МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

МФТИ

ПЛАН

- Матричные разложения
- SVD и “SVD”
- Рекомендации
- Тексты
- Особенности обозначений

МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

$$\underset{l \times n}{\boldsymbol{X}} \approx \underset{l \times k}{\boldsymbol{U}} \cdot \underset{k \times n}{\boldsymbol{V}^T}$$

$$\|\boldsymbol{X} - \boldsymbol{U} \cdot \boldsymbol{V}^T\| \to \mathbf{min}$$

МФТИ

МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

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

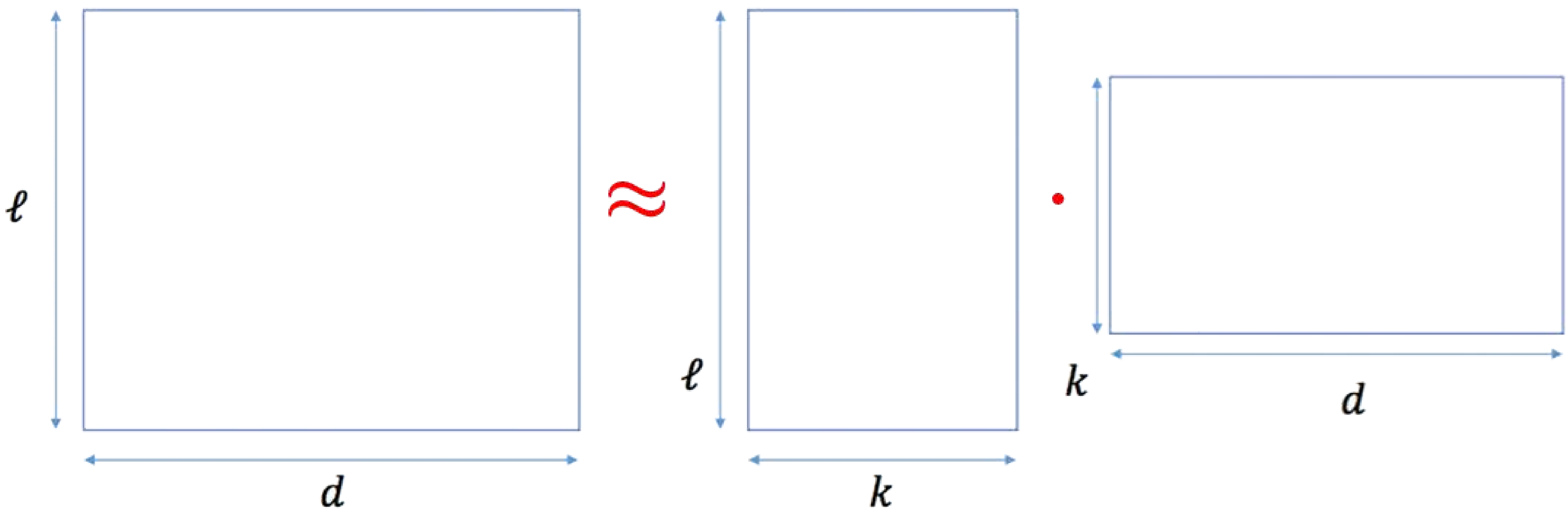
МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

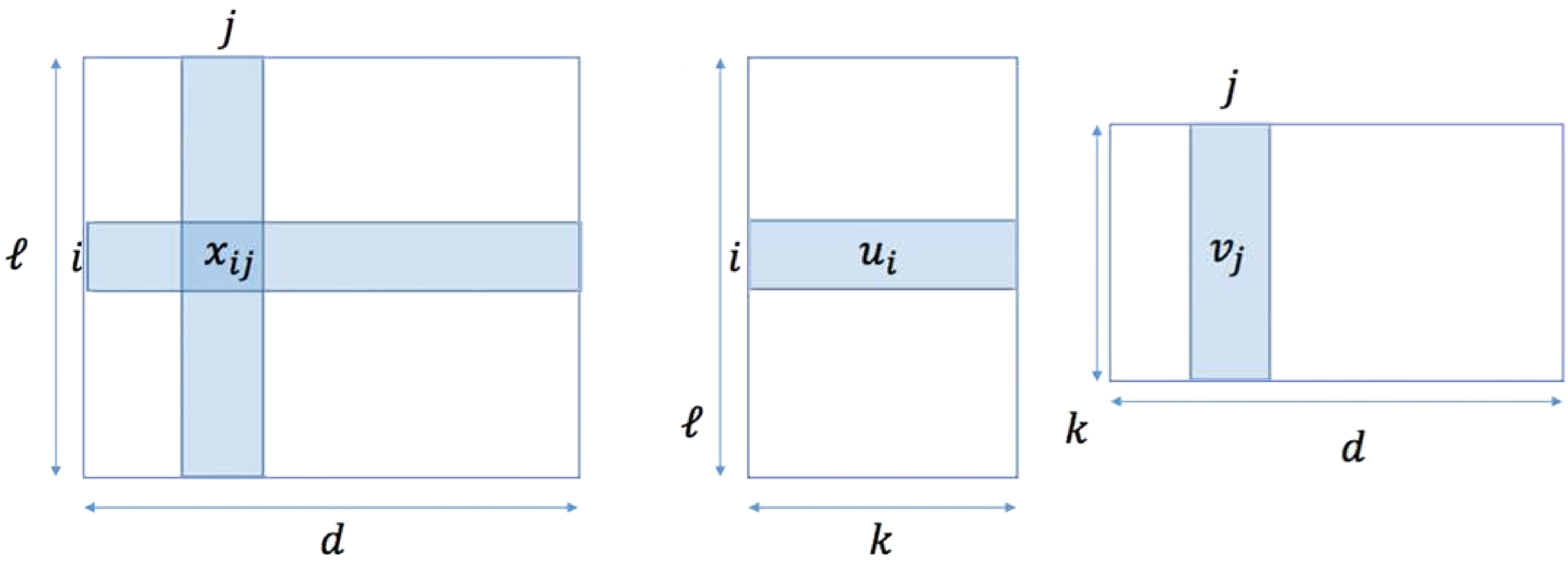
$$\sum_{i,j} (x_{ij} - \langle u_i, v_j \rangle)^2 \to \min$$

Разбираемся с обозначениями



$$X \approx U \cdot V^T$$

Разбираемся с обозначениями



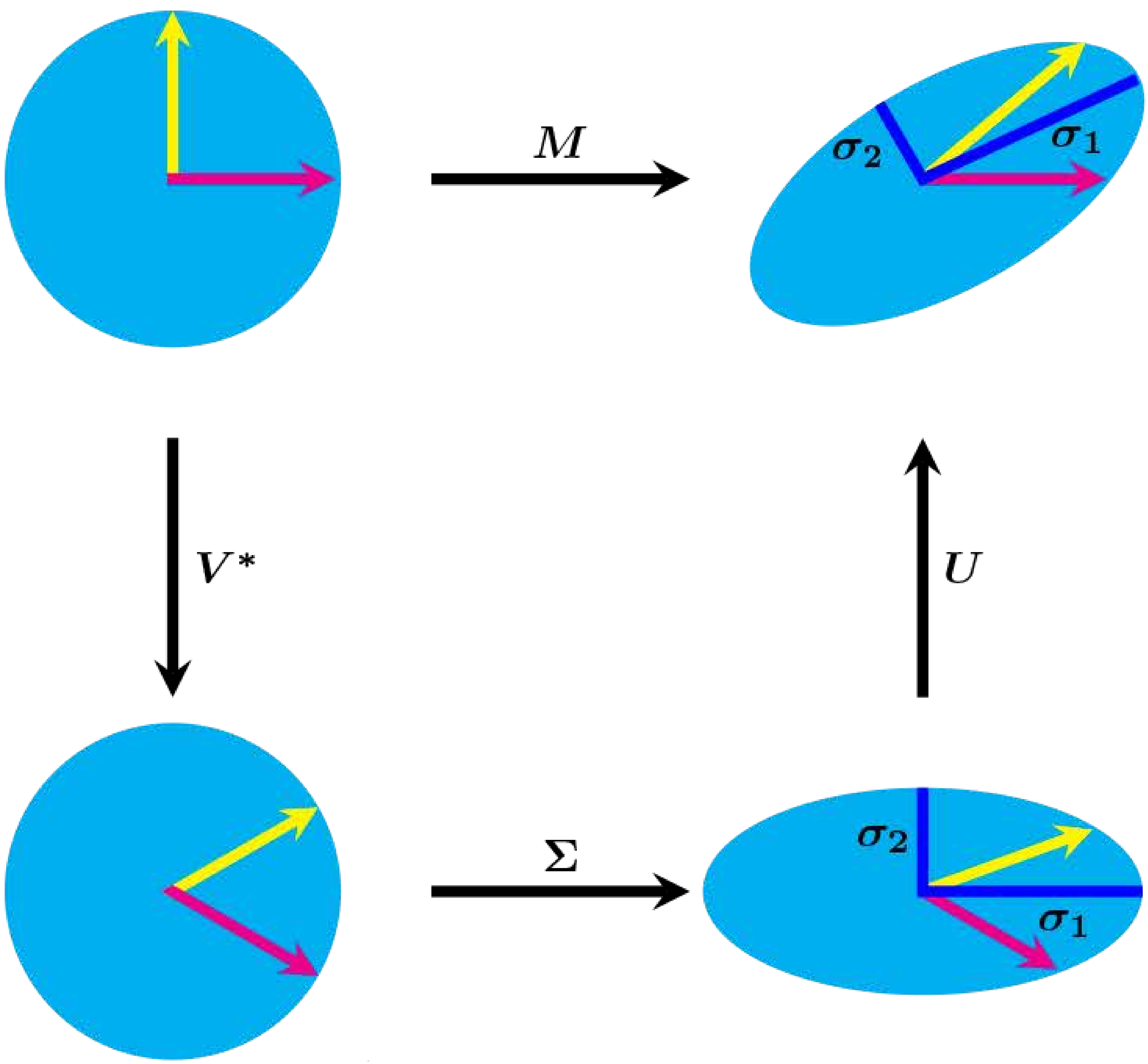
$$x_{ij} = \langle u_i, v_j \rangle$$

SVD В ЛИНЕЙНОЙ АЛГЕБРЕ

- $X = U\Sigma V^T$
- U — ортогональная
- Σ — диагональная
- V — ортогональная

SVD В ЛИНЕЙНОЙ АЛГЕБРЕ

- $X = U\Sigma V^T$



ПРИБЛИЖЕНИЕ МАТРИЦЫ С ПОМОЩЬЮ SVD

$$\underset{l \times n}{\boldsymbol{X}} \approx \underset{l \times k}{\boldsymbol{U}} \cdot \underset{k \times n}{\boldsymbol{V}^T}$$

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

ПРИБЛИЖЕНИЕ МАТРИЦЫ С ПОМОЩЬЮ SVD

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

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

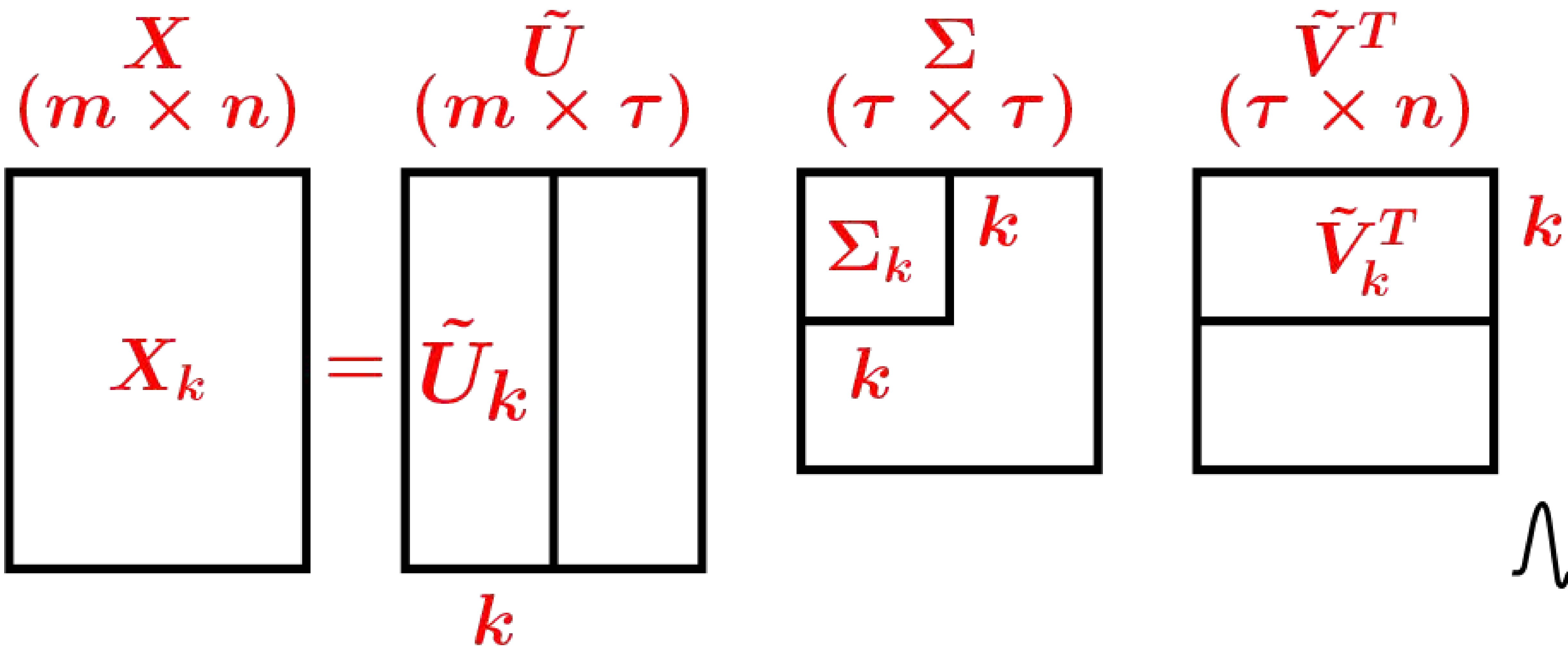
МФТИ

ПРИБЛИЖЕНИЕ МАТРИЦЫ С ПОМОЩЬЮ SVD

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

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



ПРИБЛИЖЕНИЕ МАТРИЦЫ С ПОМОЩЬЮ SVD

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

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

- $\tilde{U}_k, \Sigma_k, \tilde{V}_k$ — усеченные матрицы из SVD

$$U = \tilde{U}_k \Sigma_k \qquad V = \tilde{V}_k$$

МФТИ

Приближение матрицы с помощью SVD

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

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

- $\tilde{U}_k, \Sigma_k, \tilde{V}_k$ — усеченные матрицы из SVD

$$U = \tilde{U}_k \qquad V = \tilde{V}_k \Sigma_k$$

МФТИ

ПРИБЛИЖЕНИЕ МАТРИЦЫ С ПОМОЩЬЮ SVD

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

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

- $\tilde{U}_k, \Sigma_k, \tilde{V}_k$ — усеченные матрицы из SVD

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

МФТИ

“SVD” В МАШИННОМ ОБУЧЕНИИ

$$\underset{l \times n}{X} \approx \underset{l \times k}{U} \cdot \underset{k \times n}{V^T}$$

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

- u_i — “профили” объектов
- v_j — “профили” исходных признаков

МФТИ

МАТРИЦА РЕЙТИНГОВ И SVD

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля	5	5	2	
Вова			3	5
Коля	3	?	4	5
Петя				4
Ваня		5	3	3

МАТРИЦА РЕЙТИНГОВ И SVD

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля	5	5	2	
Вова			3	5
Коля	3	?	4	5
Петя				4
Ваня		5	3	3

МАТРИЦА РЕЙТИНГОВ И SVD

j

	Пила	Улица Вязов	Ванильное небо	1+1	
Маша	5	4	1	2	
Юля	5	5	2		
Вова			3	5	i
Коля	3	?	4	5	
Петя				4	
Ваня		5	3	3	

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

- u_i — “интересы пользователей”
- v_j — ”параметры фильмов”

МАТРИЦА ЧАСТОТ СЛОВ И SVD

	database	SQL	index	regression	likelihood	linear
d1	24	21	9	0	0	3
d2	32	10	5	0	3	0
d3	12	16	5	0	0	0
d4	6	7	2	0	0	0
d5	43	31	20	0	3	0
d6	2	0	0	18	7	16
d7	0	0	1	32	12	0
d8	3	0	0	22	4	2
d9	1	0	0	34	27	25
d10	6	0	0	17	4	23

МАТРИЦА ЧАСТОТ СЛОВ И SVD

	database	SQL	index	regression	likelihood	linear	
d1	24	21	9	0	0	3	
d2	32	10	5	0	3	0	
d3	12	16	5	0	0	0	
d4	6	7	2	0	0	0	
d5	43	31	20	0	3	0	i
d6	2	0	0	18	7	16	
d7	0	0	1	32	12	0	
d8	3	0	0	22	4	2	
d9	1	0	0	34	27	25	
d10	6	0	0	17	4	23	
					j		

МАТРИЦА ЧАСТОТ СЛОВ И SVD

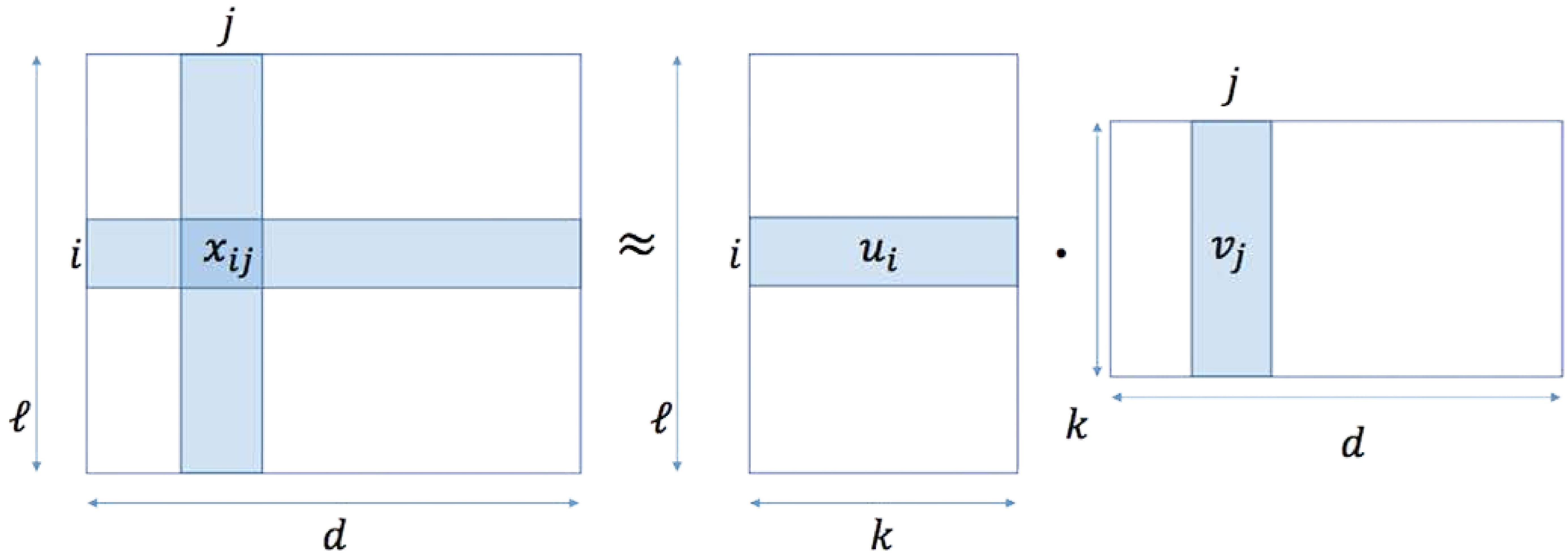
	database	SQL	index	regression	likelihood	linear	
d1	24	21	9	0	0	3	
d2	32	10	5	0	3	0	
d3	12	16	5	0	0	0	
d4	6	7	2	0	0	0	
d5	43	31	20	0	3	0	i
d6	2	0	0	18	7	16	
d7	0	0	1	32	12	0	
d8	3	0	0	22	4	2	
d9	1	0	0	34	27	25	
d10	6	0	0	17	4	23	
					j		

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

- u_i — “темы” документов
- v_j — “темы” слов

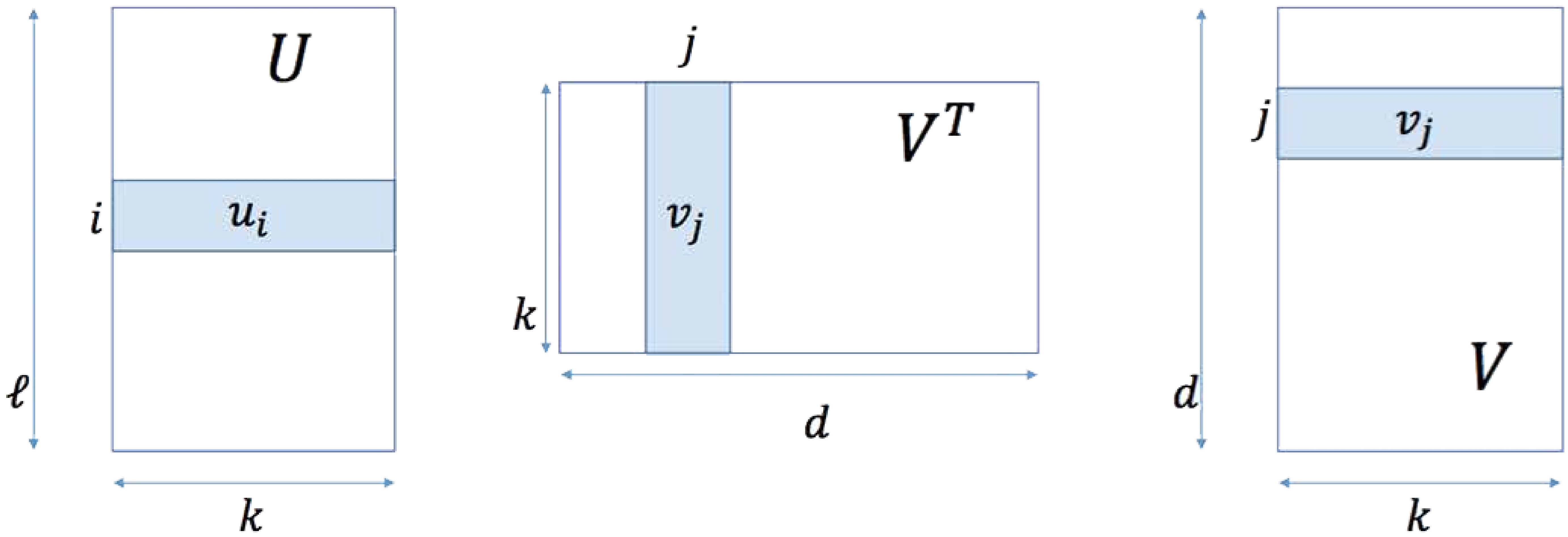
МФТИ

Еще немного про обозначения



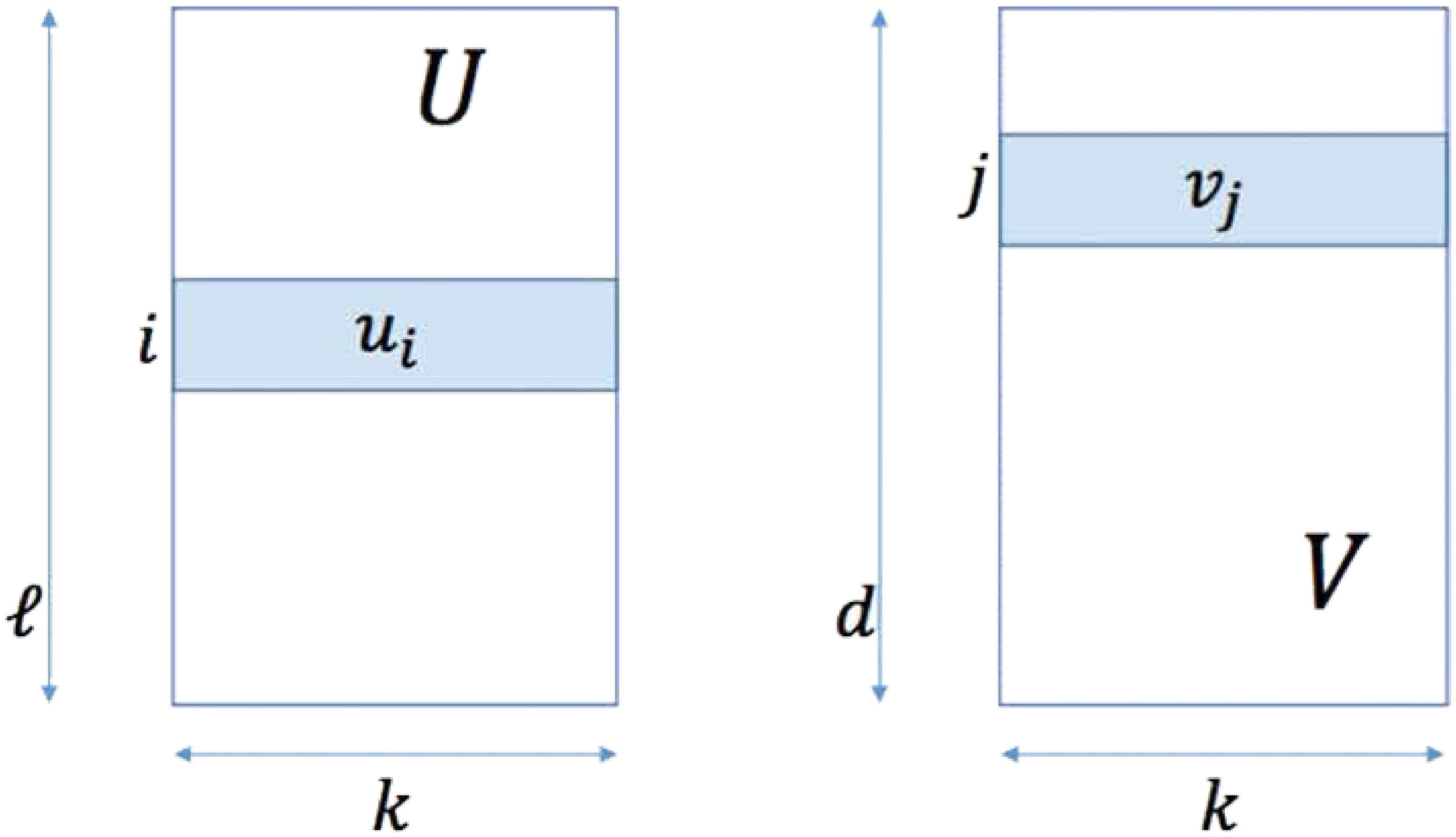
$$x_{ij} \approx \langle u_i, v_j \rangle$$

ЕЩЕ НЕМНОГО ПРО ОБОЗНАЧЕНИЯ



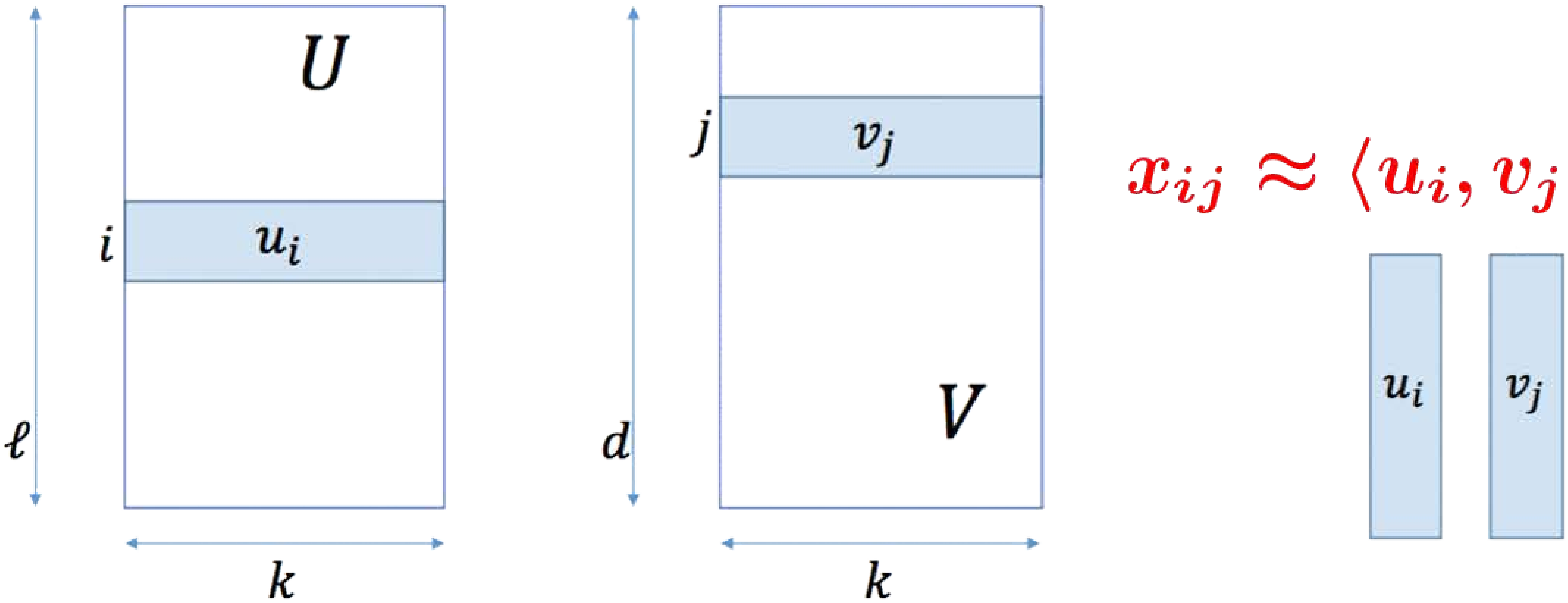
$$x_{ij} \approx \langle u_i, v_j \rangle$$

ЕЩЕ НЕМНОГО ПРО ОБОЗНАЧЕНИЯ



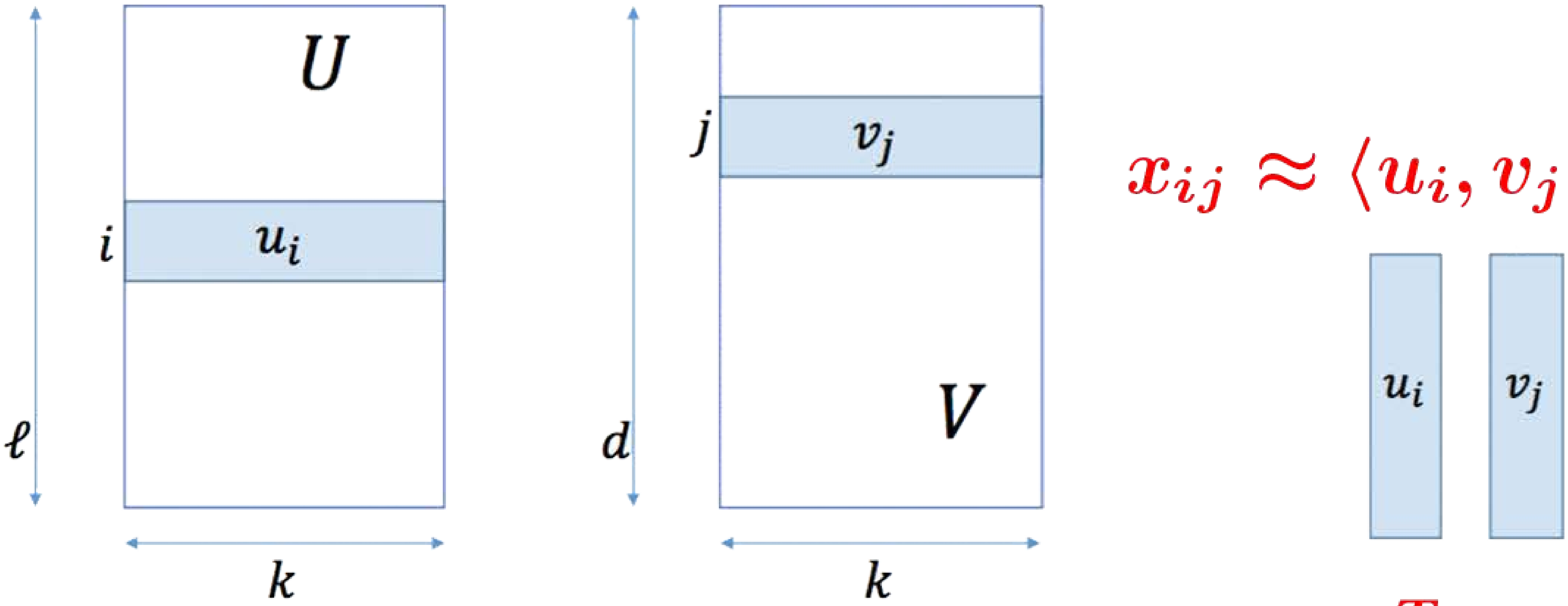
$$x_{ij} \approx \langle u_i, v_j \rangle$$

ЕЩЕ НЕМНОГО ПРО ОБОЗНАЧЕНИЯ



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

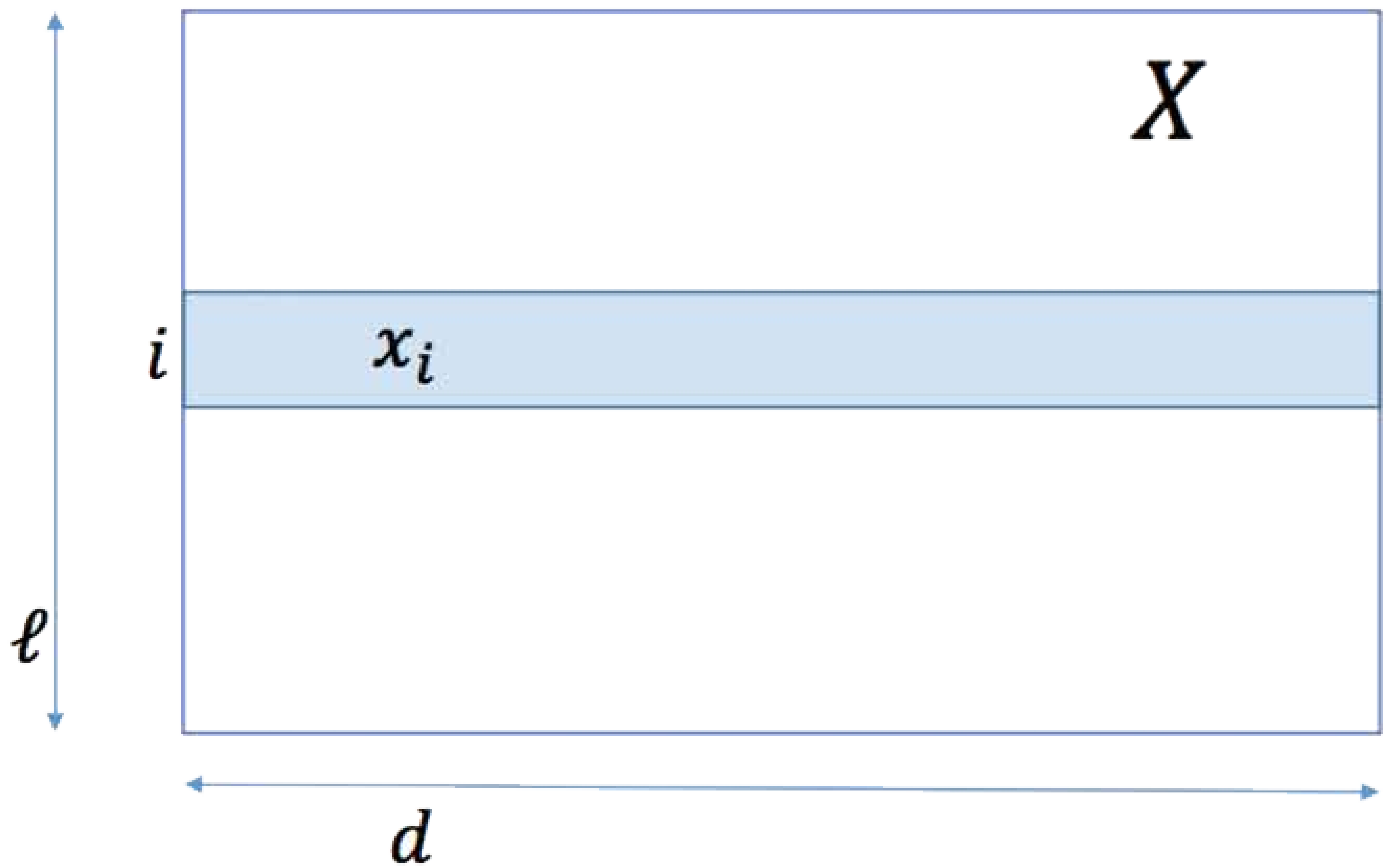
ЕЩЕ НЕМНОГО ПРО ОБОЗНАЧЕНИЯ



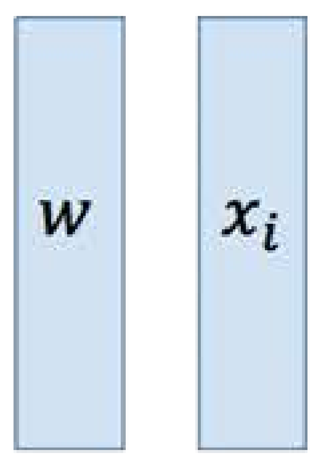
$$x_{ij} \approx \langle u_i, v_j \rangle$$

$$x_{ij} = u_i^T v_j$$

Аналогия с матрицей признаков



- В линейных моделях: $\langle \mathbf{w}, x_i \rangle = \mathbf{w}^T x_i$



КАКИЕ ОБОЗНАЧЕНИЯ ВСТРЕЧАЮТСЯ

- $X = UV^T$
- $X = PQ^T$
- $X = WH^T$
- $X = \Phi\Theta$

SGD и ALS

ПЛАН

- Постановка задачи
- Градиентный спуск
- Стохастический градиентный спуск
- ALS
- Регуляризация

ПОСТАНОВКА ЗАДАЧИ

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

Градиентный спуск (GD)

- $$Q = \sum_{i,j} (\langle u_i, v_j \rangle - x_{ij})^2 \to \min_{u_i, v_j}$$
- $$\frac{\partial Q}{\partial u_i} = \sum_{i,j} \frac{\partial}{\partial u_i} (\langle u_i, v_j \rangle - x_{i,j})^2 =$$
$$= \sum_j 2(\langle u_i, v_j \rangle - x_{ij}) \frac{\partial \langle u_i, v_j \rangle}{\partial u_i} =$$
$$= \sum_j 2(\langle u_i, v_j \rangle - x_{ij}) v_j$$
- $\varepsilon_{ij} = (\langle u_i, v_j \rangle - x_{ij})$ — ошибка на x_{ij}

МФТИ

Градиентный спуск (GD)

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

  $$\frac{\partial Q}{\partial u_i} = \sum_j 2(\langle u_i, v_j \rangle - x_{ij}) v_j$$

- $\varepsilon_{ij} = (\langle u_i, v_j \rangle - x_{ij})$ — ошибка на x_{ij}

$$\boxed{u_i^{(t+1)} = u_i^{(t)} - \gamma_t \sum_j \varepsilon_{ij} v_j}$$

Стохастический градиентный спуск (SGD)

GD:

$$u_i^{(t+1)} = u_i^{(t)} - \gamma_t \cancel{\sum_j} \varepsilon_{ij} v_j$$

$$v_j^{(t+1)} = v_j^{(t)} - \eta_t \cancel{\sum_i} \varepsilon_{ij} u_i$$

SGD:

$$u_i^{(t+1)} = u_i^{(t)} - \gamma_t \varepsilon_{ij} v_j$$

$$v_j^{(t+1)} = v_j^{(t)} - \eta_t \varepsilon_{ij} u_i$$

Для случайных i, j

ПЛЮСЫ И МИНУСЫ SGD

- Плюсы:
  - Простота реализации
  - Сходимость
- Минусы:
  - Медленно сходится
  - Сложность выбора шага градиентного спуска (γ_t и η_t)

МФТИ

ПЛЮСЫ И МИНУСЫ SGD

- Минусы:
  - Медленно сходится
  - Сложность выбора шага градиентного спуска (γ_t и η_t)
  - При константном шаге сходится очень медленно

Идея ALS

$$Q \to \min_{u_i, v_j}$$

- Повторяем до сходимости:

$$\frac{\partial Q}{\partial u_i} = 0 \Rightarrow u_i \qquad \frac{\partial Q}{\partial v_j} = 0 \Rightarrow v_j$$

Выписываем шаг в ALS

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

- $$\frac{\partial Q}{\partial u_i} = \sum_j 2(\langle u_i, v_j \rangle - x_{i,j}) v_j = 0$$
$$\sum_j v_j \langle v_j, u_i \rangle = \sum_j x_{ij} v_j$$
- $$\sum_j v_j v_j^T u_i = \sum_j x_{ij} v_j$$
- $$\left(\sum_j v_j v_j^T \right) u_i = \sum_j x_{ij} v_j$$

МФТИ

ALS: ИТОГОВЫЙ АЛГОРИТМ

- Повторяем по случайным i, j до сходимости:

$$\left(\sum_j v_j v_j^T\right) u_i = \sum_j x_{ij} v_j \Rightarrow u_i$$

$$\left(\sum_i u_i u_i^T\right) v_j = \sum_i x_{ij} u_i \Rightarrow v_j$$

(решение системы линейных уравнений)

МФТИ

Регуляризация

- $$Q = \sum_{i,j} (\langle u_i, v_j \rangle - x_{ij})^2 + \alpha \sum_i \|u_i\|^2 + \beta \sum_j \|v_j\|^2 \to \min_{u_i, v_j}$$
- α и β — небольшие положительные числа (0.001, 0.01, 0.5)

Резюме

- Постановка задачи
- Градиентный спуск
- Стохастический градиентный спуск
- ALS
- Регуляризация

ПРОГНОЗИРОВАНИЕ НЕИЗВЕСТНЫХ ЗНАЧЕНИЙ В МАТРИЦЕ

ПЛАНЫ

- Постановка задачи в рекомендациях
- Модель прогнозирования неизвестных значений
- Оптимизируемый функционал
- Добавление сдвига
- Добавление базовых предикторов

ПЛАНЫ

- Модель прогнозирования неизвестных значений
- Оптимизируемый функционал
- Добавление сдвига
- Добавление базовых предикторов
- Регуляризация по базовым предикторам

ПОСТАНОВКА ЗАДАЧИ В РЕКОМЕНДАЦИЯХ

j

		Пила	Улица Вязов	Ванильное небо	1+1
	Маша	5	4	1	2
	Юля	5	5	2	
	Вова			3	5
i	Коля	3	?	4	5
	Петя				4
	Ваня		5	3	3

Модель прогнозирования

j

		Пила	Улица Вязов	Ванильное небо	1+1
	Маша	5	4	1	2
	Юля	5	5	2	
	Вова			3	5
i	Коля	3	?	4	5
	Петя				4
	Ваня		5	3	3

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

u_i — «интересы пользователей»

v_j — «параметры фильмов»

МФТИ

ОПТИМИЗИРУЕМЫЙ ФУНКЦИОНАЛ

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

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

Сдвиг

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

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

БАЗОВЫЕ ПРЕДИКТОРЫ

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

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

Регуляризация

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

$$+\alpha \sum_i \|u_i\|^2 + \beta \sum_j \|v_j\| +$$

$$\boxed{+\gamma \sum_i {b_i^u}^2 + \delta \sum_j {b_j^v}^2} \to \min$$

Резюме

- Постановка задачи в рекомендациях
- Модель прогнозирования неизвестных значений
- Оптимизируемый функционал
- Добавление сдвига
- Добавление базовых предикторов

Резюме

- Модель прогнозирования неизвестных значений
- Оптимизируемый функционал
- Добавление сдвига
- Добавление базовых предикторов
- Регуляризация по базовым предикторам

НЕГАТИВНЫЕ ПРИМЕРЫ
И IMPLICIT РАЗЛОЖЕНИЯ

МФТИ

ПЛАН

- Рекомендации товаров
- Почему нельзя всё делать как обычно
- Implicit feedback
- Implicit matrix factorization: идея
- Implicit ALS

Рекомендации товаров

j

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

ПОЧЕМУ НУЖНО ЧТО-ТО МЕНЯТЬ

j

	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

i

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

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

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

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

МФТИ

EXPLICIT И IMPLICIT

- **Explicit feedback:** есть положительные и отрицательные примеры (например, низкие и высокие оценки фильмов, лайки и дислайки и т.д.)

- **Implicit feedback:** есть только положительные (покупки, просмотры, лайки) или только отрицательные примеры (дислайки)

IMPLICIT MATRIX FACTORIZATION

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

Сумма по всем индексам (не только по известным элементам матрицы)

- w_{ij} принимает большие значения для $x_{ij} \neq 0$ и значительно меньшие для $x_{ij} = 0$

МФТИ

IMPLICIT ALS

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

$$\mathrm{w}_{ij} = 1 + \alpha |x_{ij}| \qquad \alpha = 10, 100, 1000$$

u_i, v_j оцениваем с помощью ALS

РЕЗЮМЕ

- Рекомендации товаров
- Почему нельзя всё делать как обычно
- Implicit feedback
- Implicit matrix factorization: идея
- Implicit ALS

ВЕРОЯТНОСТНАЯ ИНТЕРПРЕТАЦИЯ

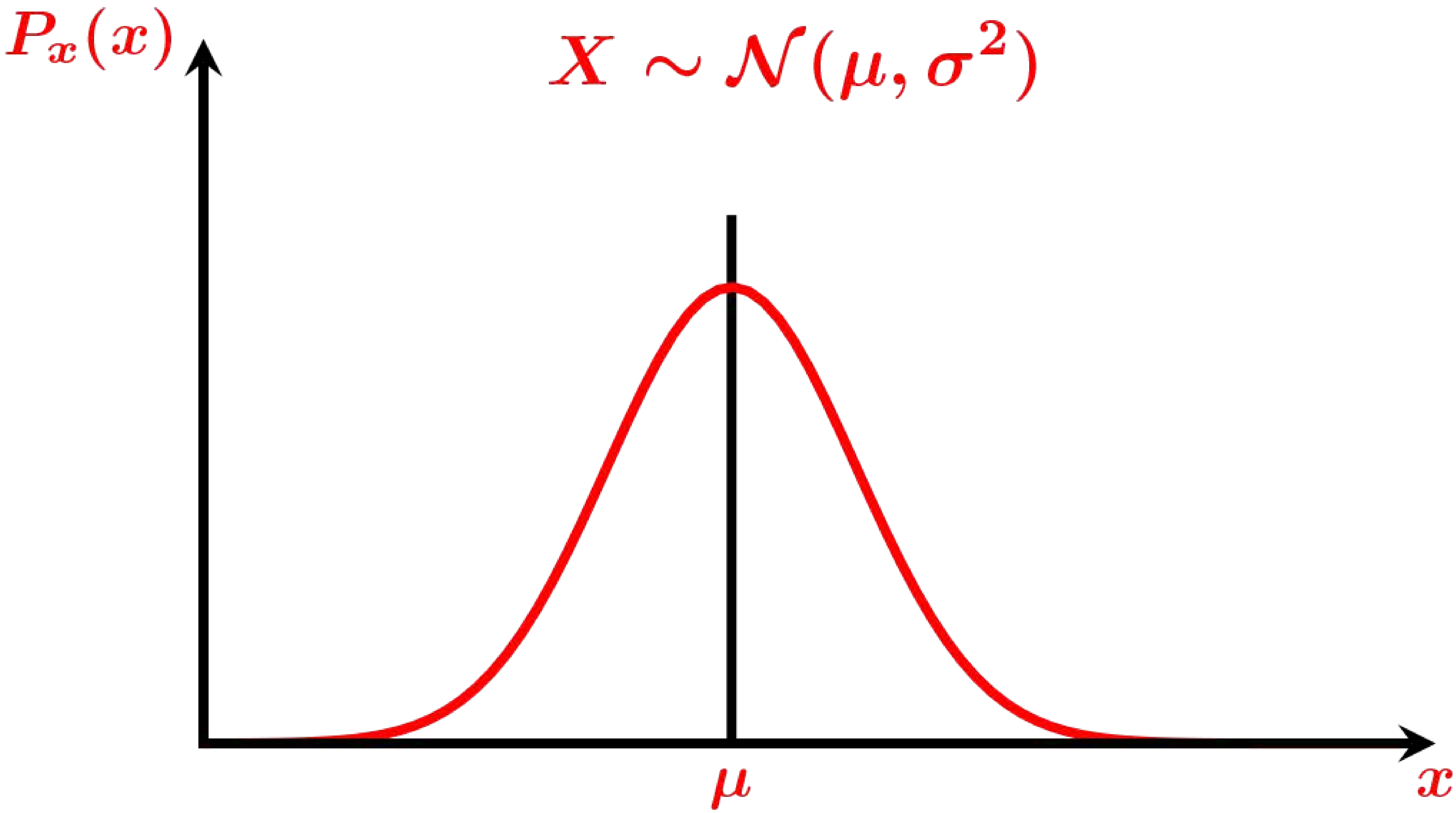
ПЛАН

- "SVD" и нормальное распределение
- Какое распределение подходит больше
- Распределение Пуассона
- Неотрицательное матричное разложение
- Другие неотрицательные матричные разложения

ПОСТАНОВКА ЗАДАЧИ В "SVD"

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

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ



- $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Связь "SVD" и нормального распределения

- $x_{ij} = \langle u_i, v_j \rangle + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$x_{ij} \sim \mathcal{N}(\langle u_i, v_j \rangle, \sigma^2)$$

- $\prod\limits_{i,j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_{ij} - \langle u_i, v_j \rangle)^2}{2\sigma^2}} \to \max$

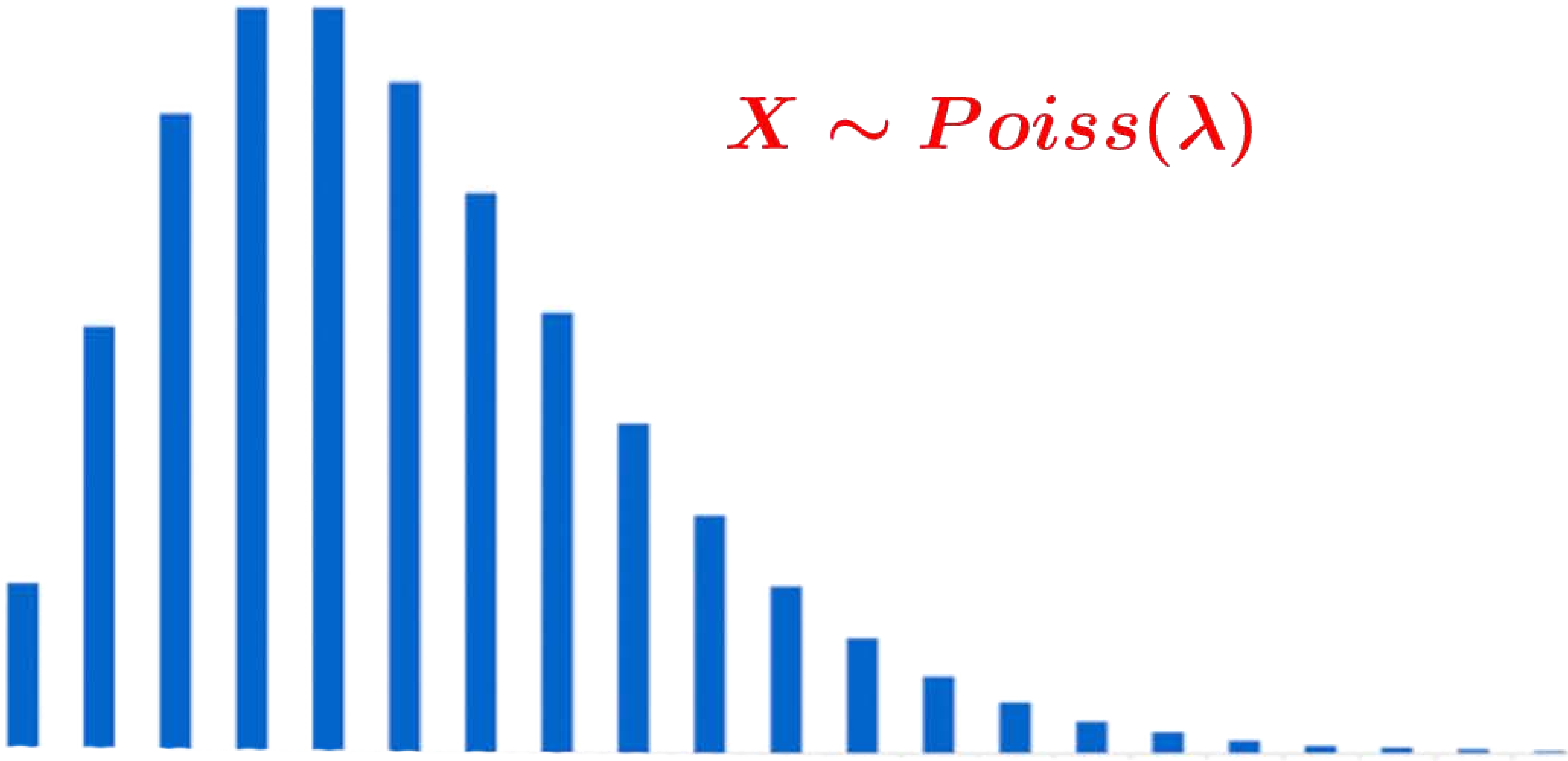
- $\sum\limits_{i,j} \frac{(x_{ij} - \langle u_i, v_j \rangle)^2}{2\sigma^2} - \frac{1}{2} \ln 2\pi\sigma^2 \to \min$

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

КАКОЕ РАСПРЕДЕЛЕНИЕ ПОДХОДИТ БОЛЬШЕ

	database	SQL	index	regression	likelihood	linear
d1	24	21	9	0	0	3
d2	32	10	5	0	3	0
d3	12	16	5	0	0	0
d4	6	7	2	0	0	0
d5	43	31	20	0	3	0
d6	2	0	0	18	7	16
d7	0	0	1	32	12	0
d8	3	0	0	22	4	2
d9	1	0	0	34	27	25
d10	6	0	0	17	4	23

Распределение Пуассона



$$X \sim Poiss(\lambda)$$

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad \mathbb{E}X = \lambda$$

МФТИ

РАСПРЕДЕЛЕНИЕ ПУАССОНА И МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

- $x_{ij} \sim Poiss\langle u_i, v_j \rangle$
- $P\left(x_{ij}\right) = \frac{\langle u_i, v_j \rangle^{x_{ij}}}{x_{ij}!} e^{-\langle u_i, v_j \rangle}$
- $\prod\limits_{i,j} \frac{\langle u_i, v_j \rangle^{x_{ij}}}{x_{ij}!} e^{-\langle u_i, v_j \rangle} \rightarrow \max$
- $\sum\limits_{i,j} \langle u_i, v_j \rangle - x_{ij} \ln \langle u_i, v_j \rangle \ln x_{ij}! \rightarrow \min$
- $\sum\limits_{i,j} \langle u_i, v_j \rangle - x_{ij} \ln \langle u_i, v_j \rangle \rightarrow \min$

МФТИ

SGD для NMF (Non-negative matrix factorization)

- $$Q = \sum_{i,j} \langle u_i, v_j \rangle - x_{ij} \ln \langle u_i, v_j \rangle \to \min$$
- $$\frac{\partial Q}{\partial u_i} = \sum_j v_j - \frac{x_{ij}}{\langle u_i, v_j \rangle} v_j =$$
$$= \sum_j \underbrace{\frac{\langle u_i, v_j \rangle - x_{ij}}{\langle u_i, v_j \rangle}}_{\tilde{\varepsilon}_{ij}} v_j \to \min$$

$\tilde{\varepsilon}_{ij}$ - “относительная ошибка прогноза”

SGD:
$$u_i^{(t+1)} = u_i^{(t)} - \gamma_t \tilde{\varepsilon}_{ij} v_j$$
$$v_j^{(t+1)} = v_j^{(t)} - \eta_t \tilde{\varepsilon}_{ij} u_i$$

ДРУГИЕ НЕОТРИЦАТЕЛЬНЫЕ МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

- Можно использовать норму Фробениуса, но добавить ограничения неотрицательности для $\boldsymbol{U}$ и $\boldsymbol{V}$:

$$Q = \sum_{i,j} (\langle u_i, v_j \rangle - x_{ij})^2 \to \min_{\substack{u_i, v_j: \\ u_{ik} \geq 0 \\ v_{jk} \geq 0}}$$

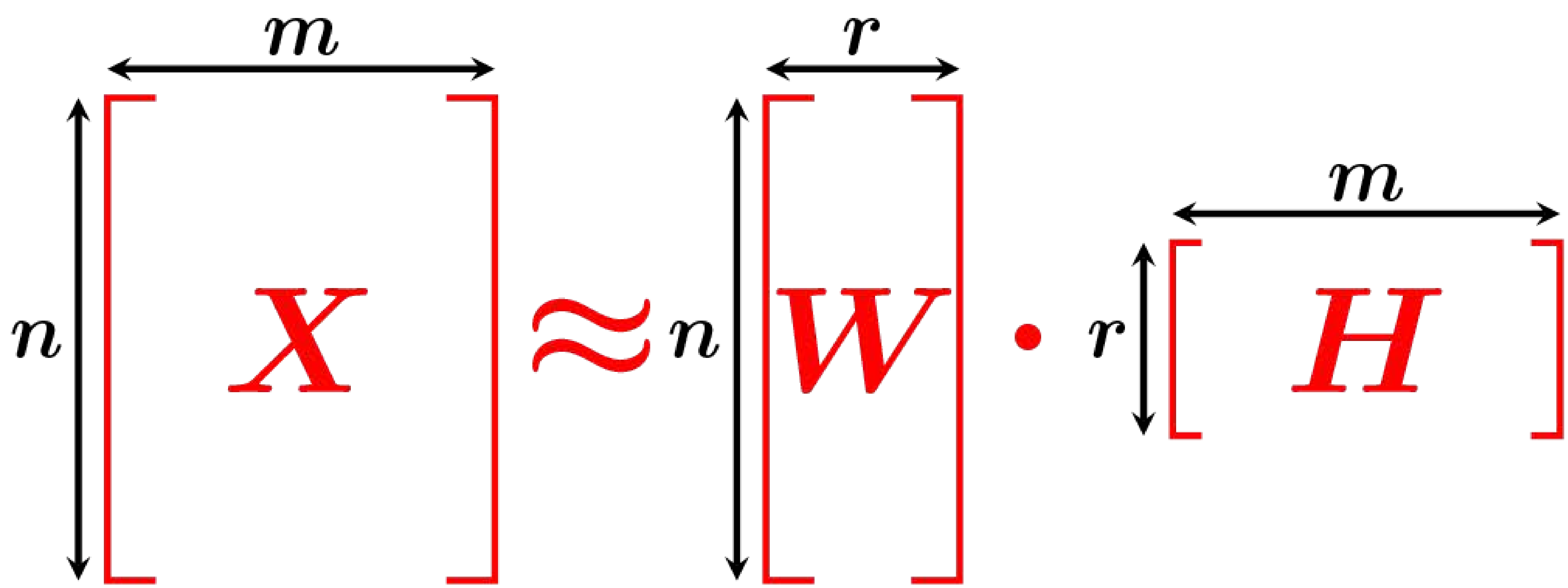
Резюме

- ″SVD″ и нормальное распределение
- Какое распределение подходит больше
- Распределение Пуассона
- Неотрицательное матричное разложение
- Другие неотрицательные матричные разложения

НЕОТРИЦАТЕЛЬНЫЕ МАТРИЧНЫЕ РАЗЛОЖЕНИЯ: ПОСТАНОВКА И РЕШЕНИЕ

МФТИ

МАТРИЧНЫЕ РАЗЛОЖЕНИЯ



$$r < \min(n, m)$$

$$(W^*, H^*) = \operatorname*{argmin}_{W,H} D(X, WH)$$

НЕОТРИЦАТЕЛЬНЫЕ МАТРИЧНЫЕ РАЗЛОЖЕНИЯ (NMF)

$$D\left(X, \hat{X}\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} d\left(x_{ij}, \hat{x}_{ij}\right)$$

- Обозначим $\hat{X} = WH$.

$$D\left(X, \hat{X}\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} d\left(x_{ij}, \hat{x}_{ij}\right)$$

- Чаще всего в качестве функции потерь берут норму Фробениуса:

$$d_F\left(x, \hat{x}\right) = \left(x - \hat{x}\right)^2$$

$$D_F\left(X, \hat{X}\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(x_{ij} - \hat{x}_{ij}\right)^2 \equiv \|X - WH\|_F^2$$

МФТИ

ПРОБЛЕМЫ NMF

- Некорректно поставлена: если W_0, H_0 — решение, то $W = W_0 Y$, $H = Y^{-1} H_0$ тоже может быть решением.
- $D(X, WH)$ не выпукла по совокупности аргументов $(W, H) \to$ блочно-покоординатные методы минимизации:
  Вход: $W^0 \geq 0$, $H^0 \geq 0$
  Цикл
  $$H \leftarrow f(X, W, H),$$
  $$W \leftarrow g(X, W, H)$$
  ($f(X, W, H) = g^T(X^T, H^T, W^T)$, поскольку задача симметрична)

МФТИ

NMF
С НОРМОЙ ФРОБЕНИУСА

- Оптимизационная задача:

$$(W^*, H^*) = \underset{W \geq 0,\, H \geq 0}{\operatorname{argmin}} \|X - WH\|_F^2$$

- Без ограничения неотрицательности решение можно было бы получить с помощью SVD.
- Базовый метод — поочерёдный градиентный спуск:

$$h_{kj} \leftarrow h_{kj} - \nu_{kj} \frac{\partial D_F}{\partial h_{kj}}$$

$$w_{ik} \leftarrow w_{ik} - \eta_{ik} \frac{\partial D_F}{\partial w_{ik}}$$

$$k = 1, \ldots, r, \quad i = 1, \ldots, n \;,$$
$$j = 1, \ldots, m$$

МФТИ

Мультипликативные обновления

- Идея: выбрать шаги градиентного спуска так, чтобы обновления стали мультипликативными.

$$\frac{\partial D_F}{\partial h_{kj}} = \sum_{i=1}^{n} w_{ik}\hat{x}_{ij} - \sum_{i=1}^{n} w_{ik}x_{ij}, \quad \nu_{kj} = \frac{h_{kj}}{\sum_{i=1}^{n} w_{ik}\hat{x}_{ij}}$$

$$h_{kj} \leftarrow h_{kj} - \frac{h_{kj}}{\sum_{i=1}^{n} w_{ik}\hat{x}_{ij}} \left(\sum_{i=1}^{n} w_{ik}\hat{x}_{ij} - \sum_{i=1}^{n} w_{ik}x_{ij} \right) =$$

$$= h_{kj} \frac{\sum_{i=1}^{n} w_{ik}x_{ij}}{\sum_{i=1}^{n} w_{ik}\hat{x}_{ij}}$$

Мультипликативные обновления

- В матричном виде:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \left(\boldsymbol{W}^T \boldsymbol{X}\right) \oslash \left(\boldsymbol{W}^T \hat{\boldsymbol{X}}\right)$$

$\otimes$ — поэлементное умножение матриц

$\oslash$ — поэлементное деление

- Функция потерь монотонно невозрастает

Мультипликативные обновления

- Методы с мультипликативными обновлениями популярны, потому что они:
  - просты в реализации;
  - хорошо масштабируются и легко приспосабливаются к работе с разреженными матрицами;
  - были предложены в самой первой работе по NMF.
- Скорость сходимости невелика, но её можно увеличить, если обновлять W и H по несколько раз подряд.

Метод попеременных наименьших квадратов

- Alternating Least Squares (ALS): на каждом шаге находится решение задачи наименьших квадратов по одной из компонент, затем проецируется на неотрицательную область.

$$H \leftarrow \max\left(\underset{H}{\operatorname{argmin}} \|X - WH\|_F^2, 0\right) = \\ = \max\left(\left(W^T W\right)^{-1} W^T X, 0\right)$$

- Метод быстрый, но грубый: итерационный процесс не сходится, функция потерь осциллирует. Можно использовать для инициализации других методов.

МФТИ

Метод попеременных неотрицательных наименьших квадратов

- Alternating Nonnegative Least Squares (ANLS): на каждом шаге точно находится покомпонентный минимум в неотрицательной области.

$$H \leftarrow \operatorname*{argmin}_{H \geq 0} \|X - WH\|_F$$

- Метод точный, но медленный: каждая итерация требует существенных вычислительных затрат. Можно использовать для уточнения решения, найденного более простыми методами.

МЕТОД ИЕРАРХИЧЕСКИХ ПОПЕРЕМЕННЫХ НАИМЕНЬШИХ КВАДРАТОВ

- Hierarchical alternating least squares (HALS): на каждом шаге точно находится минимум в неотрицательной области по столбцу w_k или строке h_k.

$$h_k \leftarrow \underset{h_k \geq 0}{\operatorname{argmin}} \|X - WH\|_F =$$

$$= \max\left(0, \frac{w_k^T X - \sum_{l \neq k} w_k^T w_l h_l}{w_k^T w_k}\right)$$

  - на каждом шаге вычислительные затраты небольшие
  - сходится быстрее метода мультипликативных обновлений
  - чувствителен к начальному приближению

Резюме

- Неотрицательное матричное разложение, постановка и проблемы задачи
- Методы решения для нормы Фробениуса

ДАЛЕЕ В ПРОГРАММЕ

- Другие функции потерь
- Инициализация

НЕОТРИЦАТЕЛЬНЫЕ МАТРИЧНЫЕ РАЗЛОЖЕНИЯ: ФУНКЦИОНАЛЫ И ИНИЦИАЛИЗАЦИЯ

МФТИ

НЕОТРИЦАТЕЛЬНЫЕ МАТРИЧНЫЕ РАЗЛОЖЕНИЯ (NMF)

$$(W^*, H^*) = \underset{W \geq 0, H \geq 0}{\operatorname{argmin}} D(X, WH)$$

Функции потерь

$$D\left(X,\hat{X}\right)=\sum_{i=1}^{n}\sum_{j=1}^{m}d\left(x_{ij},\hat{x}_{ij}\right)$$

$$d\left(x,\hat{x}\right)\geq 0,\ d\left(x,\hat{x}\right)=0\Leftrightarrow x=\hat{x}$$

Название	$d\left(x,\hat{x}\right)$
норма l_1	$d_1\left(x,\hat{x}\right)=\lvert x-\hat{x}\rvert$
норма Фробениуса	$d_F\left(x,\hat{x}\right)=\left(x-\hat{x}\right)^2$
обобщённая дивергенция Кульбака-Лейблера	$d_{KL}\left(x,\hat{x}\right)=x\ln\frac{x}{\hat{x}}-x+\hat{x}$
дивергенция Итакура-Саито	$d_{IS}\left(x,\hat{x}\right)=\ln\frac{\hat{x}}{x}+\frac{x}{\hat{x}}-1$
расстояние Хеллингера	$d_H\left(x,\hat{x}\right)=\left(\sqrt{x}-\sqrt{\hat{x}}\right)^2$

МФТИ

Функции потерь

- Часто функция потерь — это замаскированное правдоподобие: существует такая плотность $p\left(X|\hat{X}\right)$, что

$$D\left(X,\hat{X}\right) \propto -\ln p\left(X|\hat{X}\right)$$

Функция потерь	Порождающая модель
Фробениуса	аддитивная гауссовская
Кульбака-Лейблера	пуассоновская
Итакура-Саито	мультипликативная гамма

С ДИВЕРГЕНЦИЕЙ КУЛЬБАКА-ЛЕЙБЛЕРА

- Мультипликативные обновления:

$$h_{kj} \leftarrow h_{kj} \frac{\sum\limits_{i=1}^{n} w_{ik} x_{ij}}{\sum\limits_{i=1}^{n} w_{ik} \hat{x}_{ij}}$$

- Функция потерь монотонно невозрастает
- Тот же функционал, но по-другому, минимизирует метод PLSA (скоро в программе!)

МФТИ

Инициализация

- Поскольку методы сходятся локально, начальное приближение играет большую роль.
  - Случайная инициализация
  - Кластеризация
  - SVD
  - ALS
  - Мультистарт:

    1. С помощью ALS генерируются 10-20 пар матриц
    2. Делаются 10-20 итераций целевого метода на каждой паре
    3. В качестве начального приближения выбирается пара с наименьшим значением функционала

Резюме

- Функции потерь
- Инициализация

ОБРАБОТКА ПРОПУСКОВ

ПРОПУСКИ

Матрица данных: $X \in \mathbb{R}^{\ell \times d}$
Часть наблюдений пропущена.

- Можно ли заполнить пропуски?
- Если X — матрица объекты-признаки, как с ней делать классификацию или регрессию?

Случайность пропусков

- Предполагаем, что пропуски в матрице располагаются случайно.

- Пример исключений:
  - отказ респондентов отвечать на вопрос
  - Абрахам Вальд и повреждения самолётов

Отбрасывание

- Можно выбросить объекты, для которых значение хотя бы одного признака пропущено (перед этим лучше избавиться от сильно разреженных признаков).

$\Rightarrow$ информация теряется

ЗАПОЛНЕНИЕ

- По ближайшему объекту: для строки x_{i_1} находим самую похожую на неё x_{i_2} и заменяем $x_{i_1 j} = NA$ на $x_{i_2 j}$
- Заполнение средними или медианами по столбцу
- EM-алгоритм
- Матричные разложения

$\Rightarrow$ информация берётся из ниоткуда?

МФТИ

Другой способ обработки

- В некоторых задачах матрица X используется только для подсчёта величин вида $\frac{1}{\ell}X^TX$ и $\frac{1}{\ell}X^Ty$
- Можно считать их только полными парами:

$$\frac{1}{\ell}\left(X^TX\right)_{jk} = \frac{1}{\ell}\sum_{i=1}^{\ell} x_{ij}x_{ik} \approx$$

$$\approx \frac{1}{\ell_{jk}}\sum_{i=1}^{\ell} x_{ij}x_{ik}\left[x_{ij} \neq NA, x_{ik} \neq NA\right]$$

ℓ_{jk} — число полных пар

$\Rightarrow$ информация не исчезает и не появляется из ниоткуда

ЕЩЁ ДЕТАЛИ

- Пропуски в категориальных переменных удобно закодировать отдельной категорией (работает и для неслучайных пропусков!)
- Для деревьев пропуски в непрерывных признаках можно заполнить значением, сильно отличающимся от всех остальных значений признака (-99999999), тогда они будут попадать в отдельный лист
- Обработка по умолчанию в некоторых методах плохая (регрессия, отбрасывание), в некоторых — хорошая (xgboost, сложно).

Резюме

- Пропуски
- Заполнение и отбрасывание