Блок

Feature Engineering

Занятие № 5

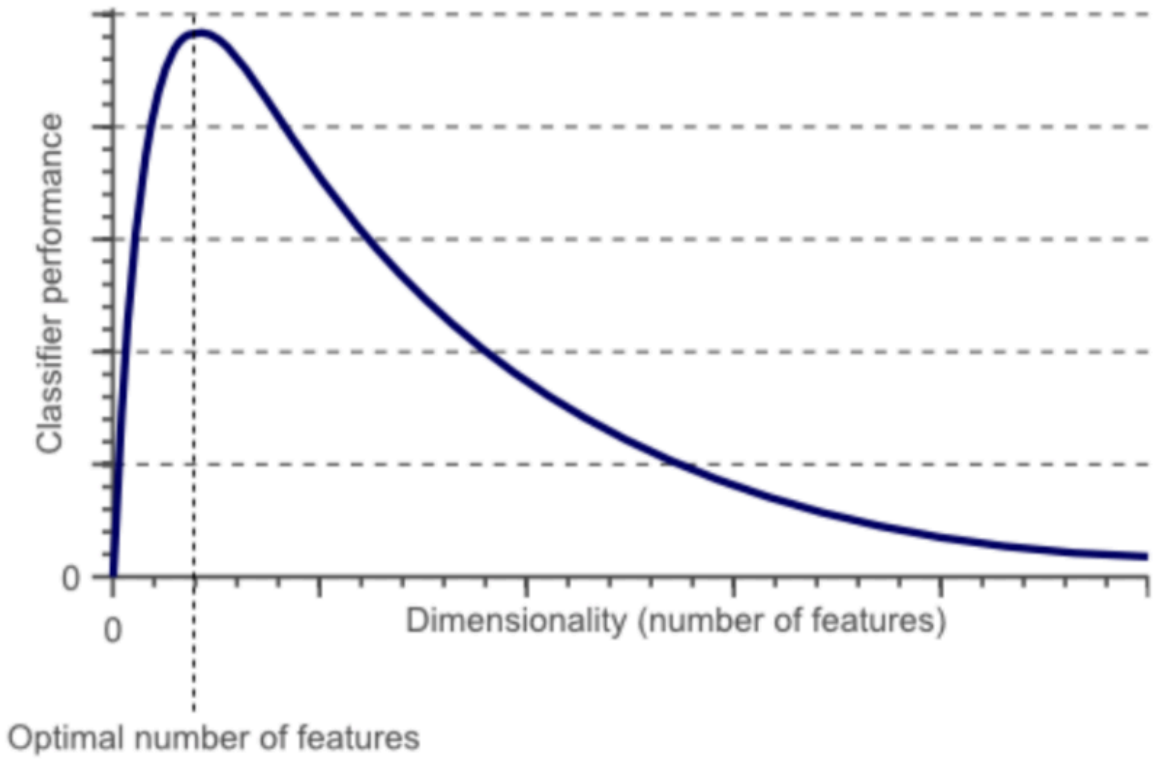
Feature Selection

Цели занятия

- Вспомним методы стастики
- Узнаем про методы декомпозиции данных
- Понять принцип PCA на практике

Зачем всё это?

ПРОКЛЯТЬЕ РАЗМЕРНОСТИ

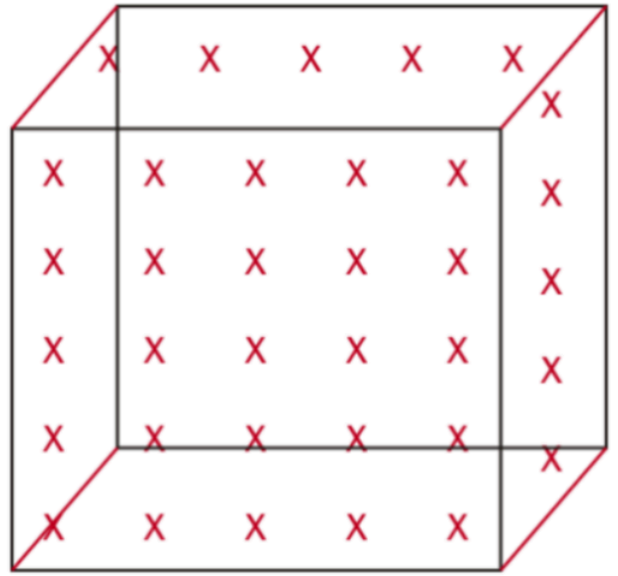


https://towardsdatascience.com/curse-of-dimensionality-2092410f3d27

ПРОКЛЯТЬЕ РАЗМЕРНОСТИ

Одно измерение - 5 точек

Два измерение - 25 точек

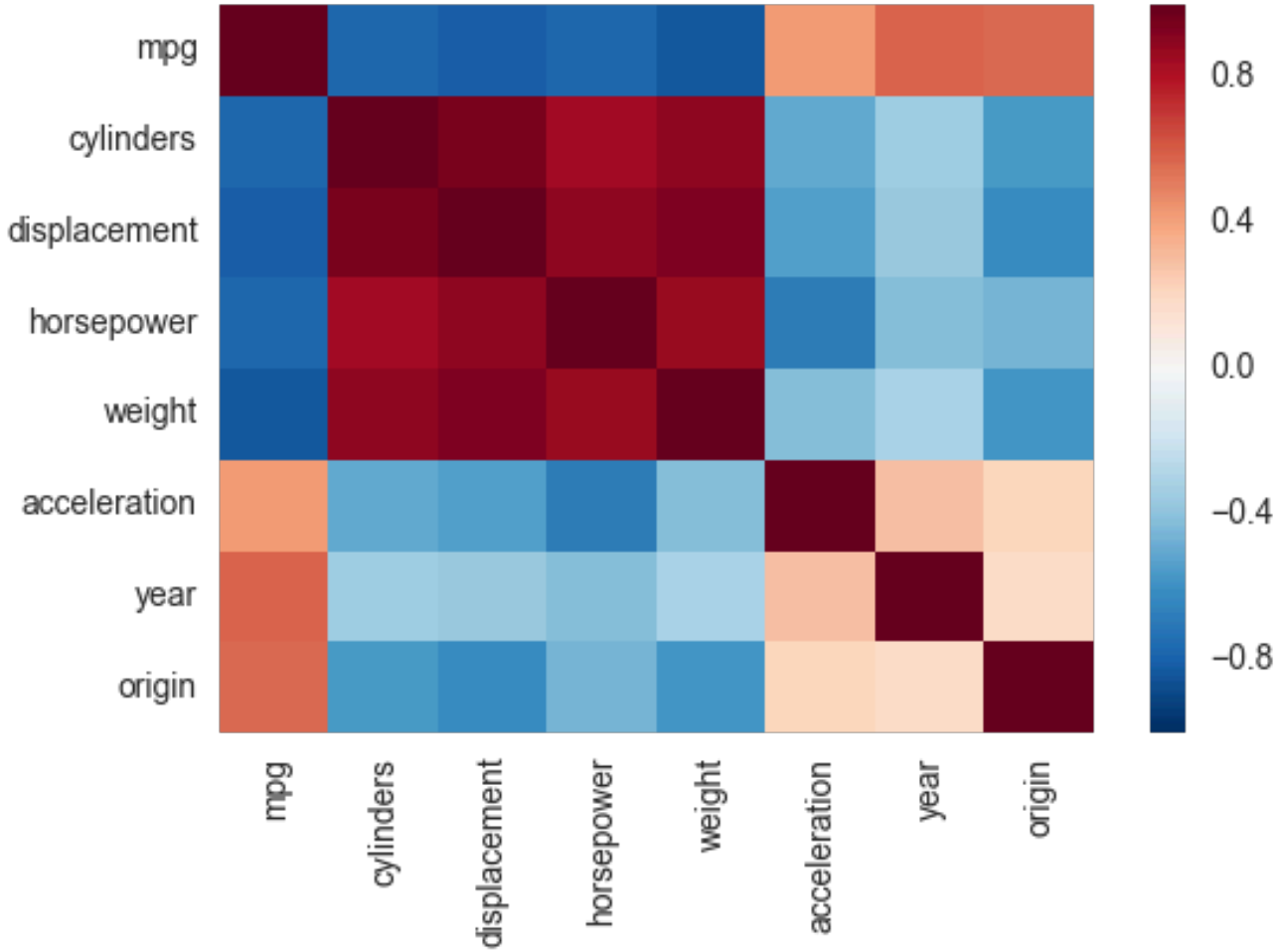


Три измерения - 125 точек

https://towardsdatascience.com/curse-of-dimensionality-2092410f3d27

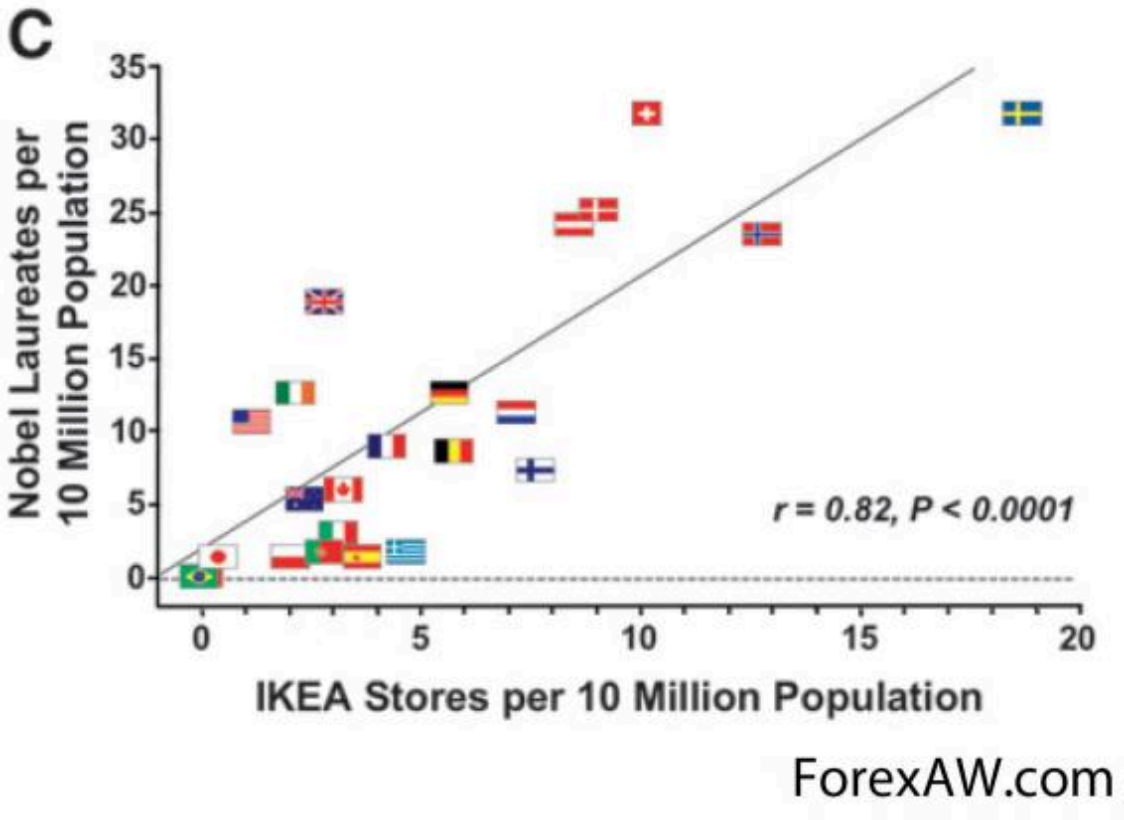
Статистика в отборе признаков

Корреляция


mpg
cylinders
displacement
horsepower
weight
acceleration
year
origin
0.8
0.4
0.0
−0.4
−0.8

Корреляция

Статистическая зависимость двух и более величин

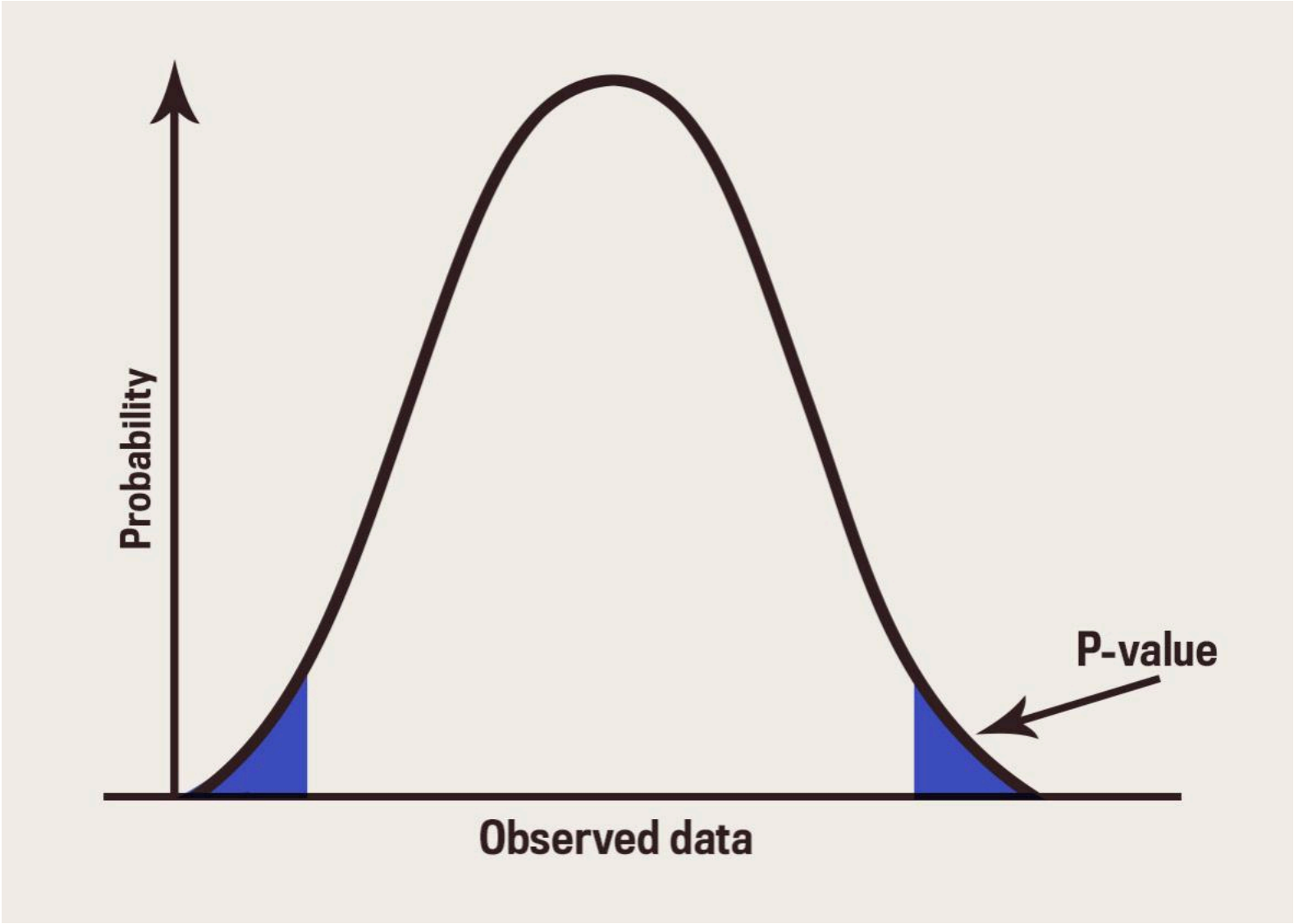


T-статистика

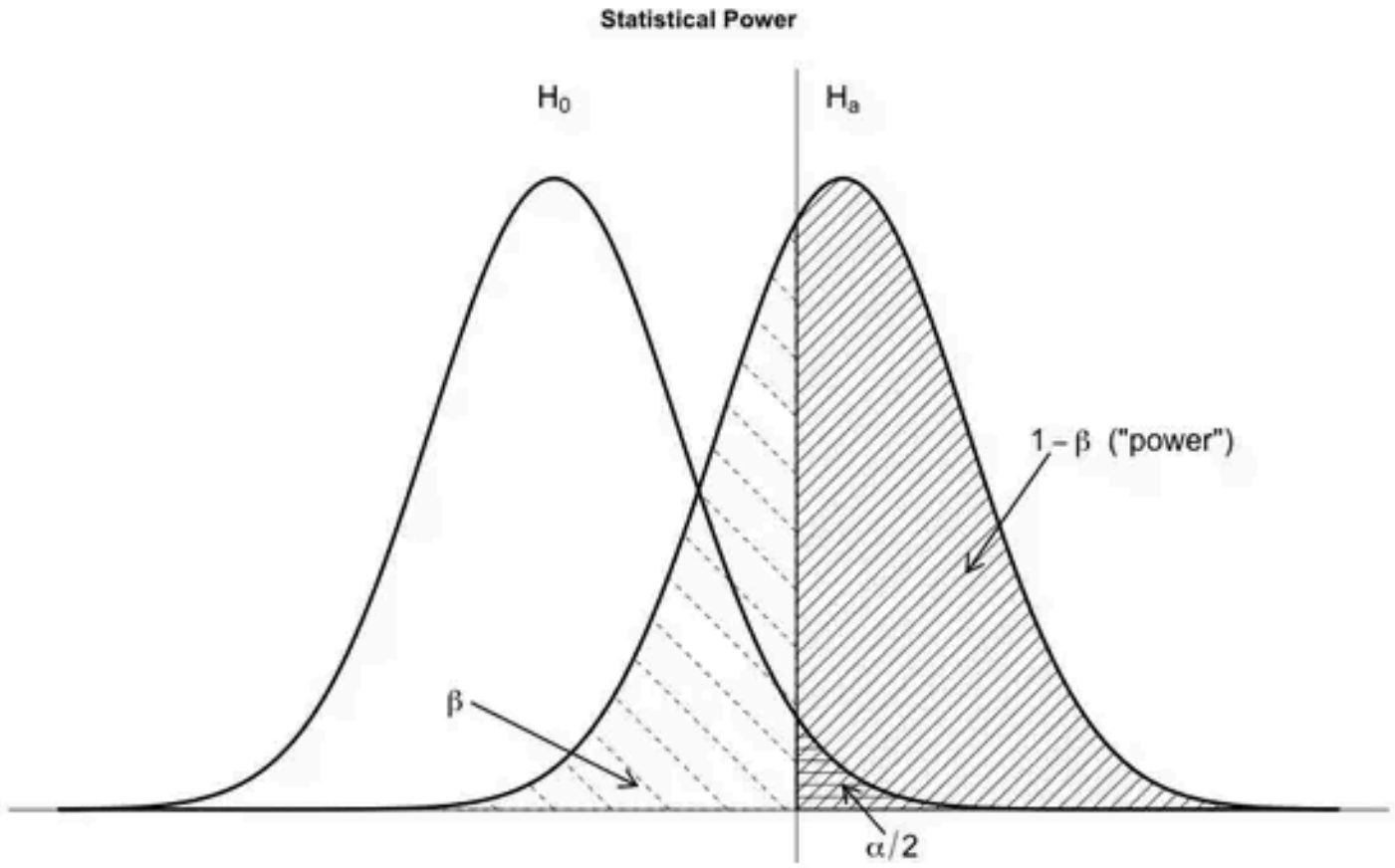
$$t = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$$

- Если между x_i и y нет зависимости, то t соответсвует t-распределению с n−2 степенями свободы
- p-value - вероятность того, что при известном распределении наблюдаемое значение ≥ |t| (**при условии, что $\beta_i = 0$**)
- Если p-value достаточно маленький (< 1%), то мы можем отклонить H_0

P-value



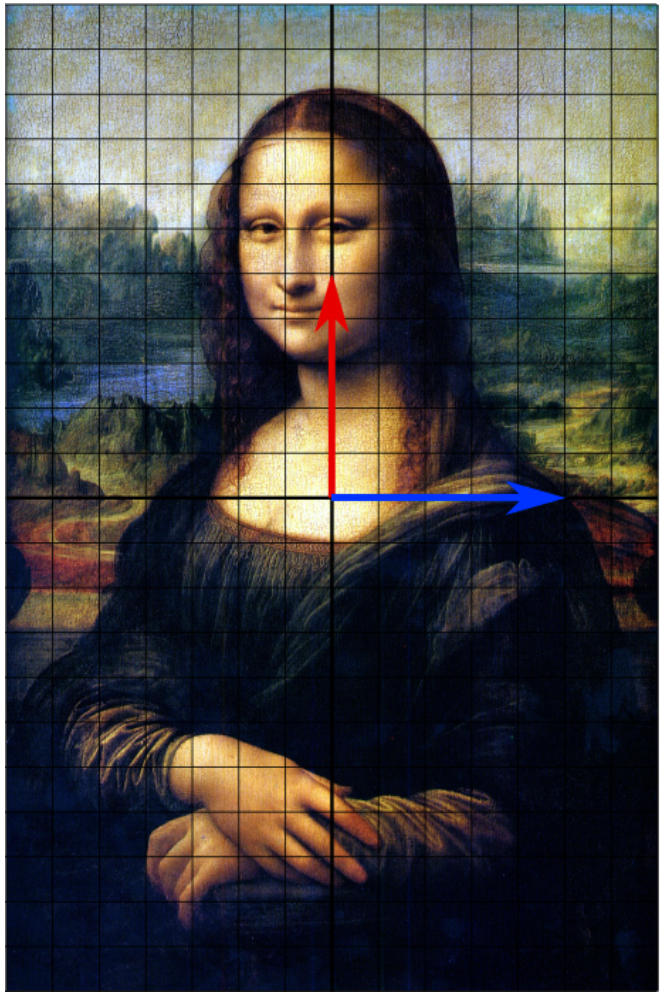
P-value

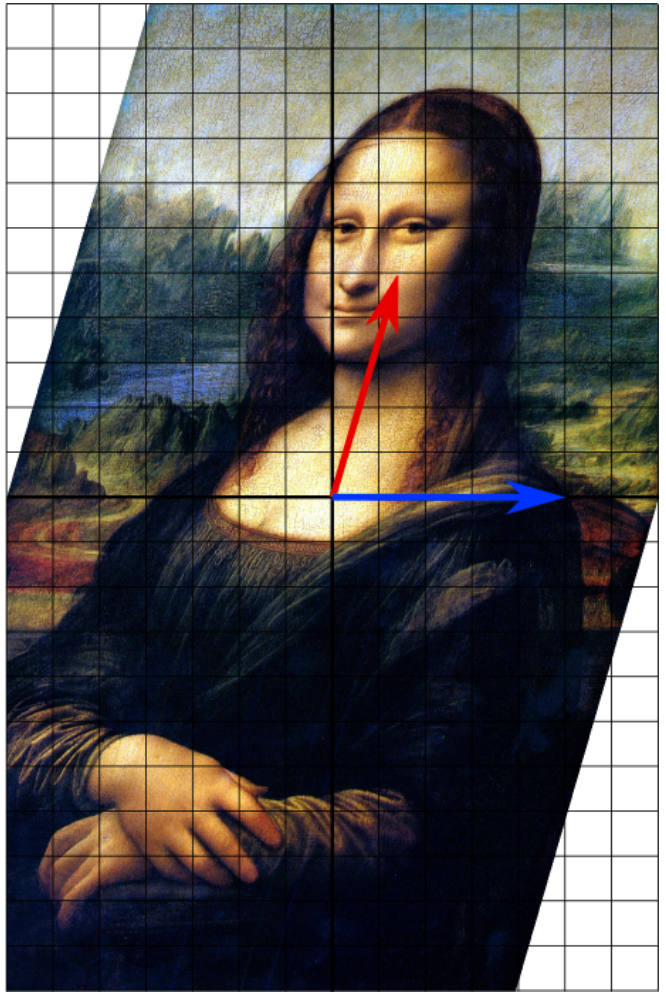


Декомпозиция данных

Собственный вектор

$$M\vec{x} = \lambda\vec{x}$$

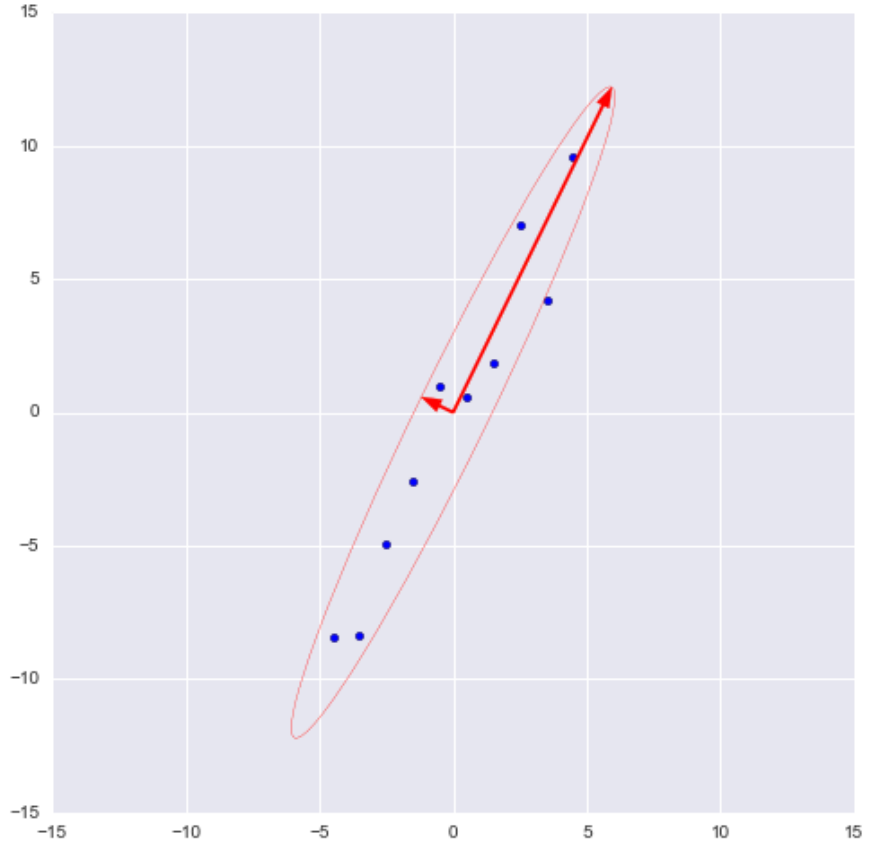




http://mathprofi.ru/sobstvennye_znachenija_i_sobstvennye_vektory.html

PCA

Зачем он нужен? Он уменьшает размерность ☺



PCA

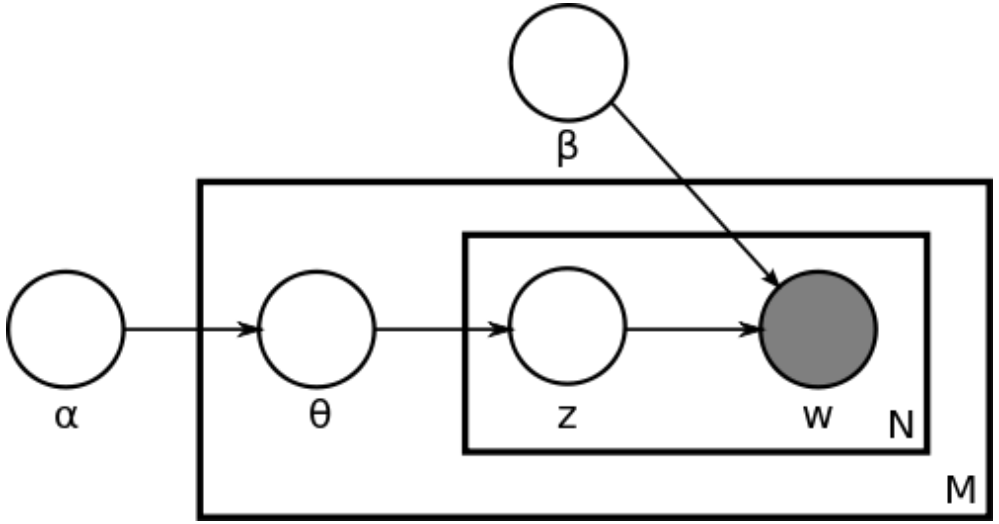
$$Cov(X_i, X_j) = E\Big[\big(X_i - E(X_i)\big) \cdot \big(X_j - E(X_j)\big)\Big] = E(X_i X_j) - E(X_i) \cdot E(X_j)$$

$$Var(X^*) = \Sigma^* = E(X^* \cdot X^{*T}) = E\Big((\vec{v}^T X) \cdot (\vec{v}^T X)^T\Big) =$$
$$= E(\vec{v}^T X \cdot X^T \vec{v}) = \vec{v}^T E(X \cdot X^T)\vec{v} = \vec{v}^T \Sigma \vec{v}$$

https://habr.com/ru/post/304214/

LDA

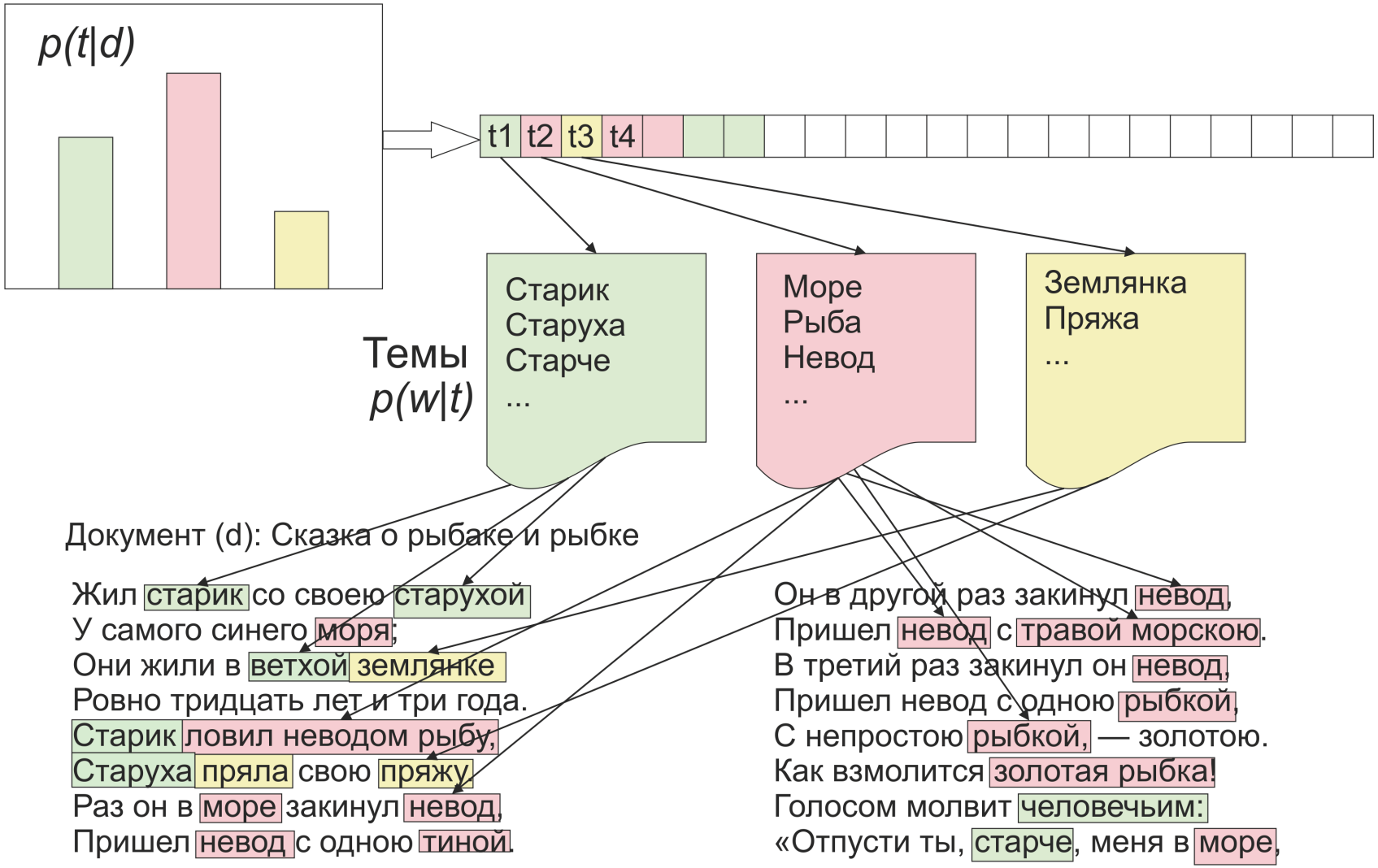
это иерархическая байесовская модель, состоящая из двух уровней:
на первом уровне – смесь, компоненты которой соответствуют «темам»;
на втором уровне – мультиномиальная переменная с априорным распределением Дирихле, которое задаёт «распределение тем» в документе.



$$p(\theta,,,N \mid \alpha,\beta) = p(N \mid \xi)p(\theta \mid \alpha)\prod_{n=1}^{N} p(z_n \mid \theta)p(w_n \mid z_n,\beta).$$

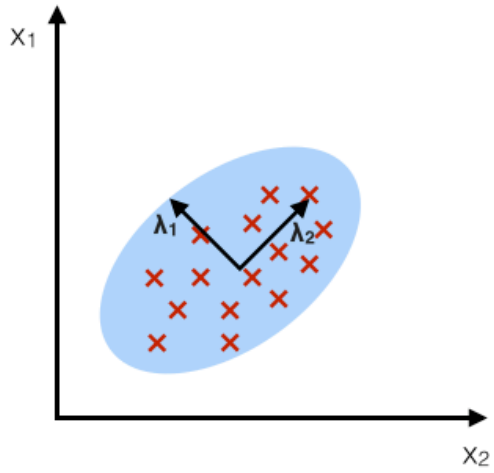
https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

LDA

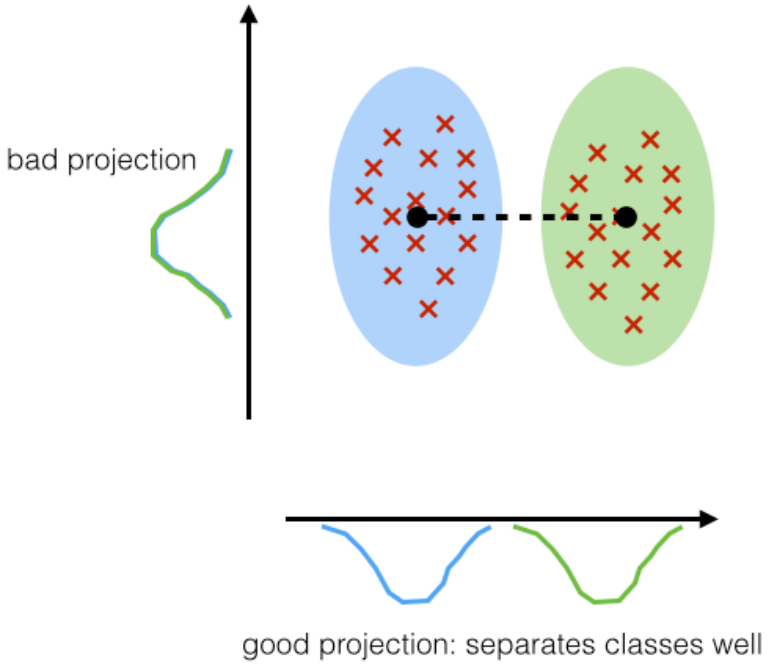

p(t|d)
t1 t2 t3 t4
Темы
p(w|t)
Старик
Старуха
Старче
...
Море
Рыба
Невод
...
Землянка
Пряжа
...
Документ (d): Сказка о рыбаке и рыбке
Жил старик со своею старухой
У самого синего моря;
Они жили в ветхой землянке
Ровно тридцать лет и три года.
Старик ловил неводом рыбу,
Старуха пряла свою пряжу.
Раз он в море закинул невод,
Пришел невод с одною тиной.
Он в другой раз закинул невод,
Пришел невод с травой морскою.
В третий раз закинул он невод,
Пришел невод с одною рыбкой,
С непростою рыбкой, — золотою.
Как взмолится золотая рыбка!
Голосом молвит человечьим:
«Отпусти ты, старче, меня в море,

Сравнение

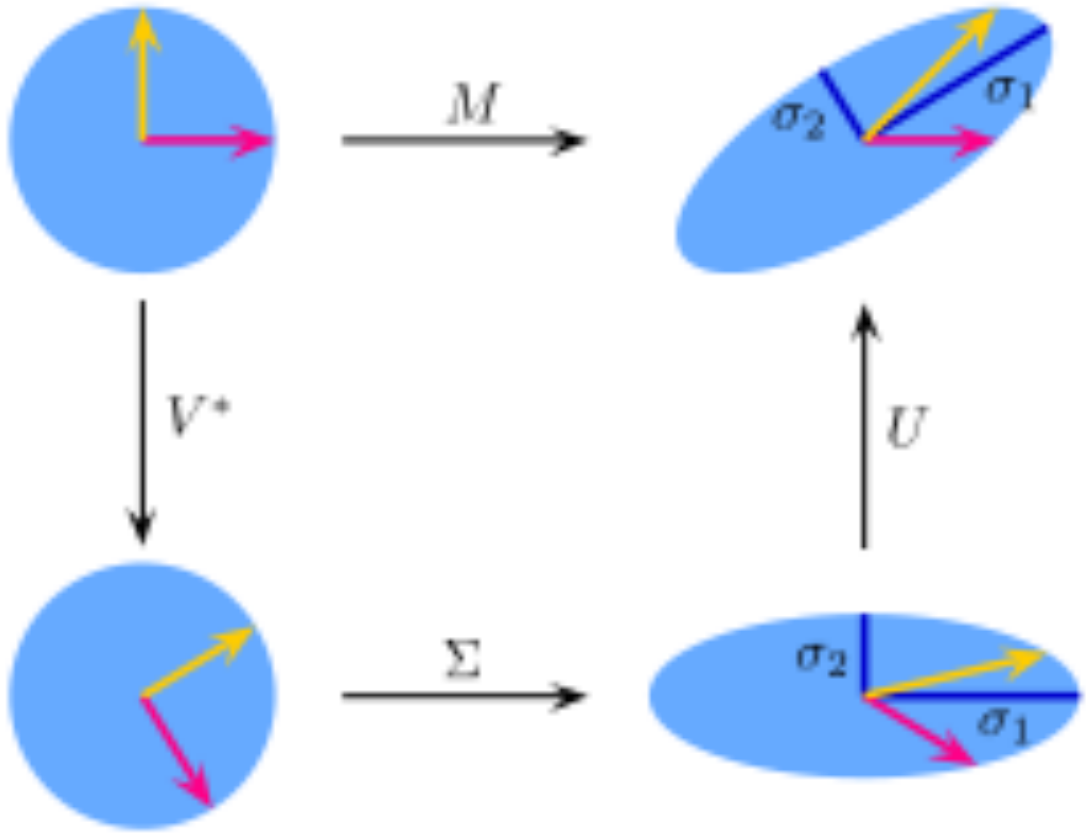
PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation

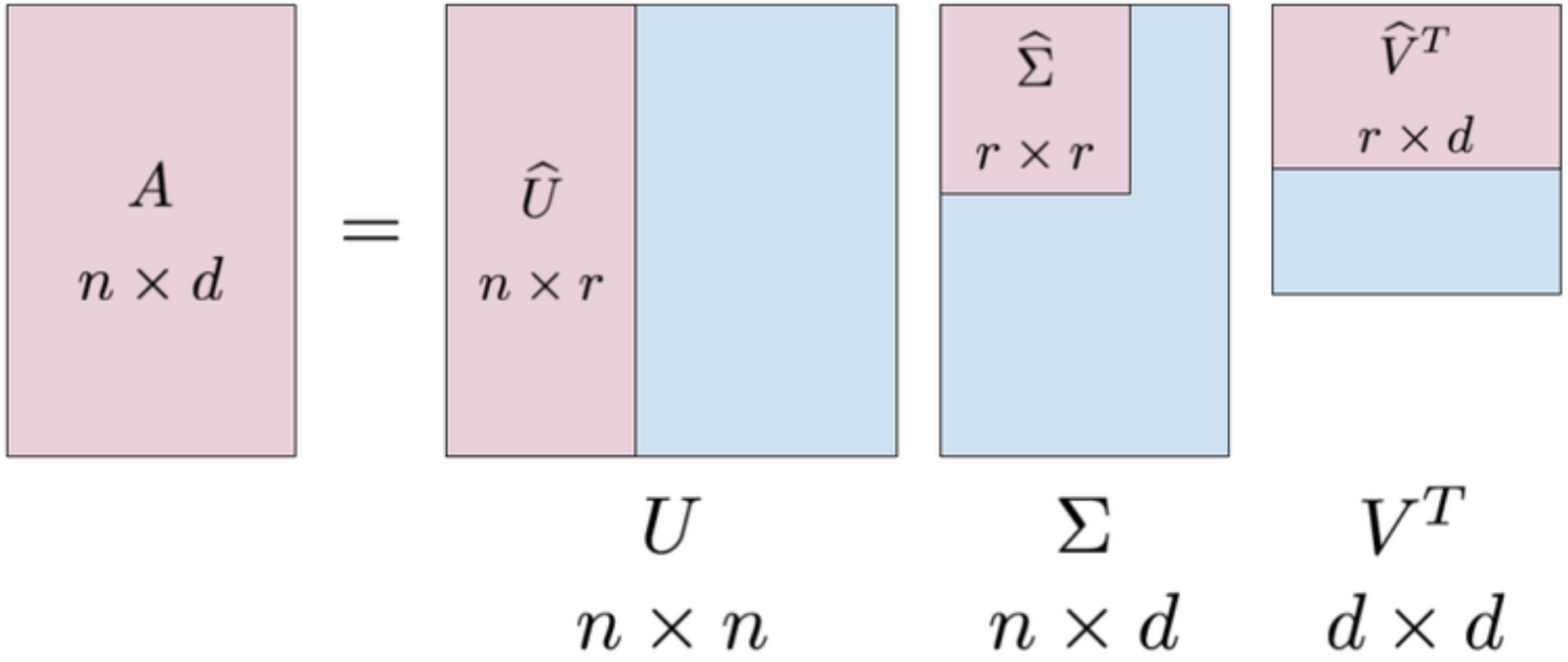


SVD



$$M = U \cdot \Sigma \cdot V^*$$

SVD


A
$n \times d$
$=$
$\widehat{U}$
$n \times r$
$\widehat{\Sigma}$
$r \times r$
$\widehat{V}^T$
$r \times d$
U
$n \times n$
Σ
$n \times d$
V^T
$d \times d$

SVD

	Трактористы	Свинарка и пастух	Once Upon a Tractor	Tractor, Love & Rock'n Roll	Babe
Вася	?	3	4	5	2
Пётр	3	5	2	2	5
Валерик	5	3		4	3
Жанночка	5	5	5		4
Петрович	2	3		2	2

mu: 2.5455953363826 1

User base: 0.7271 0.1626 0.7139 1.9097 -0.9677

Item base: 0.8450 0.6593 0.2731 0.7328 0.0354

User features:

user 0: -0.5087 -0.8326

user 1: 1.0220 1.2826

user 2: -0.9509 0.2792

user 3: 0.1031 -0.4814

user 4: 0.6095 0.0557

Item features:

item 0: -0.8368 0.2511

item 1: 1.1101 0.4120

item 2: -0.4159 -0.4073

item 3: -0.3130 -0.9115

item 4: 0.6408 1.2205

ПРАКТИЧЕСКАЯ ЧАСТЬ

ВОПРОСЫ