} = z_{cj}] + \beta_w}{\sum\limits_{w=1}^{W}\left(\sum\limits_{(d,i)\neq(c,j)} [x_{di} = w][z_{di} = z_{cj}] + \beta_w\right)}$$

Громоздкие суммы индикаторов – это счетчики, которые имеют привычный смысл, но имеют одну особенность: в них исключается позиция (c, j), для которой на данном шаге рассчитываются вероятности тем.

Итак, предположим мы сделали S сэмплов из многомерного ненормированного распределения на Z. Тогда оценить распределения ϕ_{wt} и θ_{td} можно по формулам:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{\sum_{w=1}^{W}(n_{wt} + \beta_w)} \quad \theta_{td} = \frac{n_{td} + \alpha_t}{\sum_{t=1}^{T}(n_{td} + \alpha_t)},$$

где

$$n_{wt} = \sum_{d=1}^{D}\sum_{i=1}^{N_d}[x_{di} = w]\sum_{s=1}^{S}\frac{1}{S}[z_{di}^s = t], \quad n_{td} = \sum_{i=1}^{N_d}\sum_{s=1}^{S}\frac{1}{S}[z_{di}^s = t].$$

Эти оценки можно трактовать как промежуточный вариант между случаем, когда мы знали темы каждой словопозиции точно (в самом начале лекции), и случаем, когда мы знали аналитическое распределение на темы (например, на M-шаге обучения PLSA). Здесь мы знаем S сэмплов и усредняем их, чтобы подсчитать оценку числа сопоставлений слова w теме t или темы t документу d. На практике для больших колекций часто достаточно $S = 1$.

Список литературы:

1. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, pp 50–57

2. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022

3. Griffiths T, Steyvers M (2004) Finding Scientific Topics. Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235.

4. McCallum A, Mimno DM, Wallach HM (2009) Rethinking LDA: Why Priors Matter. In: Advances in Neural Information Processing Systems 22, pp 1973–1981.

5. http://www.machinelearning.ru/wiki/images/5/57/BMMO11_9.pdf – конспект по байесовскому выводу из курса БММО.

6. http://www.machinelearning.ru/wiki/images/6/6b/BMMO11_10.pdf – конспект по методам MCMC из курса БММО.

7. http://www.machinelearning.ru/wiki/images/8/82/BMMO11_14.pdf – конспект по модели LDA из курса БММО.