Тематическое моделирование

Санкт-Петербургский государственный университет
Кафедра статистического моделирования
Семинар «Статистическое и машинное обучение»

Санкт-Петербург, 2025

Понятие тематического моделирования

Тематическое моделирование (topic modeling) — это способ построения модели коллекции текстовых документов (или **тематической модели, topic model**), которая определяет, к каким темам относится каждый из документов и какие термины (слова или словосочетания) определяют каждую тему.

Вероятностная тематическая модель (BTM) описывает каждую тему дискретным распределением на множестве терминов, а каждый документ — дискретным распределением на множестве тем.

Приложения тематического моделирования

Тематические модели применяются в решении следующих задач автоматического анализа текстов:

- Кластеризация, классификация и ранжирование текстов;
- Суммаризация и аннотация текстов;
- Тематический поиск документов;
- Фильтрация спама;
- Реализация рекомендательных систем.

Тематическое моделирование может рассматриваться как задача одновременной «мягкой» кластеризации документов по темам и слов по темам.

Возникновение LSA

Двигаясь от простого к сложному, первые методы информационного поиска сперва работали на принципе точного совпадения запроса и текста в документе. Однако такие методы сразу показали свою несостоятельность ввиду проблем синонимии и полисемии.

Решить поставленную задачу предлагалось методом **Латентно-семантического анализа (LSA)** или **Латентно-семантического индексирования (LSI)**.

Понятие LSA

Задача LSA состоит в том, чтобы спроецировать часто встречающиеся вместе термины в одно и то же измерение семантического пространства, которое имеет пониженную размерность по сравнению с оригинальной **терм-документной матрицей**, которая обычно довольно разрежена.

Элементы этой матрицы содержат веса терминов в документах, назначенные с помощью выбранной **весовой функции**. Простейшая весовая функция равна 1, если термин встретился в документе, и 0 — если не встретился.

Весовые функции в LSA

Пусть t — термин, T — множество всех терминов, d — документ, D — множество всех документов в коллекции. Классическим методом назначения весов словам является TF-IDF:

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D) \tag{1}$$

где TF (term frequency) — нормализованная частота термина в документе:

$$TF(t, d) = \frac{freq(t, d)}{\max_{w \in T} freq(w, d)} \tag{2}$$

а IDF (inverse document frequency) — обратная частота документов, включающих термин:

$$IDF(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{3}$$

Весовые функции в LSA

В дальнейшем появлялись различные модификации TF-IDF, которых теперь существует несколько десятков. Одна из популярных модификаций выглядит следующим образом:

$$TF'(t,d) = \frac{TF(t,d)}{TF(t,d) + 0,5 + 1,5\frac{len(d)}{\frac{1}{N}\sum_{d \in D} len(d)}} \tag{4}$$

$$IDF'(t,D) = \frac{\log(IDF(t,D))}{\log(|D|+1)} \tag{5}$$

LSA и сингулярное разложение

Наиболее распространенный вариант LSA основан на использовании сингулярного разложения терм-документной матрицы. Как известно, любая вещественная прямоугольная матрица может быть разложена на произведение трех матриц:

$$A = TSD^T \tag{6}$$

где матрицы T и D — ортогональные, а S — диагональная.

Оставив первые k синуглярных значений в матрице S и убрав соответсвующие столбцы матриц T и D, мы получим наилучшее приближение $\hat{A}$ терм-документной матрицы A.

Параметр k зависит от задачи и подбирается эмпирически.

LSA и сингулярное разложение

Таким образом, каждый термин или документ представляется при помощи векторов в общем пространстве размерности k (так называемом **пространстве гипотез**).

Тогда с помощью скалярного произведения можно решить следующие задачи:

- Сравнить два термина между собой в контексте их использования в данной коллекции документов.
- Сравнить два документа между собой.
- Определить степень отношения термина к документу.

Достойнства и недостатки LSA

Ключевое достоинство метода по сравнению с поиском точных соотвествий состоит в том, что проблема синонимии и полисемии частично снимается.

Недостатки же у метода следующие:

- Вычислительная сложность порядка $(N_{doc} + N_{term})^k$.
- Модель предполагает, что порядок слов в документе не важен.
- Модель не учитывает реальное вероятностное распределение слов в документах и работает так, как будто это распределение является нормальным.

Последняя проблема решается введением **вероятностного латентно-семантического анализа (pLSA)**.

Понятие pLSA

В основе pLSA лежит **аспектная модель**, которая связывает скрытые переменные тем с каждой наблюдаемой переменной — термином или документом. Таким образом, каждый документ может относиться к нескольким темам с некоторой вероятностью, что является выгодной отличительной особенностью этой модели.

Модель по прежнему опирается на так называемую **гипотезу «мешка слов»**, согласно которой порядок терминов в документе не важен для определения темы документа.

Постановка задачи pLSA

Итак, у нас есть термины $t \in T$ и документы $d \in D$, а также некоторые неизвестные заранее темы $z \in Z$.

Необходимо найти множество тем Z, распределения $p(t|z)$ для всех $z \in Z$ и распределения $p(z|d)$ для всех $d \in D$.

Распределение $p(z|d)$ является при этом удобным признаковым описанием документа в задачах информационного поиска, классификации и категоризации документов.

Модель pLSA

Совместная вероятностная модель над документами и словами может быть определена асимметрично или симметрично.

Асимметричное представление модели:

$$P(d,t) = P(d) \sum_{z \in Z} P(t|z) P(z|d) \tag{7}$$

Симметричное представление модели:

$$P(d,t) = \sum_{z \in Z} P(z) P(t|z) P(d|z) \tag{8}$$

При этом термин-наблюдение из распредления $P(d,t)$ встречается в документе d с частотой $n(d,t)$.

Модель pLSA

Можем записать функцию правдоподобия для вероятности всей коллекции документов:

$$P(D) = \prod_{t,d} P(d,t)^{n(d,t)} \tag{9}$$

Подставив симметричное представление совместной вероятностной модели над документами и словами, а затем взяв логарифм, получим:

$$\mathcal{L} = \sum_{d,t} n(d,t) \log \prod_{d,t} P(z)P(t|z)P(d|z) \tag{10}$$

Для определения оптимальных значений скрытых параметров используется EM-алгоритм.

EM-алгоритм для pLSA

На E-шаге оценивается вероятность $P(z|d,t)$:

$$P(z|d,t) = \frac{P(z)P(d|z)P(t|z)}{\sum_{z' \in Z} P(z')P(d|z')P(t|z')} \tag{11}$$

На M-шаге вычисляются $P(t|z)$, $P(d|z)$ и $P(z)$:

$$P(t|z) = \frac{\sum_d n(d,t)P(z|d,t)}{\sum_{d,t'} n(d,t')P(z|d,t')} \tag{12}$$

$$P(d|z) = \frac{\sum_t n(d,t)P(z|d,t)}{\sum_{d',t} n(d',t)P(z|d',t)} \tag{13}$$

$$P(z) = \frac{\sum_{d,t} n(d,t')P(z|d,t')}{\sum_{d,t} n(d,t)} \tag{14}$$

Недостатки pLSA

Несмотря на очевидное преимущество, которое даёт введение скрытых параметров-тем, модель обладает определёнными недостатками:

- Модель содержит большое число параметров, которое растёт линейно от числа документов. Как следствие, склонность к переобучению на больших наборах данных.
- Невозможно как-либо работать с документами, которых не было в изначальном наборе данных.

Для решения обоих проблем было предложено **латентное размещение Дирихле (LDA)**.

Понятие LDA

LDA — это генеративная модель, которая моделирует процесс порождения текстов. Ключевая идея модели состоит в том, что каждый документ принадлежит смеси распределений тем, а каждая тема — это распределение терминов по словарю.

Хотя первоначальная задача LDA — генерация текста, модель даёт осмысленные числовые признаки для текстов из коллекции. То есть после обучения модели на коллекции документов, для каждого документа можно вычислить вектор длиной K, который описывает документ в пространстве тем.

Терминология и нотация LDA

Слово — это базовая единица словаря, имеющая индекс v от 1 до V. Слово представлено вектором w; w^i — i-ый компонент вектора. Если $i = v$, то $w^i = 1$, иначе $w^i = 0$.

Документ — это последовательность из N слов $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, где w_n — n-ое слово в последовательности.

Корпус — это коллекция из M документов $\mathbf{D} = (\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M})$.

K — количество скрытых параметров-тем, которое определяется заранее.

Алгоримт для LDA

LDA предполагает следующий генеративный процесс:

1. Выбрать N из $Poisson(\xi)$.
2. Выбрать параметр θ_d из $Dir(\alpha)$.
3. Выбрать параметр ϕ_k из $Dir(\beta)$.
4. Для каждого из N слов w_n:
   1. Выбрать тему z_n из $Multinimial(\theta_d)$.
   2. Выбрать слово w_n из $Multinimial(\phi_k)$.

Ключевая задача LDA — восстановить скрытые параметры θ_d и ϕ_k (обычно с помощью сэмплирования Гиббса или вариационного вывода).

Модель LDA

Полная вероятностная модель имеет следующий вид:

$$P(w, z, \theta, \phi|\alpha, \beta) = \prod_{k=1}^{K} P(\phi_k|\beta) \prod_{d=1}^{D} P(\theta_d|\alpha) \prod_{n=1}^{N_d} P(z_{d,n}|\theta_d) P(w_{d,n}|\phi_{z_{d,n}}) \tag{15}$$

Максимизируем же мы предельное правдоподобие наблюдаемых слов:

$$P(w|\alpha, \beta) = \int \int \prod_{d=1}^{D} (P(\theta_d|\alpha) \prod_{n=1}^{N_d} \sum_{z_{d,n}} P(z_{d,n}|\theta_d) P(w_{d,n}|\phi_{z_{d,n}})) P(\phi|\beta) d\theta d\phi \tag{16}$$

LDA для извлечения признаков

Зная вектора ϕ_k (распределение слов по темам) и θ_d (распределение тем по документам), мы получаем числовое представление каждого слова из словаря и каждого документа из корпуса.

Эти вектора затем используются в других моделях (для задач кластеризации или классификации текстов). Таким образом, прямая задача LDA на практике — только лишь извлечение признаков для других моделей.

Недостатки LDA

Хотя модель менее склонна к переобучению по сравнению с pLSA, а также способна работать с документами, которых не было в корпусе, она обладает и рядом недостатков:

- Всё ещё используется гипотеза «мешка слов».
- Число тем должно быть задано заранее.
- Сэмплирование Гиббса, которое используется в большинстве программных реализаций, обладает большой вычислительной сложностью.

Ряд модификаций LDA решает проблему определения числа тем, но не гипотезы «мешка слов».

Критерии качества тематической модели

В отличие от задач классификации или регрессии в оценке качества тематических моделей нет чёткого понятия «ошибки» или «потери».

Стандартные критерии качества кластеризации типа средних внутрикластерных или межкластерных расстояний или их отношений плохо подходят для оценивания «мягкой» совместной кластеризации документов и терминов.

Критерии качества тематической модели

Наиболее распространённым критерием является **перплексия (perplexity)**. Это мера «удивлённости» модели $p(t|d)$ терминам t, наблюдаемым в документах d коллекции D, определяемая через логарифм правдоподобия:

$$Perplexity(D;p) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{t\in d} n_{d,t}\ln p(t|d)\right).$$

Чем меньше эта величина, тем лучше модель p предсказывает появление терминов t в документах d коллекции D.

Ограничение перплексии как меры качества

Минимизация перплексии эквивалентна максимизации правдоподобия модели на тестовых данных, однако этот показатель всё равно имеет ряд существенных ограничений:

- Перплексия измеряет уверенность модели в своих прогнозах, но не их достоверность.
- Перплексия плохо подходит для прямого сравнения моделей с разной архитектурой, словарём или способом токенизации.
- Переобученная модель может показывать низкую перплексию.