Решение задачи Search Results Relevance

Александр Дьяконов

Московский государственный университет имени М.В. Ломоносова (Москва, Россия)

Moscow Data Science Meetup

27 мая 2016 г., пятница

Начало события в 18.30

Задача Search Results Relevance

Похожа на классическую задачу поиска

row	id	query	product_title	product_description	median relevance	relevance variance
78	230	harleydavidson	Harley-Davidson and Philosophy (Paperback)		4	0.943
120	367	ninja turtle socks	Highpoint Kids' 5PK No Show (Assorted)	He will feel like a superhero in the Teenage Mutant Ninja Turtles 5PK No Show Socks by Highpoint.	4	0.800
6076	19557	long prom dress	Daniella Collection	This one piece dress has mesh and rhinestone detailing. This dress features long sleeves, a round neckline and a flared hem.	3	0.866
153	472	16 gb memory card	Sandisk 16GB non-HS MSD Flash Drive 3in1 - Black (SDSDQR-016G-T46A)	SanDisk Elevate 16GB non-HS MSD 3in, microSD memory card with USB adapter and SD adapter. Includes RescuePro recovery software. Memory Storage Capacity: 16GB Wired Connectivity: Micro SD Slot Features: Plug and Play Includes: MicroSD Adapter, Adapter, USB Adapter Battery no battery used	4	0.000

Задача

Дан запрос: 16 gb memory card

Ему соответствует выдача.

Элемент выдачи –

(название товара, описание)

Sandisk 16GB non-HS MSD Flash Drive 3in1 - Black (SDSDQR-016G-T46A)	SanDisk Elevate 16GB non-HS MSD 3in, microSD memory card with USB adapter and SD adapter. Includes RescuePro recovery software. Memory Storage Capacity: 16GB Wired Connectivity: Micro SD Slot Features: Plug and Play Includes: MicroSD Adapter, Adapter, USB Adapter Battery no battery used

Паре (запрос, выдача) соответствует релевантность – 1, 2, 3 или 4.

Дана ещё дисперсия релевантностей (т.к. несколько ассесоров оценивало выдачу), но не была использована.

Запомним: релевантность в обучении – медиана релевантностей, которые проставили ассесоры

Задача

Почему не классическая задача поиска?

обучение

запрос 1	выдача 1
запрос 1	выдача 2
запрос 1	выдача 3
запрос 2	выдача 1
запрос 2	выдача 2
запрос 2	выдача 3

тест

запрос 1	выдача 4
запрос 1	выдача 5
запрос 2	выдача 4
запрос 2	выдача 5
запрос 2	выдача 6

Запросы в тесте такие же как и в обучении!

Это можно использовать!

Задача

Хорошая выдача (для определённого запроса)
должна быть похожа на хорошие выдачи (этого запроса)!

Функционал качества: Quadratic Weighted Kappa

показывает согласованность порядков,
когда ответы "мера релевантности"

y = 1 1 1 2 2 3 3 3 # правильный ответ
a = 1 1 2 1 3 2 3 3 # наш ответ
	0.6666667

a = 1 1 1 2 2 3 3 3 # наш ответ	1
a = 3 3 3 2 2 1 1 1 # наш ответ	-1

Quadratic Weighted Kappa

```
E = table(y) %*% t(table(a))
O = table(y,a)
E = E/sum(E)*sum(O)
n = length(unique(y))
W = (matrix(1:n,nr=n,nc=n) -
        matrix(1:n,nr=n,nc=n,byrow = TRUE))**2/(n-1)**2
kappa = 1-sum(W*O)/sum(W*E)
```

```
a = 1 1 2 1 3 2 3 3
y = 1 1 1 2 2 3 3 3

O =
   a
y   1 2 3
  1 2 1 0
  2 1 0 1
  3 0 1 2
W =
     [,1] [,2] [,3]
[1,] 0.00 0.25 1.00
[2,] 0.25 0.00 0.25
[3,] 1.00 0.25 0.00
```

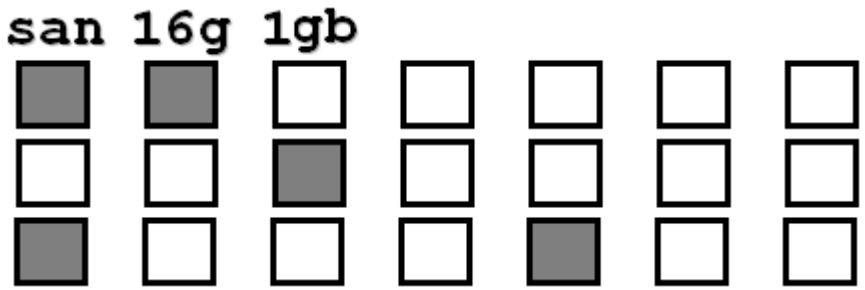
```
E =
   a
y   1 2 3
  1 9 6 9
  2 6 4 6
  3 9 6 9
E = нормализованная
   a
y        1    2     3
  1 1.125 0.75 1.125
  2 0.750 0.50 0.750
  3 1.125 0.75 1.125

Kappa =
0.6666667
```

Метод решения

1. Предобработка данных
2. Генерация признаков
3. Выбор модели / настройка
4. Ансамбли
5. Деформация ответов / решающее правило

Построение очень простой модели!

Sandisk 16GB non-HS MSD Flash Drive 3in1 - Black (SDSDQR-016G-T46A)	**Удаляем html-теги, схлопываем текст, удаляем спецсимволы** sandisk16gbnonhsmsdflashdrive3in1 blacksdsdqr016gt46a
Делаем 3-граммы san,and,ndi,dis,isk,sk1,k16,16g, 6gb,gbn ... **Кстати, для правильного слова "scandisk":** sca, can, and, ndi, dis, isk	**Переходим к модели "мешок слов" + tf-idf** 

3-граммы могут победить опечатки.

Другие подходы (2е место)

hardisk	hard drive
extenal	external
soda stream	sodastream
fragance	fragrance
16 gb	16gb
32 gb	32gb
500 gb	500gb
2 tb	2tb
shoppe	shop
refrigirator	refrigerator
assassinss	assassins
harleydavidson	harley davidson
harley-davidson	harley davidson

На практике так тоже делают – много ручной разметки.

Здесь для простоты модели не стали

Заметим, что модель становится переобученной. Для новых запросов такого ручного устранения неоднозначностей нет!

Другие подходы (1е место)

Здесь создан словарь синонимов

```
child, kid kid
bicycle, bike bike
refrigerator, fridge, freezer fridge
fragrance, perfume, cologne, eau de toilette perfume
```

Все делали стемминг (мы нет).

Признаки

Насколько похожи тексты?

Точнее: множества

san,and,ndi,dis,isk,sk1,k16,16g, 6gb,gbn ...

sca, can, and, ndi, dis, isk

cos-мера,

мощность пересечения (нормированная)

Признаки

cos(описание, запрос)

cos(описание, запрос) нормируем на max(описание, запрос) по всем описаниям этого запроса

mean(cos(описание, описание всех товаров этого запроса в обучении: оценка = 4))

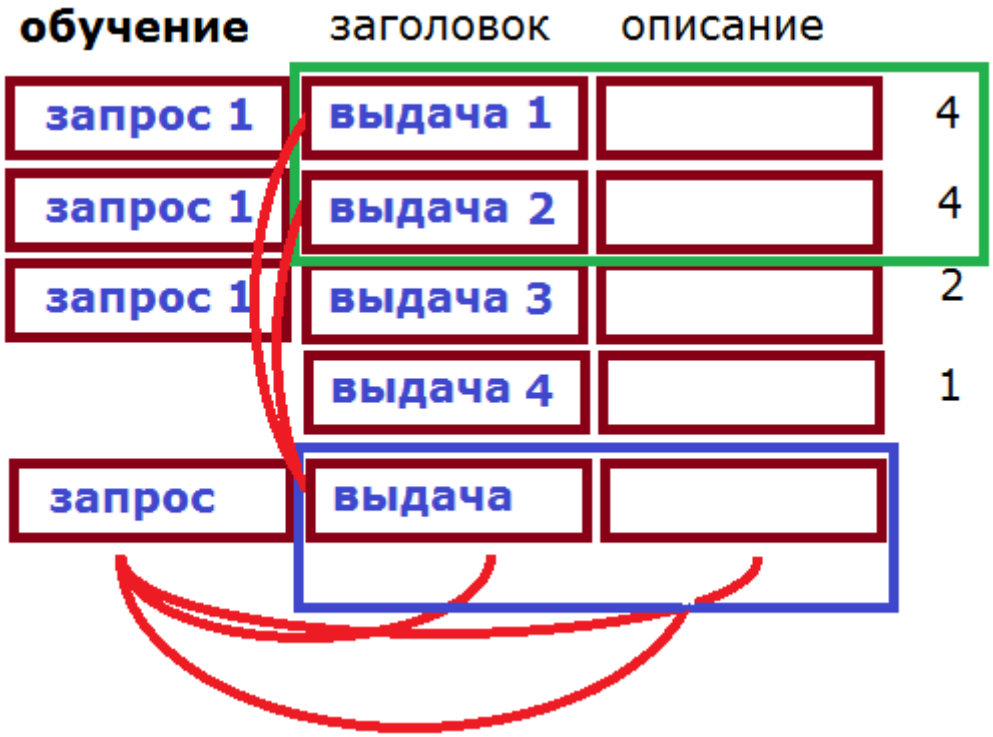
оценка товара: на нём max cos(описание, его описание)

описание =

- **само описание**
- **заголовок**
- **заголовок + описание**

Признаки

– это взаимодействия блоков данных!



Выбор модели

Изначально проблема как настраиваться:

1. Задача классификации с классами [1, 2, 3, 4]

2. Задача регрессии с метками [1, 2, 3, 4]

3. Задача регрессии с метками [1, 4, 9, 16]

4. Задачи классификации/регрессии с метками
[0, 1, 1, 1]
[0, 0, 1, 1]
[0, 0, 0, 1]

Выбор модели

Изначально проблема как настраиваться:

1. Задача классификации с классами [1, 2, 3, 4]

плохо – нет учёта порядка

2. Задача регрессии с метками [1, 2, 3, 4]

хорошо

3. Задача регрессии с метками [1, 4, 9, 16]

нужны эксперименты

4. Задачи классификации/регрессии с метками

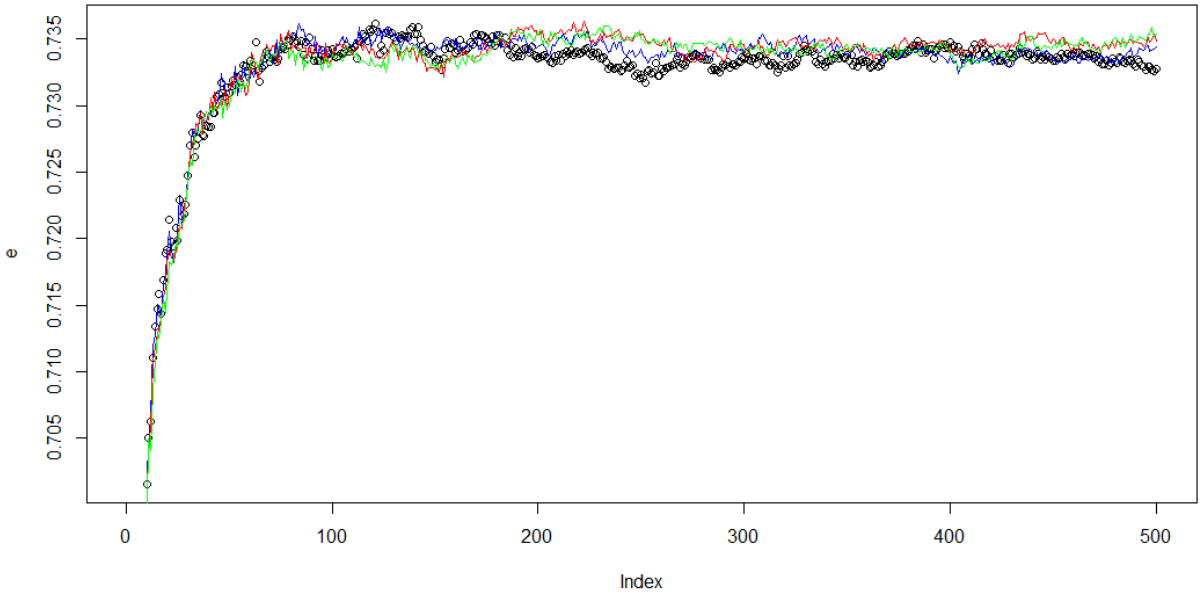
[0, 1, 1, 1]

[0, 0, 1, 1]

[0, 0, 0, 1]

Как потом использовать ответы?

Настройка модели



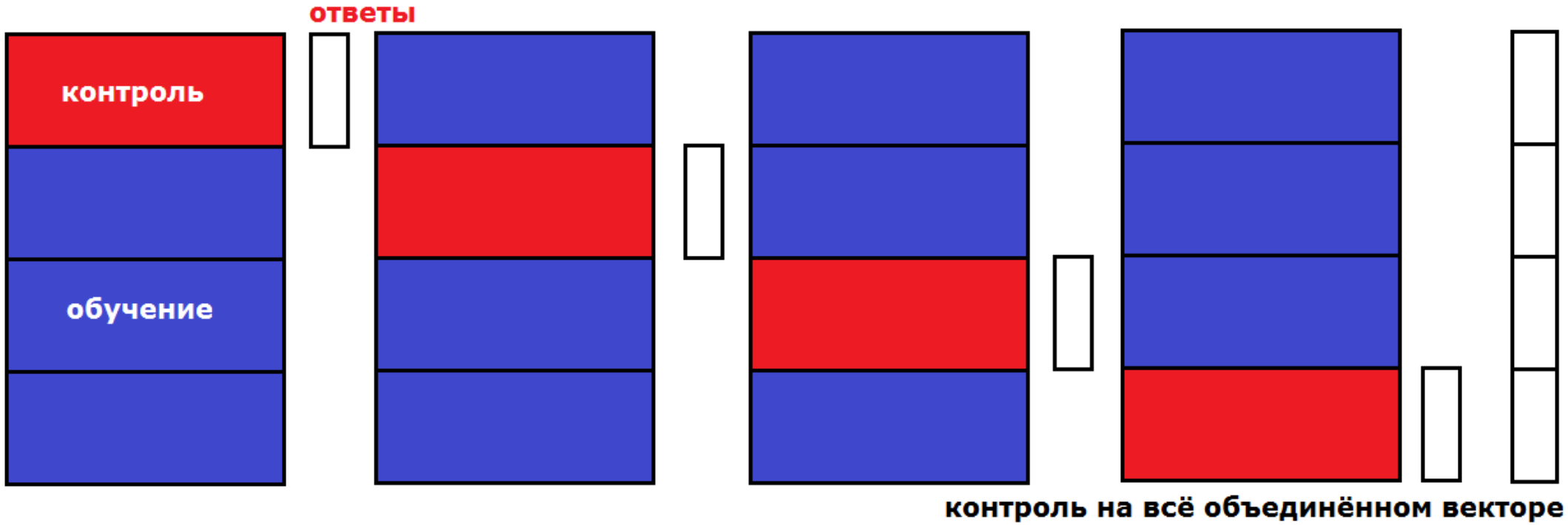

```
model <- randomForest(M1, mr, mtry=40, ntree=200, nodesize=10)
a <- predict(model, M2)
b = pmin(pmax(round(1.45*(a-3.309805)+3.309805),1),4)
```

Использовались случайные леса

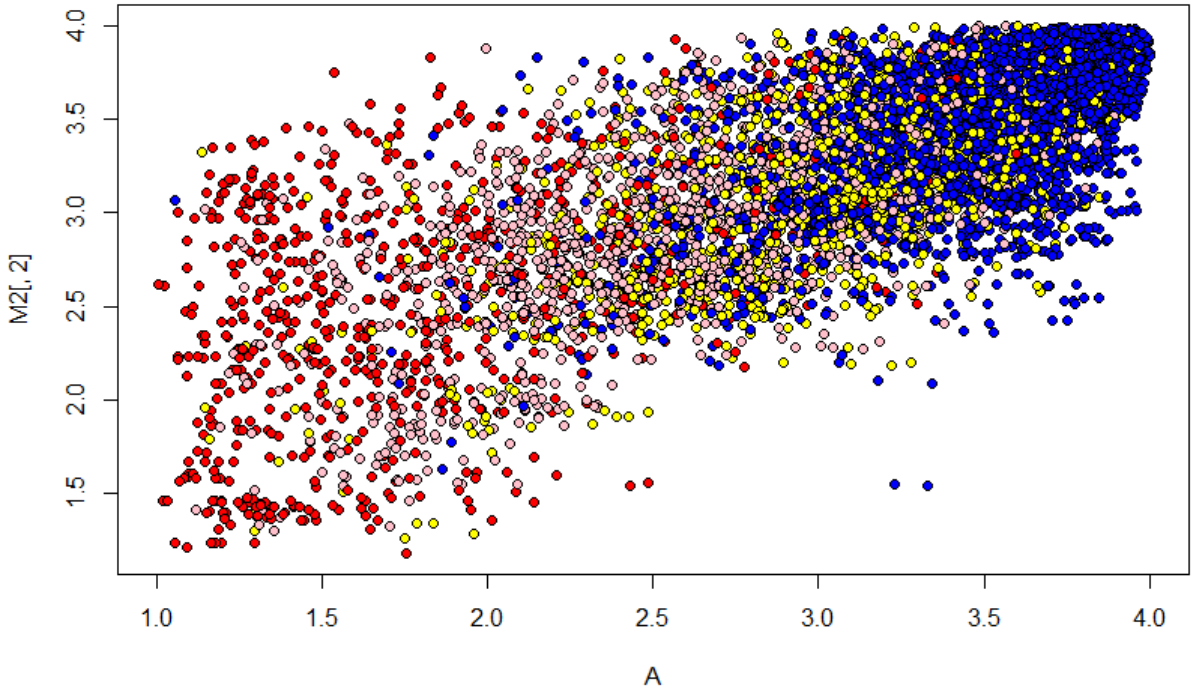
(даже не было зависимости от параметров!)

Отладка

k-fold

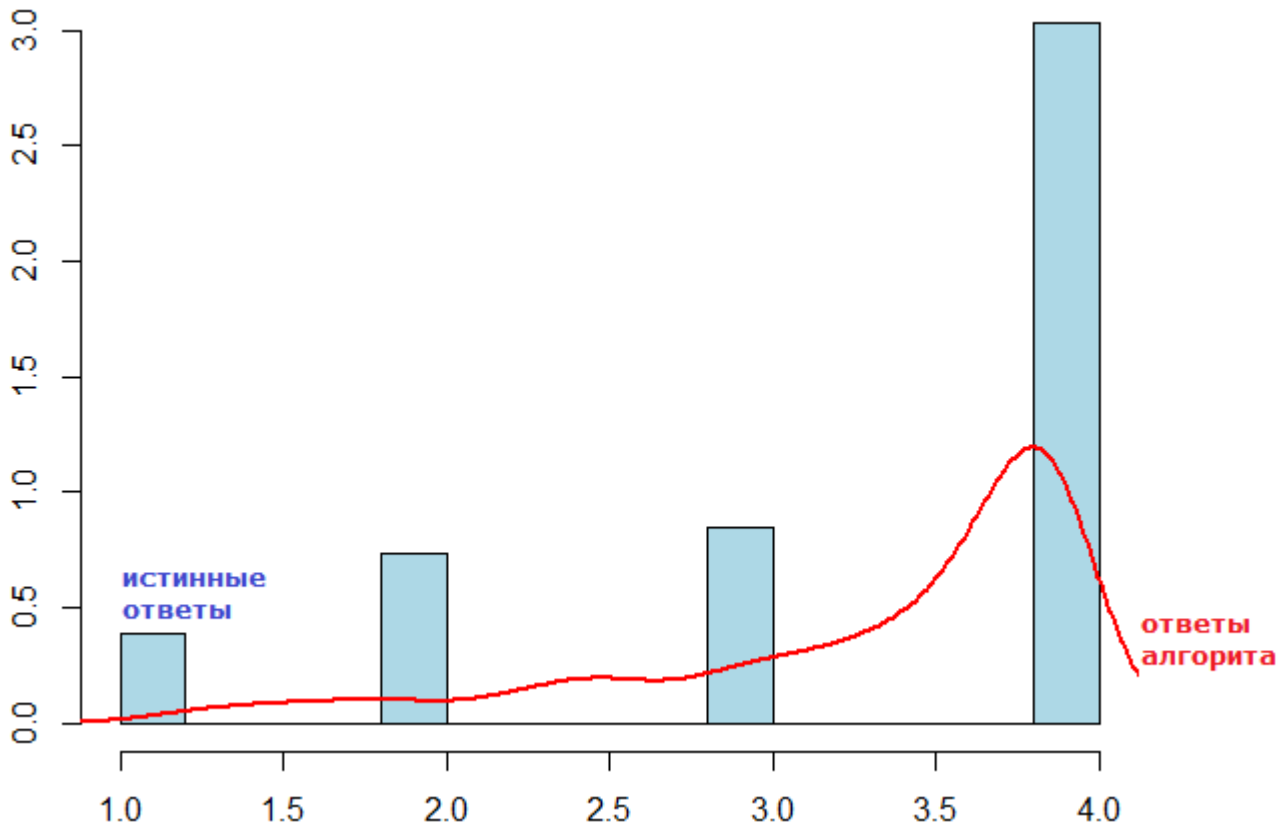


Ответы алгоритма



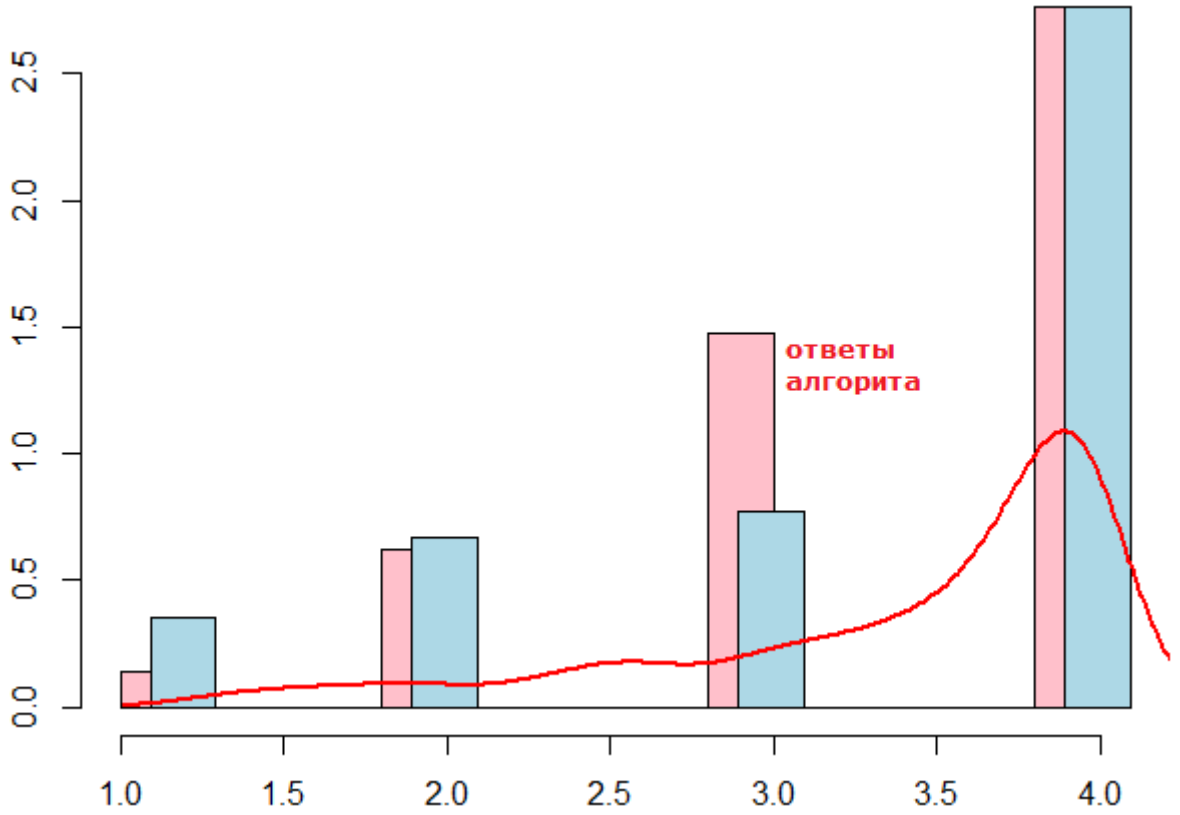
Почему медиана разных алгоритмов лучше усреднения?!

Почему нужны деформации




```
hist(train[,1], probability=TRUE, col='light blue')
points(density(a), type='l', lwd=2, col='red')
```

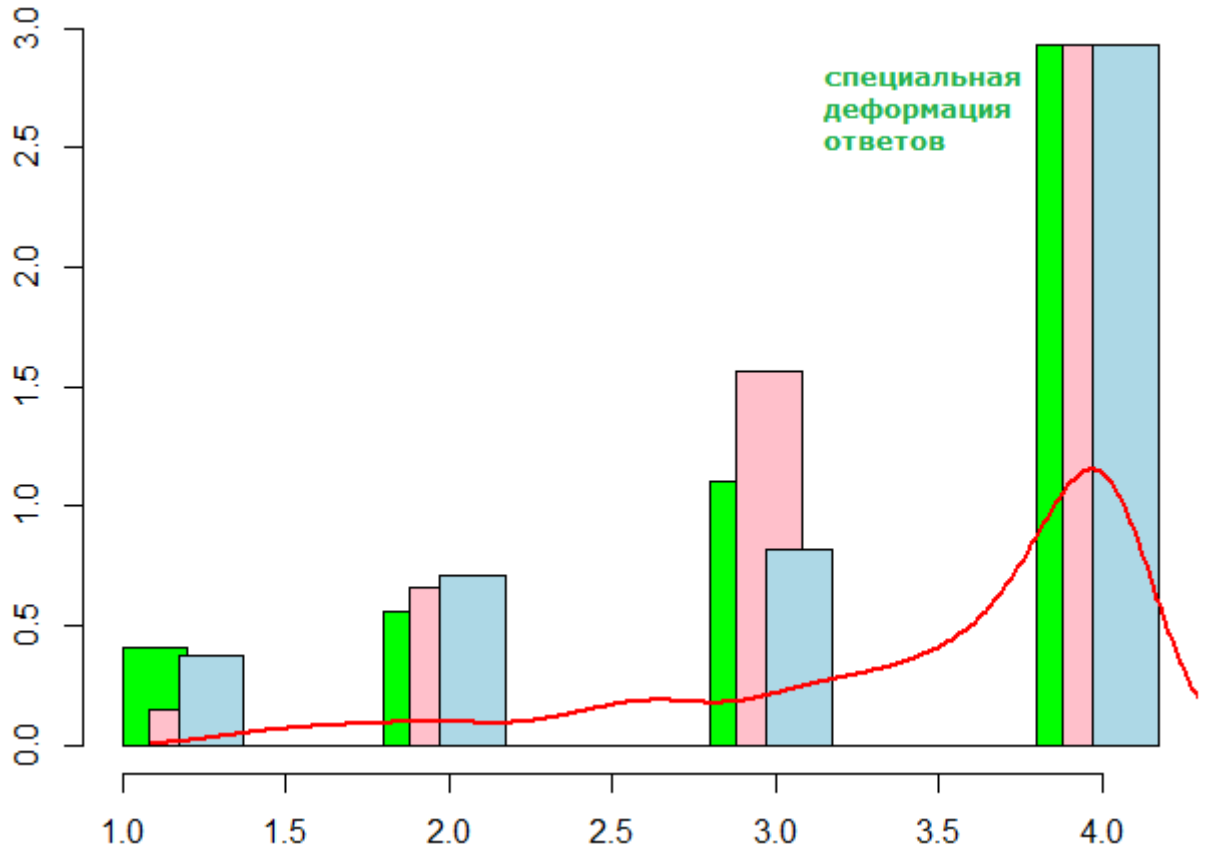
Почему нужны деформации



Если просто округлить – перекос в распределениях

```
hist(round(a), probability=TRUE, col='pink')
```

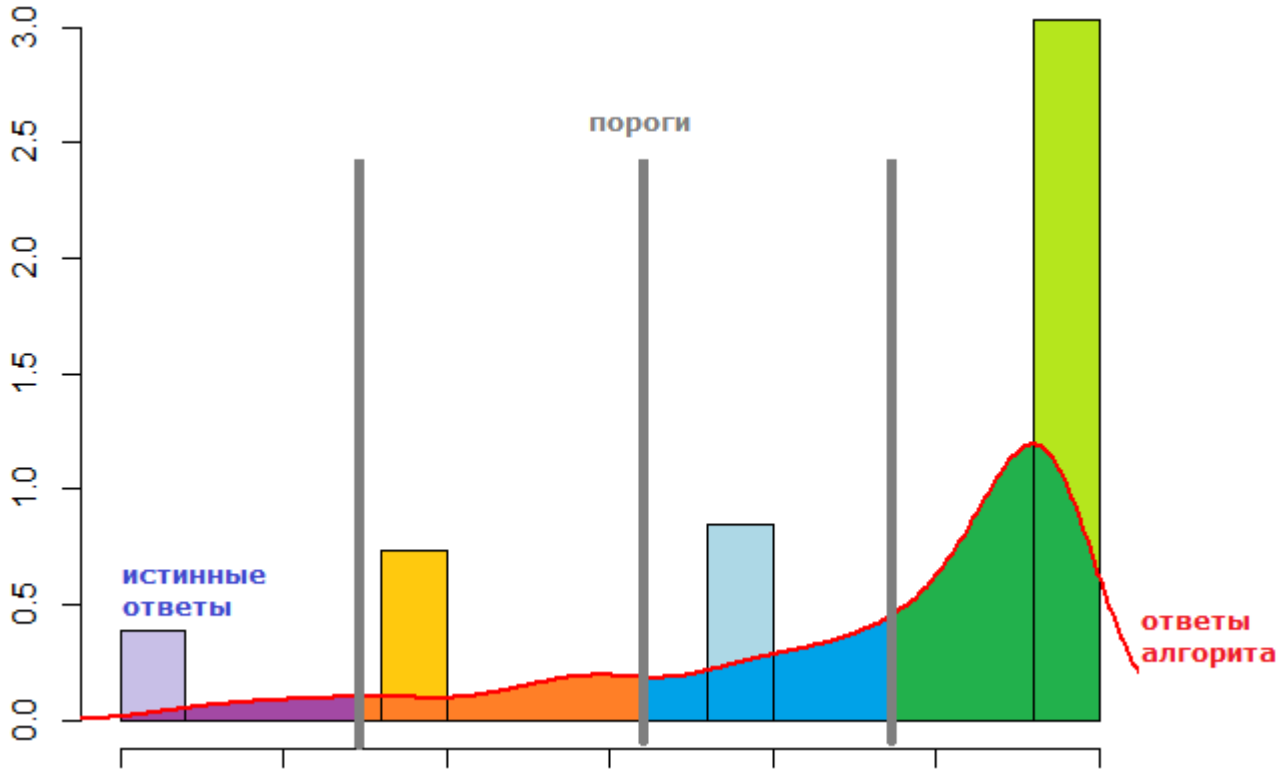
Почему нужны деформации



Специальная деформация ⇒ распределения выравниваются

```
hist(pmin(pmax(round(1.45*(a-3.309805)+3.309805),1),4), probability=TRUE, col='green')
```

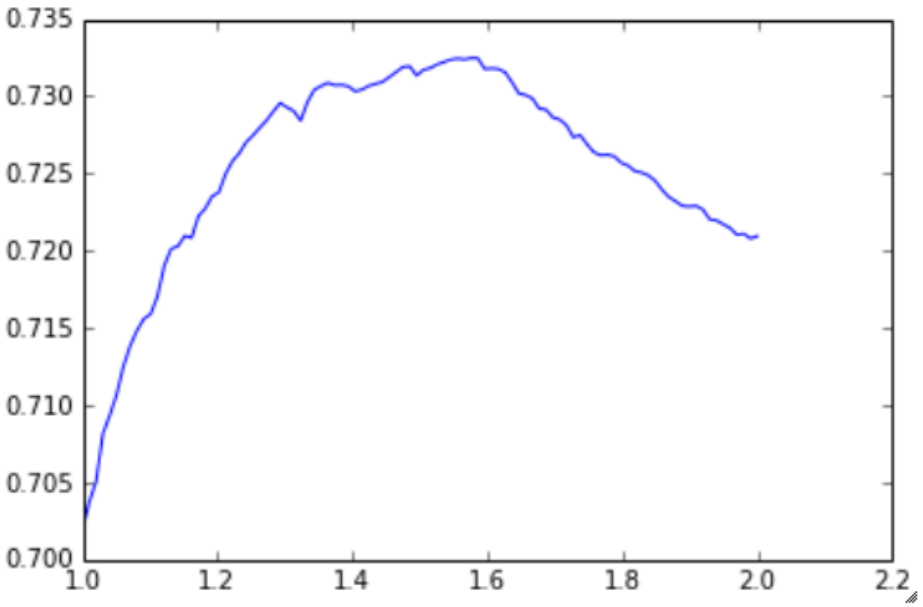
Почему нужны деформации

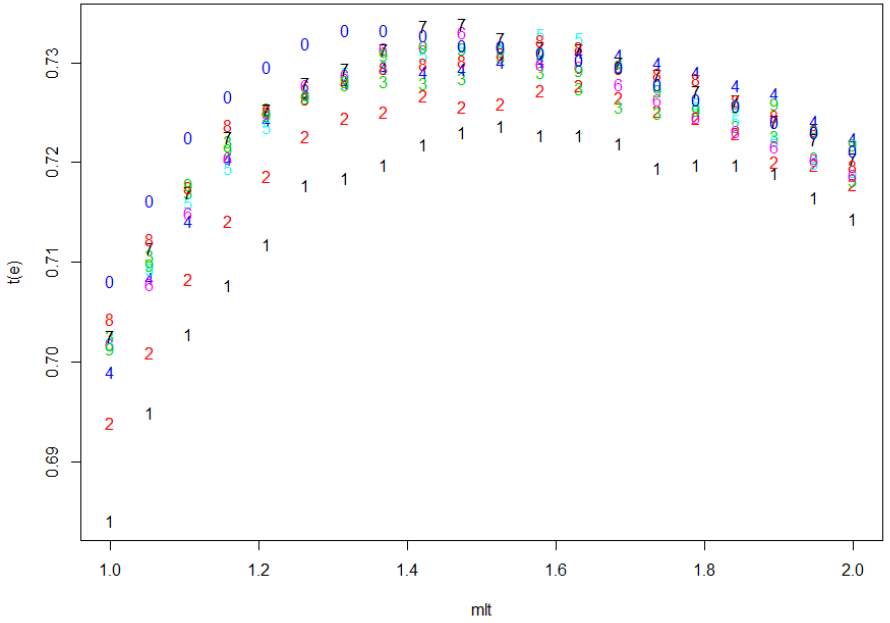


В принципе, можно просто выбирать пороги, чтобы

- **выровнять распределения**
- **повысить качество на контроле**

Почему нужны деформации





Как зависит качество от множителя

```
pmin(pmax(round(1.45*(a-3.309805)+3.309805),1),4)
```

Кстати, 3.309805 – средняя оценка.

Для справки

Хорошая выдача (рел.3-4) похожа на хорошую выдачу в обучении (3-4).

Плохая выдача (1-2) может быть не похожа на плохую выдачу!

Для справки

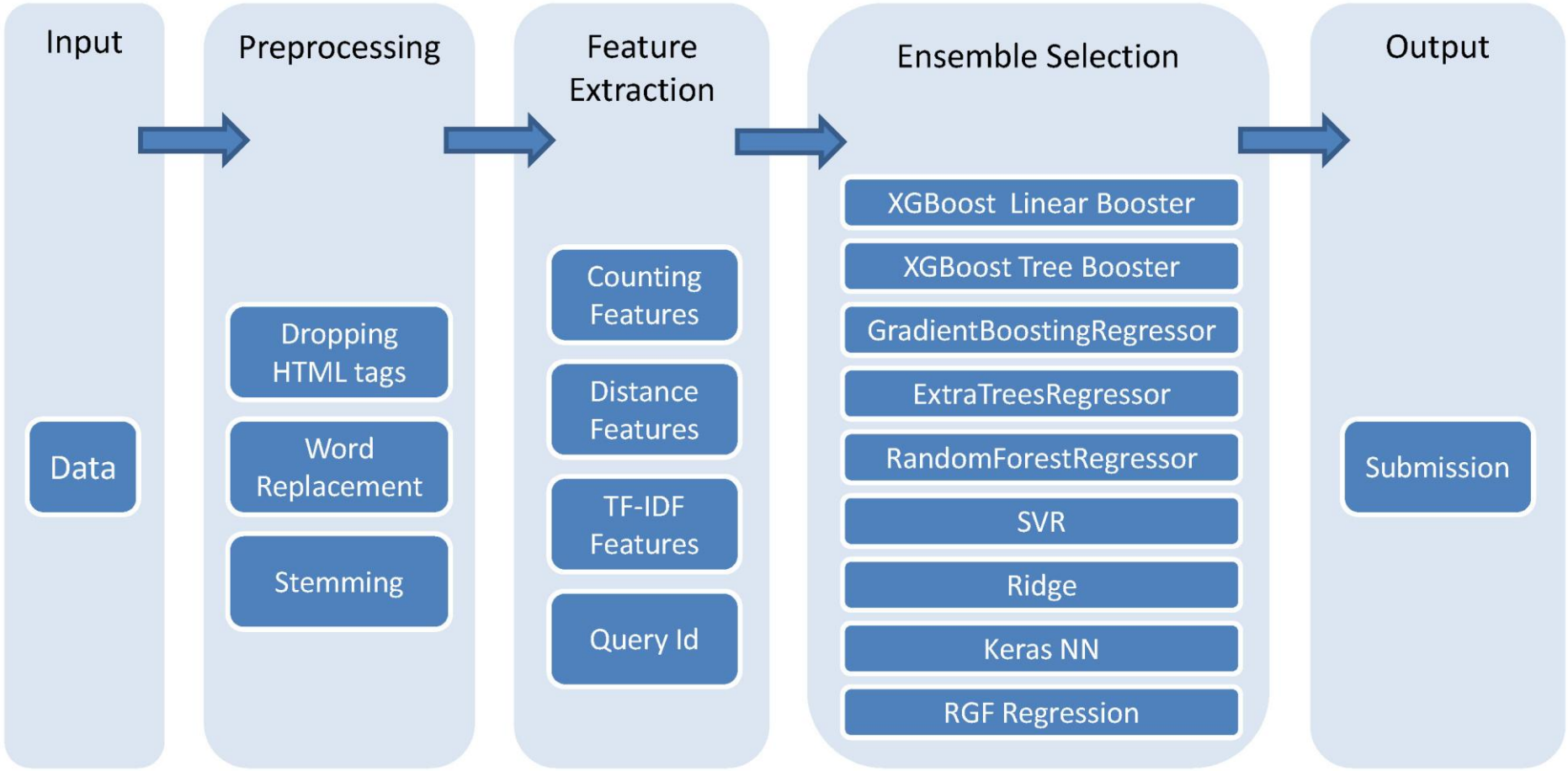
Решающее правило – переводит ответы регрессионных алгоритмов в окончательные.

Обычно

- **простое**
- **тоже требует настройки**
- **использует постановку задачи**

(непересечение классов, малое число меток у объектов и т.п.)

Решение победителя (Chenglong Chen)



https://github.com/ChenglongChen/Kaggle_CrowdFlower

Задача Search Results Relevance

Completed • $20,000 • 1,326 teams

CrowdFlower

Search Results Relevance

Mon 11 May 2015 – Mon 6 Jul 2015 (2 months ago)

#	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	Chenglong Chen ‡ *	0.72189	160	Mon, 06 Jul 2015 09:53:22
2	Mikhail & Stanislav & Dmitry ‡ *	0.71871	83	Mon, 06 Jul 2015 22:55:46 (-1.2h)
3	Quartet ‡ *	0.71861	279	Mon, 06 Jul 2015 17:24:26 (-3.3d)
4	Shize & Shail & Phil	0.71802	252	Mon, 06 Jul 2015 23:44:14
5	I love Phở Bò	0.71700	48	Mon, 06 Jul 2015 08:48:29 (-10.5h)
6	Gzs_iceberg	0.71681	122	Mon, 06 Jul 2015 14:27:09 (-9.9d)
7	YDM	0.71374	283	Mon, 06 Jul 2015 16:31:54 (-1.7h)
8	A & A & G	0.71297	229	Mon, 06 Jul 2015 19:13:28 (-25.6h)
9	ë	0.71265	96	Sun, 05 Jul 2015 21:50:24 (-5.9d)
10	**Alexander D'yakonov (PZAD, Russia)**	**0.71262**	**93**	**Mon, 06 Jul 2015 19:40:28 (-33d)**
11	SearchSearchSearch	0.71022	58	Mon, 06 Jul 2015 01:06:18 (-3.9h)
12	woshialex	0.70889	52	Thu, 02 Jul 2015 00:21:23 (-10.1h)
13	Alexander Ryzhkov (PZAD, Russia)	0.70777	64	Mon, 06 Jul 2015 22:57:31