

Большой итоговый проект

Постройте свою рекомендательную систему для фильмов.

Возьмите датасет MovieLens-100K (открытый, 100 000 оценок, ~ 1000 пользователей $\times \sim 1700$ фильмов). Постройте матрицу «пользователь–фильм» R , где $R_{ij} \in \{1, \dots, 5\}$ — оценка, поставленная пользователем i фильму j , а 0 — «не смотрел».

Важно: нули в матрице R означают *пропущенные* оценки, а не настоящие нулевые рейтинги. Прямое SVD по матрице с нулями — лишь грубый учебный baseline: непросмотренные фильмы тогда воспринимаются как фильмы с очень низкой оценкой. Более корректный вариант — обучать малоранговую модель только на известных оценках либо предварительно заполнить пропуски средними значениями пользователя/фильма и затем сравнить результат.

1. Постройте малоранговое приближение $R \approx U \Sigma \bar{V}$ с $k = 20$ компонентами через SVD.
2. Объясните содержательно, что такое строки U (профили пользователей) и столбцы V (профили фильмов) в этом приближении. Сколько чисел теперь хранит модель вместо $1000 \cdot 1700$?
3. Для нескольких реальных пользователей выпишите топ-5 фильмов, которые рекомендация модели предлагает посмотреть. Совпадают ли они с любимыми жанрами пользователя?
4. Найдите два «противоположных» фильма с наибольшим расстоянием между их столбцами V и подумайте, можно ли увидеть содержательный смысл в этой разделяющей оси (боевик vs. мелодрама? старое vs. новое?).
5. (★) Замените SVD на алгоритм PageRank, построив граф «фильмы, которые часто оценивают одинаково». Получится ли тот же топ?

! Задачи для самостоятельной работы

1. Сложите в Python 10 000 000 копий числа 0, 1. Сколько должно получиться? Сколько получается на самом деле? Почему?
2. Вычислите $f(x) = \sqrt{x+1} - \sqrt{x}$ при $x = 10^{16}$ напрямую и через формулу $f(x) = \frac{1}{\sqrt{x+1} + \sqrt{x}}$. Сравните ответы. Какой устойчив, почему?
3. Найдите вручную SVD матрицы $A = \begin{pmatrix} 30 & 04 \\ 01 & -10 \end{pmatrix}$ и U, Σ, V . Что в обоих случаях геометрически делают U, Σ, V ?
4. Возьмите свою фотографию 300×300 в градациях серого. Сделайте сжатие через SVD с рангами $k = 5, 20, 50, 100$. При каком k вы перестаёте узнавать самого себя?
5. Для игрушечного графа из рис. 0.9 прокрутите 5 итераций степенного метода *вручную*, начав с $r^{(0)} = (\frac{1}{6}, \dots, \frac{1}{6})$. Совпали ли ваши значения с программными после 5 шагов? а после 20?
6. * Покажите, что для функции $\sigma(x) = \max(0, x)$ (ReLU) константа Липшица равна 1. Тот же вопрос для сигмоиды $\sigma(x) = 1/(1 + e^{-x})$: чему равна $\text{Lip}(\sigma)$ и в какой точке производная максимальна?
7. * Найдите спектральную норму матрицы $W = \begin{pmatrix} 12 & 21 \end{pmatrix}$ двумя способами: (а) через определение $\|W\|_2 = \max_{\|x\|=1} \|Wx\|$ (с использованием собственных значений $\bar{W}W$); (б) одной итерацией степенного метода (с любым стартом). Совпали ли ответы?