Матричные разложения и их применения в ML

Настя Городилова
Ваня Пешехонов
Ира Голобродько

БПМИ191
Факультет Компьютерных Наук
НИУ ВШЭ

28 сентября 2021 г.

SVD in ML

Definition

$$\underset{N\times D}{X} = \underset{N\times N}{U} \cdot \underset{N\times D}{\Sigma} \cdot \underset{D\times D}{V^T}$$

U и V – унитарные,

$\Sigma = diag(\sigma_1, \sigma_2, \ldots) : \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_R > \sigma_{R+1} = 0$

SVD in ML

Math intuition

$$\varphi \ : \ \mathbb{R}^D \longrightarrow \mathbb{R}^N$$

$$A_\varphi = X$$

$\exists$ унитарные базисы $v_1, \ldots, v_D \in \mathbb{R}^D$ и $u_1, \ldots, u_N \in \mathbb{R}^N$:

$$\begin{gathered} \varphi(v_1) = \sigma_1 u_1, \\ \vdots \\ \varphi(v_R) = \sigma_R u_R, \\ \varphi(v_{R+1}) = 0, \\ \vdots \\ \varphi(v_D) = 0 \end{gathered}$$

SVD in ML

Probabilistic interpretation

$X = U\Sigma V^T$ – полное SVD

$$X^T X = V\Sigma^T \underbrace{U^T \cdot U}_{=E} \Sigma V^T = V\Sigma^T\Sigma V^T$$

$\implies$ в унитарном базисе из столбцов V матрица $X^T X$ имеет вид

$\mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots)$

SVD in ML

Probabilistic interpretation

Пусть $x_1, \dots, x_N \sim \mathcal{N}(\mu, C)$:

$$\rho(x_i) = \frac{1}{(2\pi)^{D/2}|C|^{1/2}} e^{-\frac{1}{2}(x_i-\mu)C^{-1}(x_i-\mu)^T}$$

$$\mu = 0 \quad \Rightarrow C = \frac{1}{N}X^T X = V\left(\frac{1}{N}\Sigma^T\Sigma\right)V^T$$

Замена координат $x = zV^T$

$$\rho(x_i) = \mathrm{const} \cdot \exp\left(-\frac{1}{2} \cdot x_i \cdot C^{-1} \cdot x_i^T\right)$$

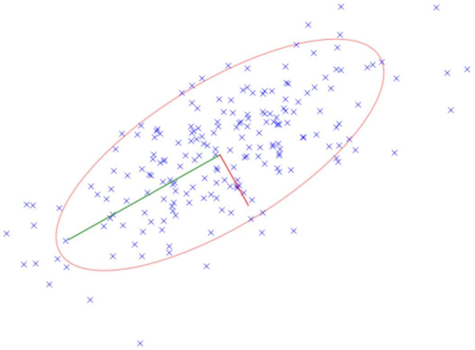
$$\rho(z_i) = \mathrm{const} \cdot \exp\left(-\frac{1}{2} \cdot z_i \underbrace{V^T \cdot V}_{=E} (n\Sigma^{-1}\Sigma^{-T}) \underbrace{V^T \cdot V}_{=E} z_i^T\right) =$$

SVD in ML

Probabilistic interpretation

$$= \text{const} \cdot \exp\left(-\frac{N}{2} z_i \Sigma^{-1} \Sigma^{-T} z_i^T\right) =$$

$$= \text{const} \cdot \exp\left(-\frac{N}{2}\left(\frac{1}{\sigma_1^2} z_{i1}^2 + \frac{1}{\sigma_2^2} z_{i2}^2 + \ldots\right)\right) = \rho(x'_{i1}) \cdot \ldots \cdot \rho(x'_{iD})$$



PCA

Допустим мы разложили матрицу в произведение - и что?

$$\underset{N\times D}{X} \sim \underset{N\times R}{B} \cdot \underset{R\times D}{C}$$

$$x_{ij} \sim \sum_{t=1}^{R} b_{it} \cdot c_{tj}$$

$$\begin{pmatrix} x_{1j} \\ \vdots \\ x_{Nj} \end{pmatrix} = \begin{pmatrix} b_{11}c_{1j} + \cdots + b_{1R}c_{Rj} \\ \vdots \\ b_{N1}c_{1j} + \cdots + b_{NR}c_{Rj} \end{pmatrix} = c_{1j} \cdot \begin{pmatrix} b_{11} \\ \vdots \\ b_{N1} \end{pmatrix} + c_{Rj} \cdot \begin{pmatrix} b_{1R} \\ \vdots \\ b_{NR} \end{pmatrix}$$

B - новая матрица объекты - признаки

C - смешивающая матрица

$$\underset{1\times D}{x_i} \sim \underset{1\times R}{z_i} \cdot \underset{R\times D}{C}$$

PCA

Мотивация описать датасет меньшим числом признаков

- Признаков слишком много
- В данных есть шум
- Признаки линейно зависимы

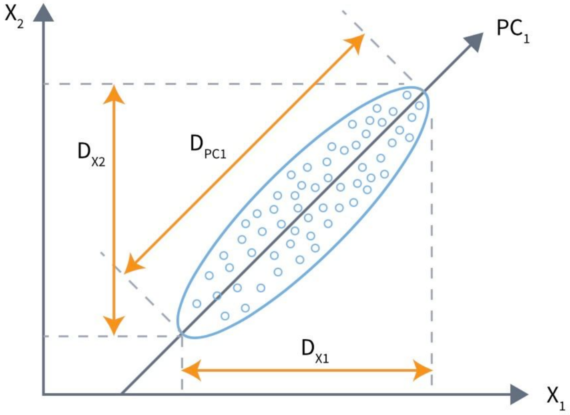
PCA

Идея метода

1. Новое признаковое пространство
2. Новые оси (главные компоненты) - ортогональны
3. Новые признаки - линейная комбинация старых
4. Сохранение большинства информации об исходных данных

PCA

Математика



SVD in PCA

$$C = \text{const} \cdot X^T X \Rightarrow C = \text{const} \cdot V\Sigma^T \underbrace{U^T \cdot U}_{=E} \Sigma V^T = \text{const} \cdot V\Sigma^T\Sigma V^T$$

То есть правые сингулярные векторы X - главные компоненты

$$\begin{tikzcd}
 & \begin{array}{c} X_{N\times D} \\ \text{центрированная матрица с исходными признаками} \end{array} \arrow[dl] \arrow[dr] & \\
C = X^T X \arrow[d] & & U\Sigma V^T \arrow[dd] \\
\begin{array}{c} W, \lambda \\ \text{собств. вектора и числа C} \end{array} \arrow[d] & & \\
\begin{array}{c} Z_K = X \cdot W_K \\ \text{сжимаем размерность} \end{array} \arrow[r] & XW = XV = U\Sigma V^T \cdot V = U\Sigma & \begin{array}{c} Z_K = U\Sigma_K \\ \text{сжимаем} \end{array} \arrow[l]
\end{tikzcd}$$

LSA

Основная идея

Матрица данных X : термы - документы
x_{ij} - частота использования терма i в документе j
Хотим:

- Уменьшить размерность, сохранив похожую структуру зависимостей, присутствующих в исходной матрице
- Получаем на выход две матрицы описывающие
  1. Темы для каждого текста
  2. Слова для каждой темы

LSA

Отличие от PCA

- Лучше работает с разреженными данными
- PCA - конкретный метод, который может применяться в задаче LSA тоже, но в LSA могут применяться и другие виды матричных разложений

LSA

SVD in LSA

x_{ij} - описывает вхождение слова i в документ j

$$\begin{tikzcd}[column sep=small] & U\Sigma V^T \arrow[dl] \arrow[d] \arrow[dr] & \\ \underset{term-concept}{U} & \Sigma & \underset{concept-document}{V^T} \end{tikzcd}$$

$$X_R = \left(U_K \Sigma_K^{\frac{1}{2}}\right)\left(\Sigma_K^{\frac{1}{2}} V_K^T\right)$$

Чего не умеет SVD?

1. Данные не всегда распределены нормально, они могут обладать сложной геометрией, но SVD будет упрямо искать эллипсоид
2. Самое важное не всегда самое масштабное
3. Новые признаки не обязаны быть хорошо интерпретируемыми
4. Выбросы почти наверняка усложнят жизнь

Применение матричных разложений в рекомендательных системах

Intro

Мотивация создания рекомендательных систем

Intro

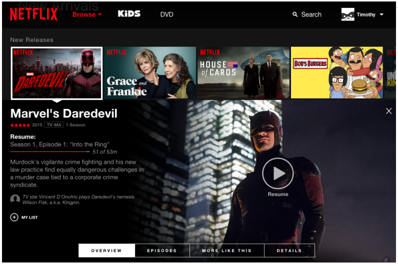
Мотивация создания рекомендательных систем

- Предлагаем пользователю услуги/товары, которые могли бы заинтересовать

Intro

Мотивация создания рекомендательных систем

- Предлагаем пользователю услуги/товары, которые могли бы заинтересовать
- Используем информацию о профиле пользователя, статистику



Intro

Формализация задачи

- Есть объекты (фильмы/товары/услуги), есть пользователи (покупатели/зрители/клиенты)

Intro

Формализация задачи

- Есть объекты (фильмы/товары/услуги), есть пользователи (покупатели/зрители/клиенты)
- Предсказываем рейтинг, который пользователь поставит объекту

Intro

Формализация задачи

- Есть объекты (фильмы/товары/услуги), есть пользователи (покупатели/зрители/клиенты)
- Предсказываем рейтинг, который пользователь поставит объекту
- Рейтинги: дискретные, непрерывные, бинарные

Intro

Фильтрация: контентная vs совместная

- Контентная фильтрация – на основе заранее составленных характеристик

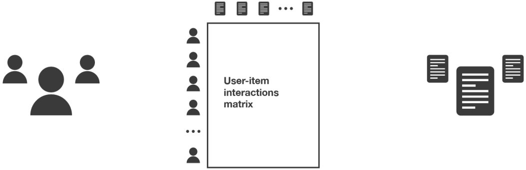
Intro

Фильтрация: контентная vs совместная

- Контентная фильтрация – на основе заранее составленных характеристик
- Совместная (коллаборативная) фильтрация – на основе информации о предыдущих действиях пользователя, а также о выборе, которые делали пользователи со схожим поведением

Intro

Совместная фильтрация и матрица user-item



Users	User-item interactions matrix	Items
suscribers	rating given by a user to a movie (integer)	movies
readers	time spent by a reader on an article (float)	articles
buyers	product clicked or not when suggested (boolean)	products
	...	

Intro

Совместная фильтрация и матрица user-item

Users	User-item interactions matrix	Items
suscribers	rating given by a user to a movie (integer)	movies
readers	time spent by a reader on an article (float)	articles
buyers	product clicked or not when suggested (boolean)	products

- Матрица user-item – разреженная

Intro

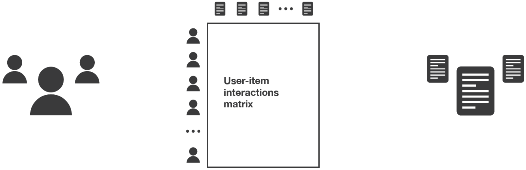
Совместная фильтрация и матрица user-item

Users	User-item interactions matrix	Items
suscribers	rating given by a user to a movie (integer)	movies
readers	time spent by a reader on an article (float)	articles
buyers	product clicked or not when suggested (boolean)	products
	...	

- Матрица user-item – разреженная
- Наша цель – <u>восполнить пропуски</u>

Intro

Совместная фильтрация и матрица user-item



Users	User-item interactions matrix	Items
suscribers	rating given by a user to a movie (integer)	movies
readers	time spent by a reader on an article (float)	articles
buyers	product clicked or not when suggested (boolean)	products

•••

- Матрица user-item – разреженная
- Наша цель – восполнить пропуски
- Netfix prize: рекомендательные системы, использующие факторизацию матрицы user-item

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

Супер-краткое напоминание по PCA

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

Супер-краткое напоминание по PCA:

- Пусть X – наш датасет. Главные компоненты – собственные векторы матрицы ковариаций X.

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

Супер-краткое напоминание по PCA:

- Пусть X – наш датасет. Главные компоненты – собственные векторы матрицы ковариаций X.
- Любой объект (вектор признаков) можно восстановить как линейную комбинацию главных компонент

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

Супер-краткое напоминание по PCA:

- Пусть X – наш датасет. Главные компоненты – собственные векторы матрицы ковариаций X.
- Любой объект (вектор признаков) можно восстановить как линейную комбинацию главных компонент

$$\begin{aligned} \text{Face 1} &= \alpha_1 \cdot \text{Creepy guy 1} + \cdots + \alpha_{400} \cdot \text{Creepy guy 400}; \\ \text{Face 2} &= \beta_1 \cdot \text{Creepy guy 1} + \cdots + \beta_{400} \cdot \text{Creepy guy 400}. \end{aligned}$$

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

- Цель метода главных компонент – выявить типичные векторы (объекты), которые мы будем называть скрытыми факторами. Объект – это сочетание скрытых факторов с характерными весами.

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

- Цель метода главных компонент – выявить типичные векторы (объекты), которые мы будем называть скрытыми факторами. Объект – это сочетание скрытых факторов с характерными весами.
- Очевидно ли, как метод главных компонент распространяется на матрице предпочтений пользователей?

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

- Цель метода главных компонент – выявить типичные векторы (объекты), которые мы будем называть скрытыми факторами. Объект – это сочетание скрытых факторов с характерными весами.
- Очевидно ли, как метод главных компонент распространяется на матрице предпочтений пользователей?

$$R = \begin{array}{cc} \begin{pmatrix} 1 & ? & 2 & ? & ? \\ ? & ? & ? & ? & 4 \\ 2 & ? & 4 & 5 & ? \\ ? & ? & 3 & ? & ? \\ ? & 1 & ? & 3 & ? \\ 5 & ? & ? & ? & 2 \end{pmatrix} & \begin{array}{c} \text{Alice} \\ \text{Bob} \\ \text{Charlie} \\ \text{Daniel} \\ \text{Eric} \\ \text{Frank} \end{array} \end{array}$$

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

- Цель метода главных компонент – выявить типичные векторы (объекты), которые мы будем называть скрытыми факторами. Объект – это сочетание скрытых факторов с характерными весами.
- Очевидно ли, как метод главных компонент распространяется на матрице предпочтений пользователей?

$$R = \begin{array}{cl} \begin{pmatrix} 1 & 1 & 2 & 2 & 1 \\ 4 & 4 & 5 & 2 & 4 \\ 2 & 2 & 4 & 5 & 3 \\ 4 & 1 & 3 & 3 & 4 \\ 4 & 1 & 2 & 3 & 2 \\ 5 & 2 & 3 & 4 & 2 \end{pmatrix} & \begin{array}{c} \text{Alice} \\ \text{Bob} \\ \text{Charlie} \\ \text{Daniel} \\ \text{Eric} \\ \text{Frank} \end{array} \end{array}$$

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

- Типичные любители каких-то жанров:

$$\text{Alice} = 10\%\ \text{Action fan} + 10\%\ \text{Comedy fan} + 50\%\ \text{Romance fan} + \cdots$$

$$\text{Bob} = 50\%\ \text{Action fan} + 30\%\ \text{Comedy fan} + 10\%\ \text{Romance fan} + \cdots$$

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

- Типичные любители каких-то жанров:

$$\text{Alice} = 10\%\ \text{Action fan} + 10\%\ \text{Comedy fan} + 50\%\ \text{Romance fan} + \cdots$$

$$\text{Bob} = 50\%\ \text{Action fan} + 30\%\ \text{Comedy fan} + 10\%\ \text{Romance fan} + \cdots$$

- А если транспонировать?

$$R^T = \begin{pmatrix} — & \text{Titanic} & — \\ — & \text{Toy Story} & — \\ & \vdots & \\ — & \text{Fargo} & — \end{pmatrix}$$

Анализируем данные с помощью PCA & SVD

PCA над матрицей рейтингов

- Типичные любители каких-то жанров:

$$\text{Alice} = 10\%\ \text{Action fan} + 10\%\ \text{Comedy fan} + 50\%\ \text{Romance fan} + \cdots$$

$$\text{Bob} = 50\%\ \text{Action fan} + 30\%\ \text{Comedy fan} + 10\%\ \text{Romance fan} + \cdots$$

- А если транспонировать?

$$R^T = \begin{pmatrix} — & \text{Titanic} & — \\ — & \text{Toy Story} & — \\ & \vdots & \\ — & \text{Fargo} & — \end{pmatrix}$$

- Типичные фильмы :)

Анализируем данные с помощью PCA & SVD

Разложение матрицы рейтингов

- Пусть $R = MU^T$ – скелетное разложение, основанное на SVD.

Анализируем данные с помощью PCA & SVD

Разложение матрицы рейтингов

- Пусть $R = MU^T$ – скелетное разложение, основанное на SVD.

$$\begin{pmatrix} & & \\ & r_{ui} & \\ & & \end{pmatrix} = \begin{pmatrix} — & p_u & — \end{pmatrix} \begin{pmatrix} & | & \\ & q_i & \\ & | & \end{pmatrix} \implies r_{ui} = p_u \cdot q_i.$$

Анализируем данные с помощью PCA & SVD

Разложение матрицы рейтингов

- Пусть $R = MU^T$ – скелетное разложение, основанное на SVD.

$$\begin{pmatrix} & & \\ & r_{ui} & \\ & & \end{pmatrix} = \begin{pmatrix} — & p_u & — \end{pmatrix} \begin{pmatrix} & | & \\ & q_i & \\ & | & \end{pmatrix} \implies r_{ui} = p_u \cdot q_i.$$

- r_{ui} – оценка пользователя u, который посмотрел фильм i

Анализируем данные с помощью PCA & SVD

Разложение матрицы рейтингов

- Пусть $R = MU^T$ – скелетное разложение, основанное на SVD.

$$\begin{pmatrix} & & \\ & r_{ui} & \\ & & \end{pmatrix} = \begin{pmatrix} — & p_u & — \end{pmatrix} \begin{pmatrix} & | & \\ & q_i & \\ & | & \end{pmatrix} \implies r_{ui} = p_u \cdot q_i.$$

- r_{ui} – оценка пользователя u, который посмотрел фильм i
- p_u – коэффициенты при главных компонентах в линейном выражении профиля пользователя

Анализируем данные с помощью PCA & SVD

Разложение матрицы рейтингов

- Пусть $R = MU^T$ – скелетное разложение, основанное на SVD.

$$\begin{pmatrix} & & \\ & r_{ui} & \\ & & \end{pmatrix} = \begin{pmatrix} - & p_u & - \end{pmatrix} \begin{pmatrix} & | & \\ & q_i & \\ & | & \end{pmatrix} \implies r_{ui} = p_u \cdot q_i.$$

- r_{ui} – оценка пользователя u, который посмотрел фильм i
- p_u – коэффициенты при главных компонентах в линейном выражении профиля пользователя

$$\text{Alice} = 10\%\ \text{Action fan} + 10\%\ \text{Comedy fan} + 50\%\ \text{Romance fan} + \cdots$$

Анализируем данные с помощью PCA & SVD

Разложение матрицы рейтингов

- Пусть $R = MU^T$ – скелетное разложение, основанное на SVD.

$$\begin{pmatrix} & & \\ & r_{ui} & \\ & & \end{pmatrix} = \begin{pmatrix} — & p_u & — \end{pmatrix} \begin{pmatrix} & | & \\ & q_i & \\ & | & \end{pmatrix} \Longrightarrow r_{ui} = p_u \cdot q_i.$$

- r_{ui} – оценка пользователя u, который посмотрел фильм i
- p_u – коэффициенты при главных компонентах в линейном выражении профиля пользователя

$$\text{Alice} = 10\% \text{ Action fan} + 10\% \text{ Comedy fan} + 50\% \text{ Romance fan} + \cdots$$

Анализируем данные с помощью PCA & SVD

SVD-разложение матрицы рейтингов

- q_i – коэффициенты при главных компонентах в линейном выражении профиля фильма

Анализируем данные с помощью PCA & SVD

SVD-разложение матрицы рейтингов

- q_i – коэффициенты при главных компонентах в линейном выражении профиля фильма

$$\text{Titanic} = 20\%\ \text{Action} + 00\%\ \text{Comedy} + 70\%\ \text{Romance} + \cdots$$

Анализируем данные с помощью PCA & SVD

SVD-разложение матрицы рейтингов

- q_i – коэффициенты при главных компонентах в линейном выражении профиля фильма

$$\text{Titanic} = 20\%\ \text{Action} + 00\%\ \text{Comedy} + 70\%\ \text{Romance} + \cdots$$

- p_u отражает близость пользователя к одному из скрытых признаков; q_i отражает близость фильма к одному из скрытых признаков.

Анализируем данные с помощью PCA & SVD

SVD-разложение матрицы рейтингов

- q_i – коэффициенты при главных компонентах в линейном выражении профиля фильма

$$\text{Titanic} = 20\%\ \text{Action} + 00\%\ \text{Comedy} + 70\%\ \text{Romance} + \cdots$$

- p_u отражает близость пользователя к одному из скрытых признаков; q_i отражает близость фильма к одному из скрытых признаков.
- $$\boxed{r_{ui} = p_u \cdot q_i = \sum_{f \in \text{latent factors}} \text{affinity of } u \text{ for } f \times \text{affinity of } i \text{ for } f}$$

Пример алгоритма рекомендательной системы

SVD для предсказания рейтинга

- Теперь перейдём к разреженным матрицам

Пример алгоритма рекомендательной системы

SVD для предсказания рейтинга

- Теперь перейдём к разреженным матрицам
- Если пытаться чем-то заполнить пропуски, получим сильное отклонение

Пример алгоритма рекомендательной системы

SVD для предсказания рейтинга

- Лучше – оптимизационная задача: ищем такие наборы векторов $\{p_u\}$, $\{q_i\}$, так что

$$\begin{cases} r_{ui} = p_u \cdot q_i \; \forall u, i \\ p_u \text{ ортогональны друг другу} \\ q_i \text{ ортогональны друг другу} \end{cases}$$

Пример алгоритма рекомендательной системы

SVD для предсказания рейтинга

- Лучше – оптимизационная задача: ищем такие наборы векторов $\{p_u\}$, $\{q_i\}$, так что

$$\begin{cases} r_{ui} = p_u \cdot q_i \ \forall u, i \\ p_u \text{ ортогональны друг другу} \\ q_i \text{ ортогональны друг другу} \end{cases}$$

- По сути ищем $$\min_{p_u, q_i:\ p_u \perp p_v;\ q_i \perp q_j} \sum_{r_{ui} \in R} (r_{ui} - p_u \cdot q_i)^2$$

Пример алгоритма рекомендательной системы

SVD для предсказания рейтинга

- Разреженный случай:

$$\min_{p_u, q_i} \sum_{r_{ui} \in R} (r_{ui} - p_u \cdot q_i)^2.$$

Пример алгоритма рекомендательной системы

SVD для предсказания рейтинга

- Разреженный случай:

$$\min_{p_u, q_i} \sum_{r_{ui} \in R} (r_{ui} - p_u \cdot q_i)^2.$$

- Решается с помощью *SGD*

Пример алгоритма рекомендательной системы

SVD для предсказания рейтинга

- Разреженный случай:

$$\min_{p_u, q_i} \sum_{r_{ui} \in R} (r_{ui} - p_u \cdot q_i)^2.$$

- Решается с помощью *SGD*

Closure

Проблемы и направления развития

- Cold-start problem

Closure

Проблемы и направления развития

- Cold-start problem
- Чтобы строить рекомендации касательно новых объектов, необходимо накопить достаточное число оценок от пользователей

Closure

Проблемы и направления развития

- Cold-start problem
- Чтобы строить рекомендации касательно новых объектов, необходимо накопить достаточное число оценок от пользователей. Чтобы рекомендовать что-то новому пользователю, нужно немного за ним понаблюдать

Closure

Проблемы и направления развития

- Cold-start problem
- Чтобы строить рекомендации касательно новых объектов, необходимо накопить достаточное число оценок от пользователей. Чтобы рекомендовать что-то новому пользователю, нужно немного за ним понаблюдать
- Rich-get-richer effect

Closure

Проблемы и направления развития

Closure

Проблемы и направления развития

- Большое количество данных

Closure

Проблемы и направления развития

- Большое количество данных
- Кросс-системная фильтрация – несколько систем обмениваются друг с другом шаблонами поведения

Closure

Проблемы и направления развития

- Большое количество данных
- Кросс-системная фильтрация – несколько систем обмениваются друг с другом шаблонами поведения
- Устойчивость к манипуляциям (robust collaborative filtering)

Источники

N. Hug. **Understanding matrix factorization for recommendation**. 2017. URL: https://nicolas-hug.com/blog/.

N. Hug. **Collaborative filtering for recommendation systems in Python**. 2017. URL: https://youtu.be/z0dx-YckFko.

R. B. Yehuda Koren и C. Volinsky. **MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS**. 2009. URL: https://datajobs.com/data-science-repo/Recommender-Systems-%5C%5BNetflix%5C%5D.pdf.

Танцы вокруг линейной регрессии

Связь линейной регрессии и задачи наименьших квадратов

Дано:

$X \in \mathbb{R}^{\ell \times d}$— матрица объекты-признаки, $y \in \mathbb{R}^{\ell}$— вектор ответов

Рассмотрим семейство линейных моделей, которые дают предсказание, равное линейной комбинации признаков:

$$\mathcal{A} = \left\{ a(x_i) = w_0 + w_1 x_{i1} + \ldots + w_d x_{id} \;\middle|\; x_i = X[i, :], (w_0, \ldots, w_d) \in \mathbb{R}^{d+1} \right\}$$

Танцы вокруг линейной регрессии

Связь линейной регрессии и задачи наименьших квадратов

Дано:

$X \in \mathbb{R}^{\ell \times d}$— матрица объекты-признаки, $y \in \mathbb{R}^{\ell}$— вектор ответов

Рассмотрим семейство линейных моделей, которые дают предсказание, равное линейной комбинации признаков:

$$\mathcal{A} = \left\{ a(x_i) = w_0 + w_1 x_{i1} + \ldots + w_d x_{id} \;\middle|\; x_i = X[i, :], (w_0, \ldots, w_d) \in \mathbb{R}^{d+1} \right\}$$

Будем искать в этом семействе модель, лучшую в смысле MSE:

$$\text{MSE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(a(x_i) - y_i\right)^2 \to \min_{a \in \mathcal{A}}$$

Танцы вокруг линейной регрессии

Связь линейной регрессии и задачи наименьших квадратов

Распишем MSE:

$$\mathrm{MSE}(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}(w_0 + w_1 x_{i1} + \ldots + w_d x_{id} - y_i)^2 =$$

$$= |x_{i0} = 1| = \frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2 =$$

$$= \frac{1}{\ell}(Xw - y)^T(Xw - y)$$

где $w = (w_0, w_1, \ldots, w_d)$ — вектор весов.

Танцы вокруг линейной регрессии

Связь линейной регрессии и задачи наименьших квадратов

Распишем MSE:

$$\mathrm{MSE}(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}(w_0 + w_1 x_{i1} + \ldots + w_d x_{id} - y_i)^2 =$$

$$= |x_{i0} = 1| = \frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2 =$$

$$= \frac{1}{\ell}(Xw - y)^T(Xw - y)$$

где $w = (w_0, w_1, \ldots, w_d)$ — вектор весов.

$$\mathrm{MSE}(a, X) \to \min_{a \in \mathcal{A}} \iff (Xw - y)^T(Xw - y) \to \min_{w \in \mathbb{R}^{d+1}}$$

Танцы вокруг линейной регрессии

Формулировка задачи наименьших квадратов

Пусть есть $X \in \mathbb{R}^{\ell \times d}$ — матрица данных, $y \in \mathbb{R}^{\ell}$ — вектор ответов, $w \in \mathbb{R}^{d}$ — неизвестный вектор. Преполагается, что $\ell \gg d$.

Танцы вокруг линейной регрессии

Формулировка задачи наименьших квадратов

Пусть есть $X \in \mathbb{R}^{\ell \times d}$ — матрица данных, $y \in \mathbb{R}^{\ell}$ — вектор ответов, $w \in \mathbb{R}^d$ — неизвестный вектор. Преполагается, что $\ell \gg d$.

Хотим найти вектор w, такой что $Xw = y$. Вообще говоря, такая система не имеет точного решения, поэтому вместо точного решения будем искать w, который является решением *задачи наименьших квадратов (задачи LS)*:

$$\|Xw - y\|_2 \to \min_{w \in \mathbb{R}^d}$$

Танцы вокруг линейной регрессии

Формулировка задачи наименьших квадратов

$$\|Xw - y\|_2 \to \min_{w \in \mathbb{R}^d}$$

Почему минимизируем именно такой функционал?

1. Это разумно;

Танцы вокруг линейной регрессии

Формулировка задачи наименьших квадратов

$$\|Xw - y\|_2 \to \min_{w \in \mathbb{R}^d}$$

Почему минимизируем именно такой функционал?

1. Это разумно;
2. Следующие задачи оптимизации эквивалентны:

$$\|Xw-y\|_2 \to \min_{w \in \mathbb{R}^d} \iff \|Xw-y\|_2^2 \to \min_{w \in \mathbb{R}^d} \iff \frac{1}{2}\|Xw-y\|_2^2 \to \min_{w \in \mathbb{R}^d}$$

Танцы вокруг линейной регрессии

Формулировка задачи наименьших квадратов

$$\|Xw - y\|_2 \to \min_{w \in \mathbb{R}^d}$$

Почему минимизируем именно такой функционал?

1. Это разумно;
2. Следующие задачи оптимизации эквивалентны:

$$\|Xw - y\|_2 \to \min_{w \in \mathbb{R}^d} \iff \|Xw - y\|_2^2 \to \min_{w \in \mathbb{R}^d} \iff \frac{1}{2}\|Xw - y\|_2^2 \to \min_{w \in \mathbb{R}^d}$$

3. $\psi(w) = \frac{1}{2}\|Xw - y\|_2^2$ дифференцируем. $\nabla\psi(w) = X^T(Xw - y)$, $w = \min \psi \implies \nabla\psi(w) = 0$;

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Пусть матрица данных $X \in \mathbb{R}^{\ell \times d}$ имеет ранг d (что это значит, если интерпретировать X как матрицу объекты-признаки?).

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Пусть матрица данных $X \in \mathbb{R}^{\ell \times d}$ имеет ранг d (что это значит, если интерпретировать X как матрицу объекты-признаки?).

Утверждение: решение задачи LS единственно, и удовлетворяет симметричной, положительно определённой системе

$$X^T X w = X^T y$$

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 0: В лоб.

Матрица $X^T X$ квадратная и невырожденная:

$$w = (X^T X)^{-1} X^T y$$

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 0: В лоб.

Матрица $X^T X$ квадратная и невырожденная:

$$w = (X^T X)^{-1} X^T y$$

Сложность: $3d^3 + 2\ell d^2 + \mathcal{O}(\ell d)$.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 0: В лоб.

Матрица $X^T X$ квадратная и невырожденная:

$$w = (X^T X)^{-1} X^T y$$

Сложность: $3d^3 + 2\ell d^2 + \mathcal{O}(\ell d)$.

Численная устойчивость: хреновая.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 1: SVD-разложение.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 1: SVD-разложение.

$X = U\Sigma V^T \implies X^T X = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$

$$V\Sigma^2 V^T w = V\Sigma U^T y \iff \Sigma V^T w = U^T y \iff w = V\Sigma^{-1} U^T y$$

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 1: SVD-разложение.

$X = U\Sigma V^T \implies X^T X = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$

$$V\Sigma^2 V^T w = V\Sigma U^T y \iff \Sigma V^T w = U^T y \iff w = V\Sigma^{-1} U^T y$$

Сложность: $9d^3 + 12\ell d^2 + \mathcal{O}(\ell d)$.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 1: SVD-разложение.

$X = U\Sigma V^T \implies X^T X = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$

$$V\Sigma^2 V^T w = V\Sigma U^T y \iff \Sigma V^T w = U^T y \iff w = V\Sigma^{-1} U^T y$$

Сложность: $9d^3 + 12\ell d^2 + \mathcal{O}(\ell d)$.

Численная устойчивость: ничё такая.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 2: QR-разложение.

Танцы вокруг линейной регрессии

QR-разложение

Для произвольной вещественной матрицы $A \in \mathbb{R}^{m \times n}$ существуют единственные матрицы $Q \in \mathbb{R}^{m \times n}$ — унитарная, и $R \in \mathbb{R}^{n \times n}$ — верхнетреугольная, такие что

$$A = QR$$

Сложность вычисления: $-\frac{2}{3}n^3 + 2mn^2 + \mathcal{O}(mn)$.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 2: QR-разложение.

$X = QR \implies X^T X = R^T Q^T Q R = R^T R$

$$R^T R w = R^T Q y \iff Rw = Qy$$

Далее формулы для w_j могут быть выписаны явно.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 2: QR-разложение.

$X = QR \implies X^T X = R^T Q^T Q R = R^T R$

$$R^T R w = R^T Q y \iff R w = Q y$$

Далее формулы для w_j могут быть выписаны явно.

Сложность: $-\frac{2}{3} n^3 + 2mn^2 + \mathcal{O}(mn)$.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 2: QR-разложение.

$X = QR \implies X^T X = R^T Q^T Q R = R^T R$

$$R^T R w = R^T Q y \iff R w = Q y$$

Далее формулы для w_j могут быть выписаны явно.

Сложность: $-\frac{2}{3}n^3 + 2mn^2 + \mathcal{O}(mn)$.

Численная устойчивость: приличная.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 3: Разложение Холецкого.

Танцы вокруг линейной регрессии

Разложение Холецкого

Для всякой симметричной положительно-определённой матрицы $A \in \mathbb{R}^{n \times n}$ существует единственная нижнетреугольная матрица $L \in \mathbb{R}^{n \times n}$, такая что матрица A представима в виде

$$A = LL^T$$

Если A — симметричная положительно-определённая матрица, то элементы матрицы L вычисляются итерационно сверху вниз, слева направо:

$$L_{11} = \sqrt{A_{11}} \qquad L_{j1} = \frac{A_{j1}}{L_{11}}$$

$$L_{ii} = \sqrt{A_{ii} - \sum_{k=1}^{i-1} L_{ik}^2} \qquad L_{ji} = \frac{A_{ji} - \sum_{k=1}^{i-1} L_{ik} L_{jk}}{L_{ii}}$$

Сложность вычисления: $\frac{1}{3} n^3 + mn^2 + \mathcal{O}(mn)$.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 3: Разложение Холецкого.

$X^T X = LL^T$. Положим $L^T w = v$. Решаем две системы:

1. $Lv = X^T y$;
2. $L^T w = v$;

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 3: Разложение Холецкого.

$X^T X = LL^T$. Положим $L^T w = v$. Решаем две системы:

1. $Lv = X^T y$;
2. $L^T w = v$;

Сложность: $\frac{1}{3} n^3 + mn^2 + \mathcal{O}(mn)$.

Танцы вокруг линейной регрессии

Задача LS. Полноранговый случай

Как решать?

$$X^T X w = X^T y$$

Способ № 3: Разложение Холецкого.

$X^T X = LL^T$. Положим $L^T w = v$. Решаем две системы:

1. $Lv = X^T y$;
2. $L^T w = v$;

Сложность: $\frac{1}{3}n^3 + mn^2 + \mathcal{O}(mn)$

Численная устойчивость: откровенно не очень.

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Теперь будем считать, что матрица данных $X \in \mathbb{R}^{\ell \times d}$ имеет ранг $r < d$.
Пусть w — решение задачи LS, а $z \in \ker X$. Тогда $w + z$ — тоже решение задачи LS.
Множество решений задачи LS для матрицы с неполным рангом

$$\chi = \left\{ w_* + z \;\middle|\; w_* = \min_w \|Xw - y\|_2, z \in \ker X \right\}$$

не выпукло.
Ясно, что среди всех $w \in \chi$ существует одно решение с минимальной нормой.

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Теперь будем считать, что матрица данных $X \in \mathbb{R}^{\ell \times d}$ имеет ранг $r < d$.

Пусть w — решение задачи LS, а $z \in \ker X$. Тогда $w + z$ — тоже решение задачи LS.

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Теперь будем считать, что матрица данных $X \in \mathbb{R}^{\ell \times d}$ имеет ранг $r < d$.

Пусть w — решение задачи LS, а $z \in \ker X$. Тогда $w + z$ — тоже решение задачи LS.

Множество решений задачи LS для матрицы с неполным рангом

$$\chi = \left\{ w_* + z \;\middle|\; w_* = \min_w \|Xw - y\|_2, z \in \ker X \right\}$$

не выпукло.

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Теперь будем считать, что матрица данных $X \in \mathbb{R}^{\ell \times d}$ имеет ранг $r < d$.

Пусть w — решение задачи LS, а $z \in \operatorname{ker} X$. Тогда $w + z$ — тоже решение задачи LS.

Множество решений задачи LS для матрицы с неполным рангом

$$\chi = \left\{ w_* + z \,\middle|\, w_* = \min_w \|Xw - y\|_2, z \in \operatorname{ker} X \right\}$$

не выпукло.

Ясно, что среди всех $w \in \chi$ существует одно решение с минимальной нормой.

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Как искать $w_* = \min\limits_{v \in \chi} \|Xv - y\|_2 : \|w_*\|_2 \to \min$?

Theorem

Пусть $X = U\Sigma V^T$. *Тогда* $w_* = \sum\limits_{i=1}^{r} \frac{u_i^T y}{\sigma_i} v_i = V\Sigma^+ U^T y$, *где*
$\Sigma^+ = diag\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right)$.

Танцы вокруг линейной регрессии

Псевдообратная

Definition

Пусть $A \in \mathbb{R}^{m\times n}$ и $A = U\Sigma V^T$ — сингулярное разложение. Матрица $A^+ \in \mathbb{R}^{n\times m}$ называется *псевдообратной матрицей к A, и определяется равенством* $A^+ = V\Sigma^+ U^T$,
$\Sigma^+ = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right)$.

Fun facts about A^+:
(1)

- $AA^+A = A$
- $A^+AA^+ = A^+$
- $(A^+A)^T = A^+A$
- $(AA^+)^T = AA^+$

(2)
AA^+ — ортогональный проектор на $\mathrm{Im}\ A$;
$E - A^+A$ — ортогональный проектор на $\ker A$;

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Theorem

Пусть $X = U\Sigma V^T$. *Тогда* $w_* = \sum_{i=1}^{r} \frac{u_i^T y}{\sigma_i} v_i = X^+ y$.

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Theorem

Пусть $X = U\Sigma V^T$. *Тогда* $w_* = \sum_{i=1}^{r} \frac{u_i^T y}{\sigma_i} v_i = X^+ y$.

Итоги теоремы:

- Всякое решение задачи LS имеет вид $w = X^+ y + (E - X^+ X)v, \ v \in \mathbb{R}^d$;

Танцы вокруг линейной регрессии

Задача LS. Случай неполного ранга

Theorem

Пусть $X = U\Sigma V^T$. *Тогда* $w_* = \sum_{i=1}^{r} \frac{u_i^T y}{\sigma_i} v_i = X^+ y$.

Итоги теоремы:

- Всякое решение задачи LS имеет вид $w = X^+ y + (E - X^+ X)v,\ v \in \mathbb{R}^d$;
- Решение $w_* = X^+ y$ имеет наменьшую 2-норму среди всех решений;

Танцы вокруг линейной регрессии

Регуляризация псевдообратной

- https://colab.research.google.com/drive/1kbURiCK4-OlZMKo1hWmLipzoGZ48Q4Ez?usp=sharing

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$
$$\nabla J_\mu(w) = X^T(Xw - y) + \mu w$$
$$\nabla J_\mu(w) = 0 \iff (X^T X + \mu E)w = X^T y$$

Обозначим $B(\mu) = (X^T X + \mu E)$. Несколько наблюдений:

1. $B(\mu)$ — симметричная матрица;
2. $B(\mu)$ — невырождена для любого $\mu > 0$;
3. Если $X^T X$ была близка к вырожденой, то с помощью слагаемого μE её можно немного "подвинуть-и сделать более численно устойчивой;
4. $B(\mu)^{-1} X^T \xrightarrow[\mu \to 0]{} X^+$;

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$
$$\nabla J_\mu(w) = X^T(Xw - y) + \mu w$$

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$

$$\nabla J_\mu(w) = X^T(Xw - y) + \mu w$$

$$\nabla J_\mu(w) = 0 \iff (X^T X + \mu E)w = X^T y$$

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$
$$\nabla J_\mu(w) = X^T(Xw - y) + \mu w$$
$$\nabla J_\mu(w) = 0 \iff (X^T X + \mu E)w = X^T y$$

Обозначим $B(\mu) = (X^T X + \mu E)$. Несколько наблюдений:

1. $B(\mu)$ — симметричная матрица;

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$
$$\nabla J_\mu(w) = X^T(Xw - y) + \mu w$$
$$\nabla J_\mu(w) = 0 \iff (X^T X + \mu E)w = X^T y$$

Обозначим $B(\mu) = (X^T X + \mu E)$. Несколько наблюдений:

1. $B(\mu)$ — симметричная матрица;
2. $B(\mu)$ — невырождена для любого $\mu > 0$;

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$

$$\nabla J_\mu(w) = X^T(Xw - y) + \mu w$$

$$\nabla J_\mu(w) = 0 \iff (X^T X + \mu E)w = X^T y$$

Обозначим $B(\mu) = (X^T X + \mu E)$. Несколько наблюдений:

1. $B(\mu)$ — симметричная матрица;
2. $B(\mu)$ — невырождена для любого $\mu > 0$;
3. Если $X^T X$ была близка к вырожденой, то с помощью слагаемого μE её можно немного "подвинуть-и сделать более численно устойчивой;

Танцы вокруг линейной регрессии

Регуляризация Тихонова (Ridge regression)

Откатимся в самое начало и будем минимизировать другой функционал:

$$J_\mu(w) = \frac{1}{2}\|Xw - y\|_2^2 + \frac{\mu}{2}\|w\|_2$$
$$\nabla J_\mu(w) = X^T(Xw - y) + \mu w$$
$$\nabla J_\mu(w) = 0 \iff (X^T X + \mu E)w = X^T y$$

Обозначим $B(\mu) = (X^T X + \mu E)$. Несколько наблюдений:

1. $B(\mu)$ — симметричная матрица;
2. $B(\mu)$ — невырождена для любого $\mu > 0$;
3. Если $X^T X$ была близка к вырожденой, то с помощью слагаемого μE её можно немного "подвинуть-и сделать более численно устойчивой;
4. $B(\mu)^{-1} X^T \xrightarrow[\mu \to 0]{} X^+$;

Ссылки на источники

URL: https://disk.yandex.ru/i/EBKCbSO-UGHEUQ.

URL: https://yadi.sk/i/74eD9NfYaACvIw.

URL: https://yadi.sk/i/5KeAH1gLuGgc2w.

Y. Eldar и G. Kutyniok. **Compressed Sensing: Theory and Applications**. янв. 2012. ISBN: 978-1107005587. DOI: 10.1017/CBO9780511794308.