РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ НА ПРАКТИКЕ

СЕЛЕЗНЕВ АРТЕМ

МЫ ОСТАНОВИЛИСЬ….

СОВМЕСТНАЯ ФИЛЬТРАЦИЯ

UID	43	11	7
1	3	-	1
2	-	4	-
3	-	-	1

Вычисляем сходство между пользователем (UID1) и остальным

UID	1
1	1
2	0
3	0.5

Сортируем по результату

UID	1
1	1
3	0.5
2	0

Топ N (количество)
Кластеризация
Установить пороговое значение (качество)

R = среднее по Топ N

Выбираем диапазон ближайших,
Вычисляем прогноз оценки

ДЕНЬ 2

Моя прошла команда рассказывала мне про матрасные факторизации для рекомендаций. Что это за матрасы?

Матричная факторизация….

ФАКТОРИЗАЦИЯ?

$$100 = 2 \times 2 \times 5 \times 5$$

Факторизация – декомпозиция (разделение) объекта

ФАКТОРИЗАЦИЯ?

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	4	-	-	-
5	-	-	1	-	4

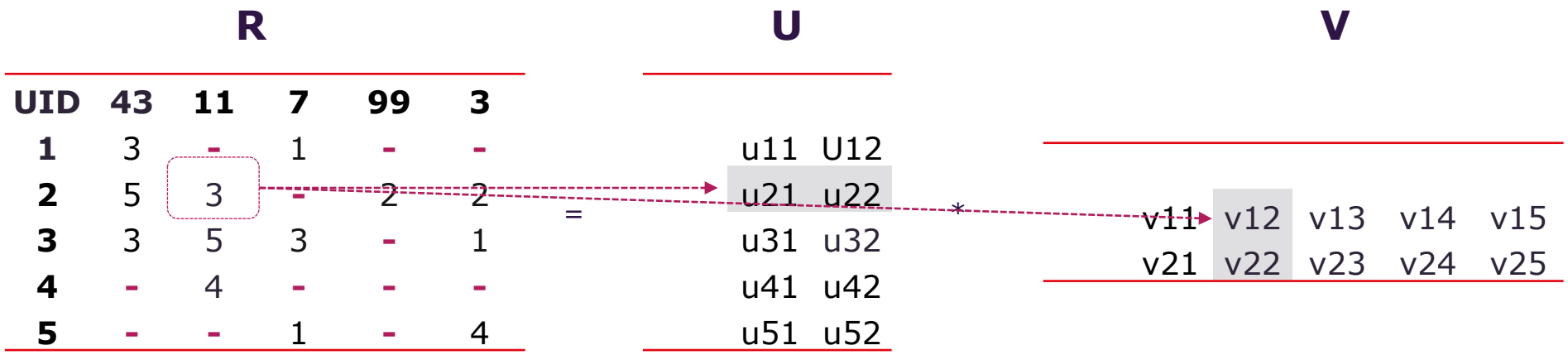
=

u11	U12
u21	u22
u31	u32
u41	u42
u51	u52

*

v11	v12	v13	v14	v15
v21	v22	v23	v24	v25

ФАКТОРИЗАЦИЯ?



ФАКТОРИЗАЦИЯ?

R – матрица рейтингов

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	4	-	-	-
5	-	-	1	-	4

=

U – признаки пользователя

u11	U12
u21	u22
u31	u32
u41	u42
u51	u52

*

Sigm - веса

Диагональная матрица весов

*

V признаки элементов

v11	v12	v13	v14	v15
v21	v22	v23	v24	v25

ФАКТОРИЗАЦИЯ?

R – матрица рейтингов

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	4	-	-	-
5	-	-	1	-	4

=

U – признаки пользователя

u11	U12
u21	u22
u31	u32
u41	u42
u51	u52

*

Sigm - веса

	43	11	7	99	3
1	1	0	0	0	0
2	0	3	0	0	0
3	0	0	5	0	0
4	0	0	0	7	0
5	0	0	0	0	9

*

V признаки элементов

v11	v12	v13	v14	v15
v21	v22	v23	v24	v25

ФАКТОРИЗАЦИЯ?

R – матрица рейтингов

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	4	-	-	-
5	-	-	1	-	4

=

Столбцы

u11	U12
u21	u22
u31	u32
u41	u42
u51	u52

*

Sigm - веса

	43	11	7	99	3
1	**1**	0	0	0	0
2	0	**3**	0	0	0
3	0	0	**5**	0	0
4	0	0	0	**7**	0
5	0	0	0	0	**9**

*

Строки

v11	v12	v13	v14	v15
v21	v22	v23	v24	v25

Веса – особые значения, которые показывают вес (в информации) для всего набора данных

Так ваши матрицы можно сокращать?
Т.е. считать будете быстрее теперь!

Можно сократить матрицу до размерности 2,
но мы будем терять информацию….

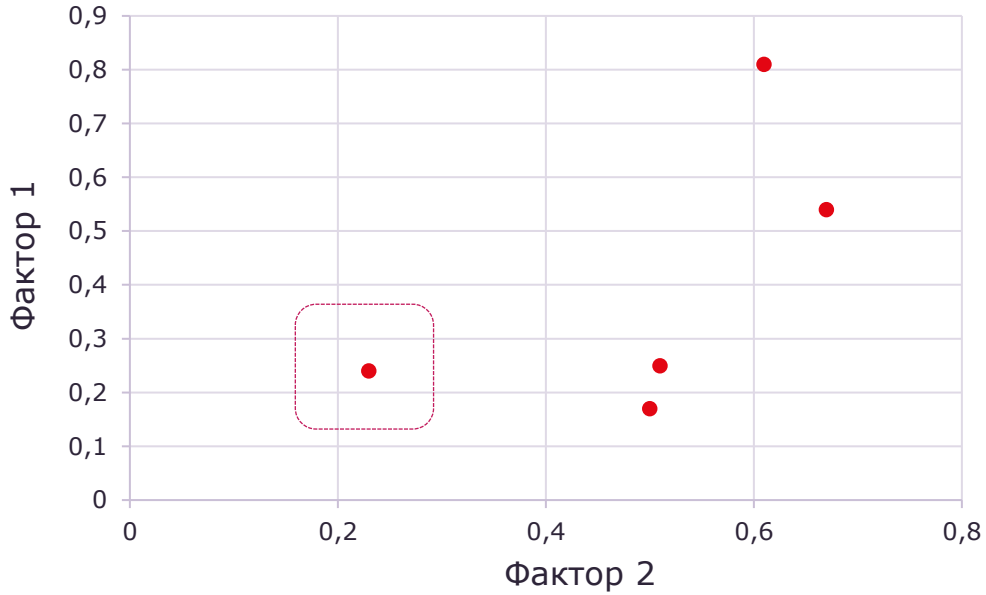
Правило – 90% информации необходимо сохранить

ФАКТОРИЗАЦИЯ

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	4	-	-	-
5	-	-	1	-	4

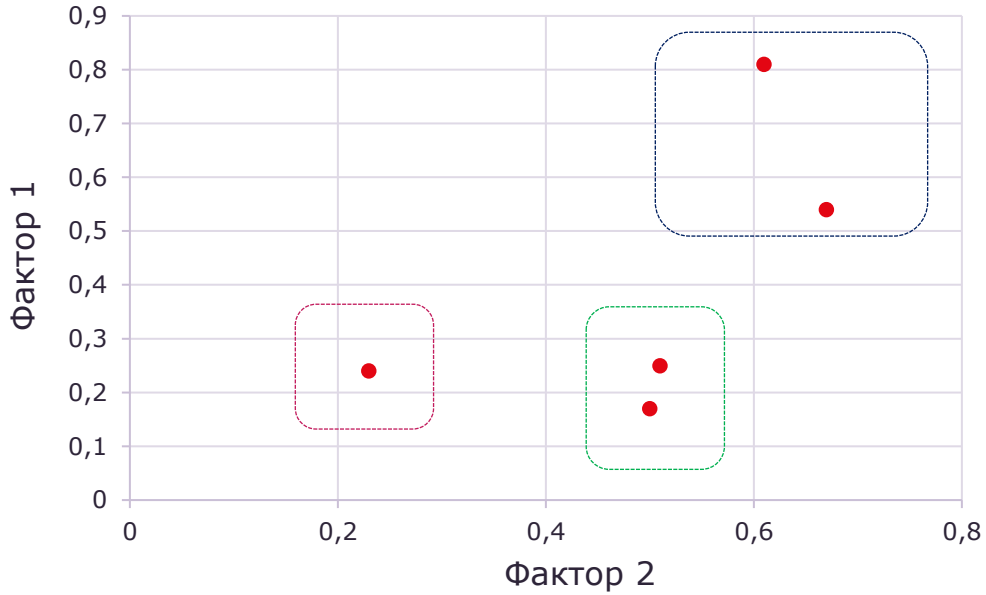
ФАКТОРИЗАЦИЯ

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	1	-	5	-
5	-	-	1	-	4



ФАКТОРИЗАЦИЯ

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	1	-	5	-
5	-	-	1	-	4



А была проблем, что пользователи без рейтингов не попадали в обучение. Что делаем теперь?

SVD позволяет добавлять новые данные…

ФАКТОРИЗАЦИЯ. НОВЫЙ ПОЛЬЗОВАТЕЛЬ

R – матрица рейтингов

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	4	-	-	-
5	-	-	1	-	4
6	-	-	1	-	-

$= r_k V^t \Sigma^{-1}$

ФАКТОРИЗАЦИЯ. НОВЫЙ ПОЛЬЗОВАТЕЛЬ

R – матрица рейтингов

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	4	-	-	-
5	-	-	1	-	4
6	-	-	1	-	-

=

R6

*

Sigm - веса

	43	11	7	99	3
1	**1**	0	0	0	0
2	0	**3**	0	0	0
3	0	0	**5**	0	0
4	0	0	0	**7**	0
5	0	0	0	0	**9**

*

V признаки элементов

v11	v12	v13	v14	v15
v21	v22	v23	v24	v25

А как я объясню причины выбора алгоритма?

SVD не интуитивно понятен, это его большой минут.

Точного ответа дать нельзя.

SVD – не лучший вариант факторизации

У меня модель показывает или хорошие или очень плохие результаты!

Кажется с рейтингам что-то не то!

«Сырые рейтинги» подходят для первых экспериментов,
попробуйте использовать:
- базисы, на основе глобального среднего…

КОРРЕКТИРОВКА РЕЙТИНГОВ В МАТРИЦЕ

Оценки не у всех одинаковые

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	1	-	5	-
5	-	-	1	-	4

КОРРЕКТИРОВКА РЕЙТИНГОВ В МАТРИЦЕ

UID	43	11	7	99	3	
1	3	-	1	-	-	2
2	5	3	-	2	2	3
3	3	5	3	-	1	3,2
4	-	1	-	5	-	3
5	-	-	1	-	4	2,5
	3,6	3	1,6	3,5	2,3	

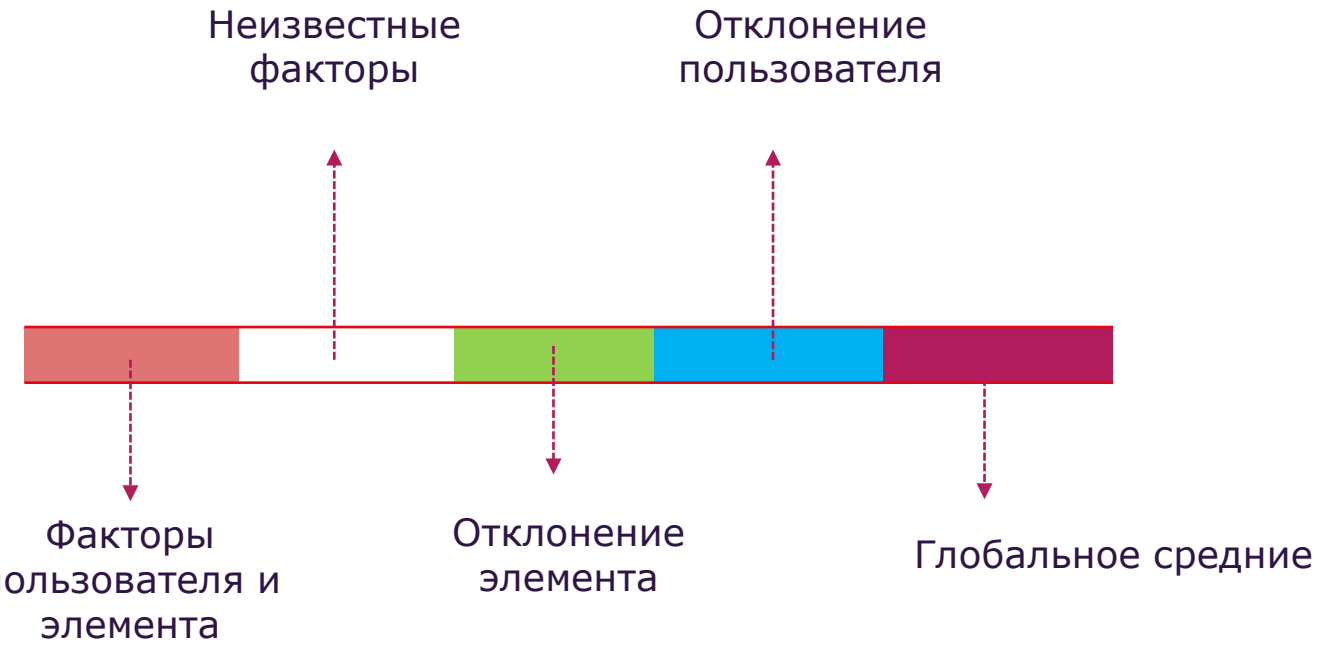
Оценки в матрице необходимо скорректировать относительно глобального среднего

Отклонение пользователя= среднее по элементу – оценка пользователя

Оценка = среднее + отклонение пользователя + отклонение элемента

КОРРЕКТИРОВКА РЕЙТИНГОВ В МАТРИЦЕ

Почему так?



КОРРЕКТИРОВКА РЕЙТИНГОВ В МАТРИЦЕ + ВРЕМЯ

UID	43	11	7	99	3	
1	3	-	1	-	-	2
2	5	3	-	2	2	3
3	3	5	3	-	1	3,2
4	-	1	-	5	-	3
5	-	-	1	-	4	2,5
	3,6	3	1,6	3,5	2,3	

Пользователь может стать из позитивного человека в недовольного

Может у него период такой=)

Добавим время в оценку

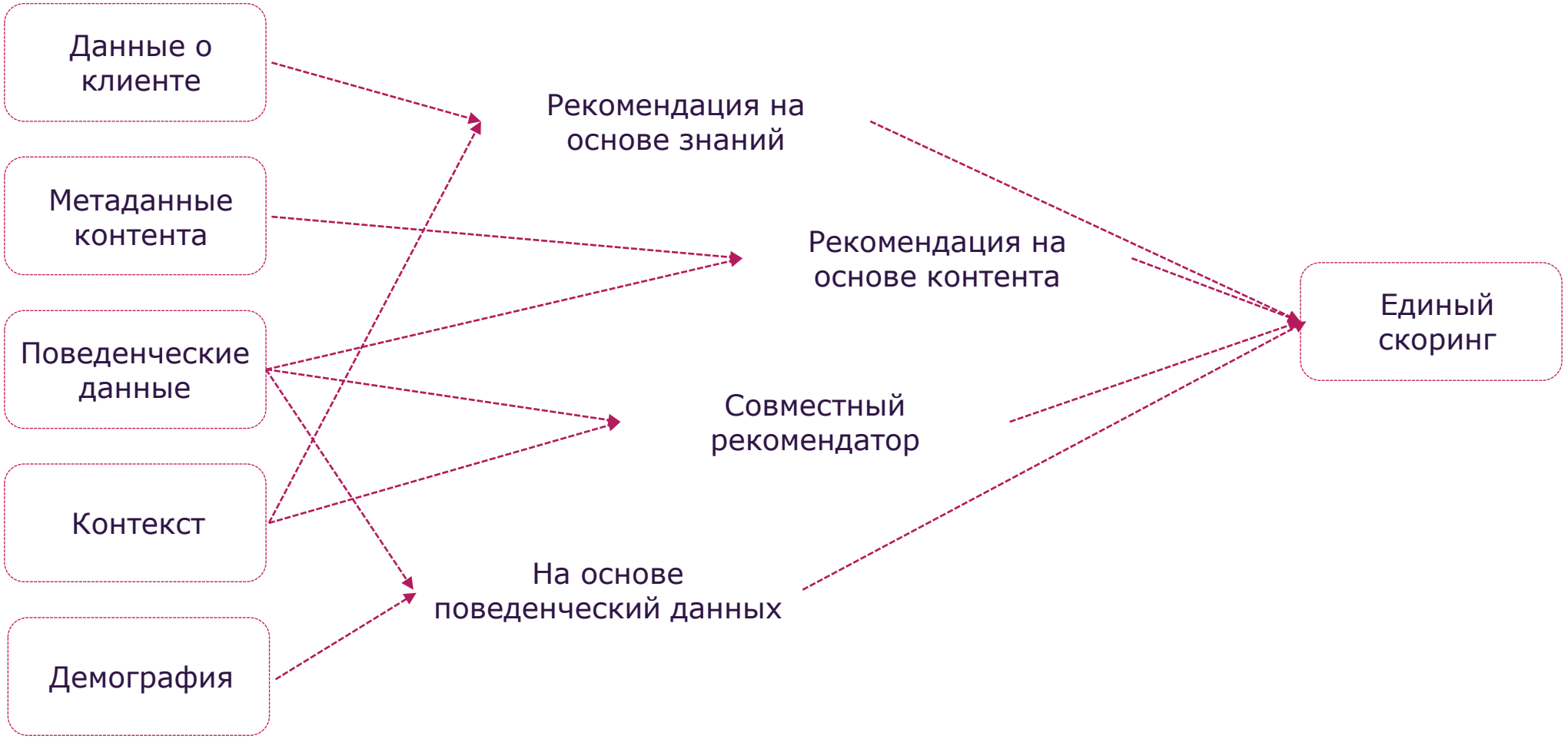
Оценка (t) = среднее + отклонение пользователя (t) + отклонение элемента (t)

t – коэф. времени от 0,01 (позднее) до 1 (недавнее)

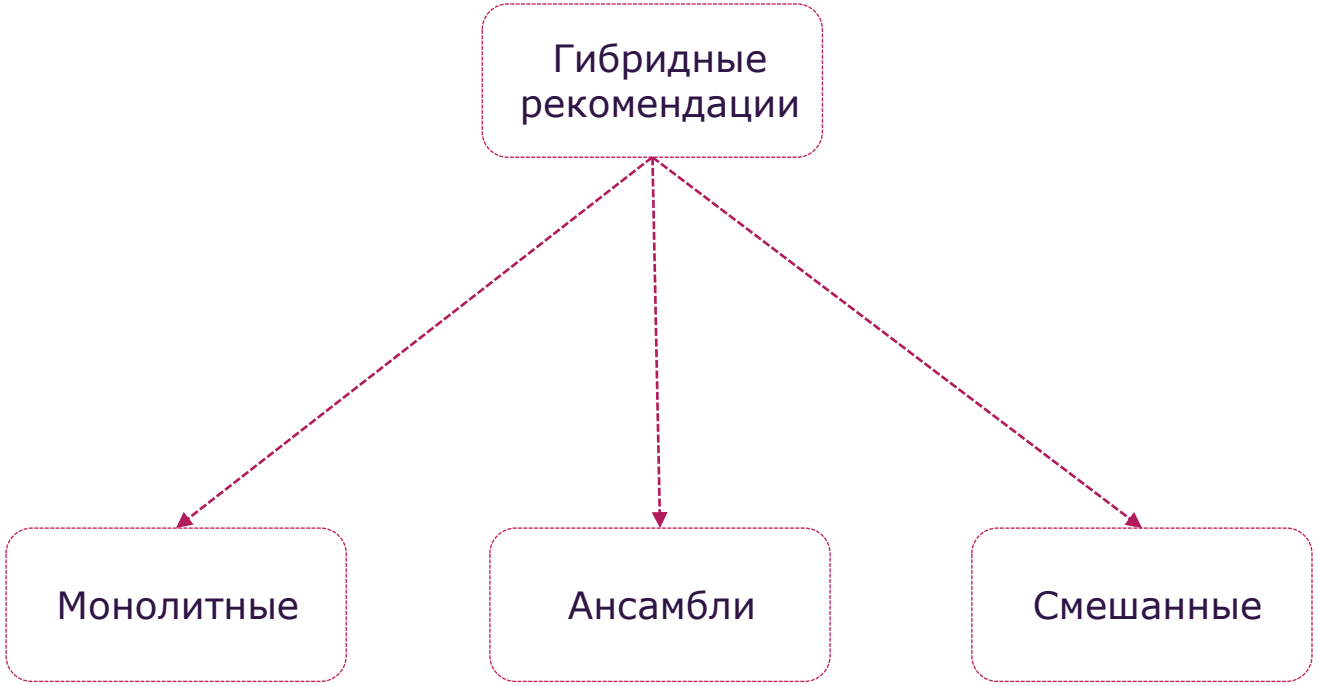
Рассмотрим на примере

ХОТИТЕ ИСПОЛЬЗОВАТЬ БОЛЬШЕ ИНФОРМАЦИИ?

ГИБРИДНЫЕ РЕКОМЕНДАЦИИ



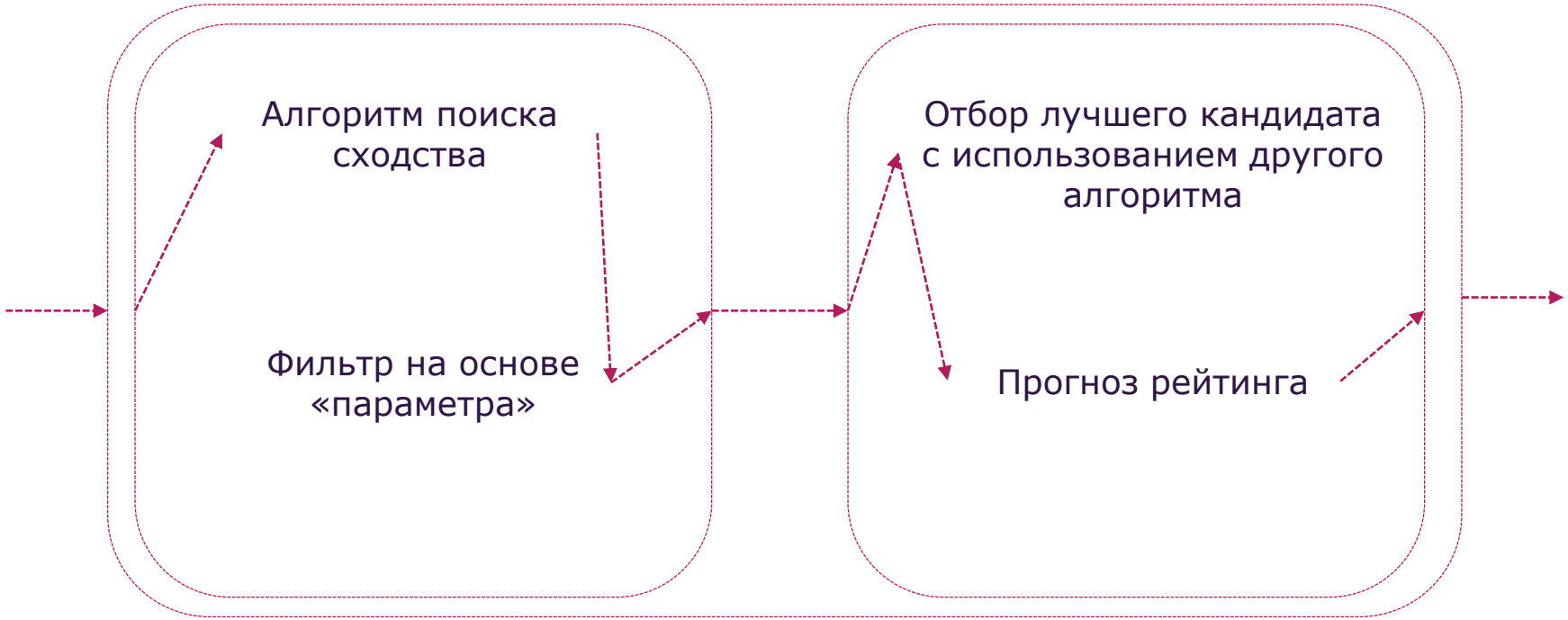
ГИБРИДНЫЕ РЕКОМЕНДАЦИИ



ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

Монолитные

Рекомендаторы, которые переиспользуют разные компоненты

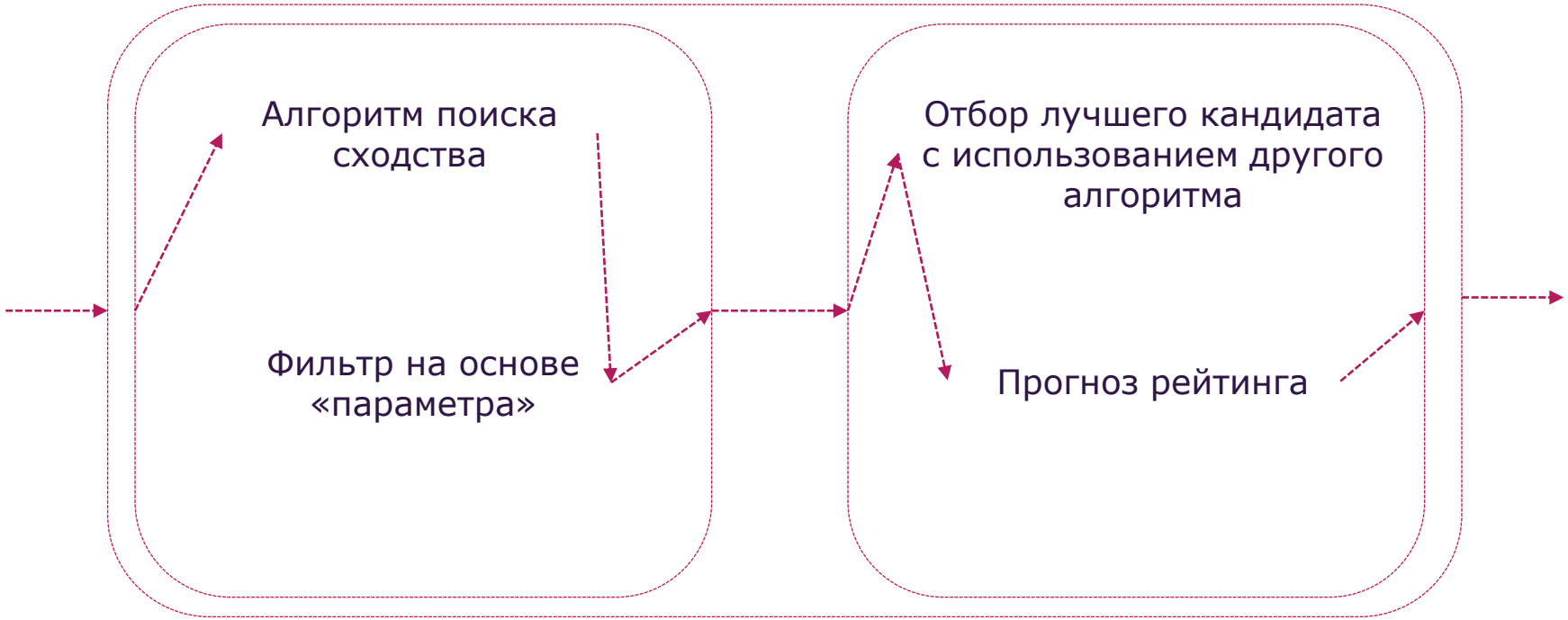


ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

Монолитные

Рекомендаторы, которые переиспользуют разные компоненты

Взять всё самое лучшее из доступного



ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

Монолитные

Рекомендаторы, которые переиспользуют разные компоненты

Взять всё самое лучшее из доступного

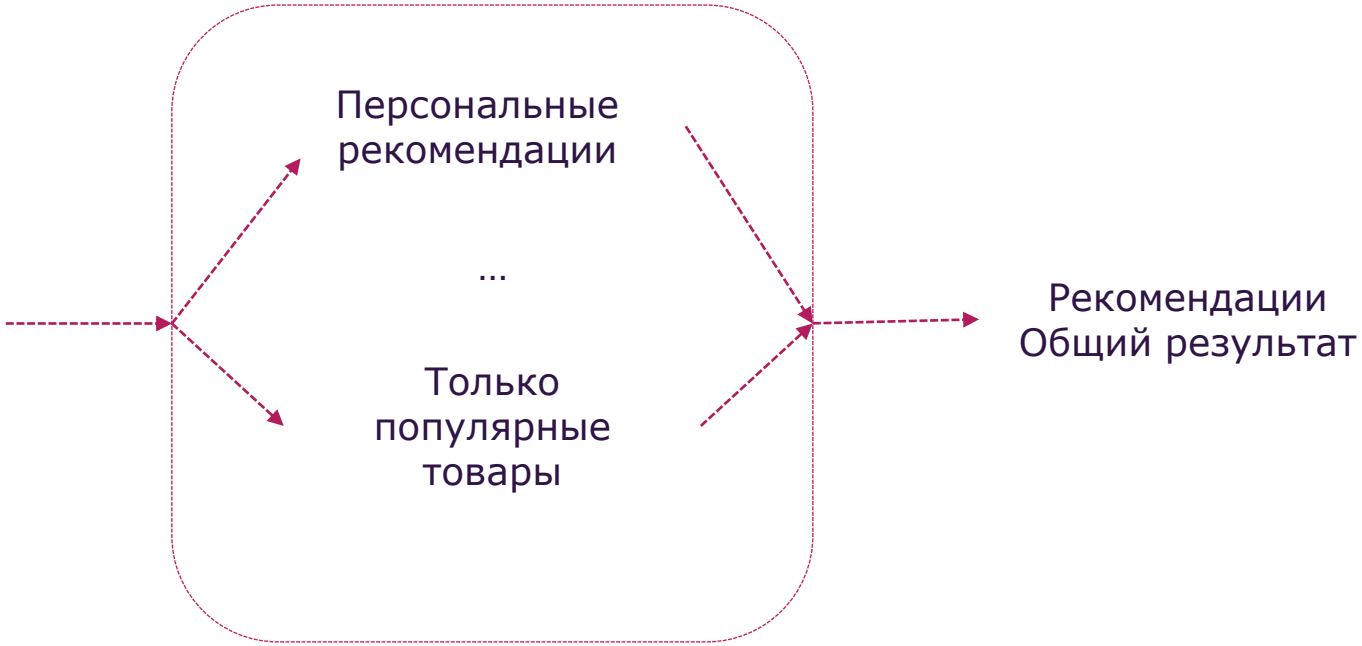
UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	1	-	5	-
5	-	-	1	-	4
New	**5**	**5**	**5**	**5**	**5**

Добавление пользователя с оценками – укрепит связь

ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

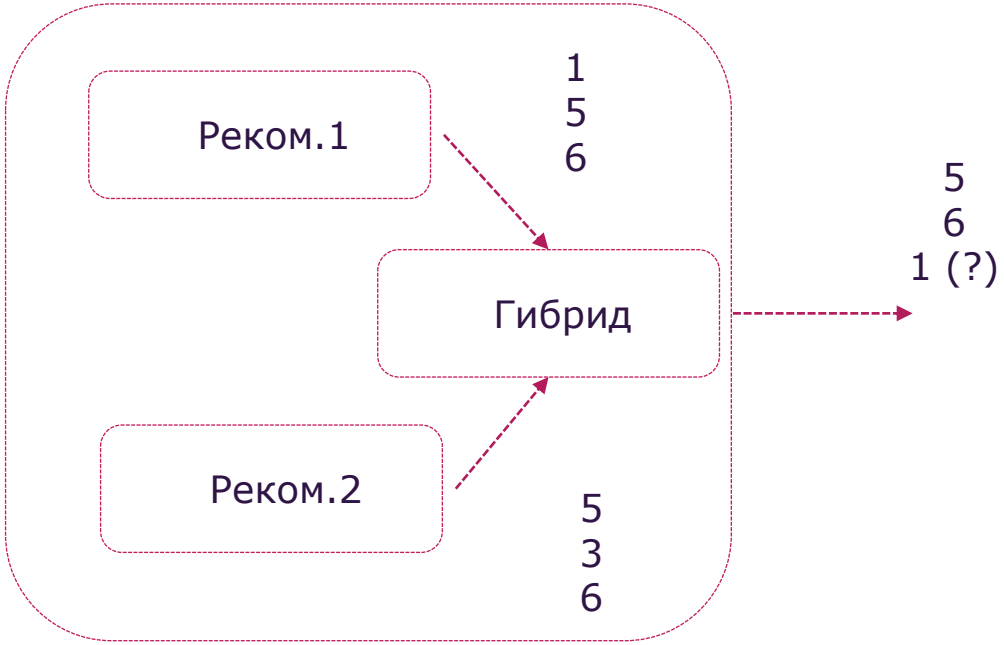
Смешанные

Смешать алгоритмы и получить общий скоринг



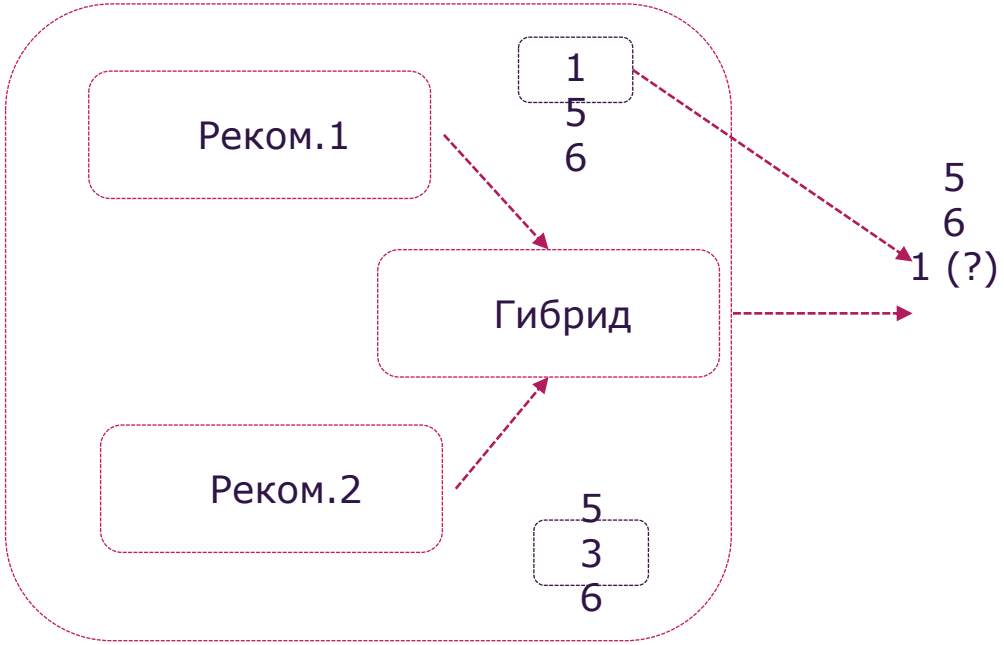
ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

Ансамбли



ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

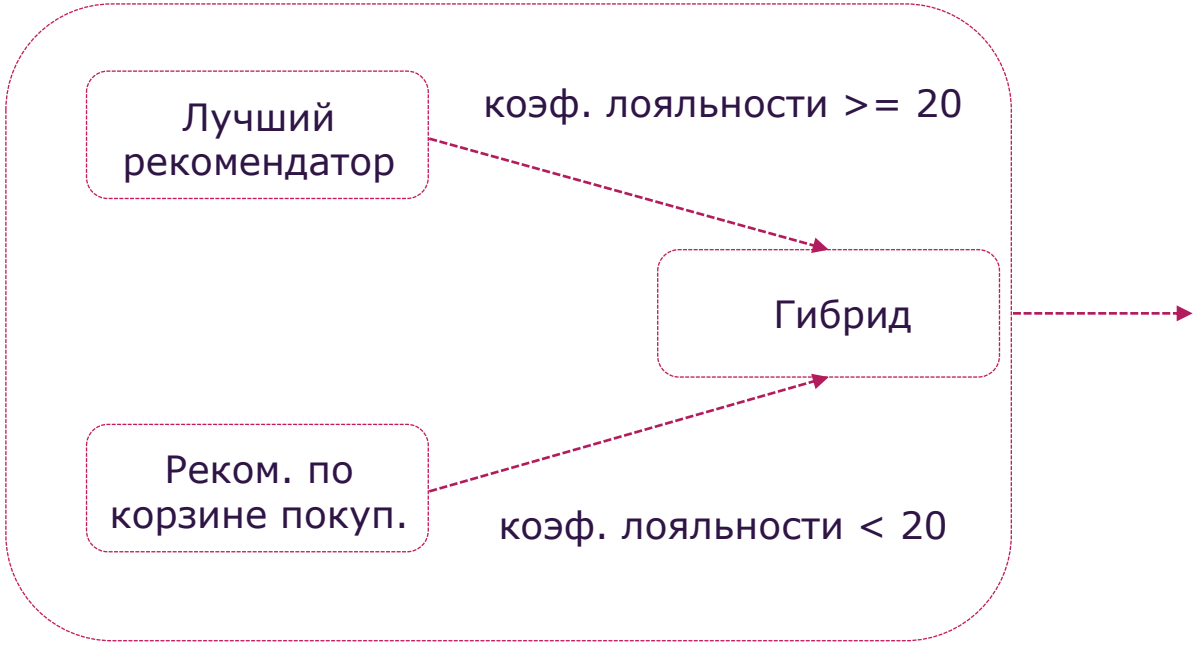
Ансамбли



ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

Ансамбли

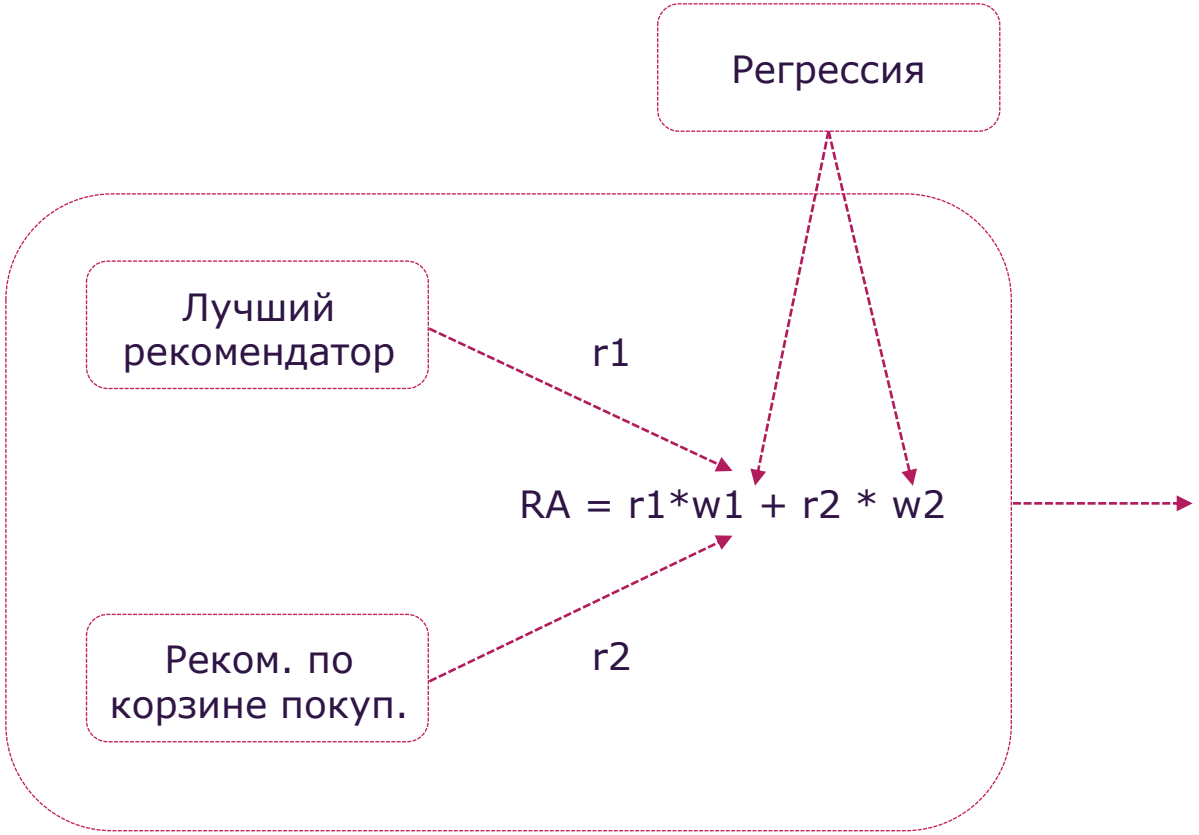
Переключаемые



ГИБРИДНЫЕ РЕКОМЕНДАЦИИ

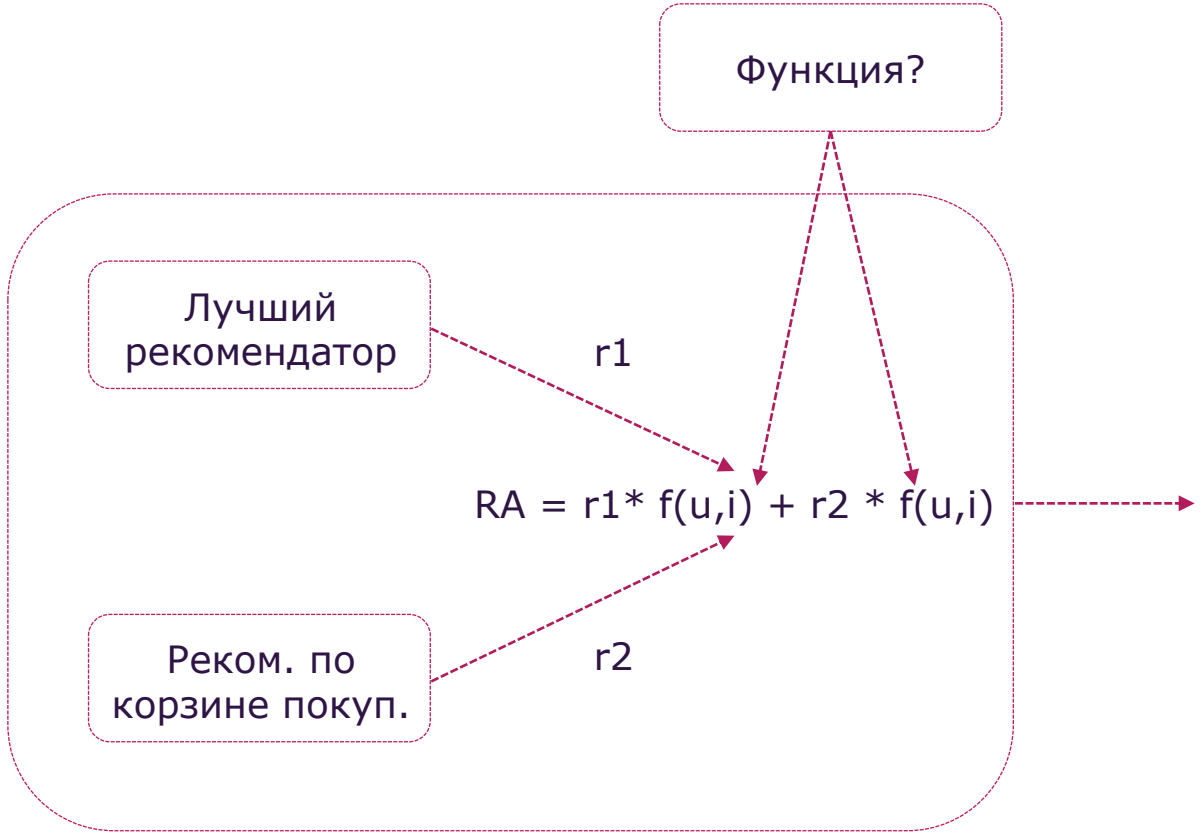
Ансамбли

Взвешенные



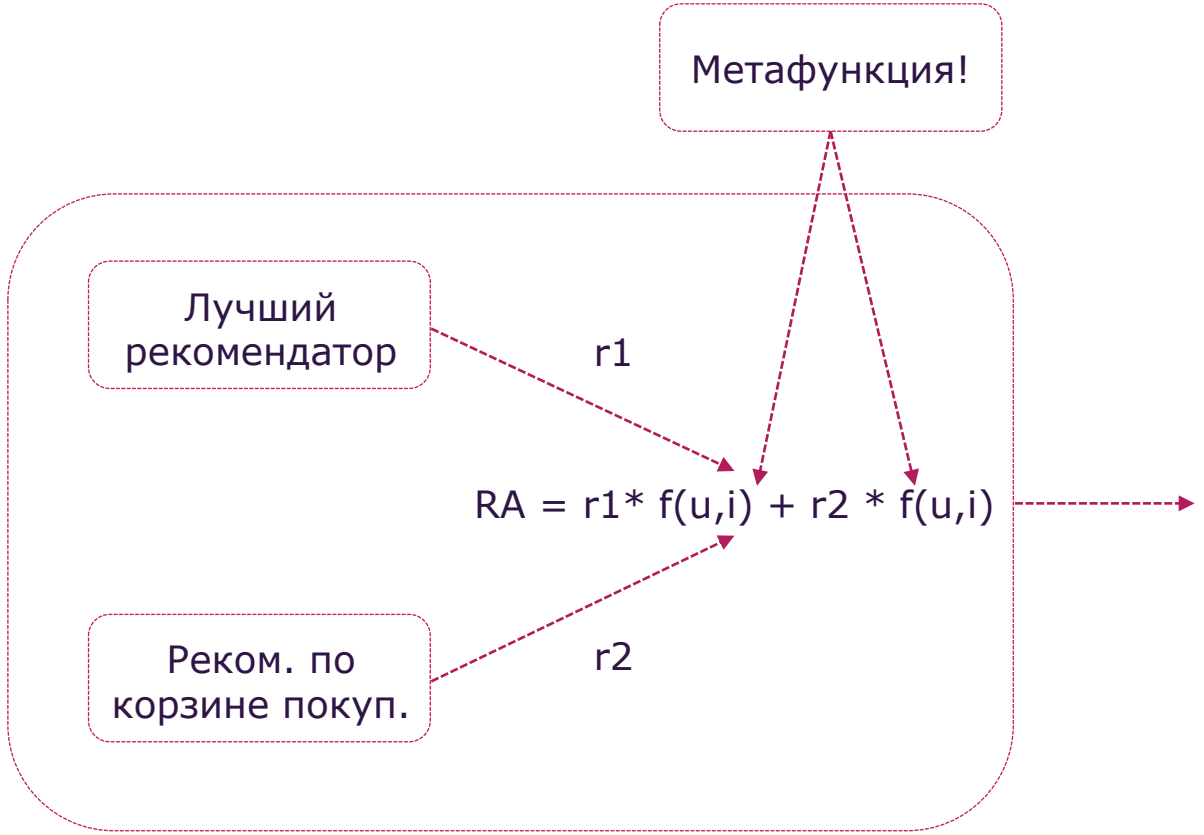
ГИБРИДНЫЕ РЕКОМЕНДАЦИИ. FWLS

Ансамбли



ГИБРИДНЫЕ РЕКОМЕНДАЦИИ. FWLS

Ансамбли



ГИБРИДНЫЕ РЕКОМЕНДАЦИИ. FWLS

Метафункции
из знаменитого
соревнования

Функция голосования
моделей
(регрессия)

Бинарная функция на
основе целевой:
- оценено более N/день

SVD на N факторах

Разные способы
расчетов средней
оценки

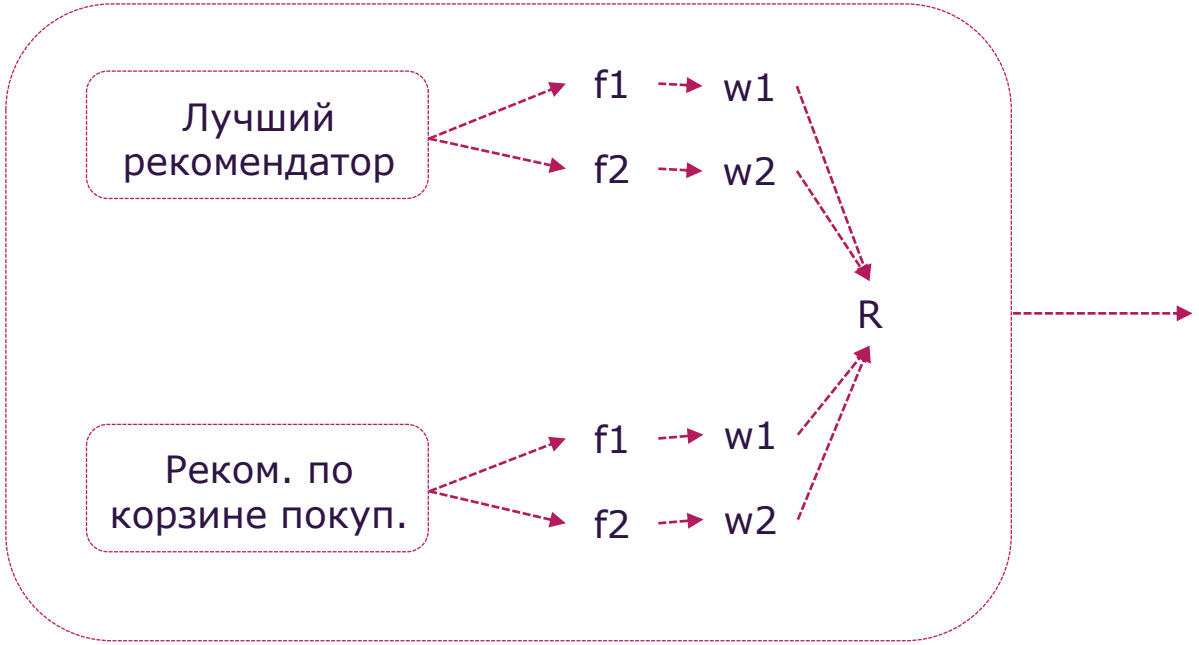
Набор стандартных
отклонений

Журналы:
- даты оценок
- количество оценок
- корреляций

Наборы пользователей
по разным параметрам и
сегментам

ГИБРИДНЫЕ РЕКОМЕНДАЦИИ. FWLS

Ансамбли



Рассмотрим на примере

Что-то стало очень сложно, можно сделать более привычными методами?
Как на счет бустингов?

Рекомендации можно и бустингами

GB ТЕОРИЯ

Рассмотрим GB

UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	1	-	5	-
5	-	-	1	-	4

UID	TARGET
1	1
2	1
3	1
4	0
5	0

UID	Фичи			
1	-	1	-	-
2	3	-	2	2
3	5	3	-	1
4	1	-	5	-
5	-	1	-	4

GB ТЕОРИЯ

Рассмотрим GB

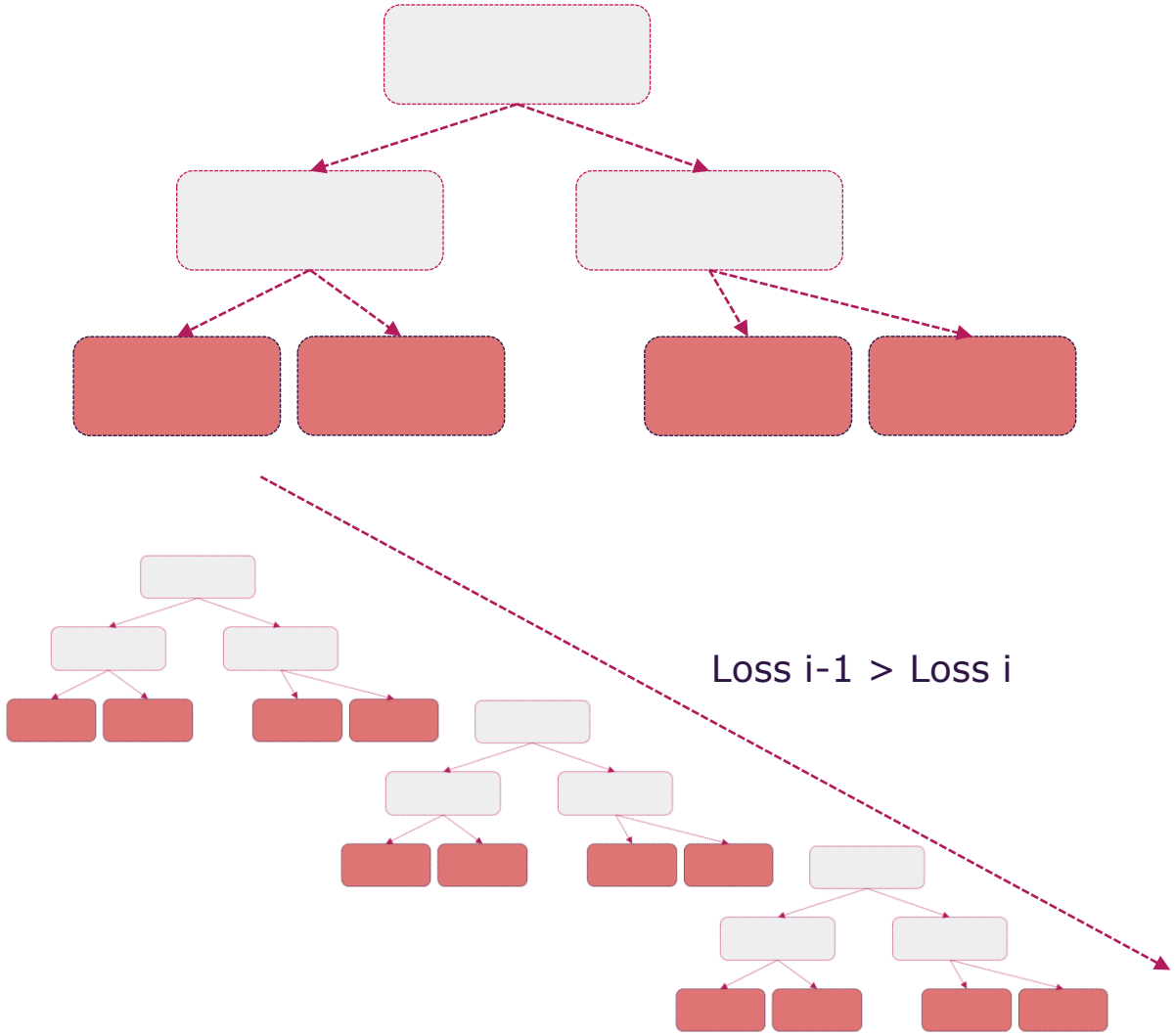
UID	43	11	7	99	3
1	3	-	1	-	-
2	5	3	-	2	2
3	3	5	3	-	1
4	-	1	-	5	-
5	-	-	1	-	4

UID	B	C	D	TARGET
1	1	1	0	1
2	2	0	2	1
3	2	1	0	1
4	2	0	5	**0**
5	1	1	0	**0**

GB ТЕОРИЯ

Решаем деревьями

UID	B	C	D	TARGET
1	1	1	0	1
2	2	0	2	1
3	2	1	0	1
4	2	0	5	**0**
5	1	1	0	**0**



GB ТЕОРИЯ

Что есть целевая?

UID	B	C	D	TARGET
1	1	1	0	1
2	2	0	2	1
3	2	1	0	1
4	2	0	5	**0**
5	1	1	0	**0**

GB ТЕОРИЯ

Что есть целевая?

UID	B	C	D	TARGET
1	1	1	0	**1**
2	2	0	2	**1**
3	2	1	0	**1**
4	2	0	5	**0**
5	1	1	0	**0**

Регрессия

TARGET	DIFF
1	1 – 0.6
1	1 – 0.6
1	1 – 0.6
0	0 – 0.6
0	0 – 0.6

Факт - Предикт

Классификация

TARGET	DIFF
1	1- Log(3/2)
1	1- Log(3/2)
1	1- Log(3/2)
0	0- Log(3/2)
0	0- Log(3/2)

Log(1/0)

GB ТЕОРИЯ

Что есть целевая?

UID	B	C	D	TARGET
1	1	1	0	**1**
2	2	0	2	**1**
3	2	1	0	**1**
4	2	0	5	**0**
5	1	1	0	**0**

Регрессия

TARGET	DIFF
1	0.4
1	0.4
1	0.4
0	– 0.6
0	– 0.6

Факт - Предикт

Классификация

TARGET	DIFF
1	0.6
1	0.6
1	0.6
0	- 0.4
0	- 0.4

Log(1/0)

GB ТЕОРИЯ

Что есть целевая?

UID	B	C	D	TARGET
1	1	1	0	**1**
2	2	0	2	**1**
3	2	1	0	**1**
4	2	0	5	**0**
5	1	1	0	**0**

Регрессия

TARGET	DIFF
1	0.4
1	0.4
1	0.4
0	– 0.6
0	– 0.6

Факт - Предикт

Классификация

TARGET	DIFF
1	0.6
1	0.6
1	0.6
0	- 0.4
0	- 0.4

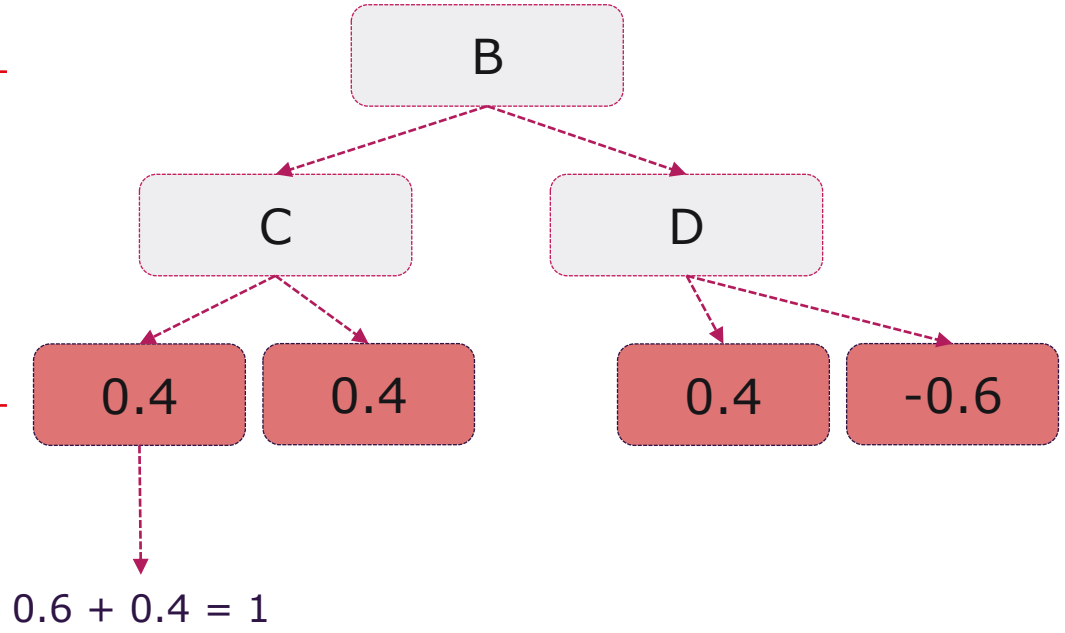
Log(3/2)

0.4 == threshold

GB ТЕОРИЯ

100% успех?

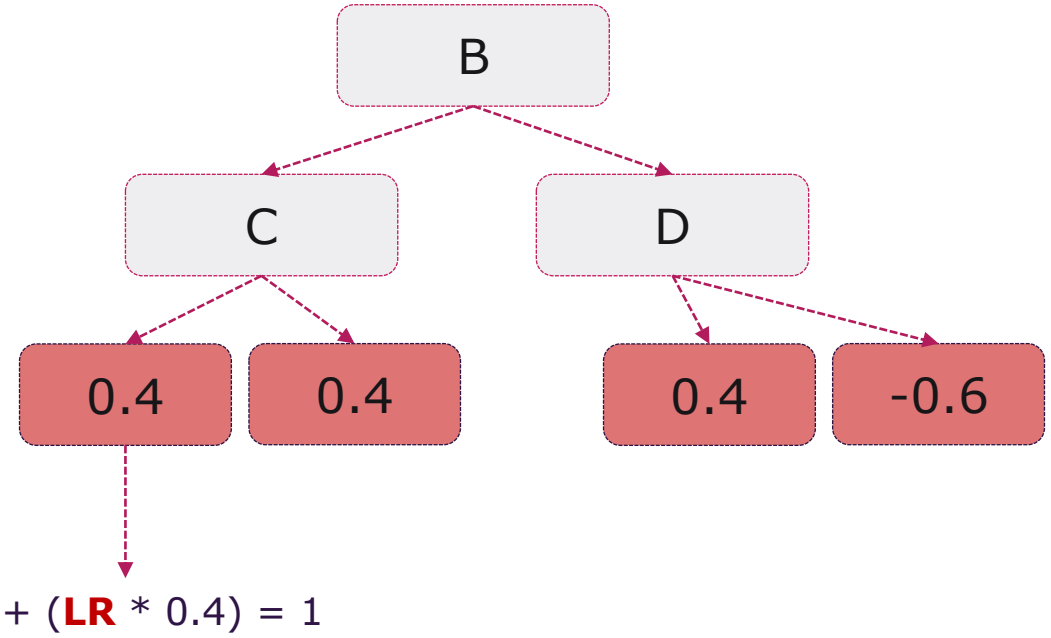
UID	B	C	D	TARGET	DIFF
1	1	1	0	**1**	0.4
2	2	0	2	**1**	0.4
3	2	1	0	**1**	0.4
4	2	0	5	**0**	– 0.6
5	1	1	0	**0**	– 0.6



GB ТЕОРИЯ

Learning rate

UID	B	C	D	TARGET	DIFF
1	1	1	0	**1**	0.4
2	2	0	2	**1**	0.4
3	2	1	0	**1**	0.4
4	2	0	5	**0**	– 0.6
5	1	1	0	**0**	– 0.6

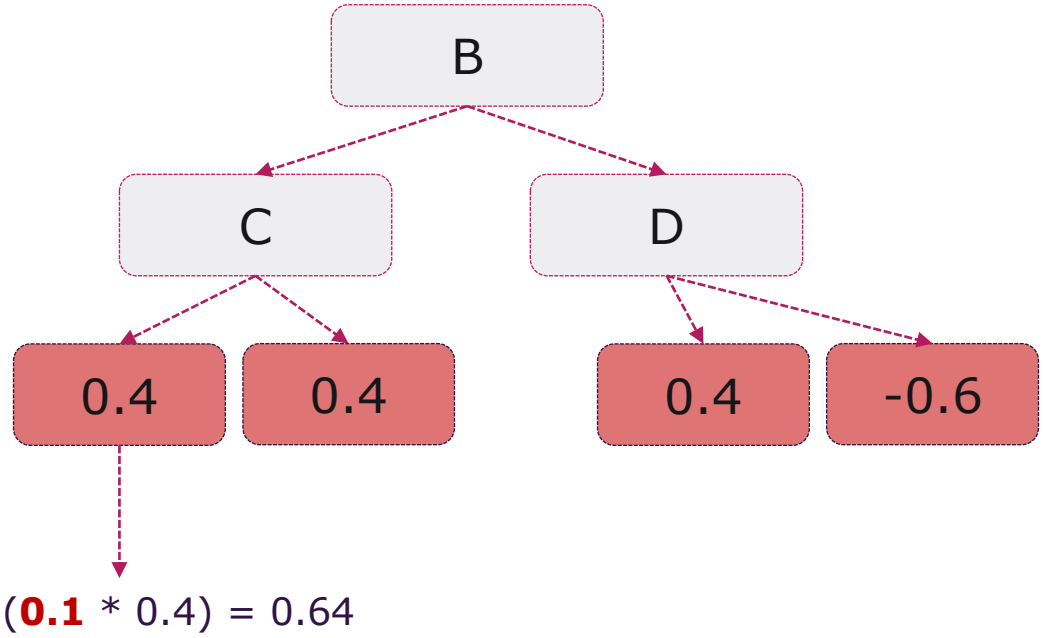


GB ТЕОРИЯ

Learning rate

0.1

UID	B	C	D	TARGET	DIFF
1	1	1	0	**1**	0.4
2	2	0	2	**1**	0.4
3	2	1	0	**1**	0.4
4	2	0	5	**0**	– 0.6
5	1	1	0	**0**	– 0.6



GB ТЕОРИЯ

Learning rate

0.1

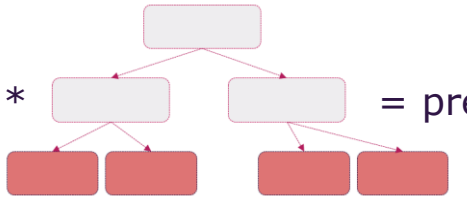
UID	B	C	D	TARGET	DIFF	DIFF2
1	1	1	0	**1**	0.4	0.36
2	2	0	2	**1**	0.4	0.36
3	2	1	0	**1**	0.4	0.36
4	2	0	5	**0**	– 0.6	-0.54
5	1	1	0	**0**	– 0.6	-0.54

0.6 + (**0.1** * 0.4) = 0.64

GB ТЕОРИЯ

Предикт

UID	B	C	D	TARGET	DIFF	DIFF2
1	1	1	0	**1**	0.4	0.36
2	2	0	2	**1**	0.4	0.36
3	2	1	0	**1**	0.4	0.36
4	2	0	5	**0**	– 0.6	-0.54
5	1	1	0	**0**	– 0.6	-0.54

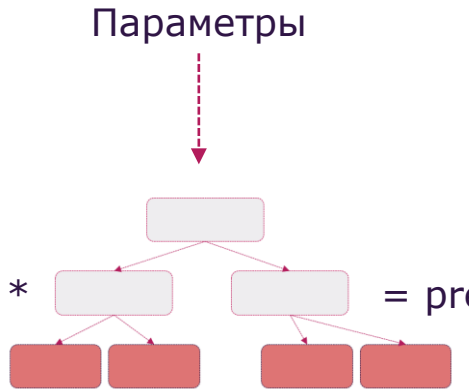
F(x) = F(x)t-1 + LR *  = predict

GB ТЕОРИЯ

Предикт

UID	B	C	D	TARGET	DIFF	DIFF2
1	1	1	0	**1**	0.4	0.36
2	2	0	2	**1**	0.4	0.36
3	2	1	0	**1**	0.4	0.36
4	2	0	5	**0**	– 0.6	-0.54
5	1	1	0	**0**	– 0.6	-0.54

F(x) = F(x)t-1 + LR *



= predict

GB ТЕОРИЯ

Но при классификации по другому

UID	B	C	D	TARGET	DIFF
1	1	1	0	**1**	0.6
2	2	0	2	**1**	0.6
3	2	1	0	**1**	0.6
4	2	0	5	**0**	- 0.4
5	1	1	0	**0**	- 0.4

0.4 + (**LR** * **Gamma**)

GB ТЕОРИЯ

Но при классификации по другому

UID	B	C	D	TARGET	DIFF
1	1	1	0	**1**	0.6
2	2	0	2	**1**	0.6
3	2	1	0	**1**	0.6
4	2	0	5	**0**	- 0.4
5	1	1	0	**0**	- 0.4

0.4 + (**LR** * **Gamma**)

$$\frac{\text{Факт - P}}{\text{P * (1 – P)}}$$

GB ТЕОРИЯ

Но при классификации по другому

UID	B	C	D	TARGET	DIFF
1	1	1	0	**1**	0.6
2	2	0	2	**1**	0.6
3	2	1	0	**1**	0.6
4	2	0	5	**0**	- 0.4
5	1	1	0	**0**	- 0.4

0.4 + (**LR** * **Gamma**)

$$\frac{0.6 - 0.4}{0.4 * (1 - 0.4)}$$

GB ТЕОРИЯ

Но при классификации по другому

UID	B	C	D	TARGET	DIFF
1	1	1	0	**1**	0.6
2	2	0	2	**1**	0.6
3	2	1	0	**1**	0.6
4	2	0	5	**0**	- 0.4
5	1	1	0	**0**	- 0.4

0.4 + (**LR** * **Gamma**)

$$\frac{0.6 - 0.4}{0.4 * (1 - 0.4)}$$

0.83

GB ТЕОРИЯ

Но при классификации по другому

UID	B	C	D	TARGET	DIFF	Gamma
1	1	1	0	**1**	0.6	0.83
2	2	0	2	**1**	0.6	0.83
3	2	1	0	**1**	0.6	0.83
4	2	0	5	**0**	- 0.4	-3.3
5	1	1	0	**0**	- 0.4	-3.3

0.4 + (**LR** * **Gamma**)

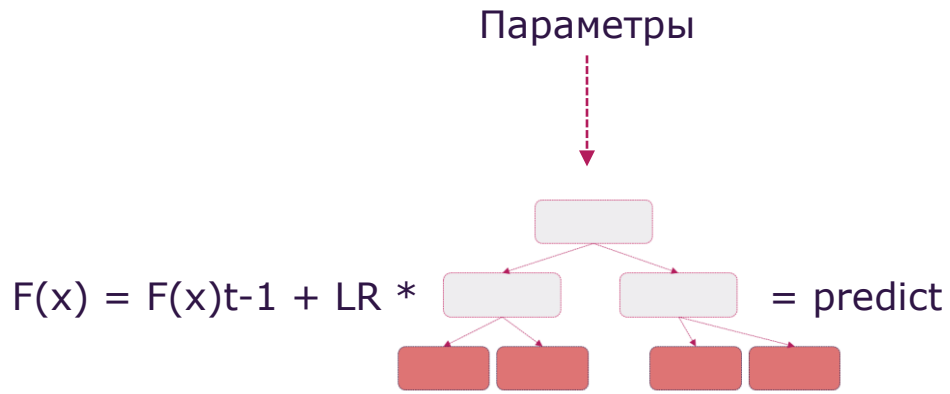
$$\frac{0.6 - 0.4}{0.4 * (1 - 0.4)}$$

0.83

GB ТЕОРИЯ

Но при классификации по другому

UID	B	C	D	TARGET	DIFF	Gamma	Predict
1	1	1	0	**1**	0.6	0.83	0.483
2	2	0	2	**1**	0.6	0.83	0.483
3	2	1	0	**1**	0.6	0.83	0.483
4	2	0	5	**0**	- 0.4	-3.3	0,07
5	1	1	0	**0**	- 0.4	-3.3	0,07



ССЫЛКИ С ДОП.МАТЕРИАЛОМ

ДОПОЛНИТЕЛЬНЫЙ МАТЕРИАЛ