НЕТОЛОГИЯ
групп

РЕКОМЕНДАЦИИ НА ОСНОВЕ СКРЫТЫХ ФАКТОРОВ

Артур Сапрыкин

ПЛАН ЗАНЯТИЯ

ПЛАН ЗАНЯТИЯ

1

SVD и снижение размерности

2

ALS для explicit feedback

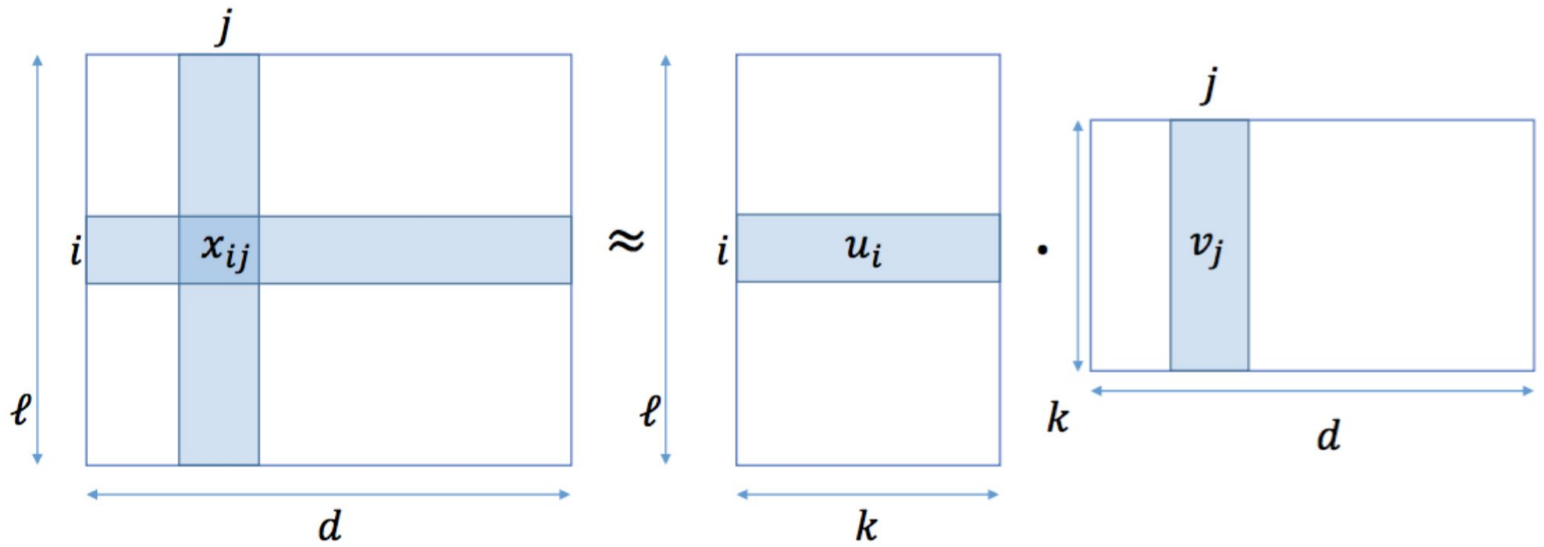
3

ALS для implicit feedback

SVD И СКРЫТЫЕ ФАКТОРЫ

Приближение матрицы с помощью SVD

$$X_{l,n} \approx U_{l,k} \cdot V^T_{k,n}$$



Поиск минимума

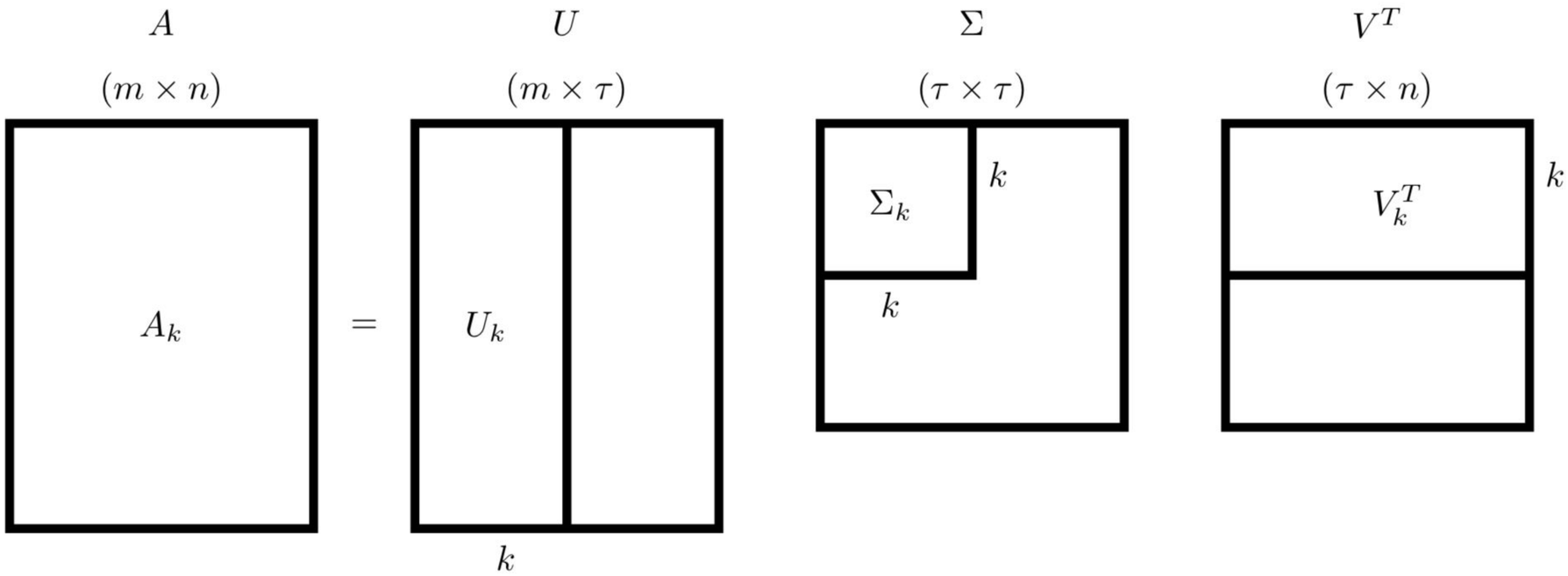
$$\|X - U \cdot V^T\| \to \min$$

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$ - Норма Фробениуса

$$\sum_{i,j} (x_{ij} - \langle u_i, v_j \rangle)^2 \to \min$$

Сингулярное разложение

$$X = \tilde{U}\Sigma\tilde{V}^T$$



Наилучшее решение по норме Фробениуса

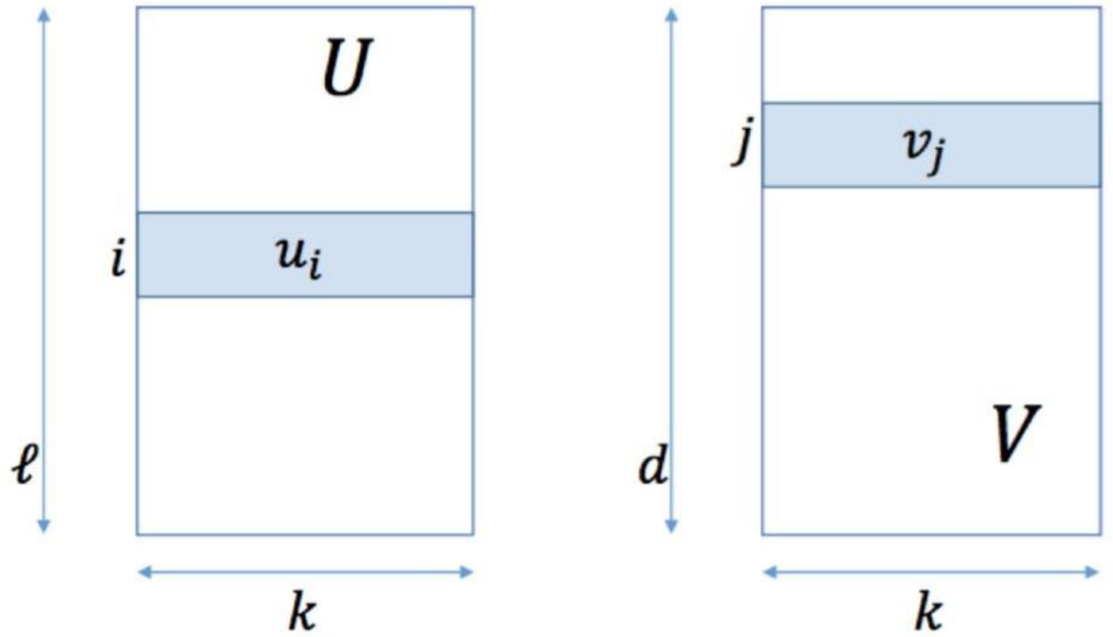
$$U = \tilde{U}_k \Sigma_k, \quad V = \tilde{V}_k,$$

$$U = \tilde{U}_k, \quad V = \tilde{V}_k \Sigma_k$$

$$U = \tilde{U}_k \sqrt{\Sigma_k}, \quad V = \tilde{V}_k \sqrt{\Sigma_k}.$$

SVD порождает скрытые факторы

$$\sum_{i,j} (x_{ij} - \langle u_i, v_j \rangle)^2 \to \min,$$



Резюме по SVD

- линейная модель предпочтений (рейтингов)
- “скрытые” факторы нельзя интерпретировать
- плохо параллелится на больших данных

ALS

Минимизация функционала

$$Q = \sum_{i,j}(\langle u_i, v_j \rangle - x_{ij})^2 \to \min_{u_i, v_j}$$

Минусы SGD

- сходится медленно
- неясно, как выбирать шаг для спуска

Метод ALS

- Считаем первую производную, приравниваем к 0 и вычисляем новое значение

$$\frac{\partial Q}{\partial u_i} = 0$$

- Считаем вторую производную, приравниваем к 0 и вычисляем новое значение

$$\frac{\partial Q}{\partial v_i} = 0$$

Использование регуляризации

$$Q = \sum_{i,j}(\langle u_i, v_j\rangle - x_{ij})^2 + \alpha \sum_i \|u_i\|^2 + \beta \sum_j \|v_j\|^2 \to \min_{u_i, v_j},$$

где α и β — небольшие положительные числа (0.001, 0.01, 0.5)

Резюме по ALS

- каждая итерация параллелится на множество подзадач
- каждая подзадача выпукла
- в каждой подзадаче [количество факторов] параметров

Чем ALS лучше, чем SGD

- ALS на каждой итерации уменьшает функцию ошибки
- ALS - разновидность координатного спуска
- не нужно выбирать learning rate

SVD ДЛЯ EXPLICIT FEEDBACK

Добавление дополнительных параметров

$$x_{ij} \approx \mu + \langle u_i, v_j \rangle,$$

$$\sum_{i,j} (\mu + \langle u_i, v_j \rangle - x_{ij})^2 \to \min$$

Добавление дополнительных параметров

$$x_{ij} \approx \mu + b_i^u + b_j^v + \langle u_i, v_j \rangle$$

$$\sum_{i,j} (\mu + b_i^u + b_j^v + \langle u_i, v_j \rangle - x_{ij})^2 \to \min$$

Регуляризация

$$\sum_{i,j}(\mu + b_i^u + b_j^v + \langle u_i, v_j \rangle - x_{ij})^2 + \alpha \sum_i \|u_i\|^2 + \beta \sum_j \|v_j\| + \gamma \sum_i {b_i^u}^2 + \delta \sum_j {b_j^v}^2 \to \min$$

ПРАКТИКА

ПРАКТИКА

SVD в surprise

Задача - рекомендации на главной странице сервиса в разделе “Персональная подборка”

Что делать?

1. Датасет тот же ml-latest
2. Использовать SVD из surprise
3. Взять любого пользователя и посмотреть на результаты предсказаний
4. Поэкспериментировать с количеством скрытых факторов и количеством эпох

Сколько есть времени?

15 минут

IMPLICIT ALS

Рекомендации товаров

	Вечернее платье	Поднос для писем	iPhone 6s	Шуба D&G
Маша	1		1	
Юля	1	1		1
Вова		1	1	
Коля	1	?	1	
Петя		1	1	
Ваня			1	1

Проблемы

- присутствуют только положительные примеры
- нет товаров, про которые точно известно, что пользователь их никогда не купит
- невозможно понять - не купил потому что не увидел или потому что не понравилось

Такой же подход

$$x_{ij} = 1 \approx \langle u_i, v_j \rangle$$

$$\sum_{i,j:x_{ij}\neq 0} (\langle u_i, v_j \rangle - x_{ij})^2 \to \min$$

$$u_i = \frac{1}{\sqrt{d}}(1 \dots 1)$$

$$v_j = \frac{1}{\sqrt{d}}(1 \dots 1)$$

Новая функция ошибок

$$\sum_{i,j} w_{ij}(\langle u_i, v_j \rangle - x_{ij})^2 \to \min$$

$$w_{ij} = 1 + \alpha|x_{ij}|$$

Обычно подбирается порядок значения: 10, 100, 100

Новая функция ошибок

- было - суммирование только по наблюдаемым парам (user, item)
- стало - суммирование **по всем возможным** парам (user, item)

ПРАКТИКА

ПРАКТИКА

Implicit ALS RS в библиотеке implicit

Задача - персональная подборка музыки для last.fm

Что делать?

1. Датасет - last.fm-360k - https://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-360K.html
2. Использовать ALS из implicit
3. Заполните профиль/выберете пользователя и получите для него персональную подборку

Сколько есть времени?

25 минут

ВОПРОСЫ