

Part 1

Introduction

Problems statements

- Text document classification by topics
- Text documents semantic similarities
- Sentiment analysis
- Author detection
- Text clustering
- Text document summarization

Text classification: features

- Dataset example: 20news_groups
- E-mails on 20 different topics
- Let's try to differ “**auto**” and “**politics.mideast**”

Text classification: features

- Example 1:

From: carl_f_hoffman@cup.portal.com
Newsgroups: rec.autos
Subject: 1993 Infiniti G20
Message-ID: <78834@cup.portal.com>
Date: Mon, 5 Apr 93 07:36:47 PDT
Organization: The Portal System (TM)
Lines: 26

I am thinking about getting an Infiniti G20.
In consumer reports it is ranked high in many
catagories including highest in reliability index for compact cars.
Mitsubishi Galant was second followed by Honda Accord).

A couple of things though:

- 1) In looking around I have yet to see anyone driving this
car. I see lots of Honda's and Toyota's.

Text classification: features

- Example 2:

From: Bob.Waldrop@f418.n104.z1.fidonet.org (Bob Waldrop)
Subject: Celebrate Liberty! 1993
Message-ID: <1993Apr5.201336.16132@dsd.es.com>
Followup-To: talk.politics.misc

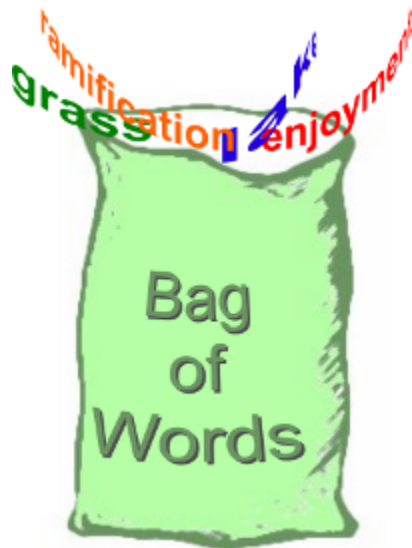
Announcing. . . Announcing. . . Announcing. . . Announcing. . .

CELEBRATE LIBERTY!
1993 LIBERTARIAN PARTY NATIONAL CONVENTION
AND POLITICAL EXPO

THE MARRIOTT HOTEL AND THE SALT PALACE
SALT LAKE CITY, UTAH

INCLUDES INFORMATION ON DELEGATE DEALS!
(Back by Popular Demand!)

Bag-of-words



the world of

TOTAL

» All About The Company

- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage

all about the company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

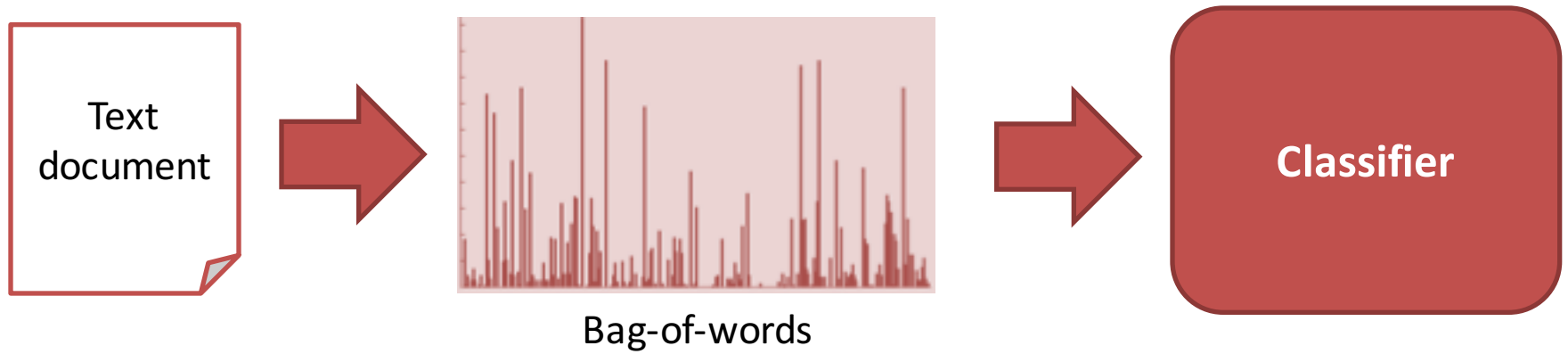
At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Basin complement already solid positions in Europe, Africa, and the U.S.

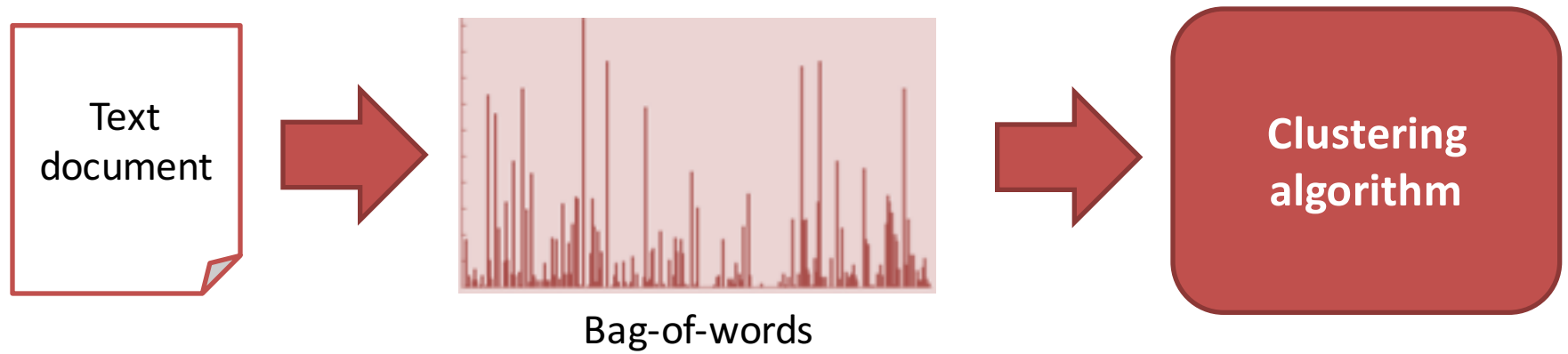
Our growing specialty chemicals sector adds balance and profit to the core energy business.

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

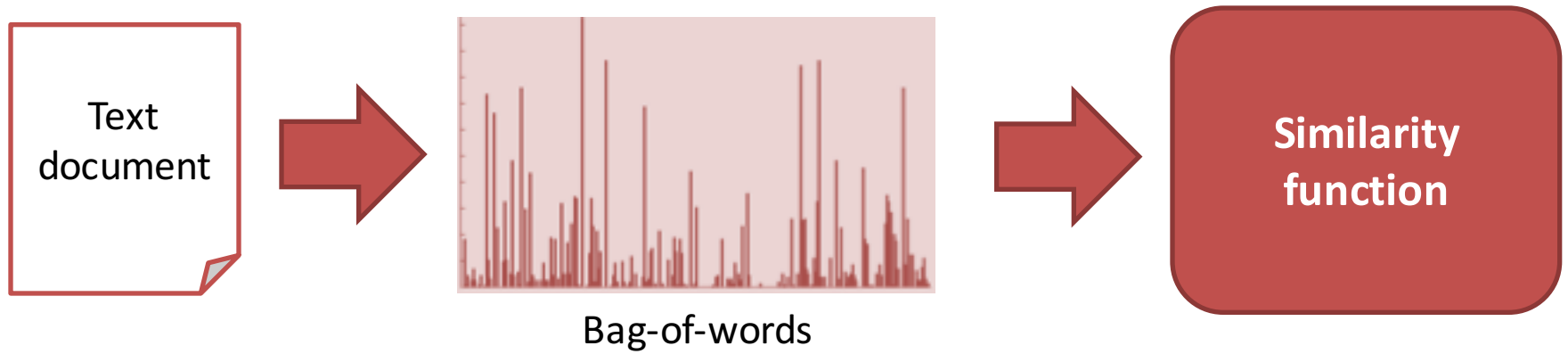
Text classification: an overview



Text clustering: an overview



Semantic similarities



$$CS(d_1, d_2) = \frac{\langle d_1, d_2 \rangle}{\sqrt{\langle d_1, d_1 \rangle \langle d_2, d_2 \rangle}}$$

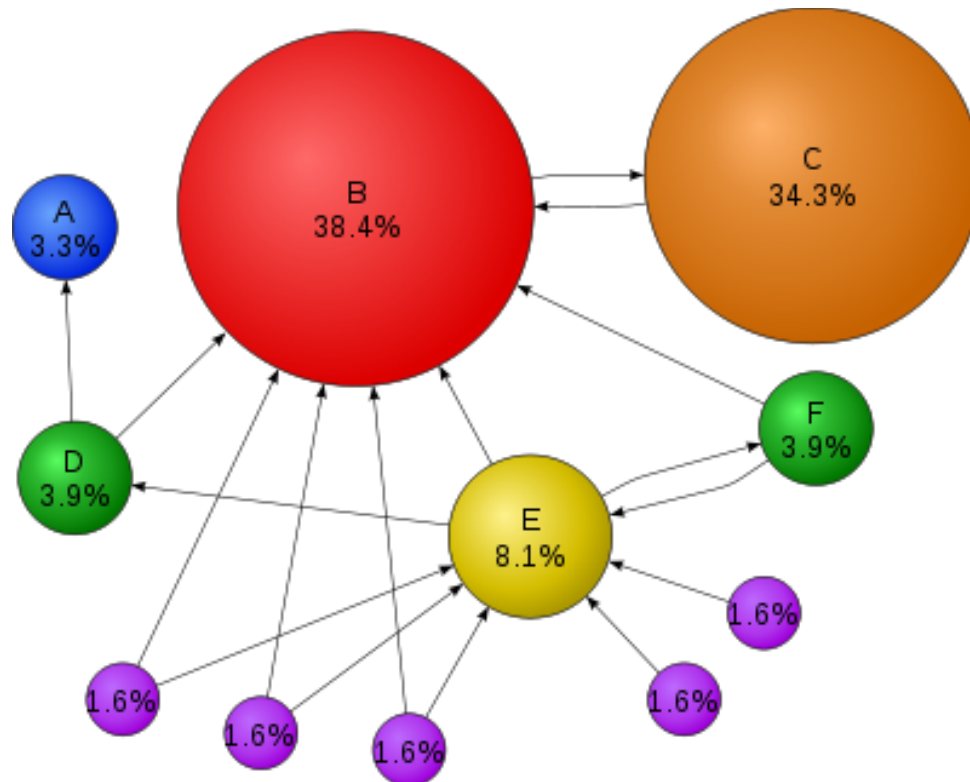
Text summarization: baseline

- Let's look on the sentences of the initial text document as separate text documents
- Cosine similarity between initial text document and sentence is sentence rang (importance)
- Summary is the list of sentences with the highest rangs

Text summarization: TextRank

Günes Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization.



PageRank

Sentiment analysis example



“I bought a Motorola phone two weeks ago. Everything was good initially. The voice was clear and the battery life was long, although it is a bit bulky. Then, it stopped working yesterday.”

- Objective and subjective statements
- General description (everything was good) and aspects description (battery life, voice, ...)

Sentiment analysis applications

- For customer: product review analysis
- For organizations: alternative for focus-groups and polls
- Politics: electors opinion mining
- Movies: revenue forecasting
- Trading: experts opinions mining and rates forecasting

SA: key difficulties

“Our sentiment analysis is as bad as everyone else’s.”

- User generated texts differ from reviewed texts
- User’s vocabulary depends on background (age, gender, education)
- Emotional meaning of words depends on describing object
- Sarcasm
- Every site with responses offers it’s own template of response

Sentiment analysis levels

- Document

Everything was good initially. Then, it stopped working yesterday.

- Sentence

Everything was good initially. Then, it stopped working yesterday.

- Aspect

I bought a Motorola phone two weeks ago. The voice was clear and the battery life was long

Discussion: simple sentiment analysis

Named Entity Recognition

Task: tag named entities in the text.

Adams and Platt are both injured and will miss England's opening World Cup qualifier against Moldova on Sunday.

<**PER**>Adams</**PER**> and <**PER**>Platt</**PER**> are both injured and will miss <**LOC**>England</**LOC**>'s opening <**EVENT**>World Cup</**EVENT**> qualifier against <**LOC**>Moldova</**LOC**> on <**DAY**>Sunday</**DAY**>.

Named Entity Recognition

- Supervised methods:
 - HMM (Hidden Markov Model) – fast
 - MEMM (Maximum Entropy Markov Model) – slow
 - CRF (Conditional Random Fields) – very accurate, but extremely slow

Part 2

Semantic analysis: word2vec

Distributive semantics



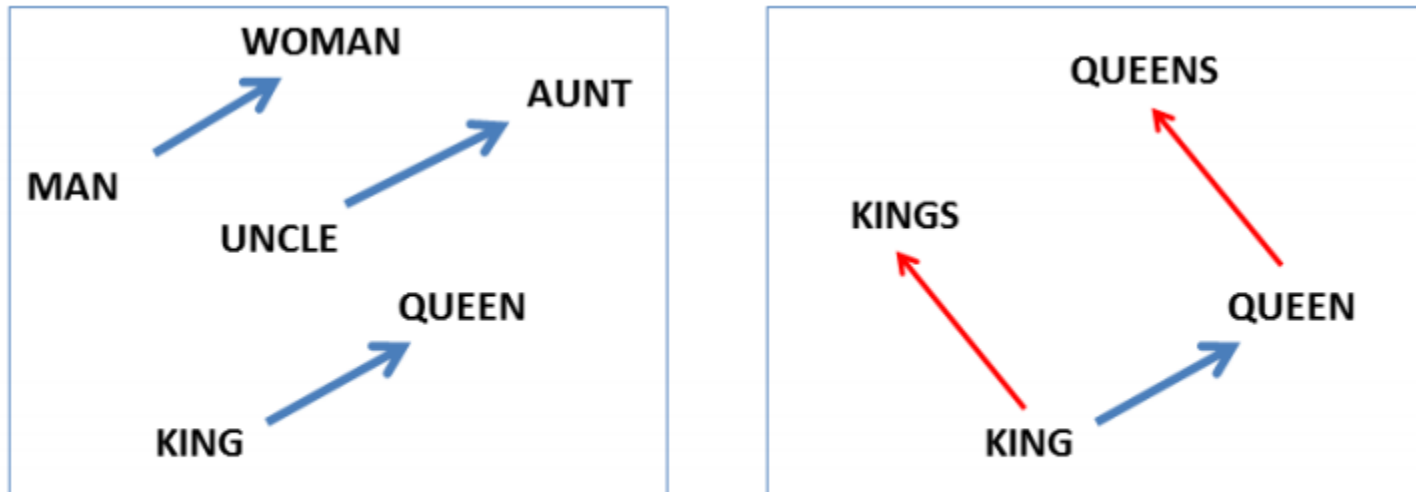
Term-context matrix

	C1	C2	C3	C4	C5	C6	C7
dog	5	0	11	2	2	9	1
cat	4	1	7	1	1	7	2
bread	0	12	0	0	9	1	9
pasta	0	8	1	2	14	0	10
meat	0	7	1	1	11	1	8
mouse	4	0	8	0	1	8	1

Term-context matrix

	dog	cat	computer	animal	mouse
dog	0	4	0	2	1
cat	4	0	0	3	5
computer	0	0	0	0	3
animal	2	3	0	0	2
mouse	1	5	3	2	0

The most popular word2vec example



(Mikolov et al., NAACL HLT, 2013)

Word2vec objective

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

Word2vec objective

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)]$$

Word2vec objective

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)]$$

Negative sampling

Word2vec objective

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)]$$

$$P_D(c) = \frac{\#(c)}{|D|}$$

Word2vec objective

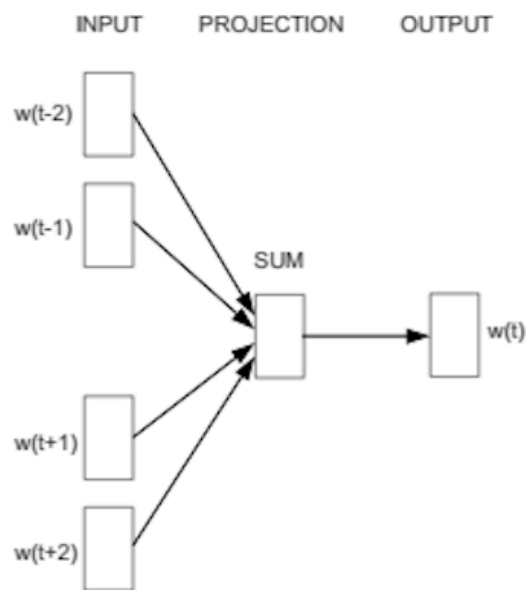
$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)]$$

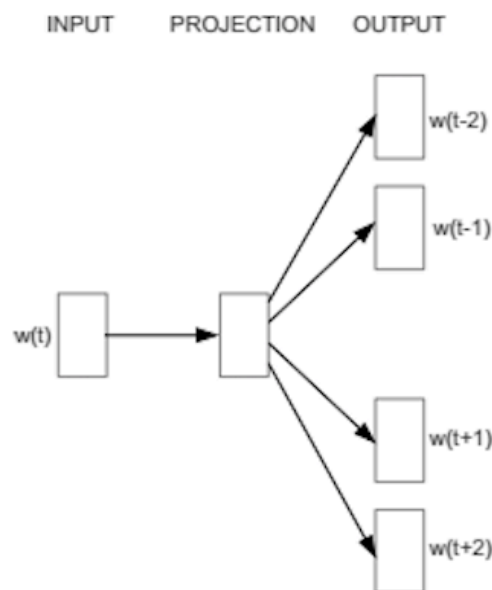
$$P_D(c) = \frac{\#(c)}{|D|}$$

$$\ell = \sum_{w \in V_W} \sum_{c \in V_C} \#(w, c) (\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} [\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

Picture from the original paper



CBOW

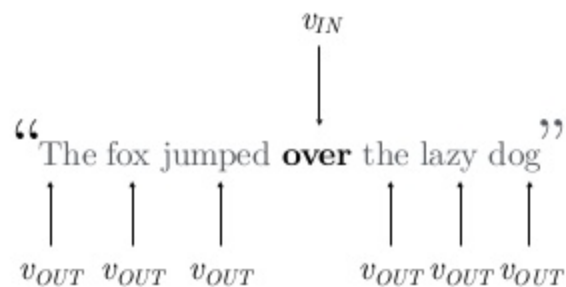


Skip-gram

CBOW and Skip-gram

SkipGram

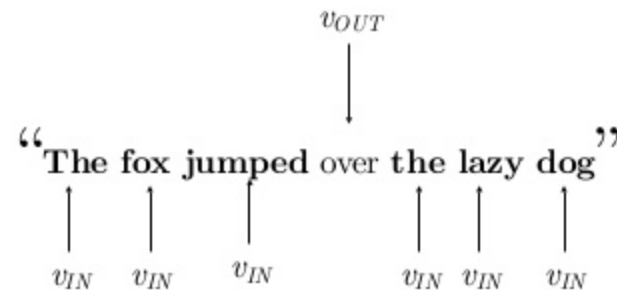
Guess the context
given the word



Better at syntax.
(this is the one we went over)

CBOW

Guess the word
given the context



~20x faster.
(this is the alternative.)

word2vec and supervised learning

- Bag of words – too many features
- Mean word2vec-vector in a text is a compact feature representation
- With word2vec dimension 100-500 one can fit ensembles of decision trees

Bonus slides:
Matrix factorizations

Matrix factorization

$$\begin{matrix} X & \approx & U & \cdot & V^T \\ l \times n & & l \times k & & k \times n \end{matrix}$$

Matrix factorizations

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

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

Matrix factorizations

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

$$\left\| X - U \cdot V^T \right\| \rightarrow \min$$

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

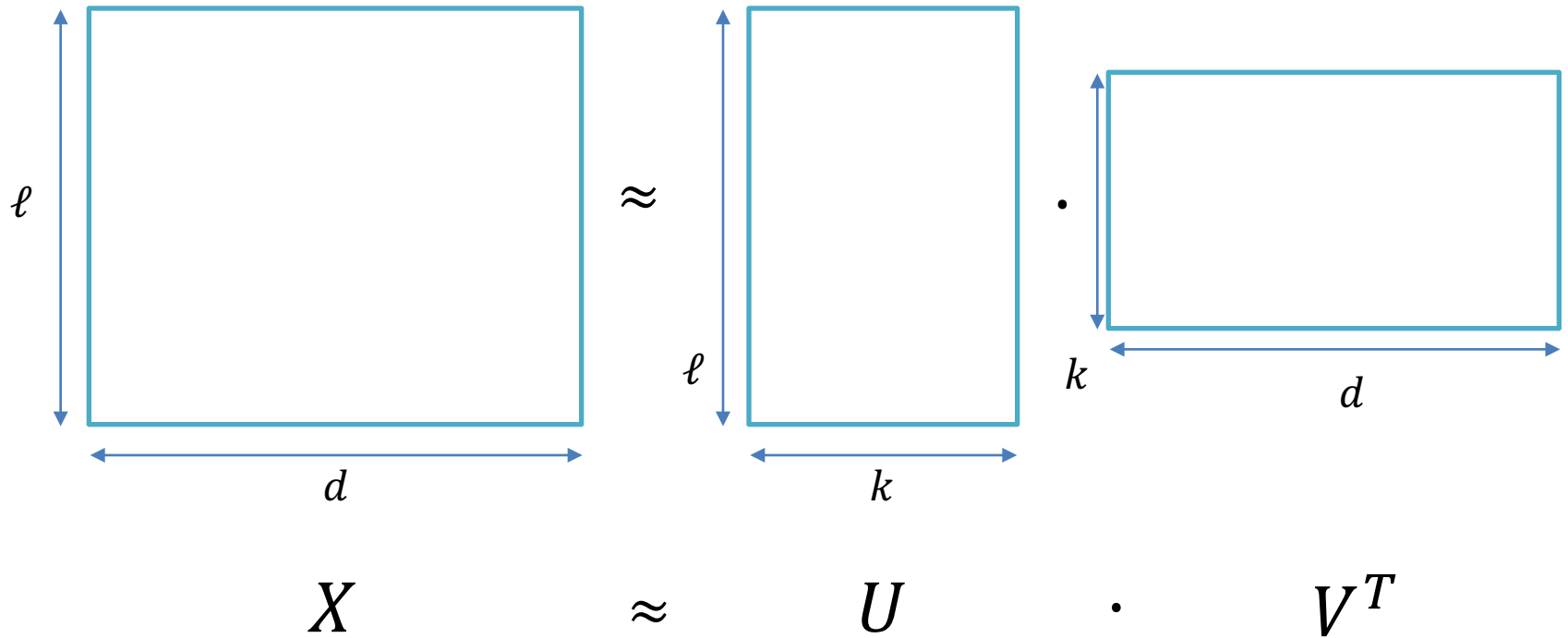
Matrix factorizations

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

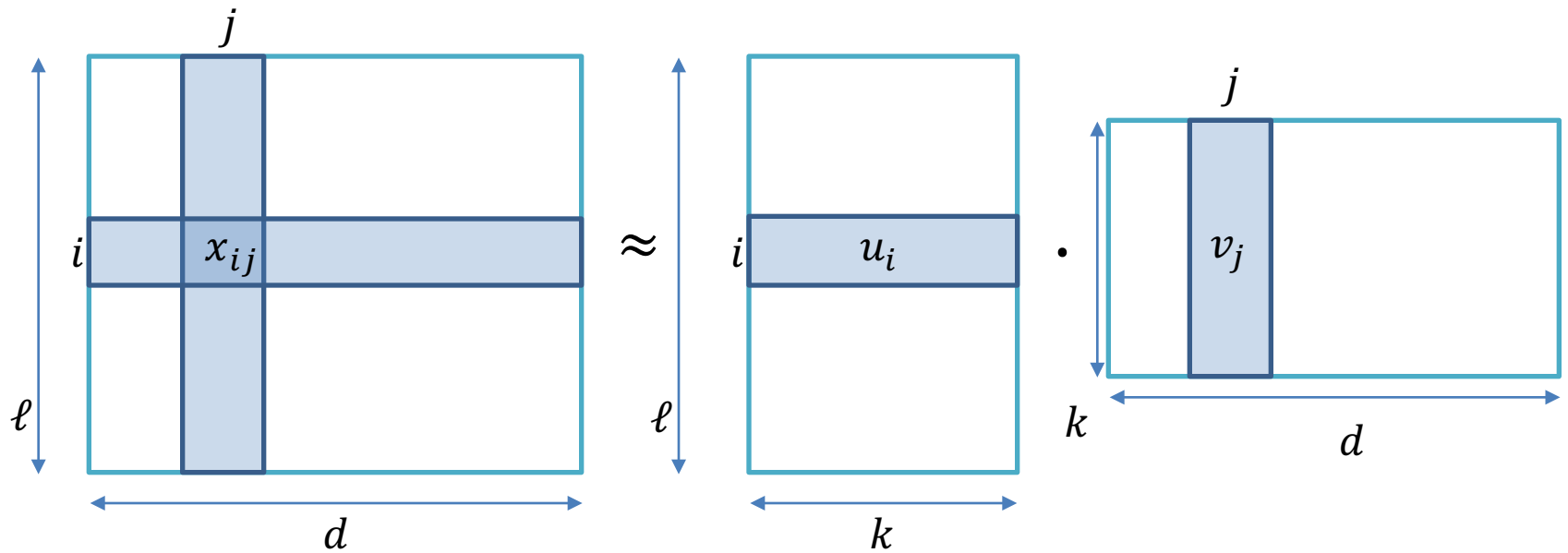
$$||X - U \cdot V^T|| \rightarrow \min$$

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

Notation



Notation



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

Singular Vector Decomposition in algebra

$$X = U\Sigma V^T$$

U - *orthogonal*

Σ - *diagonal*

V - *orthogonal*

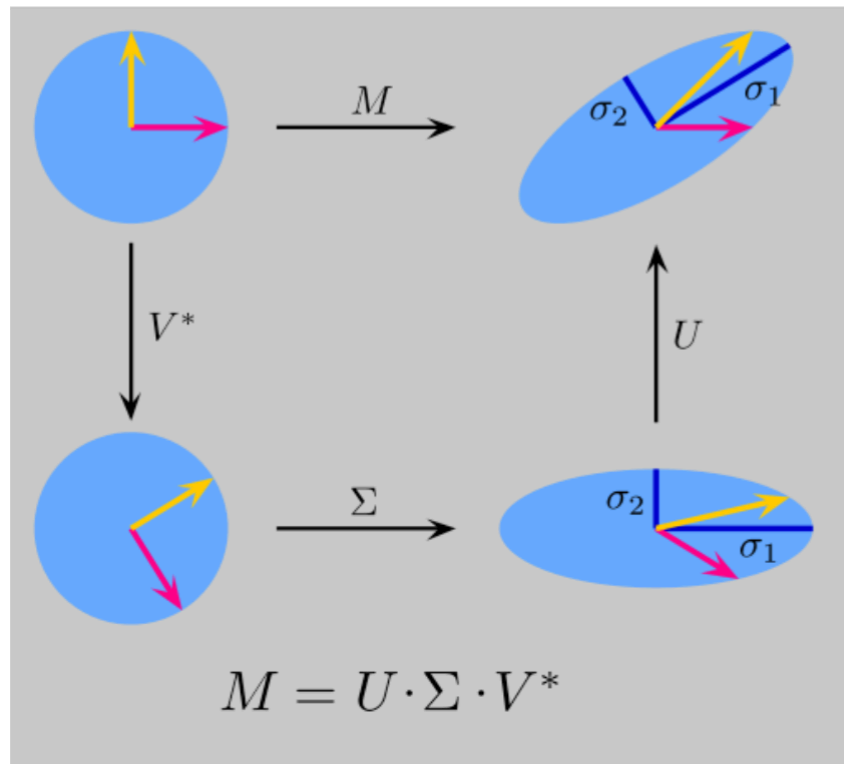
Singular Vector Decomposition in algebra

$$X = U\Sigma V^T$$

U - orthogonal

Σ - diagonal

V - orthogonal



SVD for matrix approximation

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

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

SVD for matrix approximation

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

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

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

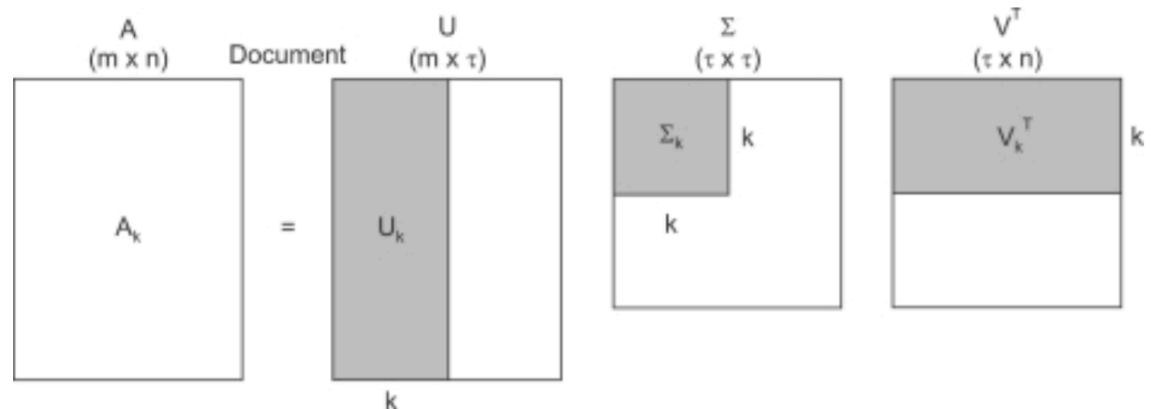
SVD for matrix approximation

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

$l \times n \quad l \times k \quad k \times n$

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

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



SVD for matrix approximation

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

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

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

$\tilde{U}_k, \Sigma_k, \tilde{V}_k$ - truncated SVD matrixes

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

SVD for matrix approximation

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

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

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

$\tilde{U}_k, \Sigma_k, \tilde{V}_k$ - truncated SVD matrixes

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

SVD for matrix approximation

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

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

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

$\tilde{U}_k, \Sigma_k, \tilde{V}_k$ - truncated SVD matrixes

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

“SVD” in Machine Learning

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

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

u_i - samples “descriptions”

v_j - features “descriptions”

Movie ratings and SVD

	Пила	Улица Вязов	Ванильное небо	1+1
Maria	5	4	1	2
Julia	5	5	2	
Vladimir			3	5
Nikolay	3		4	5
Peter				4
Ivan		5	3	3

Movie ratings and SVD

i	j			
	Пила	Улица Вязов	Ванильное небо	1+1
Maria	5	4	1	2
Julia	5	5	2	
Vladimir			3	5
Nikolay	3		4	5
Peter				4
Ivan		5	3	3

Movie ratings and SVD

<i>i</i>	<i>j</i>			
	Saw	Nightmare on Elm Street	Vanilla Sky	The Intouchables
Maria	5	4	1	2
Julia	5	5	2	
Vladimir			3	5
Nikolay	3	?	4	5
Peter				4
Ivan		5	3	3

u_i - “user interests”

v_j - “movies parameters”

$$x_{ij} \approx \langle u_i, v_j \rangle = \sum_{k=1}^K u_{ik} v_{jk}$$

Word frequencies and 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

Word frequencies and SVD

j

	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
i 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

Word frequencies and SVD

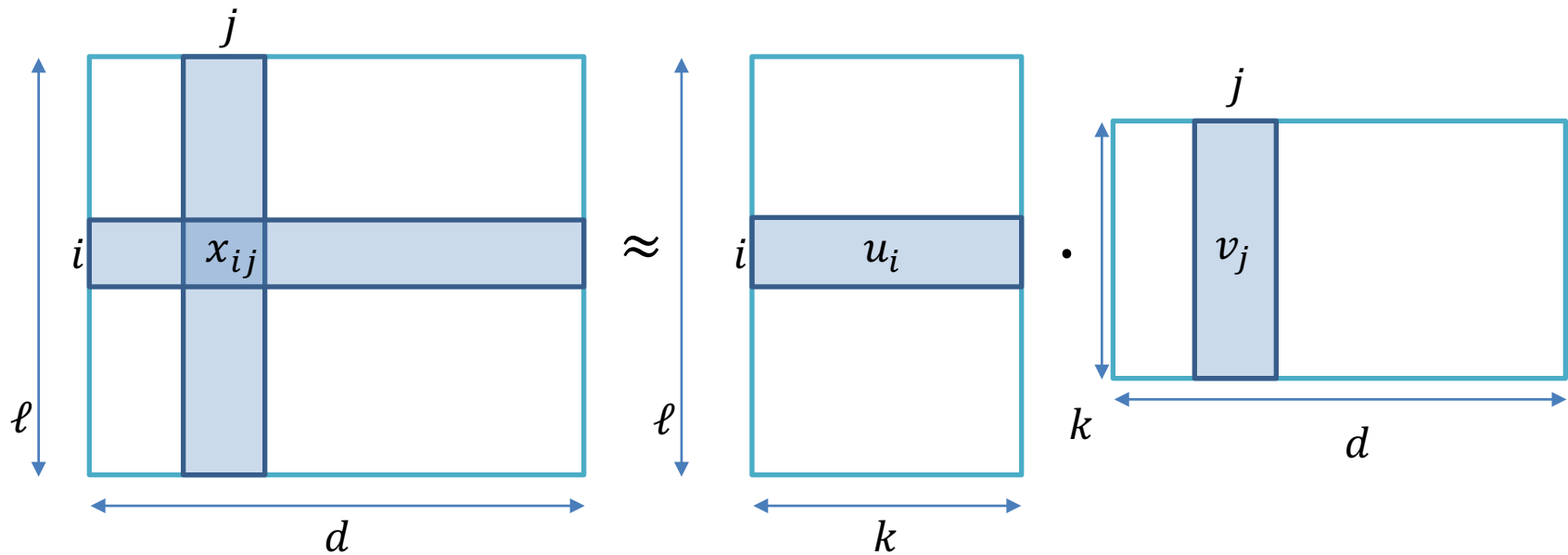
<i>i</i>	<i>j</i>					
	database	SQL	index	regression	likelihood	linear
d1	24	21	9	0	0	3
d2	32	10	5	0	3	0
d3	12	16	5	0	0	0
d4	6	7	2	0	0	0
d5	43	31	20	0	3	0
d6	2	0	0	18	7	16
d7	0	0	1	32	12	0
d8	3	0	0	22	4	2
d9	1	0	0	34	27	25
d10	6	0	0	17	4	23

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

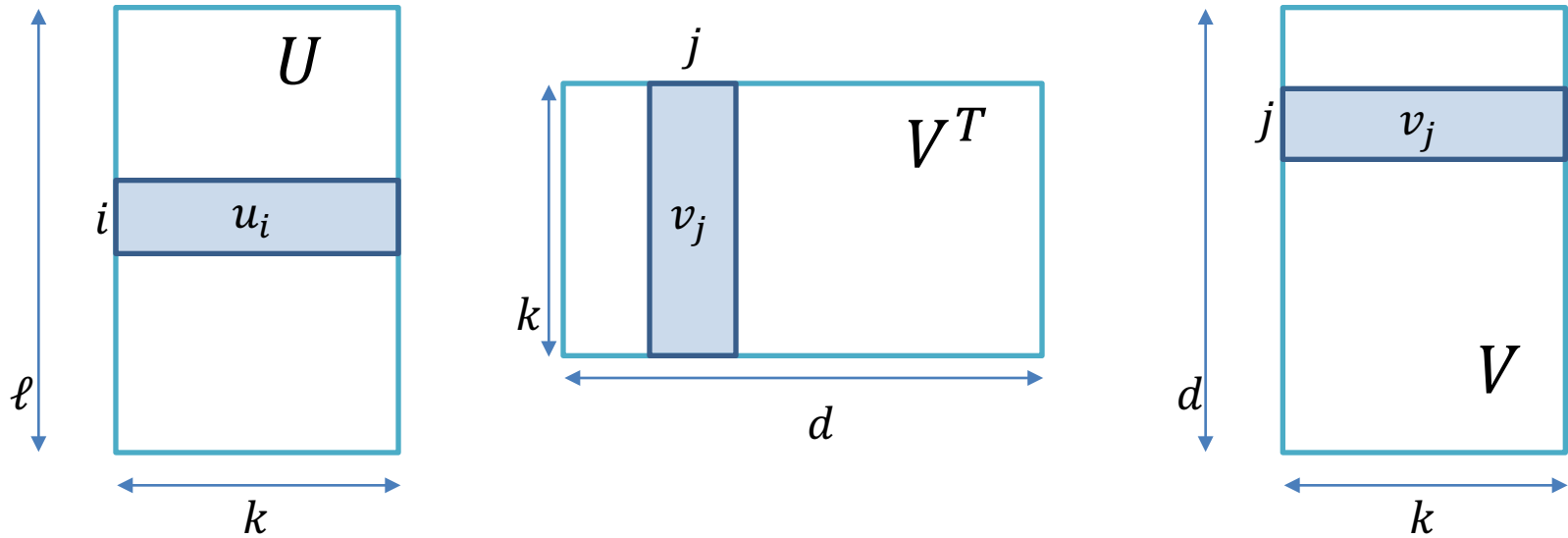
u_i - «ТЕМЫ» ДОКУМЕНТОВ

v_j - «ТЕМЫ» СЛОВ

A little bit more about notations

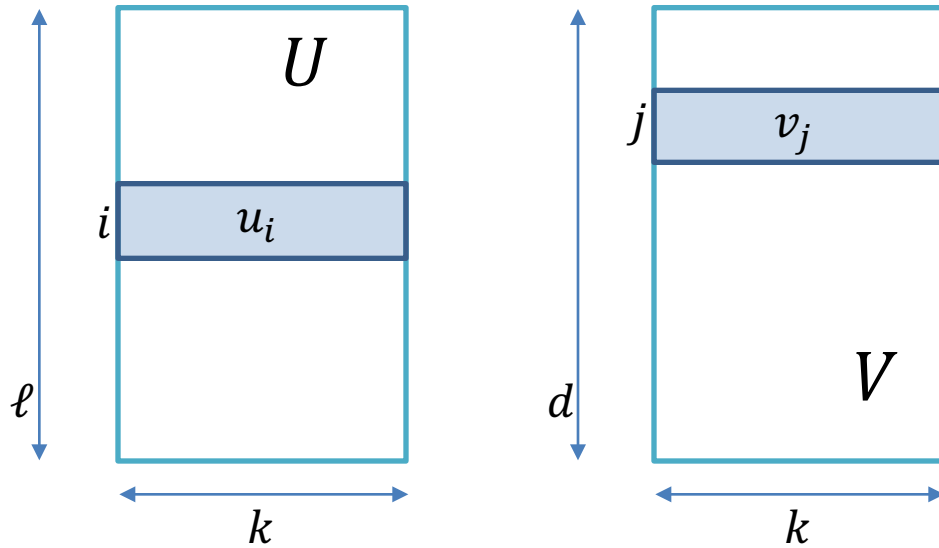


A little bit more about notations



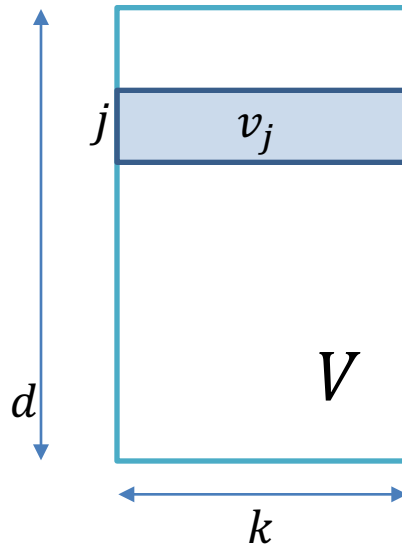
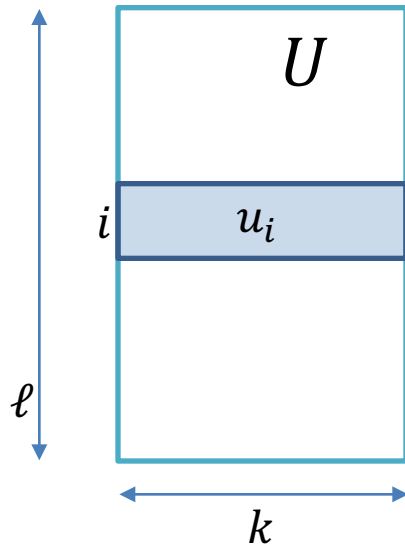
$$x_{ij} \approx \langle u_i, v_j \rangle$$

A little bit more about notations



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

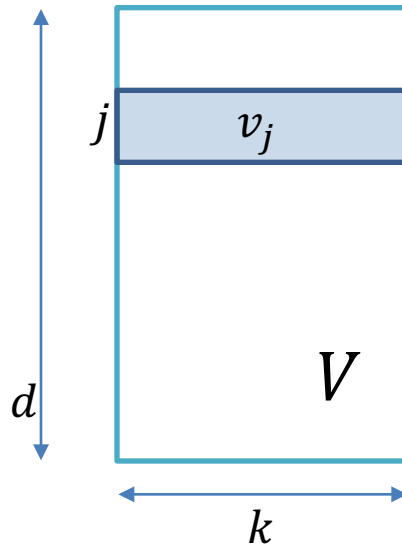
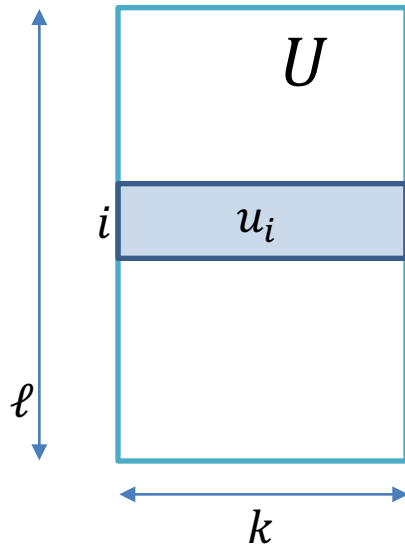
A little bit more about notations



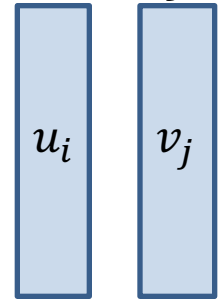
$$x_{ij} \approx \langle u_i, v_j \rangle$$

Diagram showing two vertical blue bars representing vectors u_i and v_j .

A little bit more about notations

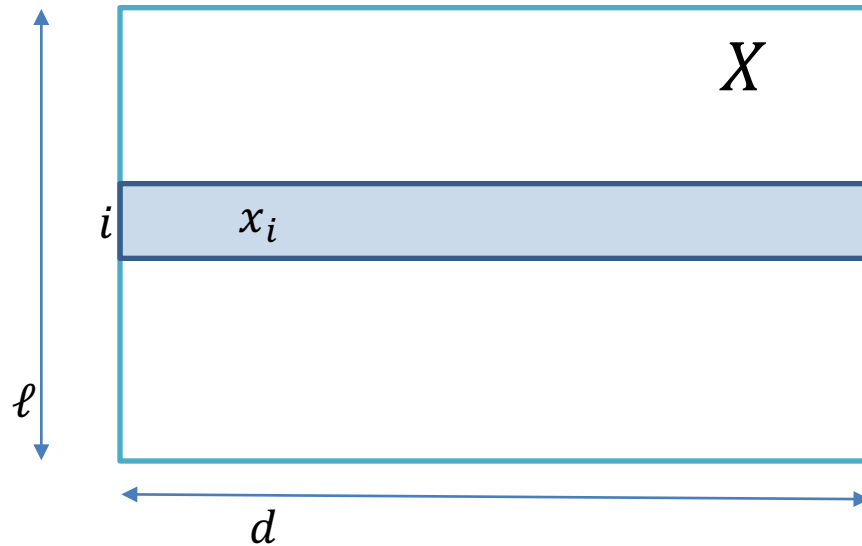


$$x_{ij} \approx \langle u_i, v_j \rangle$$



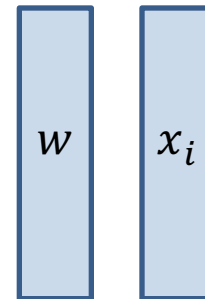
$$x_{ij} \approx u_i^T v_j$$

A little bit more about notations



In linear models:

$$\langle w, x_i \rangle = w^T x_i$$



Popular notations

$$X \approx UV^T$$

$$X \approx PQ^T$$

$$X \approx WH$$

$$X \approx \Phi\Theta$$

Problem formulation in SVD

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

Gradient Decent (GD)

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

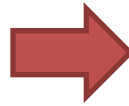
$$\begin{aligned} \frac{\partial Q}{\partial u_i} &= \sum_{\tilde{i}, j} \frac{\partial}{\partial u_{\tilde{i}}} (\langle u_{\tilde{i}}, v_j \rangle - x_{\tilde{i}j})^2 = \sum_j 2(\langle u_i, v_j \rangle - x_{ij}) \frac{\partial \langle u_i, v_j \rangle}{\partial u_i} = \\ &= \sum_j 2(\langle u_i, v_j \rangle - x_{ij}) v_j \quad \varepsilon_{ij} = (\langle u_i, v_j \rangle - x_{ij}) - \text{error on } x_{ij} \end{aligned}$$

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

Stochastic Gradient Decent (SGD)

GD:

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



SGD:

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

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

For random i, j

SGD: pros and cons

- + Simple
- + Converge (usually)
- Slow
- Needs good steps choosing (γ_t and η_t)
- Extremely slow with constant step

Alternating Least Squares (concept)

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

Iteratively repeat:

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

First step in ALS

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

$$\frac{\partial Q}{\partial u_i} = \sum_j 2(\langle u_i, v_j \rangle - x_{ij})v_j = 0 \quad \sum_j v_j \langle v_j, u_i \rangle = \sum_j x_{ij}v_j$$

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

ALS: algorithm

Repeat for random (i, j) until converge:

$$\left(\sum_j v_j v_j^T \right) u_i = \sum_j x_{ij} v_j \quad \Rightarrow \quad u_i \quad (\text{Least Squares method})$$

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

Regularization

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

α and β - small positive constants (0.001, 0.01, 0.05)

Prediction model

<i>i</i>	<i>j</i>			
	Saw	Nightmare on Elm Street	Vanilla Sky	The Intouchables
Maria	5	4	1	2
Julia	5	5	2	
Vladimir			3	5
Nikolay	3	?	4	5
Peter				4
Ivan		5	3	3

u_i - “user interests”

v_j - “movies parameters”

$$x_{ij} \approx \langle u_i, v_j \rangle = \sum_{k=1}^K u_{ik} v_{jk}$$

Minimizing function

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

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

Bias

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

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

Baseline predictors

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

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

Regularization

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

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

Recommendations

j

	Вечернее платье	Поднос для писем	iPhone 6s	Шуба D&G
i Maria	1		1	
Julia	1	1		1
Vladimir		1	1	
Nikolay	1	?	1	
Peter		1	1	
Ivan			1	1

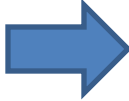
Recommendations in retail

i	j			
	Dress	Boots	Jeans	T-shirt
	Maria	1	1	
	Julia	1		1
	Vladimir	1	1	
	Nikolay	1	1	
	Peter	1	1	
Ivan			1	1

Why something is wrong?

	<i>j</i>			
	Dress	Boots	Jeans	T-shirt
Maria	1		1	
Julia	1	1		1
Vladimir		1	1	
Nikolay	1	?	1	
Peter		1	1	
Ivan			1	1

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

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


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

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

Explicit и implicit

- **Explicit feedback:** positive and negative examples (f.ex. high and low movie ratings, likes and dislikes and so on)
- **Implicit feedback:** only positive feedback (purchases, clicks, likes) or only negative feedback

Implicit matrix factorization

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



Sum over the all indices (not only
indices of known matrix elements –
suppose unknown elements are equal
to zero)

w_{ij} is high for $x_{ij} \neq 0$
and rather low for $x_{ij} = 0$

Implicit ALS

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

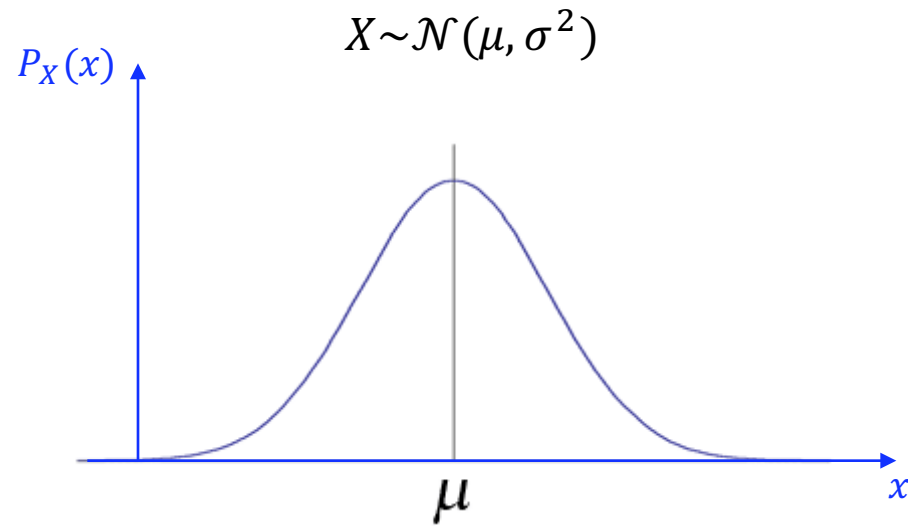
$$w_{ij} = 1 + \alpha |x_{ij}| \quad \alpha = 10, 100, 1000$$

Fitting u_i, v_j with ALS

Problem formulation in SVD

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

Normal distribution



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

SVD and normal distribution

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

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

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

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

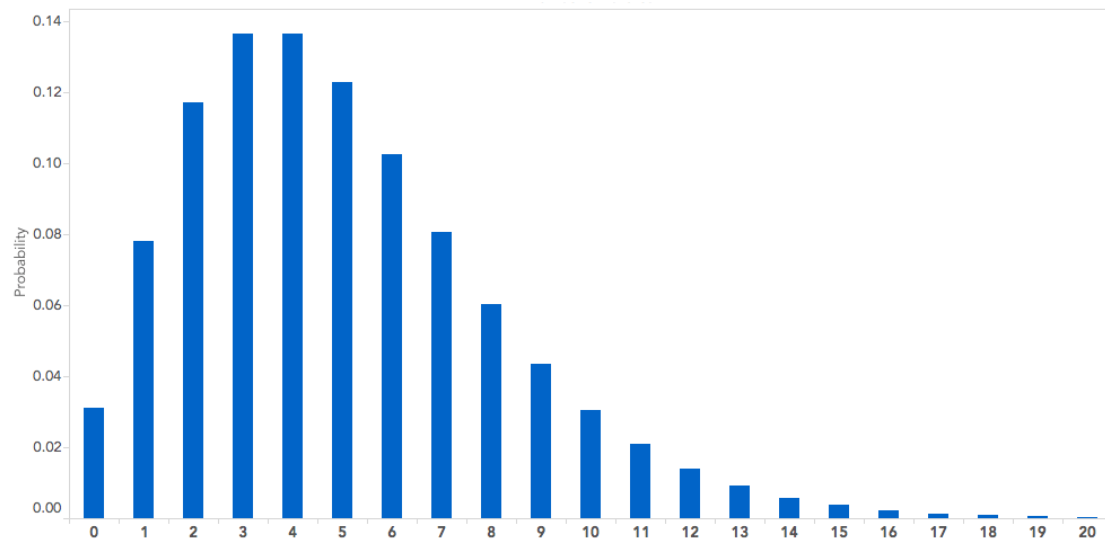
$$\sum_{i,j} (x_{ij} - \langle u_i, v_j \rangle)^2 \rightarrow \min$$

What distribution describes this data better?

	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

Poisson distribution

$$X \sim \text{Poiss}(\lambda)$$



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

Poisson distribution and matrix factorization

$$x_{ij} \sim \text{Pois}(\langle u_i, v_j \rangle) \quad P(x_{ij}) = \frac{\langle u_i, v_j \rangle^{x_{ij}}}{x_{ij}!} e^{-\langle u_i, v_j \rangle}$$

$$\prod_{i,j} \frac{\langle u_i, v_j \rangle^{x_{ij}}}{x_{ij}!} e^{-\langle u_i, v_j \rangle} \rightarrow \max$$

$$\sum_{i,j} \langle u_i, v_j \rangle - x_{ij} \ln \langle u_i, v_j \rangle + \ln x_{ij}! \rightarrow \min$$

$$\sum_{i,j} \langle u_i, v_j \rangle - x_{ij} \ln \langle u_i, v_j \rangle \rightarrow \min$$

SGD for NMF (Non-negative matrix factorization)

$$Q = \sum_{i,j} \langle u_i, v_j \rangle - x_{ij} \ln \langle u_i, v_j \rangle \rightarrow \min$$

$$\frac{\partial Q}{\partial u_i} = \sum_j v_j - \frac{x_{ij}}{\langle u_i, v_j \rangle} v_j = \sum_j \underbrace{\frac{\langle u_i, v_j \rangle - x_{ij}}{\langle u_i, v_j \rangle}}_{\tilde{\epsilon}_{ij} \text{ - relative error}} v_j \rightarrow \min$$

SGD:

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

Another non-negative factorizations

One can use for NMF Frobenius norm with restrictions on U and V:

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

Recap

- Matrix factorization
- SGD and ALS
- Implicit matrix factorizations
- Probabilistic interpretation