Приложение сингулярного разложения к заполнению пропусков в табличных данных

Воркожоков Максим

Высшая школа экономики

4 октября 2025

План

1. Зачем это нужно?
2. Наивный подход
3. Выпуклая оптимизация
4. Онлайн SVD
5. Second Section

Постановка задачи

Представим, что у нас есть некоторая матрица $M \in \mathbb{R}^{n \times k}$, но она не заполнена до конца: мы видим только элементы $m_{ij} : (i, j) \in \Omega$. Задача - по имеющимся данным максимально точно восстановить матрицу M.
Формально: Введём оператор проекции:

$$P_\Omega(X)_{ij} = \begin{cases} x_{ij}, & \text{if } (i, j) \in \Omega \\ 0, & \text{otherwise} \end{cases}$$

Тогда нужно найти такую матрицу $X \in \mathbb{R}^{n \times k}$, что $||P_\Omega(X - M)||_F \to \min\limits_X$

Некоторые предположения

1. Матрица M низкоранговая.
2. Количество наблюдений достаточно велико:

Пусть M – квадратная матрица $n \times n$ и имеет ранг r. Тогда в матрице M имеется $2nr - r^2$ степеней свободы. Если $|\Omega| < 2nr - r^2$, то точное восстановление матрицы невозможно. Но, вообще говоря, для достаточной эффективности алгоритма нужно порядка $nr \log n$ доступных значений.

Поэтому задача сводится к поиску матрицы X ранга $\leqslant r$, наиболее точно приближающей матрицу M.

Применения в жизни

- К матрице с незаполненными пропусками неудобно применять всякие известные разложения — если удалить все строки с пропусками, то может потеряться большая часть данных, поэтому удобно приблизительно их заполнить (например, заполнение данных в социологических опросах или финансовой статистике).
- Система рекомендаций (Netflix, Amazon, etc) — на основании заполнения пропущенных данных (оценок на фильмы/товары) пользователям предлагаются рекомендации. Такая система называется collaborative filtering.
- Computer vision — восстановление повреждённых фрагментов изображений

Низкоранговое приближение

1. **Инициализация.** Как-нибудь заполним пропуски в матрице M: например, нулями или средним по столбцу значением. Это будет матрица $Y^{(0)}$.
2. **Разложение.** Применим SVD к X_0, получим

$$Y^{(0)} = U^{(0)}\Sigma^{(0)}(V^{(0)})^\top$$

Рассмотрим приближение матрицей ранга k:

$$Y_k^{(0)} = U_k^{(0)}\Sigma_k^{(0)}(V_k^{(0)})^\top$$

3. **Заполним пропуски.** Если $(i,j) \in \Omega$, оставим m_{ij}, иначе возьмём y_{ij}. Получим матрицу $X^{(0)}$.
4. Повторяем шаги 1-3, пока $||X^{(i+1)} - X^{(i)}||$ не станет достаточно малым.

Проблемы такого подхода

1. Произвольный выбор k — чтобы получить наиболее точное приближение, нужно "угадать" ранг матрицы M.
2. Множество матриц X ранга $\leqslant r$ — невыпуклое, поэтому задача оптимизации — невыпуклая. Алгоритм может привести к локальному, но не глобальному минимуму.

Выпуклая оптимизация: идея

- Идея: использовать "ядерную норму" (nuclear norm) $\|X\|_* = \sum_i \sigma_i(X)$ в качестве мягкого ограничения на ранг.
- Сформулируем задачу так:

$$\frac{1}{2}\|P_\Omega(M - X)\|_F^2 + \lambda\|X\|_* \to \min_X$$

  где $\lambda > 0$ — параметр регуляризации.
- Такая задача выпуклая, следовательно, глобальный минимум гарантирован.

Soft-Impute (Emmanuel J. Candes and Terence Tao, 2009): шаги алгоритма

1. Заполняем пропуски нулями или средним по столбцу, получим матрицу $Y^{(k)}$
2. Применяем SVD: $Y^{(k)} = U^{(k)} \Sigma^{(k)} (V^{(k)})^\top$.
3. Применяем ограничение к сингулярным числам:

$$\sigma_i' = \max(\sigma_i - \lambda, 0)$$

4. Обновляем матрицу: $X^{(k+1)} = U^{(k)} (\Sigma^{(k)})' (V^{(k)})^\top$.
5. Повторяем, $||X^{(k+1)} - X^{(k)}||$ не станет мало.

Нюансы

- Гарантированная сходимость к глобальному минимуму из-за выпуклости оптимизационной задачи.
- Ранговая структура формируется естественно за счёт λ, но выбор λ всё ещё произволен, как и выбор ранга r.
- Более устойчив к шуму, чем итеративный SVD.

Онлайн SVD: идея

- Для больших матриц M обновление приведёт к долгому пересчёту полного SVD. Идея — использовать онлайн обновление.
- Онлайн SVD обновляет приближение данных при изменениях.

Онлайн SVD: алгоритм Incremental SVD (Matthew Brand, 2002)

1. Имеем аппроксимацию заполненной матрицы $M = U\Sigma V^\top$
2. Пришла новая матрица (новые столбцы или новые строки) $C \in \mathbb{R}^{m\times c}$.
3. Пусть $L = U^\top C, H = (I - UU^\top)C$.
4. Ортогонализируем H, получая матрицу J, положим $K = J^T H$.
5. Составляем матрицу:
$$Q = \begin{pmatrix} \Sigma & L \\ 0 & K \end{pmatrix}$$
   Получим
$$[M|C] = [U|J]Q\begin{pmatrix} V & 0 \\ 0 & I \end{pmatrix}^\top$$
   Вычисляем SVD для $Q = U_0\Sigma_0 V_0^\top$. Итого
$$[M|C] = [U|J]U_0\Sigma_0 V_0^\top$$
6. Время работы $O((n+m)r^2 + mc^2)$. Оптимизацией ортогонализации может быть снижено до $O(nmr)$.

Blocks of Highlighted Text

In this slide, some important text will be highlighted because it's important. Please, don't abuse it.

Block

Sample text

Alertblock

Sample text in red box

Examples

Sample text in green box. The title of the block is "Examples".

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table: Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

Smith, J. (2012).
Title of the publication.
Journal Name, 12(3):45–678.

The End