Семинары по матричным разложениям

Евгений Соколов
sokolov.evg@gmail.com

13 мая 2016 г.

1 Матричные разложения

§1.1 Постановка задачи

Пусть дана выборка $X^\ell = (x_1, \dots, x_\ell)$, объекты из которой описываются вещественными векторами признаков из $\mathbb{R}^d$. Если признаков очень много, то зачастую прибегают к понижению размерности. Для этого можно попытаться найти r базисных объектов $h_1, \dots, h_r \in \mathbb{R}^d$(которые не обязательно входят в выборку), а каждый объект из выборки представить как линейную комбинацию базисных объектов:

$$x_i = \sum_{k=1}^{r} w_{ik} h_k.$$

Такое представление можно записать в матричном виде. Если X — матрица размера $\ell \times d$, где каждая строчка соответствует признаковому описанию одного объекта, то получаем

$$\underbrace{X}_{\ell \times d} = \underbrace{W}_{\ell \times r} \underbrace{H}_{r \times d}. \tag{1.1}$$

Строки матрицы W можно использовать в качестве новых, низкоранговых признаков; строки матрицы H представляют собой набор наиболее «характерных» объектов, через которые можно выразить всю выборку. Сама же задача поиска матриц W и H называется задачей *матричного разложения*.

Матричные разложения имеют ряд интересных интерпретаций в анализе данных [1].

1. Пусть каждый объект — это фотография лица. Число признаков равно числу пикселей на фотографии, а значение каждого признака соответствует интенсивности одного из пикселей. Если матрицы W и H в разложении (1.1) будут неотрицательными, то строк матрицы H будут содержать r «базисных лиц», которые в сумме могут дать любую фотографию из выборки. Как правило, базисные объекты представляют собой изображения носа, глаз, губ и других частей лица, которые в сумме дают полноценную фотографию.

2. Пусть каждый объект — это текстовый документ, число признаков равно размеру словаря, и каждый признак равен числу вхождений соответствующего слова в документ. Если матрицы W и H будут неотрицательными, а сумма элементов в каждой их строке будет равна единице, то мы получим так называемое *тематическое разложение*. Каждый строка матрицы H будет соответствовать распределению слов в одной из тем, а каждая строка матрицы W — распределению тем в документе.

3. Пусть каждый объект — это пользователь музыкального сайта, число признаков равно числу треков в коллекции сайта, а значение признака равно числу прослушиваний данным пользователем данной композиции. В этом случае каждая из размерностей $1, \dots r$ может интерпретироваться как музыкальный жанр, строки матрицы W — как векторы интересов пользователей, описывающие их интерес к каждому из жанров, а столбцы матрицы H — как векторы, описывающие принадлежность композиций к жанрам. Интерес i-го пользователя к j-му треку при этом оценивается как скалярное произведение вектора интересов пользователя на жанровый вектор трека — типичный подход к решению задач *построения рекомендаций*.

§1.2 Функционалы качества

Наиболее распространенным функционалом в матричных разложениях является норма Фробениуса:

$$\|X - WH\|^2 = \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{d} (X_{ij} - [WH]_{ij})^2.$$

Данный функционал соответствует предположению о нормальности шума в данных. Это предположение не всегда адекватно — например, если все элементы матриц X, W и H неотрицательны, то нормальное распределение вряд ли может быть применено для моделирование шума, поскольку оно предполагает, что величины могут принимать любые значения на вещественной оси. Также элементы матрицы X могут быть натуральными (например, число просмотров фильма), и в этом случае нормальное распределение также является неадекватным.

В тематическом моделировании, где элементы матрицы X имеют смысл числа вхождений слова в документ, в качестве функции потерь используется дивергенция Кульбака-Лейблера:

$$\sum_{i=1}^{\ell} \sum_{j=1}^{d} \left(X_{ij} \log \frac{X_{ij}}{[WH]_{ij}} - X_{ij} + [WH]_{ij} \right).$$

Данная функция потерь соответствует модели с пуассоновским шумом.

В физических задачах применяется дивергенция Итакуры-Сайто:

$$\sum_{i=1}^{\ell} \sum_{j=1}^{d} \left(\frac{X_{ij}}{[WH]_{ij}} - \log \frac{X_{ij}}{[WH]_{ij}} - 1 \right),$$

которая имеет смысл расстояния между спектрами двух сигналов и соответствует модели с мультипликативным гамма-шумом.

Одним из наиболее общих семейств функционалов являются AB-дивергенции [2]:

$$D_{\mathrm{AB}}^{(\alpha,\beta)}(X, WH) = -\frac{1}{\alpha\beta}\sum_{i=1}^{\ell}\sum_{j=1}^{d}\left(X_{ij}^{\alpha}[WH]_{ij}^{\beta} - \frac{\alpha}{\alpha+\beta}X_{ij}^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}[WH]_{ij}^{\alpha+\beta}\right),$$

где $\alpha \neq 0$, $\beta \neq 0$, $\alpha + \beta \neq 0$. Легко видеть, что при $\alpha = \beta = 1$ AB-дивергенция совпадает с нормой Фробениуса, при $\alpha = 1$, $\beta = 0$ — с KL-дивергенцией, а при $\alpha = -1$, $\beta = 1$ — с дивергенцией Итакуры-Сайто.

Для простоты везде далее для простоты мы будем обсуждать только разложения с точки зрения нормы Фробениуса.

§1.3 Структура матрицы

Стандартные методы матричных разложений вроде SVD или PCA не делают никаких предположений о структуре матриц W и H. Однако зачастую возникают задачи, в которых требуется построить разложение, удовлетворяющее некоторым специальным свойствам. Рассмотрим некоторые примеры.

1. Строки матрицы H — базисные векторы — должны быть независимыми. Такое требование возникает, например, при распознавании речи нескольких одновременно говорящих людей. Речь каждого из них является независимым сигналом, а итоговая запись является линейной комбинацией таких сигналов. В качестве меры независимости могут использоваться совместная информация, моменты третьего и четвертого порядков и т.д. Класс методов, решающих данную задачу, называется *анализом независимых компонент* (Independent Component Analysis, ICA).

2. Матрицы W и/или H должны быть разреженными. В этом случае применяются методы вроде разреженного SVM, основанные на L_1-регуляризации.

3. Матрицы W и H должны быть покомпонентно неотрицательными. Выше мы столкнулись с двумя важными примерами, в которых неотрицательность матриц приводила к интересным интерпретациям разложения. К сожалению, требование неотрицательности существенно усложняет задачу и требует гораздо более трудоемких методов. Подробнее об этом речь пойдет ниже.

§1.4 Неотрицательные матричные разложения

Задача построения неотрицательного матричного разложения является существенно более сложной по сравнению с задачей обычного разложения по норме Фробениуса. На это есть несколько причин:

- Данная задача является NP-полной. Как правило, методы построения неотрицательных разложений гарантируют лишь сходимости к стационарной точке функционала.

- Данная задача является некорректно поставленной: если $X = WH$ — искомое разложение, то для любой неотрицательной невырожденной матрицы Q разложение $X = (WQ)(Q^{-1}H)$ будет давать такое же значение функционала. Чтобы добиться единственности решения, нужно вводить дополнительные требования — разреженность, ортогональность столбцов одной из матриц и т.д.

§1.5 Методы построения неотрицательных матричных разложений

Существует большое число методов для построения неотрицательных разложений: мультипликативный градиентный спуск, Alternating Least Squares, Alternating Nonnegative Least Squares и т.д. О данных методах можно прочитать, например, в обзоре [1]. Все они схожи тем, что по очереди оптимизируют матрицы W и H, реализуя стратегию блочно-покоординатного спуска. Мы остановимся на одном из самых простых и самых мощных методов — Hierarchical Alternating Least Squares (HALS). В нем по очереди оптимизируется каждый элемент матриц W и H при фиксированных остальных элементах. Пусть мы настраиваем матрицу W (формулы для H выводятся аналогично). Поизведем оптимизацию только по элементу W_{ik}; в этом случае задача примет вид

$$
\begin{aligned}
W_{ik} &= \arg\min_{w \geqslant 0} \sum_{i'=1}^{\ell} \sum_{j=1}^{d} \left(X_{i'j} - \sum_{m=1}^{r} W_{im} H_{mj} \right)^2 = \\
&= \arg\min_{w \geqslant 0} \sum_{j=1}^{d} \left(X_{ij} - \sum_{m=1}^{r} W_{im} H_{mj} \right)^2 = \\
&= \arg\min_{w \geqslant 0} \sum_{j=1}^{d} \Bigl(\underbrace{X_{ij} - \sum_{m \neq k} W_{im} H_{mj}}_{g_j} - w H_{kj} \Bigr)^2 = \\
&\arg\min_{w \geqslant 0} \sum_{j=1}^{d} \left(g_j - w H_{mj} \right)^2.
\end{aligned}
$$

Запишем лагранжиан полученной задачи:

$$L = \sum_{j=1}^{d} (g_j - w H_{mj})^2 - \lambda w.$$

Дифференцируя его по w и приравнивая к нулю, получаем

$$w \sum_{j=1}^{d} H_{kj}^2 - \sum_{j=1}^{d} g_j H_{kj} = \lambda.$$

Согласно условию дополняющей нежесткости $\lambda w = 0$. Рассмотрим два случая:

- $\lambda = 0$: отсюда получаем, что
$$w = \frac{\sum_{j=1}^{d} g_j H_{kj}}{\sum_{j=1}^{d} H_{kj}^2}.$$

- $w = 0$: тогда $\lambda = -\sum g_j H_{kj}$. Согласно условиям Куна-Таккера, двойственная переменная λ должна быть неотрицательной. Из этого вытекает неравенство

$$\sum_{j=1}^{d} g_j H_{kj} \leqslant 0.$$

Таким образом, либо $\sum g_j H_{kj}$ больше нуля и мы просто приравниваем w этой сумме, либо они меньше нуля, и тогда $w = 0$. Получаем формулу пересчета W_{ik}:

$$W_{ik} = \max\left(0, \sum_{j=1}^{d} g_j H_{kj}\right).$$

§1.6 Разделимые неотрицательные разложения

Хотя в общем случае задача неотрицательного матричного разложения является NP-полной, существует широкий класс подзадач, допускающий эффективное решение — *разделимые матричные разложения* (Separable NMF). Неотрицательная матрица X называется r-разделимой, если у нее найдется такое подмножество $\tilde{X}$ из r столбцов, что для некоторой неотрицательной матрицы H выполнено

$$X = \tilde{X} H.$$

Иными словами, матрица разделима, если все ее столбцы линейно (и неотрицательно) выражаются через подмножество столбцов. Это предположение нередко является разумным — например, в задаче тематического моделирования оно означает, что для каждой темы найдется хотя бы один документ, относящейся только к ней и ни к какой другой.

Оказывается, что для разделимой матрицы задача построения неотрицательного разложения может быть сформулирована как задача линейного программирования. Это в свою очередь означает, что для нее можно гарантировать поиск решения за полиномиальное время. Более того, это верно и для *почти разделимых матриц* — матриц, отличающихся от разделимых на некоторую шумовую матрицу.

§1.7 Факторизационные машины

Рассмотрим признаковое пространство $\mathbb{R}^d$. Допустим, что целевая переменная зависит от парных взаимодействий между признаками. В этом случае представляется разумным строить полиномиальную регрессию второго порядка:

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x_j + \sum_{j_1=1}^{d} \sum_{j_2=j_1+1}^{d} w_{j_1 j_2} x_{j_1} x_{j_2}.$$

Данная модель состоит из $d(d-1)/2 + d + 1$ параметров. Если среди признаков есть категориальные с большим числом категорий (например, идентификатор пользователя), то после их бинарного кодирования число параметров станет слишком большим. Чтобы решить проблему, предположим, что вес взаимодействия признаков j_1 и j_2 может быть аппроксимирован произведением низкоразмерных скрытых векторов v_{j_1} и v_{j_2}, характеризующих эти признаки. Мы получим модель, называемую *факторизационной машиной* (factorization machine, FM) [3]:

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x_j + \sum_{j_1=1}^{d} \sum_{j_2=j_1+1}^{d} \langle v_{j_1}, v_{j_2} \rangle x_{j_1} x_{j_2}.$$

Благодаря описанному трюку число параметров снижается до $dr + d + 1$, где r — размерность скрытых векторов.

Данная модель является обобщением моделей с матричными разложениями. Выше мы обсуждали пример построения рекомендаций песен пользователям — интерес пользователя к песне оценивался как скалярное произведение некоторых скрытых векторов. Эту задачу можно сформулировать как задачу построения регрессии с двумя категориальными признаками: идентификатором пользователя и идентификатором композиции. Целевым признаком является число прослушиваний композиции пользователем. Для некоторого подмножества пар (пользователь, композиция) мы знаем число прослушиваний; для остальных мы хотим его восстановить. После бинаризации признаков мы получим, что факторизационная машина оценивает целевую переменную как произведение скрытых векторов пользователя и композиции — иными словами, она строит разложение матрицы прослушиваний X.

Существует несколько методов настройки факторизационных машин, из которых наиболее совершенным считается метод Монте-Карло на основе марковских цепей; реализацию можно найти в библиотеке libFM.

FFM. Недавно было предложено расширение факторизационных машин, позволившее авторам победить в конкурсах Criteo и Avazu по предсказанию кликов по рекламным объявлениям. В обычных факторизационных машинах у каждой переменной имеется всего один скрытый вектор, отвечающий за взаимодействие с остальными переменными. Допустим, что переменные можно некоторым образом сгруппировать — например, в задаче рекомендации музыкальных альбомов в бинарном векторе, отвечающем за композиции, будет стоять несколько единиц, соответствующих всем композициям из альбома. Все единицы из этого вектора можно объединить в одну группу. Расширим модель, введя для каждого признака разные скрытые векторы для взаимодействия с разными группами:

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x_j + \sum_{j_1=1}^{d} \sum_{j_2=j_1+1}^{d} \langle v_{j_1, f_{j_2}}, v_{j_2, f_{j_1}} \rangle x_{j_1} x_{j_2},$$

где f_{j_1} и f_{j_2} — индексы групп признаков x_1 и x_2. Данная модель носит название *field-aware factorization machines* (FFM) [4].

Список литературы

[1] *Gillis N.* (2014). The Why and How of Nonnegative Matrix Factorization. // http://arxiv.org/abs/1401.5226v2

[2] *Cichocki A. et. al.* (2011). Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. // Entropy 2011, 13(1), 134-170.

[3] *Rendle, S.* (2012). Factorization machines with libFM. // ACM Trans. Intell. Syst. Technol. 3, 3, Article 57.

[4] Field-aware Factorization Machines:
http://www.csie.ntu.edu.tw/~r01922136/slides/ffm.pdf