Устойчивость алгоритмов обучения вероятностных тематических моделей

автор задачи: *Воронцов Константин Вячеславович*, voron@forecsys.ru
консультант: *Потапенко Анна Александровна*, anya_potapenko@mail.ru

12 декабря 2014 г.

1 Предварительные сведения

1.1 Вероятностная тематическая модель коллекции текстов

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) терминов. Известно число n_{dw} вхождений каждого из терминов w в каждый документ $d \in D$. Предполагается, что каждое слово в каждом документе связано с некоторой неизвестной (латентной) темой t из множества тем T. Обычно число тем $|T|$ много меньше $|D|$ и $|W|$. Задача заключается в том, чтобы выявить латентные темы по наблюдаемой коллекции D.

Вероятностная тематическая модель описывает вероятность появления терминов в документе:

$$p(w \mid d) = \sum_{t \in T} \phi_{wt}\, \theta_{td}, \tag{1}$$

где $\phi_{wt} = p(w \mid t)$ — неизвестное распределение на множестве терминов, определяющее тему t; $\theta_{td} = p(t \mid d)$ — неизвестное распределение на множестве тем, определяющее тематическую структуру документа d.

В *вероятностном латентном семантическом анализе* PLSA [4] для обучения модели (1) по коллекции документов D максимизируется логарифм правдоподобия при ограничениях нормировки и неотрицательности:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Phi, \Theta}; \tag{2}$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{3}$$

Задача (2), (3) является некорректно поставленной, поскольку имеет в общем случае бесконечно много решений. Действительно, в правой части (1) записано произведение *матрицы терминов тем* $\Phi = (\phi_{wt})_{W \times T}$ и *матрицы тем документов* $\Theta = (\theta_{td})_{T \times D}$. Матричное произведение определено с точностью до невырожденного линейного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$. Следствием неединственности является неустойчивость любых численных методов оптимизации, используемых для решения данной задачи, в том числе EM-алгоритма.

Известны различные условия единственности неотрицательных матричных разложений [2, 5, 6, 3]. В частности, разложение может быть единственным при сильной разреженности матриц Φ и Θ. Однако остаются не ясны вопросы: какой должна быть структура разреженности, т. е. как должны быть расположены нулевые элементы в матрицах Φ и Θ; какие дополнительные ограничения необходимо на них наложить, чтобы решение стало устойчивым; возможно ли связать эти ограничения с требованиями интерпретируемости (понятности) тем.

Исследовательские задания направлены на поиск ответов на эти вопросы.

1.2 EM-алгоритм

Для решения оптимизационной задачи (2), (3) используется итерационный процесс, называемый EM-алгоритмом [1]. Столбцы матриц Φ и Θ инициализируются случайными распределениями. На каждой итерации выполняются два шага.

На E-шаге для всех терминов в документах (d, w) по формуле Байеса вычисляются условные распределения $p(t \mid d, w)$. Вводятся вспомогательные переменные n_{dwt}, интерпретируемые как число вхождений термина w в документ d по теме t:

$$n_{dwt} = n_{dw} p(t \mid d, w), \quad p(t \mid d, w) = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}. \tag{4}$$

На M-шаге суммирование переменных n_{dwt} по d, w, t даёт частотные оценки искомых условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d \in D} n_{dwt}, \qquad n_t = \sum_{w \in W} n_{wt},$$
$$\theta_{td} = \frac{n_{dt}}{n_d}, \qquad n_{dt} = \sum_{w \in d} n_{dwt}, \qquad n_d = \sum_{t \in T} n_{dt},$$

или, в краткой записи через знак пропорциональности $\propto$:

$$\phi_{wt} \propto n_{wt}, \qquad \theta_{td} \propto n_{dt}. \tag{5}$$

Система уравнений (4), (5) определяет стационарную точку оптимизационной задачи (2), (3). Существуют различные способы построения итерационного процесса по формулам (4), (5): рациональный, стохастический, онлайновый [1].

1.3 Регуляризованный EM-алгоритм

Допустим, что наряду с правдоподобием (2) требуется максимизировать ещё k критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, k$, называемых *регуляризаторами*. Для решения задачи многокритериальной оптимизации будем максимизировать линейную комбинацию критериев с неотрицательными *коэффициентами регуляризации* τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta), \qquad L(\Phi, \Theta) + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}. \tag{6}$$

Данная задача по-прежнему решается с помощью EM-алгоритма, но вместо (5) используется модифицированная формула M-шага [1]:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+, \qquad \theta_{td} \propto \left(n_{dt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+, \tag{7}$$

где $(z)_+ = \max\{z, 0\}$ — операция положительной срезки.

2 Постановка задачи

Задания предполагают выполнение вычислительных экспериментов на синтетических (модельных) данных. Каждый эксперимент заключается в многократном восстановлении синтетических (то есть искусственно сгенерированных и потому известных экспериментатору) «истинных» матриц Φ_0, Θ_0 при различных значениях некоторой выбранной характеристики задачи.

Отклонение восстановленных распределений $p(i\,|\,j)$ от исходных $p_0(i\,|\,j)$ измеряется средним расстоянием Хеллингера

$$H(p, p_0) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\Big(\sqrt{p(i\,|\,j)} - \sqrt{p_0(i\,|\,j)}\Big)^2},$$

как для самих матриц Φ, Θ, так и для их произведения $\Phi\Theta$:

$$D_\Phi(\Phi, \Phi_0) = H(\Phi, \Phi_0);$$
$$D_\Theta(\Theta, \Theta_0) = H(\Theta, \Theta_0);$$
$$D_{\Phi\Theta}(\Phi\Theta, \Phi_0\Theta_0) = H(\Phi\Theta, \Phi_0\Theta_0).$$

Алгоритм 2.1. Восстановление стохастического матричного разложения.

1. сгенерировать синтетические матрицы Φ_0 и Θ_0;
2. сгенерировать коллекцию D;
3. EM-алгоритм: восстановить по коллекции D матрицы Φ и Θ;
4. венгерский алгоритм: найти соответствие между темами в (Φ_0, Θ_0) и (Φ, Θ);
5. вычислить расстояния D_Φ, D_Θ, $D_{\Phi\Theta}$;

Генерация реалистичных синтетических данных должна учитывать гипотезы разреженности, слабой корреляции тем, наличия фоновых тем. Возможно взять в качестве Φ_0, Θ_0 результат тематического моделирования реальной коллекции.

Рекомендуемый способ сгенерировать коллекцию с минимальным уровнем шума:

$$n_{dw} = \operatorname{pround}\Big(n_d \sum_{t\in T} \phi_{wt}\,\theta_{td}\Big),$$

где pround — функция вероятностного округления (число r округляется до $\lfloor r \rfloor$ с вероятностью $1 - \{r\}$ и до $\lceil r \rceil$ с вероятностью $\{r\}$), n_d — заданные длины документов. Использование обычного округления вместо вероятностного приводит к несоответствию сгенерированных данных модели (особенно в результате округлений к нулю или единице). Использование формулы без округления приводит к полному отсутствию шума и существованию точного матричного разложения, что эквивалентно нереалистичным предположениям об идеальной точности модели и бесконечной длине каждого документа. Тем не менее, некоторые эксперименты лучше начинать без округления, чтобы проверить, сходятся ли расстояния $D_{\Phi\Theta}$ к нулю.

Задания предполагают построение графиков зависимости всех трёх расстояний от выбранной характеристики задачи.

На этапе тестирования модели строятся зависимости перплексии и расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ от номера итерации в EM-алгоритме. Это необходимо для выяснения необходимого числа итераций при заданных параметрах задачи.

Также строятся зависимости расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем от выбранных параметров синтетической задачи или метода тематического моделирования. Тема считается восстановленной, если для неё найдено взаимно однозначное соответствие с исходной темой.

3 Творческие задания

В заданиях не всегда точно указано, какие именно зависимости надо строить, чтобы проверить сформулированные гипотезы. Предполагается, что студент учится самостоятельно додумывать постановку экспериментов. Особенно поощряется формулирование и проверка дополнительных собственных гипотез.

Задание 1. Проверяется гипотеза, что чем выше разреженность матриц Φ и Θ, тем выше устойчивость решения.

Вторая гипотеза: регуляризатор разреживания повышает устойчивость решения при условии, что (Φ_0, Θ_0) разрежены.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем от разреженности (доли ненулевых элементов) исходных матриц.

Это задание повторяет эксперимент В. Глушаченкова, описанный в [1].

Задание 2. Проверяется гипотеза, что чем выше различность тем (столбцов Φ), тем выше устойчивость решения.

Вторая гипотеза: регуляризатор декоррелирования повышает устойчивость решения при условии, что темы (столбцы Φ_0) слабо коррелированы.

Третья гипотеза: регуляризаторы разреживания и декоррелирования вместе ещё сильнее повышают устойчивость решения при условии, что темы разрежены и слабо коррелированы.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем от средней корреляции между темой и ближайшей к ней темой в исходной матрице Φ.

Задание 3. Проверяется гипотеза, что регуляризатор разреживания строк матрицы Θ способен устойчиво определять истинное число тем $|T|$, если начинать EM-алгоритм с заведомо избыточного числа тем.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$, числа восстановленных тем, а также перплексии и числа ненулевых тем ($n_t > 0$) от номера итерации в EM-алгоритме.

Задание 4. Проверяется гипотеза, что наличие фоновых тем не сильно мешает восстановить основные предметные темы.

Вторая гипотеза: при наличии плотных фоновых тем и слабо коррелированных разреженных предметных тем в (Φ_0, Θ_0) совместное применение регуляризаторов разреживания, сглаживания и декорреляции повышает устойчивость решения.

Третья гипотеза: ошибка при назначении числа фоновых тем не сильно влияет на результат.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем (только по предметным темам), от доли фоновых тем в документах коллекции. Затем, при фиксированной реалистичной доле фоновых тем (30%–70%) исследуется зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ от средней корреляции между темой и ближайшей к ней темой в исходной матрице Φ.

Задание 5. Проверяется гипотеза, что добавление в коллекцию «виртуальных документов», каждый из которых содержит список топовых слов одной темы, повышает устойчивость решения (применяется регуляризатор для частичного обучения [1]).

Вторая гипотеза: достаточно лишь небольшого числа топовых слов.

Третья гипотеза: достаточно задать топовые слова лишь небольшой части тем.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем от числа виртуальных документов и суммарного числа слов в виртуальных документах.

Задание 6. Проверяется гипотеза, что если темы сгруппированы по классам, и для некоторых документов известны их классы, то это повышает устойчивость решения (число тем порядка 100, число классов порядка 20. Выборка делится на обучающую и контрольную. Классификации контрольных документов не используются для моделирования контрольных документов, но используются для оценивания качества их классификации.

Вторая гипотеза: ковариационный регуляризатор для категоризации текстов [1] лучше других моделей классификации.

Третья гипотеза: в задаче частичного обучения достаточно небольшого числа размеченных документов.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем, а также качества классификации контрольной выборки документов от доли документов обучающей выборки, для которых известны их классы.

Задание 7. Проверяется гипотеза, что если матрицы Φ_0, Θ_0 разрежены, темы слабо коррелированы и число тем велико (скажем, $|T| = 1000$), то при использовании много меньшего числа тем в EM-алгоритме (скажем, $|T| = 100$) многие темы восстанавливаются, но каждый раз разные в зависимости от случайного начального приближения.

Вторая гипотеза: при использовании регуляризаторов разреживания и декоррелирования восстановленные темы получаются более близкими к исходным.

Третья гипотеза: при добавлении регуляризатора сглаживания фоновых тем большее число предметных тем восстанавливаются правильно.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем от разреженности (доли ненулевых элементов) исходных матриц.

Задание 8. Проверяется гипотеза, что регуляризатор отбора тем, который разреживает строки матрицы Θ, предпочитает оставлять более разреженные темы, чем их выпуклые комбинации. Для этого в исходную матрицу Φ_0 добавляются столбцы — выпуклые комбинации разреженных и слабо коррелированных основных столбцов матрицы Φ, в матрицу Θ_0 добавляются строки.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$ и числа восстановленных тем от числа линейно зависимых тем, добавленных к основным темам.

Задание 9. Проверяется гипотеза, что если матрицы Φ_0, Θ_0 разрежены, темы слабо коррелированы и число тем не велико (скажем, $|T| = 100$), но в ЕМ-алгоритме число тем изначально велико ($|T'| = 1000$ и выше), то в результате совместного применения регуляризаторов разреживания, декоррелирования и строкового разреживания исходные темы восстанавливаются более устойчиво.

Строится зависимость расстояний D_Φ, D_Θ, $D_{\Phi\Theta}$, числа отобранных тем и числа восстановленных тем от номера итерации и от начального числа тем $|T'|$.

Список литературы

[1] *Воронцов К. В.* Вероятностное тематическое моделирование. — 2014. http://www.MachineLearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf.

[2] *Donoho D., Stodden V.* When does non-negative matrix factorization give a correct decomposition into parts? // Advances in Neural Information Processing Systems 2003 / Ed. by S. Thrun, L. Saul, B. Schölkopf. — Cambridge, MA: MIT Press, 2004. http://www-stat.stanford.edu/ donoho/Reports/2003/NMFCDP.pdf.

[3] *Gillis N.* Sparse and unique nonnegative matrix factorization through data preprocessing // *Journal of Machine Learning Research.* — 2012. — Vol. 13, no. 1. — Pp. 3349–3386.

[4] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.

[5] *Lauerberg H., Christensen M., Pumbley M., Hansen L., Jensen S.* Theorems on positive data: On the uniqueness of nmf // *Computational Intelligence and Neuroscience.* — 2008. — Vol. 2008. — P. 10.

[6] *Schachtner R., Pöppel G., Lang E. W.* Towards unique solutions of non-negative matrix factorization problems by a determinant criterion // *Digital Signal Processing.* — 2011. — Vol. 21, no. 4. — Pp. 528–534.