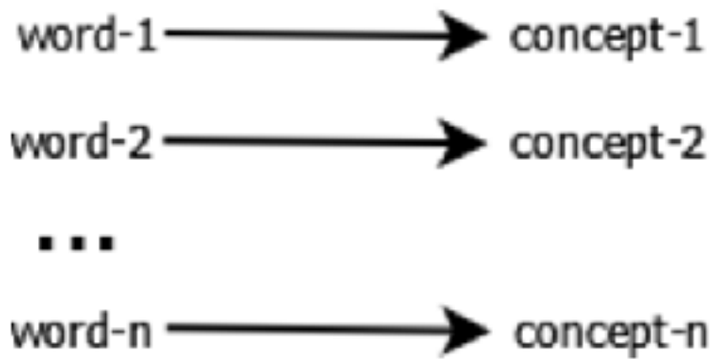
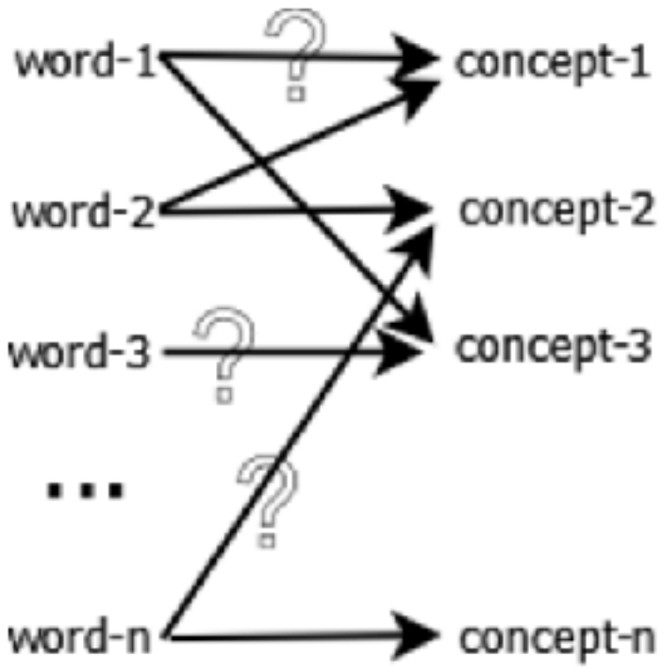
Матричные разложения

LSA, NMF, Cholesky

LSA: Latent semantic analysis

- Документ — вектор
- Слова-синонимы
- Похожие тексты
- Рекомендательные системы





LSA: построение матрицы

- c1: Human machine interface for ABC computer applications
- c2: A survey of user opinion of computer system response time
- c3: The EPS user interface management system
- c4: System and human system engineering testing of EPS
- c5: Relation of user perceived response time to error measurement
- m1: The generation of random, binary, ordered trees
- m2: The intersection graph of paths in trees
- m3: Graph minors IV: Widths of trees and well-quasi-ordering
- m4: Graph minors: A survey

LSA: term-document matrix

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

LSA

$$A = USV^T$$

A — term-document matrix
U, V — ортогональные матрицы
S — диагональная матрица

$$A_k = U_k S_k V_k^T$$

Матрица с первыми k сингулярными числами

U =

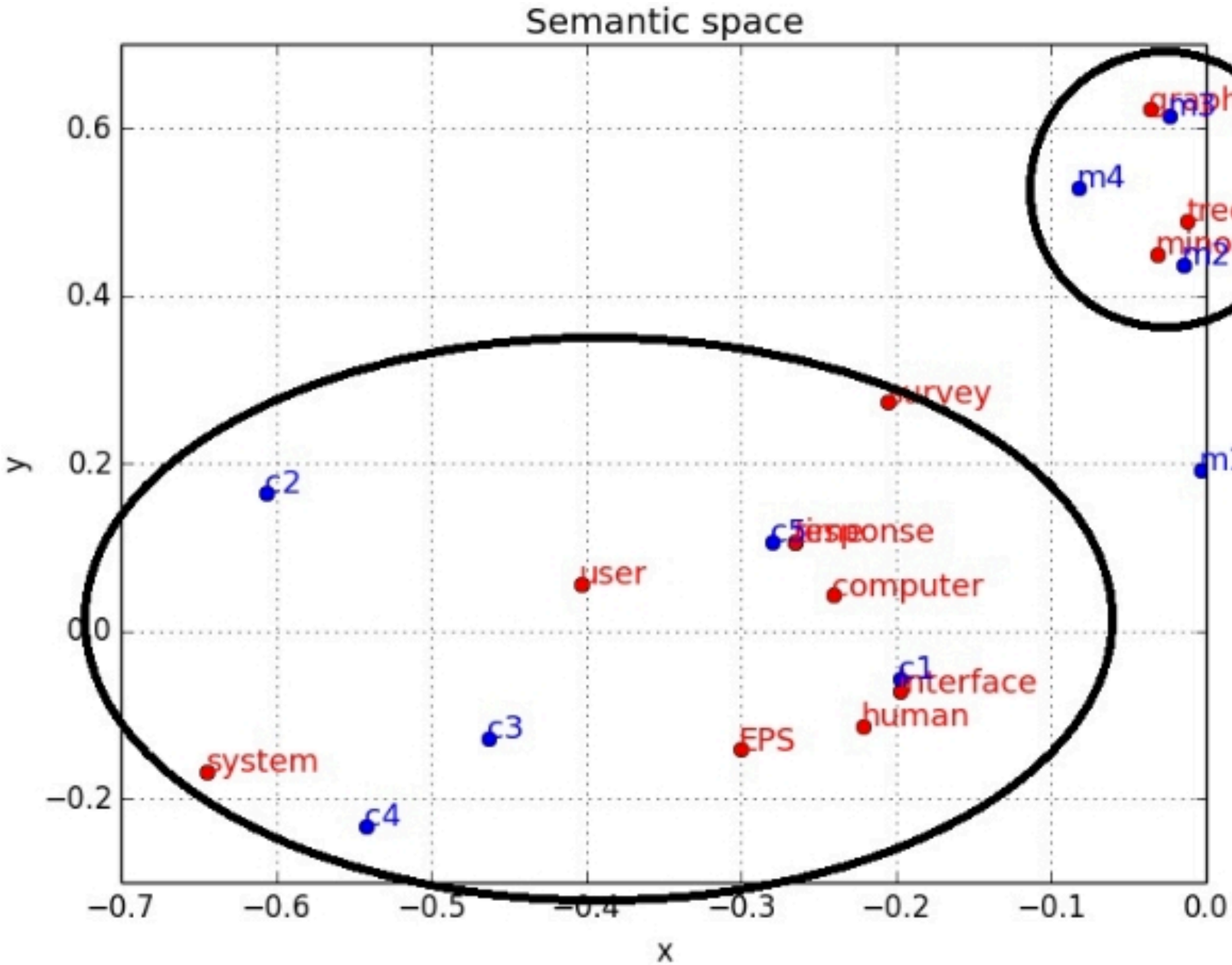
human	-0,221	-0,113	0,289	-0,415	-0,106	-0,341	-0,523	0,06	0,407
interface	-0,198	-0,072	0,135	-0,552	0,282	0,496	0,07	0,01	0,109
computer	-0,24	0,043	-0,164	-0,595	-0,107	-0,255	0,302	-0,062	-0,492
user	-0,404	0,057	-0,338	0,099	0,332	0,385	-0,003	0	-0,012
system	-0,644	-0,167	0,361	0,333	-0,159	-0,207	0,166	-0,034	-0,271
response	-0,265	0,107	-0,426	0,074	0,08	-0,17	-0,283	0,016	0,054
time	-0,265	0,107	-0,426	0,074	0,08	-0,17	-0,283	0,016	0,054
EPS	-0,301	-0,141	0,33	0,188	0,115	0,272	-0,033	0,019	0,165
survey	-0,206	0,274	-0,178	-0,032	-0,537	0,081	0,467	0,036	0,579
trees	-0,013	0,49	0,231	0,025	0,594	-0,392	0,288	-0,255	0,225
graph	-0,036	0,623	0,223	0,001	-0,068	0,115	-0,16	0,681	-0,232
minors	-0,032	0,451	0,141	-0,009	-0,3	0,277	-0,339	-0,678	-0,183

V =

3,34	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	2,54	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	2,35	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	1,65	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,51	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	1,31	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,85	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,56	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,36

WT =

-0,197	-0,606	-0,463	-0,542	-0,279	-0,004	-0,015	-0,024	-0,082
-0,056	0,166	-0,127	-0,232	0,107	0,193	0,438	0,615	0,53
0,11	-0,497	0,208	0,57	-0,505	0,098	0,193	0,253	0,079
-0,95	-0,029	0,042	0,268	0,15	0,015	0,016	0,01	-0,025
0,046	-0,206	0,378	-0,206	0,327	0,395	0,349	0,15	-0,602
-0,077	-0,256	0,724	-0,369	0,035	-0,3	-0,212	0	0,362
-0,177	0,433	0,237	-0,265	-0,672	0,341	0,152	-0,249	-0,038
0,014	-0,049	-0,009	0,019	0,058	-0,454	0,762	-0,45	0,07
0,064	-0,243	-0,024	0,084	0,262	0,62	-0,018	-0,52	0,454
c1	c2	c3	c4	c5	m1	m2	m3	m4


Semantic space
graph
m3
m4
trees
minors
m2
survey
m1
c2
response
user
computer
c1
interface
human
c3
EPS
system
c4
0.6
0.4
0.2
0.0
−0.2
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
x
y

	human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors
human	1,00	0,50	0,50	0,00	0,58	0,00	0,00	0,50	0,00	0,00	0,00	0,00
interface	0,50	1,00	0,50	0,41	0,29	0,00	0,00	0,50	0,00	0,00	0,00	0,00
computer	0,50	0,50	1,00	0,41	0,29	0,50	0,50	0,00	0,50	0,00	0,00	0,00
user	0,00	0,41	0,41	1,00	0,47	0,82	0,82	0,41	0,41	0,00	0,00	0,00
system	0,58	0,29	0,29	0,47	1,00	0,29	0,29	0,87	0,29	0,00	0,00	0,00
response	0,00	0,00	0,50	0,82	0,29	1,00	1,00	0,00	0,50	0,00	0,00	0,00
time	0,00	0,00	0,50	0,82	0,29	1,00	1,00	0,00	0,50	0,00	0,00	0,00
EPS	0,50	0,50	0,00	0,41	0,87	0,00	0,00	1,00	0,00	0,00	0,00	0,00
survey	0,00	0,00	0,50	0,41	0,29	0,50	0,50	0,00	1,00	0,00	0,41	0,50
trees	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,67	0,41
graph	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,41	0,67	1,00	0,82
minors	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,41	0,82	1,00

	human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors
human	1	0,99	0,8	0,82	0,98	0,65	0,65	1	0,17	-0,43	-0,4	-0,39
interface	0,99	1	0,86	0,88	1	0,74	0,74	1	0,29	-0,32	-0,29	-0,28
computer	0,8	0,86	1	1	0,91	0,98	0,98	0,82	0,73	0,2	0,23	0,25
user	0,82	0,88	1	1	0,92	0,97	0,97	0,84	0,71	0,17	0,2	0,21
system	0,98	1	0,91	0,92	1	0,8	0,8	0,98	0,38	-0,23	-0,19	-0,18
response	0,65	0,74	0,98	0,97	0,8	1	1	0,68	0,86	0,4	0,43	0,44
time	0,65	0,74	0,98	0,97	0,8	1	1	0,68	0,86	0,4	0,43	0,44
EPS	1	1	0,82	0,84	0,98	0,68	0,68	1	0,2	-0,4	-0,37	-0,36
survey	0,17	0,29	0,73	0,71	0,38	0,86	0,86	0,2	1	0,81	0,83	0,84
trees	-0,43	-0,32	0,2	0,17	-0,23	0,4	0,4	-0,4	0,81	1	1	1,00
graph	-0,4	-0,29	0,23	0,2	-0,19	0,43	0,43	-0,37	0,83	1	1	1,00
minors	-0,39	-0,28	0,25	0,21	-0,18	0,44	0,44	-0,36	0,84	1	1	1,00

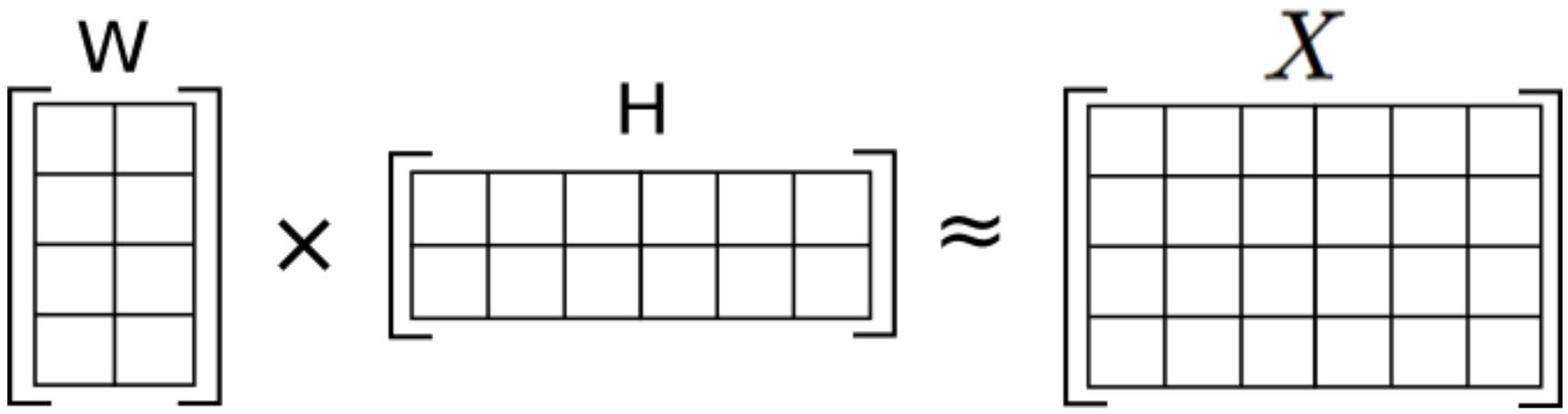
LSA: преимущества

- Одно пространство слов и документов
- Уменьшение размерности
- Широкий взгляд на вещи
- Применение без обучения

LSA: недостатки

- Нормальное распределение
- Многозначность слов
- Зависимость от SVD

NMF: Non-negative matrix factorization



- Интерпретация результата
- Снижение размерности

NMF: проблемы

Некорректно поставлена: если W_0 и H_0 — решения, то

$$W = W_0 Y$$

$$H = Y^{-1} H_0$$

тоже являются решениями

NMF: проблемы

Функция потерь имеет вид:

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

где X — искомая матрица.

$\widetilde{X} = WH$ — неотрицательное разложение.

Функция потерь невыпукла.

Решение: блочно-покоординатные методы оптимизации

NMF: поочередный градиентный спуск

Оптимизационная задача:

$$(W^*, H^*) = argmin||X - WH||_F^2$$

Сначала уменьшаем одну матрицу, потом другую:

$$h_{kj} \to h_{kj} - \upsilon \frac{\delta D_F}{\delta h_{kj}}$$

$$w_{kj} \to w_{kj} - \eta \frac{\delta D_F}{\delta w_{kj}}$$

NMF: мультипликативное обновление

Идея: выбрать шаг градиентного спуска, который не меняет знак.

$$\frac{\delta D_F}{\delta h_{kj}} = \sum_{i=1}^{n} w_{ik}\widetilde{x}_{ij} - \sum_{i=1}^{n} w_{ik}x_{ij},$$

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

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

Функция монотонно невозрастает

NMF: ALS

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

$$X \leftarrow \max\left(\underset{Y \in \mathbb{R}^{k \times n}}{\operatorname{argmin}} \|P - AY\|_F, 0\right) =$$

$$= \max\left(\left(A^T A\right)^{-1} A^T P, 0\right).$$

- ~~Дешево и сердито~~ Быстро и грубо
- Может и не сойтись
- Функция потерь осциллирует

NMF: HASL

Hierarchical alternating least squares (HALS): на каждом шаге находится точный минимум в неотрицательной области по столбцу A или строке X.

Пусть фиксированы все переменные, кроме x_{kj}.

$$x_{kj} \leftarrow \underset{x_{kj} \geqslant 0}{\operatorname{argmin}} \|P - AX\|_F = \max\left(0, \frac{A_{:k}^T P_{:j} - \sum_{l \neq k} A_{:l}^T A_{:l} x_{lj}}{A_{:k}^T A_{:k}}\right).$$

Это решение не зависит от других x в k-й строке, поэтому можно находить точный минимум сразу для всей строки:

$$X_{k:} \leftarrow \underset{X_{k:} \geqslant 0}{\operatorname{argmin}} \|P - AX\|_F = \max\left(0, \frac{A_{:k}^T P - \sum_{l \neq k} A_{:k}^T A_{:l} X_{l:}}{A_{:k}^T A_{:k}}\right).$$

NMF: HASL

- На каждом шаге небольшие вычислительные затраты
- Сходится быстрее мультипликативного обновления
- Чувствителен к начальному приближению

Cholesky

$$A = LL^T$$

A — положительно определенная матрица

L — нижнетреугольная матрица со сторого положительными элементами на диагонали

Cholesky

- Для решения систем линейных уравнений
- Для обращения матриц