Санкт-Петербургский государственный университет

Математико-механический факультет

Погорелов Петр Глебович

Лекционные материалы по теме «Вероятностные алгоритмы тематического моделирования документов: Probabilistic Latent Semantic Analysis и Latent Dirichlet Allocation»

Учебный курс «Теория Байесовский сетей»

Руководитель:
д. ф.-м. н., профессор Тулупьев А.Л.

Санкт-Петербург
2018

SAINT PETERSBURG STATE UNIVERSITY

Mathematics&Mechanics Faculty

Pogorelov Petr

Lecture notes on «Probabilistic algorithms of topic modelling: Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation»

Theory of Bayesian Networks

Scientific supervisor:
Ph.D., Professor , Alexander Tulupyev

Saint Petersburg
2018

Содержание

Введение

Цель настоящей работы – введение в модели тематического моделирования для слушателей курса «Теория байесовых сетей», которые раньше не сталкивались с данным классом методов. Для достижения этой цели были поставлены задачи:

1. сформулировать постановку задачи тематического моделирования,
2. дать описание алгоритма Probabilistic Latent Semantic Analysis, привести пример использования,
3. дать описание алгоритма Latent Dirichlet Allocation, привести пример использования.

Тематическое моделирование – одна из задач обработки естественных языков (NLP) без учителя, очень близкая к идее мягкой кластеризации. Большинство моделей тематического моделирования строятся на следующих предпосылках:

- каждый документ отписывается некоторым набором (смесью) тем;
- каждая отдельная тема описывается некоторым набором токенов (слов).

Темы – это некоторые скрытые переменные, которые присутствуют в наборе данных, но наблюдать их влияние напрямую – нельзя. Соответственно, задача тематического моделирования – извлечь эти скрытые переменные и проанализировать их влияние на имеющийся набор данных.

На текущий момент, существует большое количество алгоритмов тематического моделирования. Важно отметить, что область их применения не ограничивается сферой NLP, они в той же степени эффективны и в задачах биоинформатики, физики и др. Однако наиболее популярны они, все таки, именно в задачах NLP.

Модели тематического моделирования

Наиболее известный – латентный семантический анализ (или LSA). Данный метод идейно близок к методу главных компонент, который так же ищет скрытые переменные (главные компоненты) в тексте. Но если в случае PCA решается задача поиска некоторой матрицы P, произведение матрицы факторов на которую порождает набор некоррелированных переменных меньшей размерности (как раз – скрытые переменные). То в случае LSA – раскладывается специальная матрица, построенная на основе TF-IDF статистики, рассчитываемой по корпусу документов. В обоих случаях на промежуточных этапах (диагонализация) используется алгоритм SVD .

Алгоритм PLSA представляет собой вероятностную версию алгоритма LSA, имеющую более интуитивную интерпретацию. Алгоритм LDA – представляет собой усовершенствованную версию PLSA алгоритма, который менее подвержен проблеме переобучения за счет уменьшения количества параметров в модели. Далее будет представлен разбор обоих методов вероятностного тематического моделирования.

Модель PLSA

Предположим, что имеется набор документов $d \in \mathcal{D}$, набор слов $w \in \mathcal{W}$ и набор тематик $z \in \mathcal{Z}$. Обозначим степень релевантности документа d топику z как $P(d \mid z)$. Данную величину можно трактовать как вероятность того, что документ d относится к топику z, либо сколько процентов содержания документа d относится к топику z. Аналогично, обозначим степень релевантности слова w топику z как $P(w \mid z)$. Данную величину можно трактовать как вероятность того, что слово w можно встретить в документах, которые полностью посвящены тематике z. Далее формализуем процесс генерации документа в контексте модели PLSA.

- выбираем документ $d \in \mathcal{D}$ с вероятностью $P(d)$,
- генерируем слово w_i для выбранного документа d по следующим правилам
  - выбираем тематику $z \sim P(z \mid d)$ – мульт-е р-е.
  - выбираем слово $w \sim P(w \mid z)$ – мульт-е р-е.

Далее опишем процесс оценки вероятностей $P(z \mid d)$ и $P(w \mid z)$. Предположим, что переменные w и d условно независимы относительно переменной z [1].

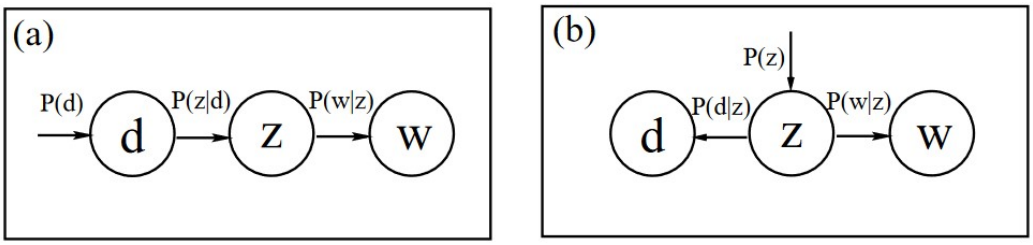

Рис. 1: Эквивалентные графические модели параметризации совместного распределения пары (w,d) с учетом новой скрытой переменной z [1].

Вероятность наблюдения пары (w,d) можно выразить следующим образом: $P(w,d) = P(d)P(w \mid d)$, или, что эквивалентно (помним об условной независимости переменных w и d относительно z)

$$P(d,w) = P(w,d) = \sum_{z}^{\mathcal{Z}} P(z)P(w \mid z)P(d \mid z)$$

Данное выражение полностью эквивалентно следующему

$$P(d,w) = \sum_{z}^{\mathcal{Z}} \frac{P(z)P(w \mid z)P(d)P(z \mid d)}{P(z)} = P(d) \sum_{z}^{\mathcal{Z}} P(w \mid z)P(z \mid d)$$

Таким образом, была формализована модель порождения документов для PLSA. Оценить её неизвестные параметры можно с помощью алгоритма оценки максимального правдоподобия. В выражении фигурирует функция $q(w,\ d)$, которая оценивает, сколько раз некоторый символ встретился в документе. $q(d)$ – выражает длину документа в словах. Построим функцию правдоподобия.

$$L = \prod_{w}^{\mathcal{W}} \prod_{d}^{\mathcal{D}} P(d,w)^{q(w,d)} = \prod_{w}^{\mathcal{W}} \prod_{d}^{\mathcal{D}} \left(P(d) \sum_{z}^{\mathcal{Z}} P(w \mid z) P(z \mid d) \right)^{q(w,d)},$$

$$\log L = \sum_{w}^{\mathcal{W}} \sum_{d}^{\mathcal{D}} q(w,d) \left(\log P(d) + \log \sum_{z}^{\mathcal{Z}} P(w \mid z) P(z \mid d) \right) =$$

$$\sum_{d}^{\mathcal{D}} q(d) \left(\log P(d) + \sum_{w}^{\mathcal{W}} \frac{q(w,d)}{q(w)} \left(\log \sum_{z}^{\mathcal{Z}} P(w \mid z) P(z \mid d) \right) \right)$$

Теперь необходимо некоторым образом максимизировать лог - правдоподобие по параметрам $P(w \mid z)$, $P(z \mid d)$:

$$p^*(z \mid d), p^*(w \mid z) = \underset{P(z \mid d),\ P(w \mid z)}{\operatorname{argmax}} logL$$

Величина $P(d)$ никак не влияет на оптимизацию, поскольку постоянна для данного набора документов, и в случае, если все документы в выборке уникальны, выражается как $P(d) = \frac{1}{(|D|)}$. Авторы оригинальной статьи предлагают использовать для оценки неизвестных параметров EM – алгоритм. Подробно с данным алгоритмом и его применением к текущей задаче можно ознакомиться в [1]. Помимо этого, был подготовлен блокнот Jupyter с иллюстрацией механизма работы данного алгоритма. Блокнот предоставляется в комплекте с лекционными материалами.

Далее будет рассмотрен пример использования PLSA на наборе данных “20 Newsgroup Dataset” . Это один из наиболее популярных наборов данных для тестирования различных NLP моделей. Он состоит из 20 тысяч документов, каждый из которых посвящен одной из 20 тематик и содержит текст дискуссии на специализированном форуме.

Рассмотрим наиболее релевантные слова для каждой категории. Для этого необходимо отсортировать слова по вероятностям для первой категории $P(w \mid z = 0)$ и второй категории $P(w \mid z = 1)$ в порядке возрастания. В рамках данной задачи выборка была ограничена двумя тематиками. Первая – специализированный форум по программированию в среде Windows, вторая – религиозный форум. Всего в наборе 1192 документов, словарь ограничен 1000 наиболее часто встречающимися словами (за исключением общих стоп-слов английского языка).

По ключевым ключевым первой тематики [2] (*god, will, people, jesus*) легко заметить, что все они относятся к религиозной теме. Ключевые слова второй тематики (*com, windows, file*) имеют непосредственное отношение к вычислительным системам.

topic_0	god	edu	can	one	will	re	people	subject	lines	jesus
topic_1	com	edu	window	can	subject	lines	file	organization	use	mit

Рис. 2: Наиболее часто встречающиеся слова в двух тематиках, выделенных алгоритмом PLSA.

Модель LDA

Модель Latent Dirichlet Allocation, предложенная Blei D., Ng A., Jordan M. [[2]] решает ту же самую задачу, что и PLSA, но исходит из других предпосылок о процессе порождения документов. Авторы статьи отмечают, что PLSA имеет серьезный недостаток: необходимость расчета оценок $|Z| * (|D| + |W|)$ параметров мультиномиальных распределений (де-факто мультинулли). Легко заметить, что это число линейно возрастает с увеличением документов в обучающей выборке. Таким образом, модель PLSA может быть подвержена оверфиттингу. Авторы предлагают подход, позволяющий устранить зависимость числа параметров от объема корпуса документов. Формализуем процесс генерации документа моделью LDA [3]:

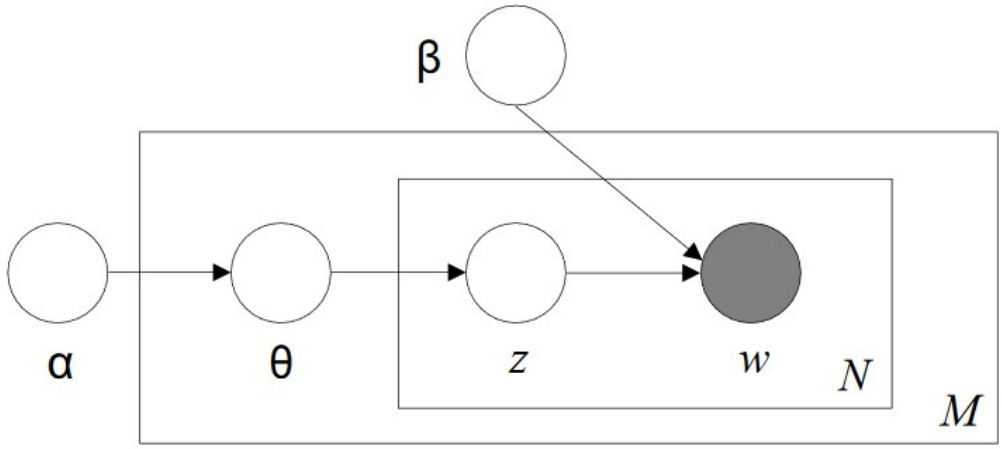

Рис. 3: Графическое представление модели LDA

- выбираем длину документа N (авторы предлагают генерировать $N \sim Poisson(\xi)$),
- генерируем вектор вероятностей $\theta \sim Dir(\alpha)$,
- генерируем слова для документа длины N:
  - выбираем тематику $z_i \sim Multinomial(\theta)$,
  - выбираем слово $w_i \sim Multinomial(\beta_{z_i})$.

В данном случае вектор θ, полученный путем семплирования из распределения Дирихле с параметром α представляет собой долю, на которую тематика z представлена в тексте. Вектор β_{z_i} (строка матрицы β) представляет собой относительную важность слов в тематике z_i (вероятность встретить слово в данной тематике). Распределение Дирихле в

данном случае выступает в роли сопряженного априорного распределения для z, поскольку его очень удобно использовать при переоценивании параметров мультиномиального распределения.

Оценка коэффициентов модели LDA сводится к оценке параметров α и β. Авторы работы предлагают использовать для этого алгоритм **«Variational Bayes»**. С работой данного алгоритма можно ознакомиться в оригинальной статье.

Далее к рассмотрению предлагается пример использования модели LDA на наборе данных “20 Newsgroup Dataset”. Для чистоты эксперимента выборка была ограничена двумя тематиками, как и в примере использования PLSA. Первая – специализированный форум по программированию в среде Windows, вторая – религиозный форум. Точно так же используется 1000 наиболее часто встречаемых слов из набора 1192 документов. [4]

Легко заметить, что алгоритм LDA выделил те же самые слова, что и PLSA, но придал им другой уровень значимости. В частности, слово “jesus” в религиозной тематике алгоритма LDA имеет больший вес, нежели тот, что ей придал алгоритм PLSA.

topic_0	god	edu	can	one	will	people	subject	re	jesus	lines
topic_1	edu	com	can	window	subject	lines	file	organization	use	mit

Рис. 4: Наиболее часто встречающиеся слова в двух тематиках, выделенных алгоритмом LDA

Заключение

В рамках настоящей работы было сформулировано понятие тематического моделирования, обозначены предпосылки, на которых базируются алгоритмы решающие данный класс задач. Было произведено подробное описание и примеры использования модели Probabilistic Latent Semantic Analysis, на базе которой строится большинство алгоритмов тематического моделирования. В т.ч. Latent Dirichlet Allocation, Additive Regularization Topic Modelling [[3]]. Также была кратко рассмотрена модель LDA и приведен пример ее использования для задачи выделения тематик в корпусе документов.

Список литературы

[1] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis // Machine Learning. 2001. Т. 42, № 1. С. 177–196.

[2] Blei D. Ng A. Jordan M. Latent Dirichlet Allocation // Journal of machine Learning research. 2003. Т. 3, № 3. С. 993–1022.

[3] К.В. Воронцов. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 455, № 3. С. 268–271.