Матричные разложения

1. Матричные разложения

Иногда бывает удобно представить матрицу как произведение других матриц, обладающих определенными свойствами:

$$X = AB \quad \text{или} \quad X = ABC.$$

Одним из вариантов такого представления является *спектральное разложение матриц*. Если матрица X симметрична, то ее можно представить в следующем виде:

$$X = S^T \cdot D \cdot S,$$

где матрица S — ортогональная, а матрица D — диагональная:

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_i \end{pmatrix}.$$

Элементы матрицы D неотрицательны и являются собственными числами матрицы X.

Часто встречаются функции вида:

$$f(y) = y^T X y,$$

где X — симметричная матрица. Такая функция $f(y)$ называется *квадратичной формой*.

Если воспользоваться спектральным разложением матрицы, то есть:

$$f(y) = y^T \cdot S^T \cdot D \cdot S \cdot y = (S \cdot y)^T D (S \cdot y),$$

а затем провести замену $z = S \cdot y$, то квадратичная форма примет более простой вид:

$$f(z) = z^T D z = \sum_{i=1}^{n} \lambda_i z_i^2.$$

В полученном виде анализировать функцию намного проще.

Другим примером является *сингулярное разложение матриц*:

$$X = UDV,$$

где матрицы U и V — ортогональные, а матрица D — диагональная.

Геометрический смысл данного разложения заключается в следующем. Исходная матрица задает некоторое линейное преобразование, которое затем представляется в виде совокупности ортогонального преобразования (например, поворота), растяжения вдоль осей и еще одного поворота.

Сингулярное разложение матриц часто используется в алгоритмах машинного обучения. Один из примеров применения — *рекомендательные системы*.

2. Приближение матрицей меньшего ранга

Матрица задает некоторое отображение, а ее ранг задает размерность образа пространства, являясь мерой «сложности» матрицы. **Ранг матрицы** — это максимальное количество линейно независимых столбцов или строк матрицы. Другими словами, рангом матрицы является максимальный размер подматрицы с ненулевым определителем.

Из определения следует, что если матрица X имеет размер $m \times n$, то:

$$\text{rg}(X) \leqslant \min(m, n).$$

Рассматривается произведение двух матриц, причем матрица A имеет размер $m \times k$, а матрица B — $k \times n$. Если k меньше m и n, то можно провести оценку неравенством:

$$X = AB, \quad \text{rg}(A) \leqslant k, \quad \text{rg}(B) \leqslant k \quad \Rightarrow \quad \text{rg}(X) \leqslant k.$$

Приближение матрицей меньшего ранга может применяться, например, исходя из предположения, что матрица X на самом деле сложнее, чем должна быть, и желательно приблизить её более простой матрицей. Один из способов такого приближения заключается в представлении матрицы как произведения двух других таким образом, чтобы ранг произведения был не больше, чем k:

$$X = X' = UV^T.$$

Если матрица U размера $m \times k$, а матрица V размера $n \times k$, то наилучшее приближение по норме:

$$\left|\left|X - UV^T\right|\right| \to \min.$$

Существует множество способов определить норму матрицы, одним из которых является *норма Фробениуса*:

$$||X||_F = \sqrt{\sum_{i,j} x_{ij}^2}.$$

С использованием данной нормы задача примет следующий вид:

$$U, V = \underset{U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{n \times k}}{\text{argmin}} \sum_{i,j} \left(x_{ij} - u_i^T v_j\right)^2.$$

В качестве примера можно рассмотреть *преобразование признаков*. Пусть X — матрица признаков объектов, тогда U — матрица новых признаков. При $k < n$ преобразование признаков понижает размерность пространства. При этом по матрице U с максимальной возможной точностью восстанавливаются исходные признаки X.

Другим примером является *задача рекомендаций*. Пусть X — матрица с оценками x_{ij}, поставленными пользователем i фильму j. Некоторые значения матрицы неизвестны. Можно рассмотреть следующую модель:

$$x_{ij} \approx \hat{x}_{ij} = u_i v_j,$$

где u_i отражает интересы пользователя, а v_j — признаковое описание фильма. Далее предлагается настроить u_i и v_j на известных x_{ij}, а неизвестные спрогнозировать. В результате будут рекомендованы фильмы, для которых спрогнозирована высокая оценка. Задача принимает вид:

$$U, V = \underset{U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{n \times k}}{\text{argmin}} \sum_{i,j:\, x_{i,j} \neq 0} \left(x_{ij} - u_i^T v_j\right)^2.$$

3. SVD и низкоранговые приближения

Приближение матрицы X матрицей более низкого ранга выглядит следующим образом:

$$\hat{X} = \underset{\text{rg}(\hat{X} \leqslant k)}{\text{argmin}} ||X - \hat{X}||.$$

В случае матричного разложения:

$$\hat{X} = UV^T, \quad U \in \mathbb{R}^{m \times k}, \quad V \in \mathbb{R}^{n \times k}.$$

SVD (сингулярное разложение матриц) имеет вид:

$$X_k = U_k D_k V_k^T.$$

Такая матрица X_k обладает некоторыми особыми свойствами. Из матрицы U берутся k столбцов, из матрицы D — квадрат размера $k \times k$, из матрицы V — k строк (см. рис. 1). В результате перемножение выбранных матриц получается матрица X_k.

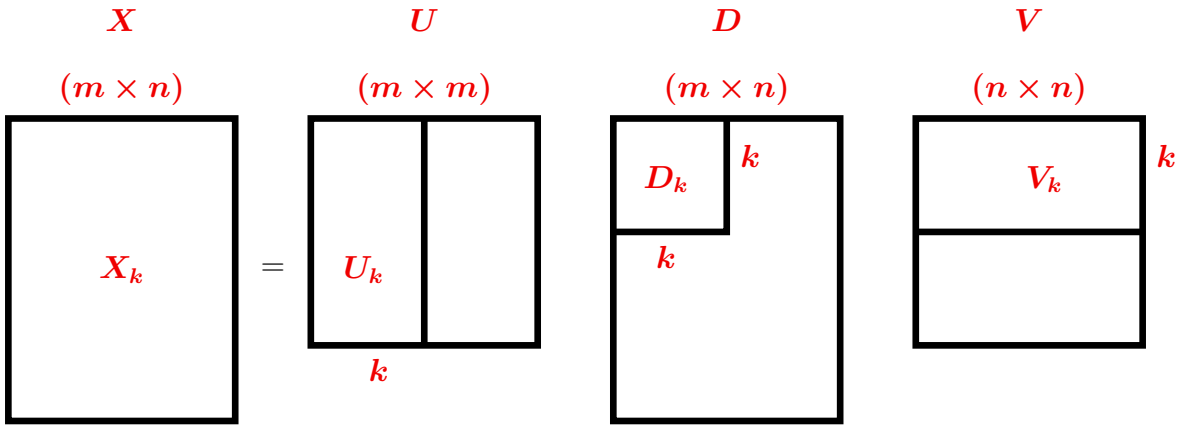

Рис. 1.

Оказывается, что матрица X_k будет наилучшим приближением матрицы X по норме Фробениуса:

$$\hat{X}_k = \underset{\text{rg}(\hat{X}) \leqslant k}{\text{argmin}} ||X - \hat{X}||_F,$$

где норма разности вычисляется следующим образом:

$$||X - \hat{X}||_F = \sqrt{\sum_{i,j} (x_{ij} - \hat{x}_{ij})^2}.$$

В рекомендательных системах также ищется наилучшее приближение. Значит, можно взять матрицы из SVD, распределив диагональную матрицу между двумя другими, и получившийся результат использовать как матричное разложение в рекомендательных системах.

Первым недостатком такого метода является тот факт, что сделать SVD — небыстрая операция. Второй недостаток заключается в том, что SVD делают для

всей матрицы, но в случае рекомендательных систем некоторые элементы неизвестны. Однако, рассматриваемое свойство SVD привело к распространению термина «SVD» в рекомендательных системах, хотя в них используется разложение на две матрицы.

Есть несколько вариантов, как именно сделать рекомендации. Первый вариант является не очень правильным по качеству, но очень простым. Делается SVD для исходной матрицы предпочтений. Матрица $U_k D_k$ используется как матрица профилей пользователей, а матрица V_k — как матрица профилей фильмов. Произведение профилей будет прогнозом оценки фильма.

Другой вариант является более правильным: не используя SVD, необходимо подобрать U и V, минимизируя функционал.

Необходимо заметить, что в SVD можно по-разному распределить диагональную матрицу между двумя другими. Пусть $X = AB$, тогда домножение на RR^{-1} в середине не меняет итоговое произведение, следовательно, матрицы:

$$A' = AR, \quad B' = R^{-1}B$$

тоже будут решением задачи, если матрица R не портит свойства матриц A и B. Это демонстрирует неоднозначность разложения.