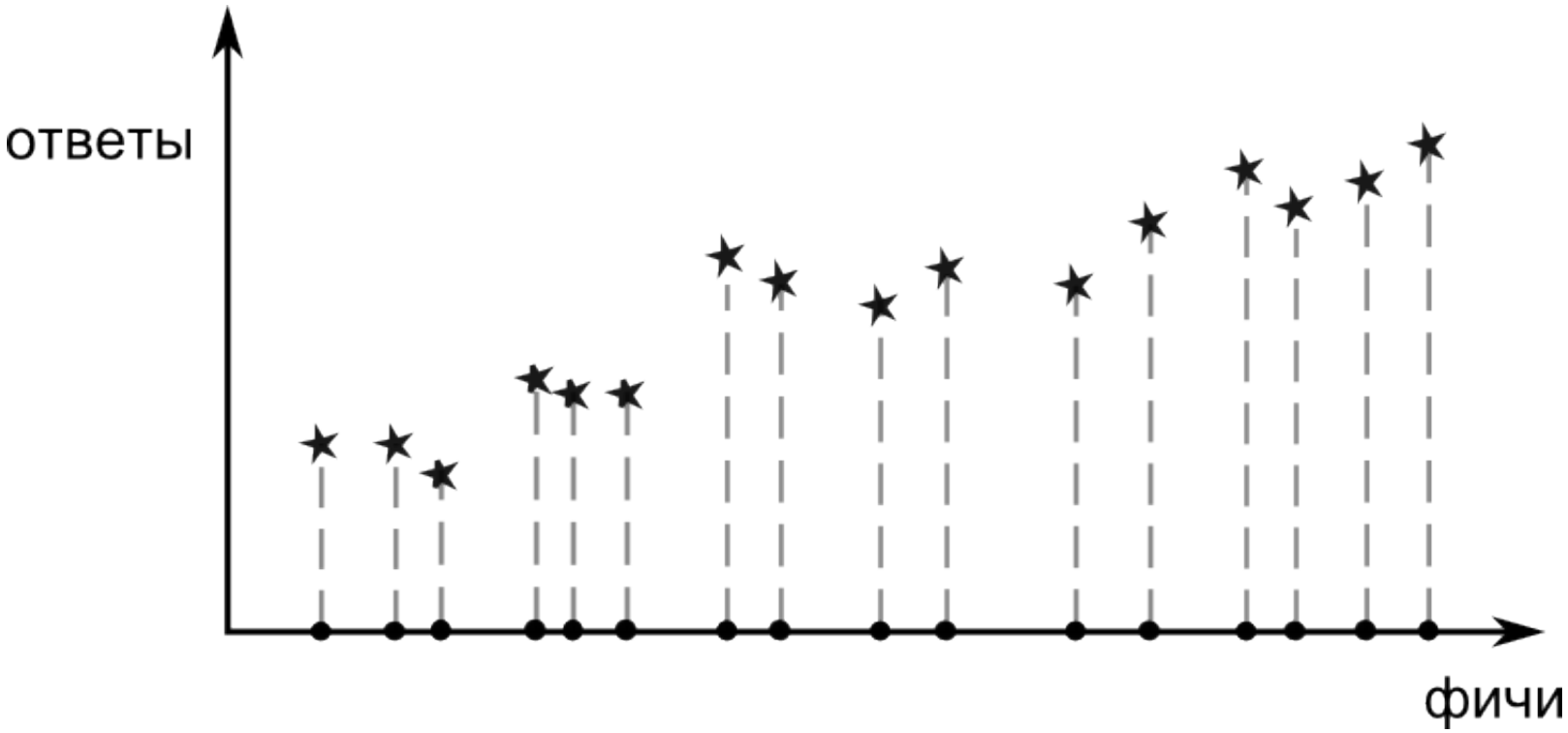
Задачи машинного обучения

Обучение по прецедентам

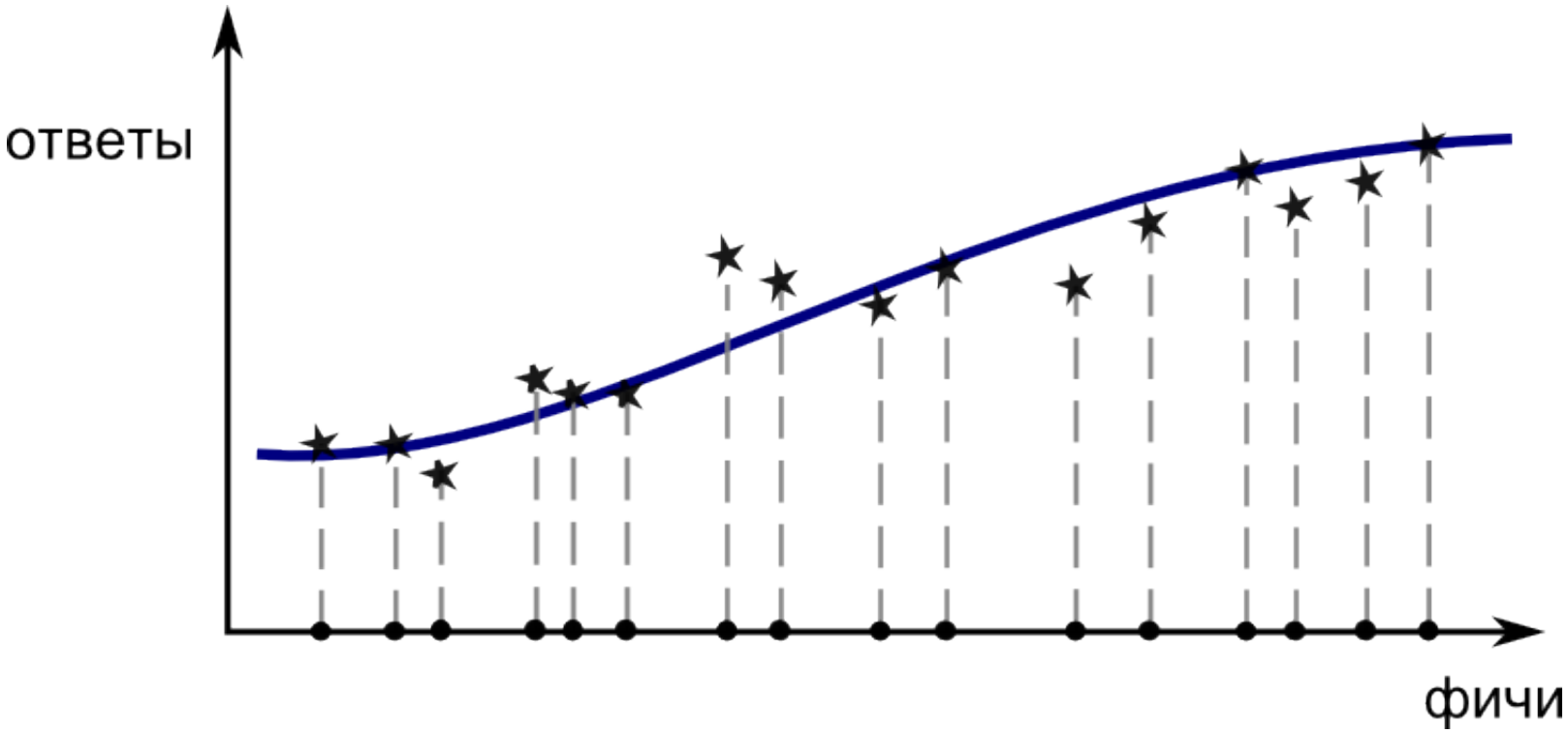
Задачи машинного обучения

Обучение по прецедентам



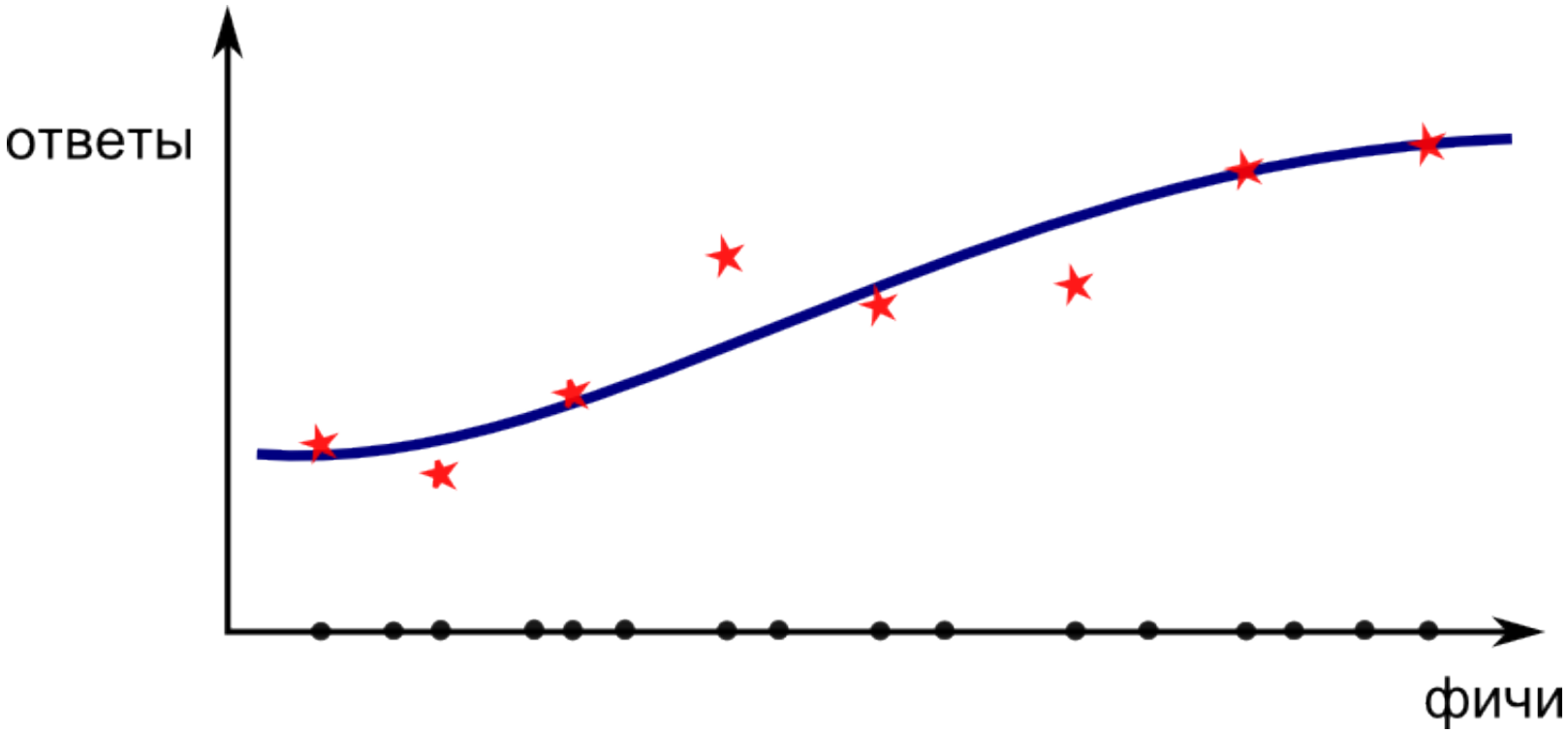
Задачи машинного обучения

Обучение по прецедентам



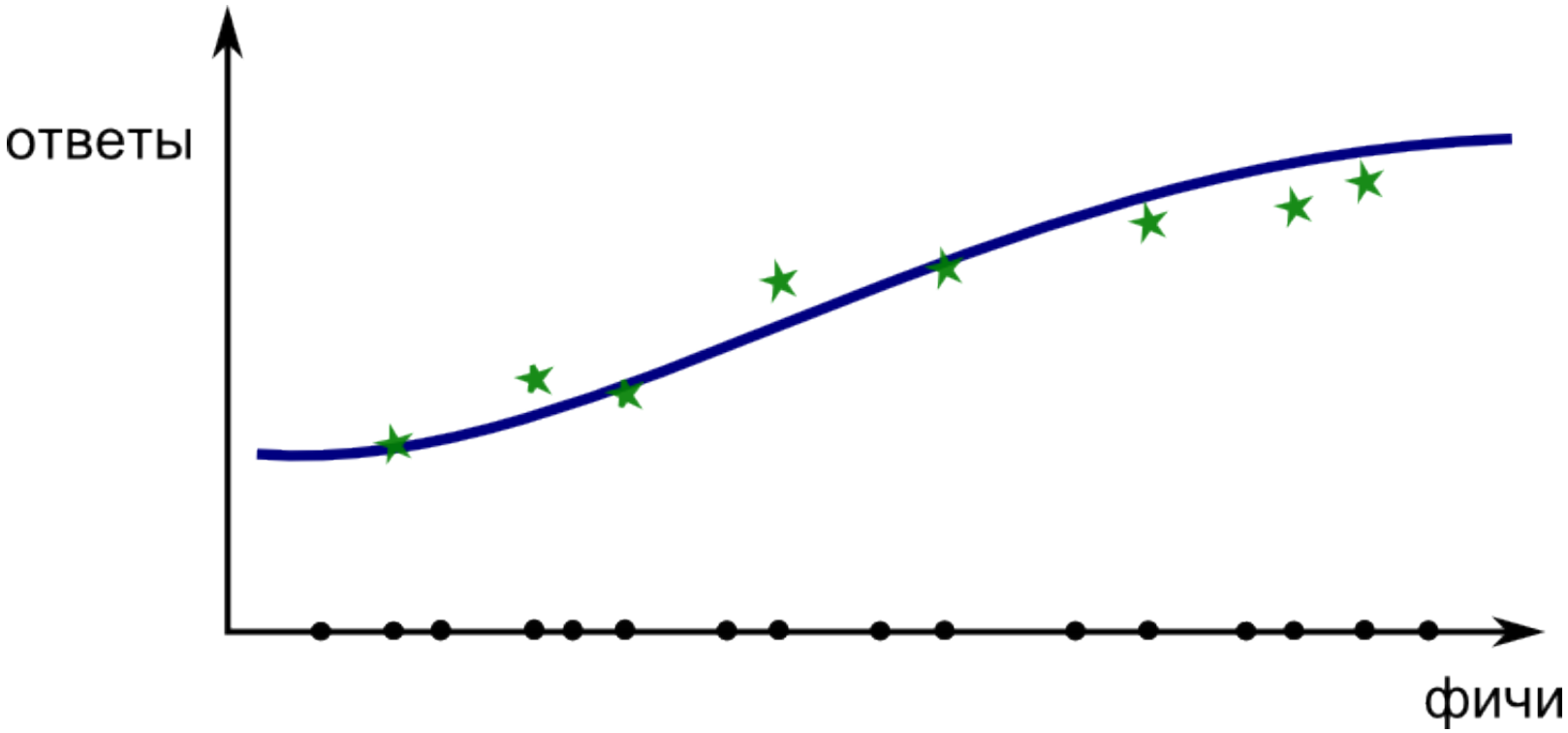
Задачи машинного обучения

Обучение по прецедентам

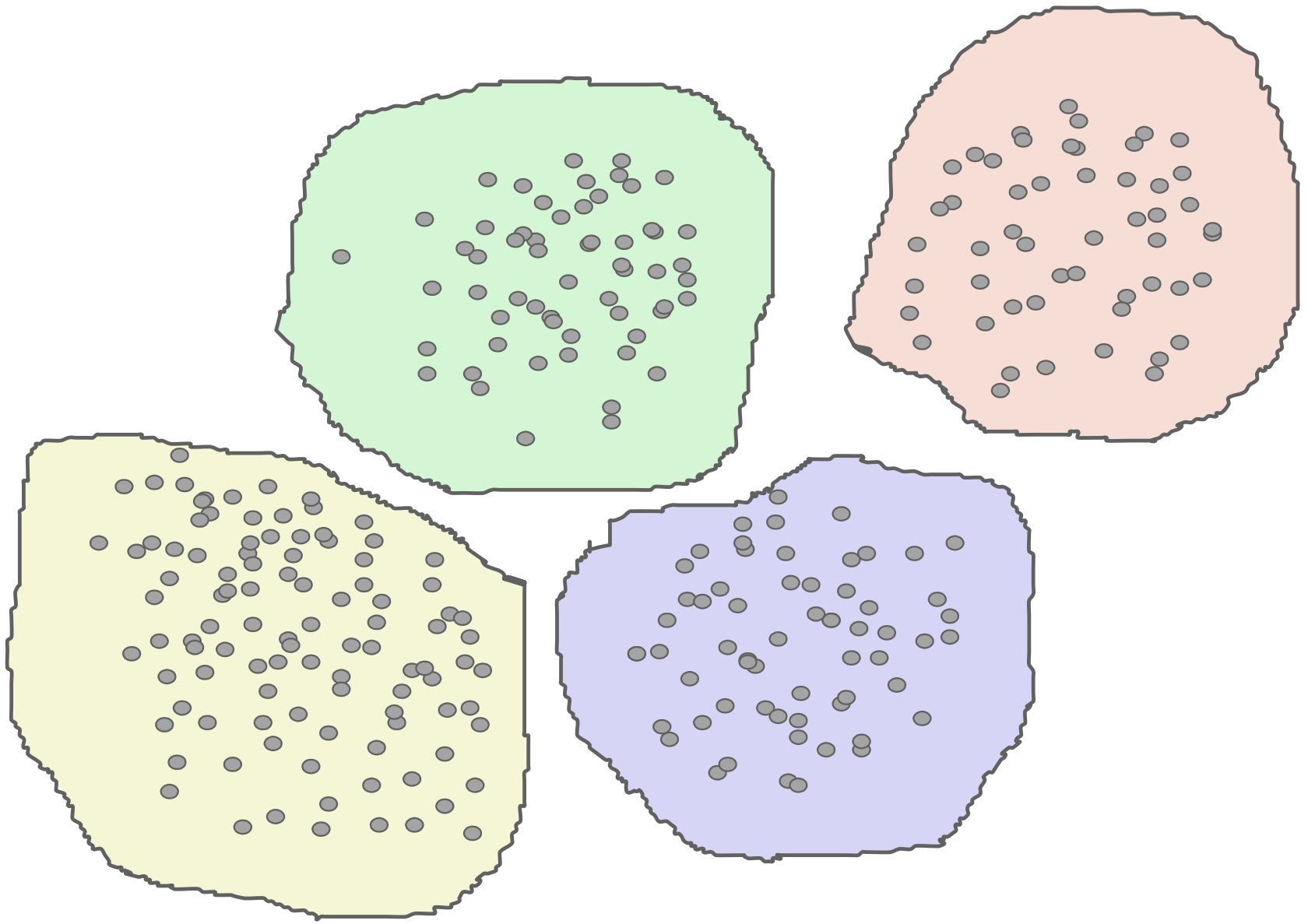


Задачи машинного обучения

Обучение по прецедентам



Кластеризация



Масштабы

1. <1 000 000 объектов

Масштабы

1. <1 000 000 объектов

   На одной машине в оперативной памяти

Масштабы

1. <1 000 000 объектов
   На одной машине в оперативной памяти

2. 1 000 000 000 объектов

Масштабы

1. <1 000 000 объектов
   На одной машине в оперативной памяти

2. 1 000 000 000 объектов
   Вычисления на кластере машин

Масштабы

1. <1 000 000 объектов
   На одной машине в оперативной памяти

2. 1 000 000 000 объектов
   Вычисления на кластере машин

3. 1 000 000 000 000 объектов

Масштабы

1. <1 000 000 объектов
   На одной машине в оперативной памяти

2. 1 000 000 000 объектов
   Вычисления на кластере машин

3. 1 000 000 000 000 объектов
   Миллионы машин. P2P вычисления

Задачи с порядком 1 млрд объектов

- Learning to rank
  - 1 млрд сайтов (2014)
  - Объем текстовых данных ~500 Тб

Задачи с порядком 1 млрд объектов

- Learning to rank
  - 1 млрд сайтов (2014)
  - Объем текстовых данных ~500 Тб
- Recommendations
  - Месячная аудитория рунета 300 млн человек
  - 400 Тб логов в месяц

Задачи с порядком 1 трлрд объектов

- Поиск искусственных сигналов в данных радиотелескопа NASA (SETIHOME)

Задачи с порядком 1 трлрд объектов

- Поиск искусственных сигналов в данных радиотелескопа NASA (SETIHOME)

- Моделирование белков

Построение модели. Семплирование

- Выделить 0.1% - 1% данных

Построение модели. Семплирование

- Выделить 0.1% - 1% данных
- Построить модель при помощи классических библиотек

Построение модели. Семплирование

- Выделить 0.1% - 1% данных
- Построить модель при помощи классических библиотек
- **В 90% случаев это работает!**

Проблемы семплирования

- Неправильное выделение семпла
  - У нас 1 млрд строк лога
  - Берем семпл s=1%
  - И строим предсказание T от фич A,B,C

Проблемы семплирования

- Неправильное выделение семпла
  - У нас 1 млрд строк лога
  - Берем семпл s=1%
  - И строим предсказание T от фич A,B,C

$$P_{sample}(T) = \frac{n(T)\cdot s}{N\cdot s} = \frac{n(T)}{N} = P(T)$$

Проблемы семплирования

- Неправильное выделение семпла
  - У нас 1 млрд строк лога
  - Берем семпл s=1%
  - И строим предсказание T от фич A,B,C

$$P_{sample}(T) = \frac{n(T)\cdot s}{N\cdot s} = \frac{n(T)}{N} = P(T)$$

$$P_{sample}(A) = \frac{n(A)\cdot s}{N\cdot s} = \frac{n(A)}{N} = P(A)$$

Проблемы семплирования

- Неправильное выделение семпла
  - У нас 1 млрд строк лога
  - Берем семпл s=1%
  - И строим предсказание T от фич A,B,C

$$P_{sample}(T) = \frac{n(T)\cdot s}{N\cdot s} = \frac{n(T)}{N} = P(T)$$

$$P_{sample}(A) = \frac{n(A)\cdot s}{N\cdot s} = \frac{n(A)}{N} = P(A)$$

$$P_{sample}(A\cup T) = \frac{n(A\cup T)\cdot s^2}{N\cdot s} = \frac{n(A\cup T)\cdot s}{N} = P(A\cup T)\cdot s$$

Проблемы семплирования

Недостаток статистики

- В полных данных

Проблемы семплирования

Недостаток статистики

- В полных данных

$N = 100\,000\,000$ $n(T) = 1000$ $n(A) = 200$ $n(A \cup T) = 100$

Проблемы семплирования

Недостаток статистики

- В полных данных

$N = 100\,000\,000 \qquad n(T) = 1000 \qquad n(A) = 200 \qquad n(A \cup T) = 100$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

Проблемы семплирования

Недостаток статистики

- В полных данных

$N = 100\,000\,000 \qquad n(T) = 1000 \qquad n(A) = 200 \qquad n(A \cup T) = 100$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

$$conf.\,interval = P \pm 1.96 \frac{\sqrt{(P \cdot (1-P))}}{\sqrt{n(A)}} = 0.5 \pm 0.08$$

Проблемы семплирования

Недостаток статистики

- В полных данных

$N = 100\,000\,000 \qquad n(T) = 1000 \qquad n(A) = 200 \qquad n(A \cup T) = 100$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

$$conf.\,interval = P \pm 1.96 \frac{\sqrt{(P \cdot (1-P))}}{\sqrt{n(A)}} = 0.5 \pm 0.08$$

- То в случайном семпле может оказаться

Проблемы семплирования

Недостаток статистики

- В полных данных

$N = 100\,000\,000 \qquad n(T) = 1000 \qquad n(A) = 200 \qquad n(A \cup T) = 100$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

$$conf.\,interval = P \pm 1.96 \frac{\sqrt{(P \cdot (1-P))}}{\sqrt{n(A)}} = 0.5 \pm 0.08$$

- То в случайном семпле может оказаться

$N = 1\,000\,000 \qquad n(T) = 10 \qquad n(A) = 2 \qquad n(A \cup T) = 1$

Проблемы семплирования

Недостаток статистики

- В полных данных

$N = 100\,000\,000 \qquad n(T) = 1000 \qquad n(A) = 200 \qquad n(A \cup T) = 100$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

$$conf.\,interval = P \pm 1.96 \frac{\sqrt{(P \cdot (1-P))}}{\sqrt{n(A)}} = 0.5 \pm 0.08$$

- То в случайном семпле может оказаться

$N = 1\,000\,000 \qquad n(T) = 10 \qquad n(A) = 2 \qquad n(A \cup T) = 1$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

Проблемы семплирования

Недостаток статистики

- В полных данных

$N = 100\,000\,000 \qquad n(T) = 1000 \qquad n(A) = 200 \qquad n(A \cup T) = 100$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

$$conf.\,interval = P \pm 1.96 \frac{\sqrt{(P \cdot (1-P))}}{\sqrt{n(A)}} = 0.5 \pm 0.08$$

- То в случайном семпле может оказаться

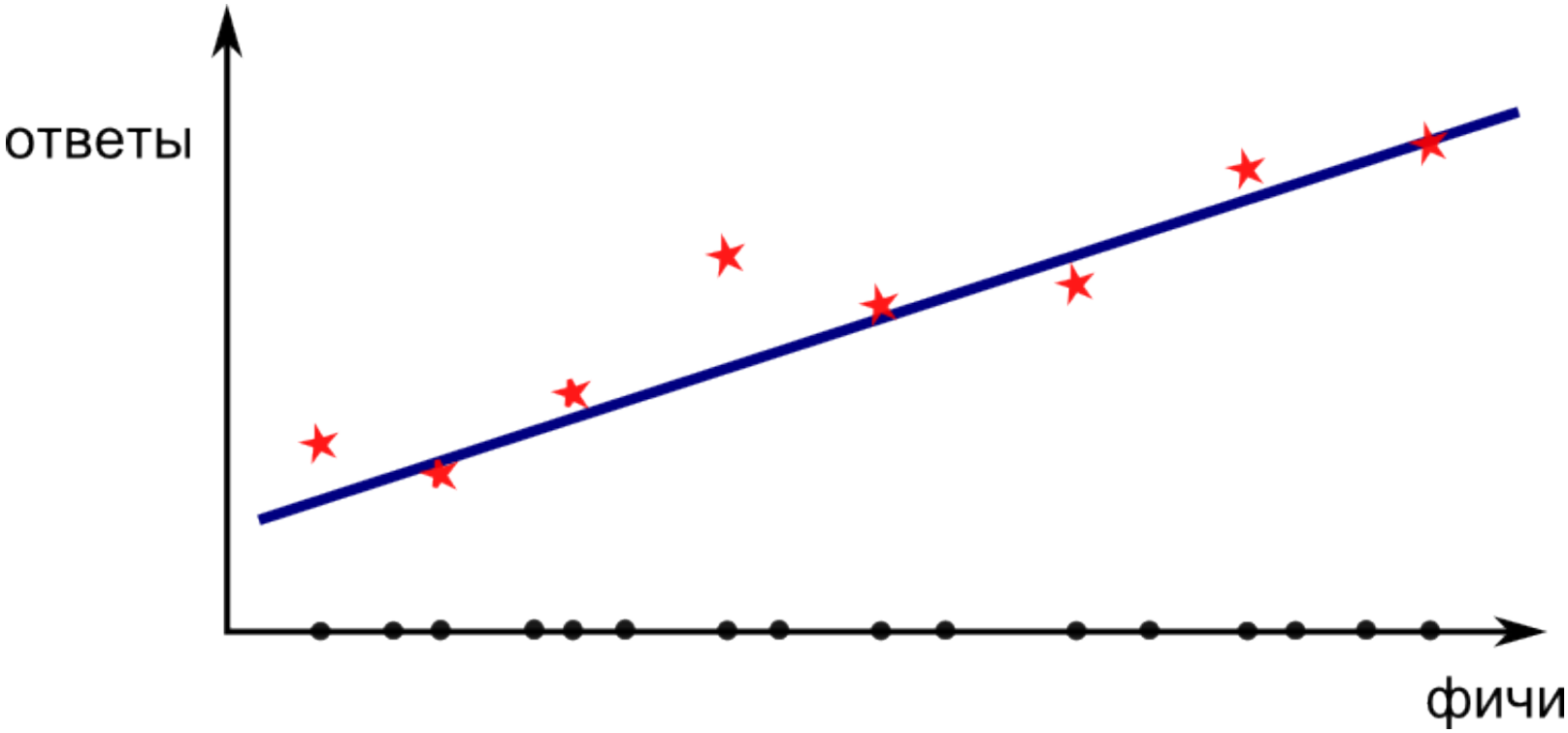
$N = 1\,000\,000 \qquad n(T) = 10 \qquad n(A) = 2 \qquad n(A \cup T) = 1$

$$P(T|A) = \frac{n(A \cup T)}{n(A)} = 0.5$$

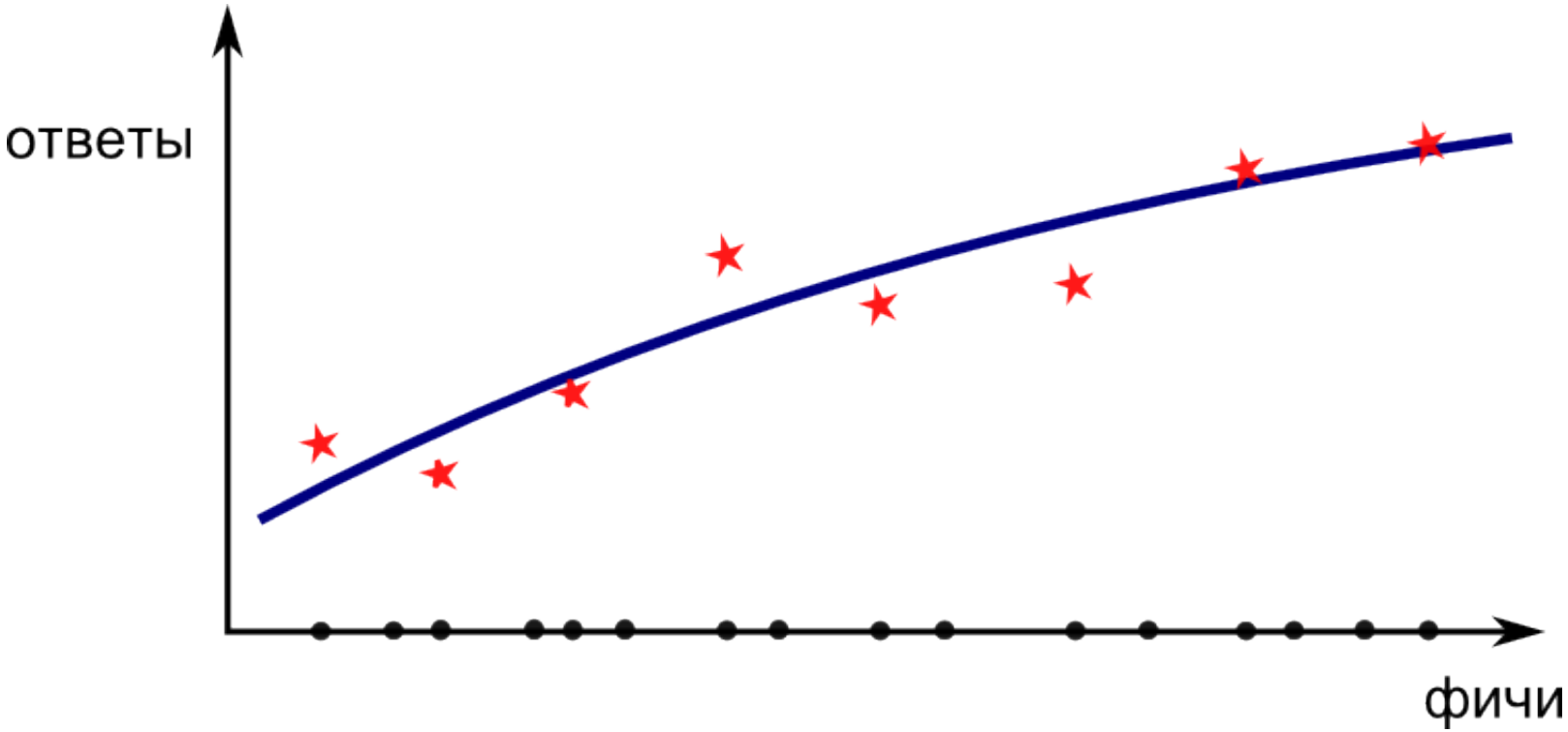
$$conf.\,interval = P \pm 1.96 \frac{\sqrt{(P \cdot (1-P))}}{\sqrt{n(A)}} = 0.5 \pm 0.69$$

Проблемы семплирования: Переобучение

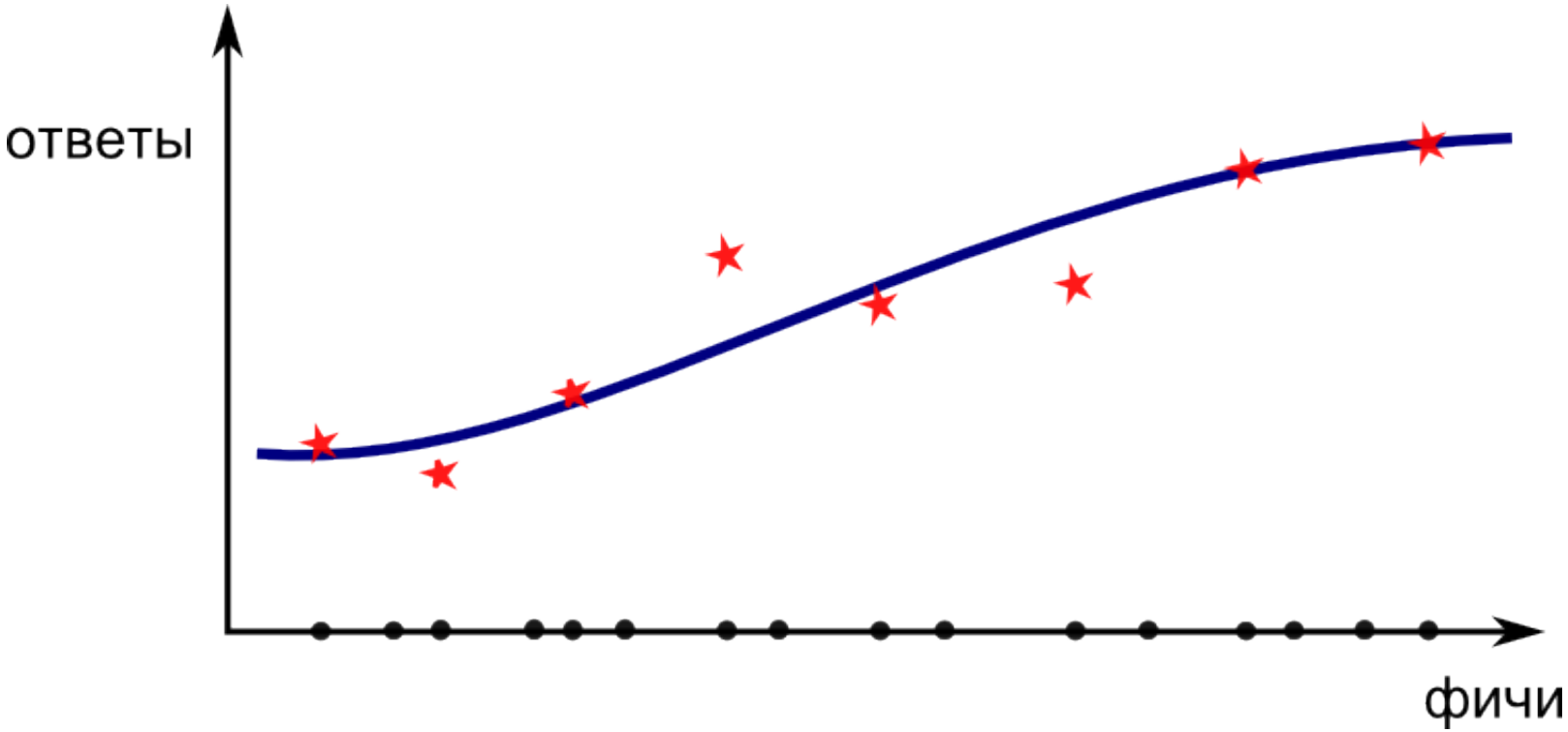
Проблемы семплирования: Переобучение



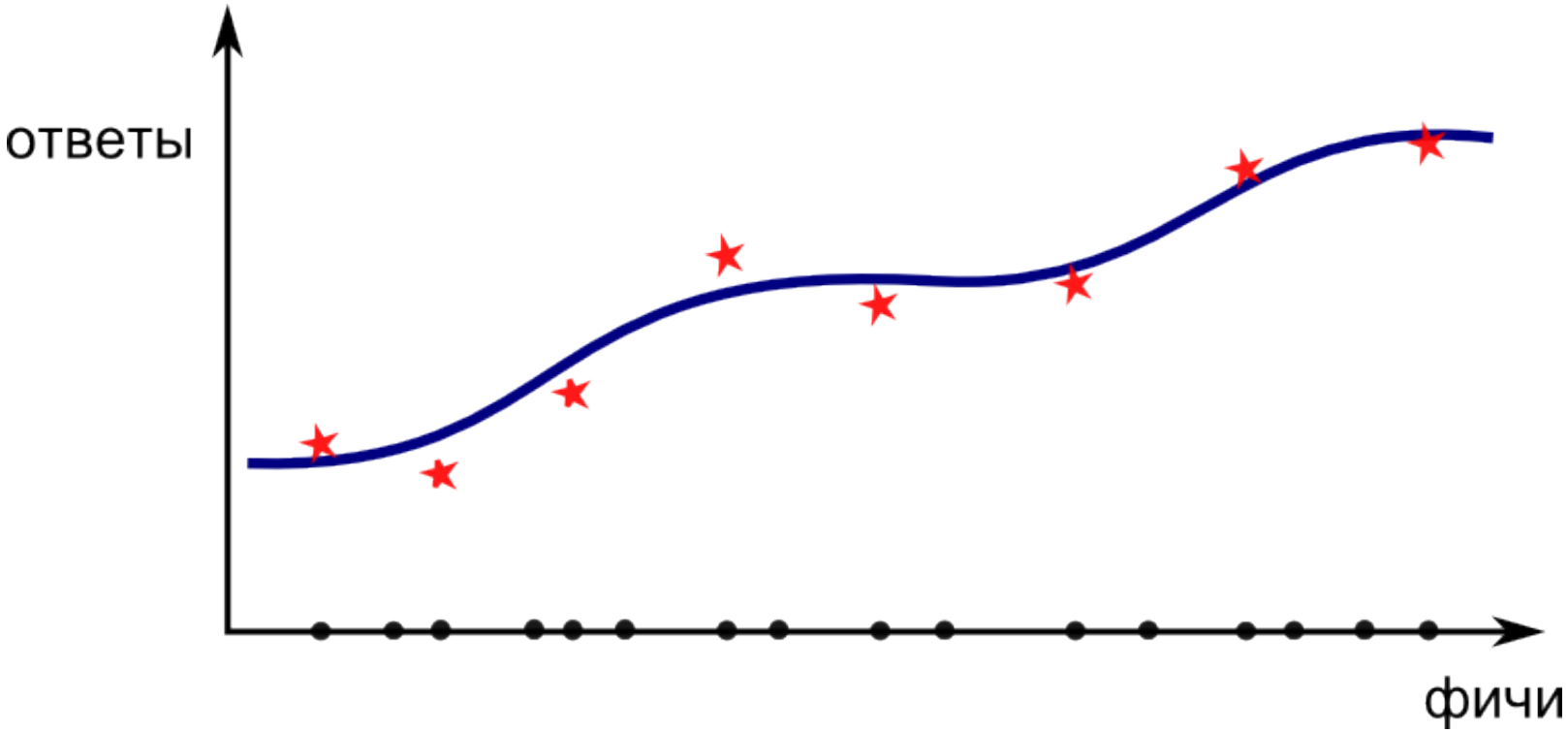
Проблемы семплирования: Переобучение



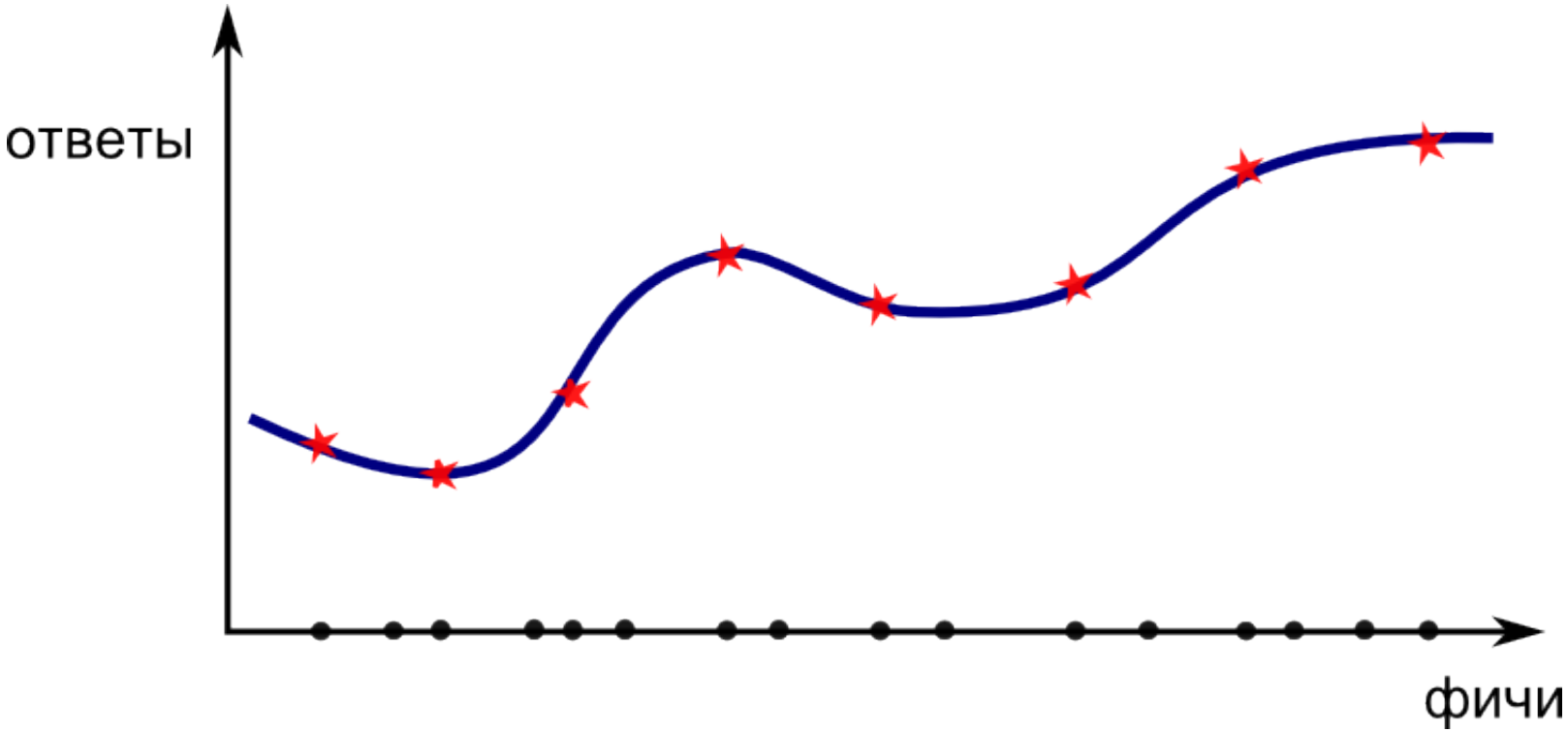
Проблемы семплирования: Переобучение



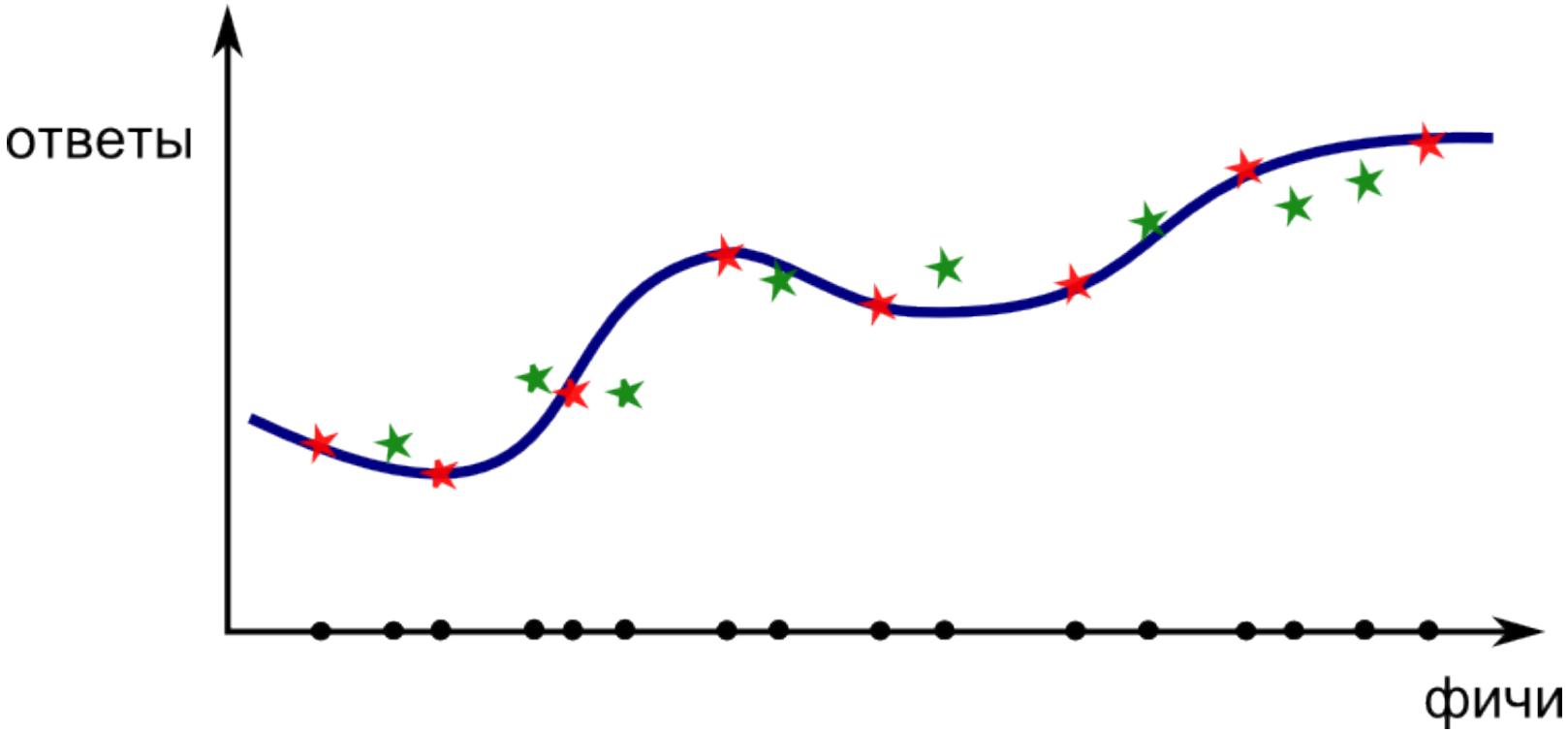
Проблемы семплирования: Переобучение



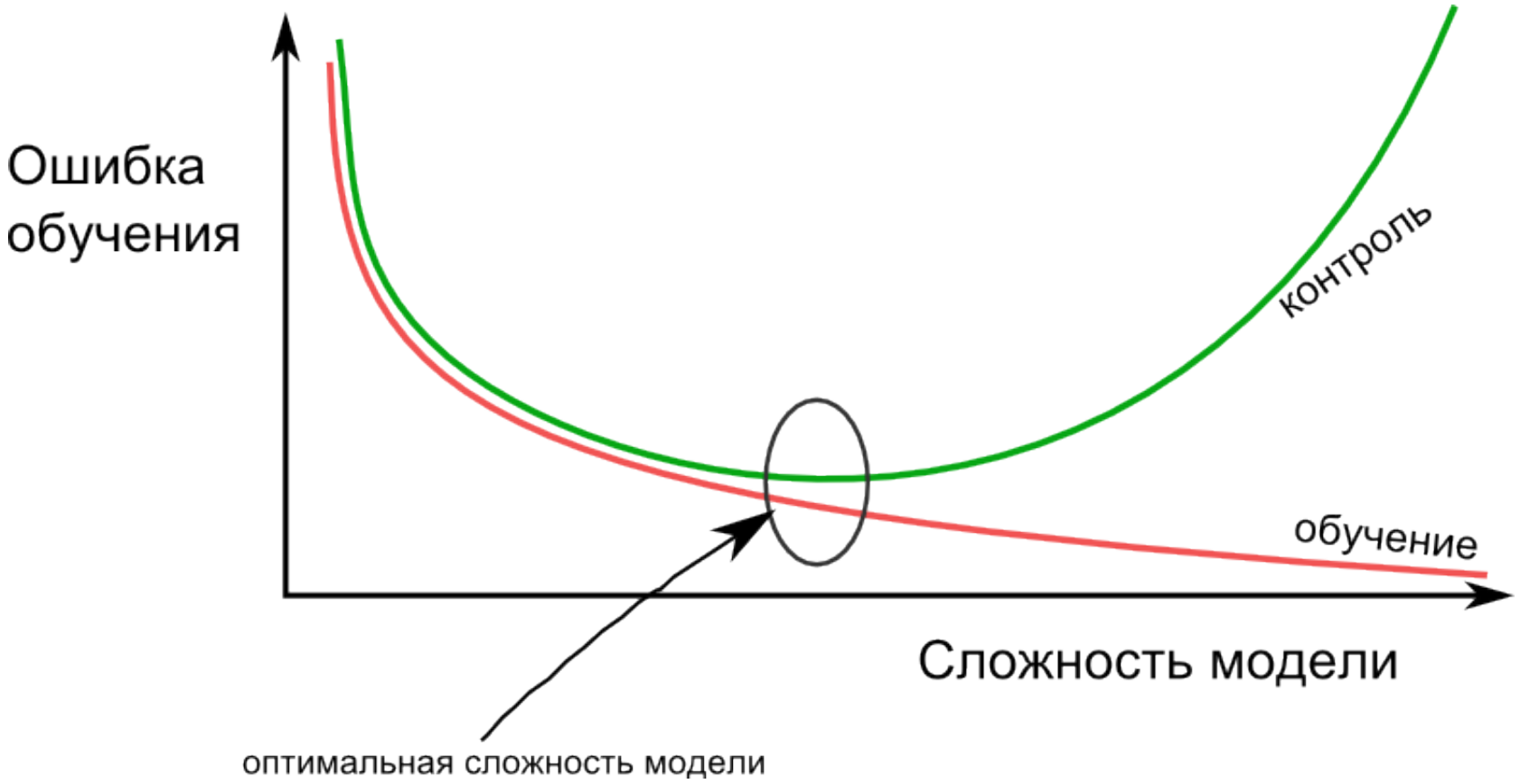
Проблемы семплирования: Переобучение



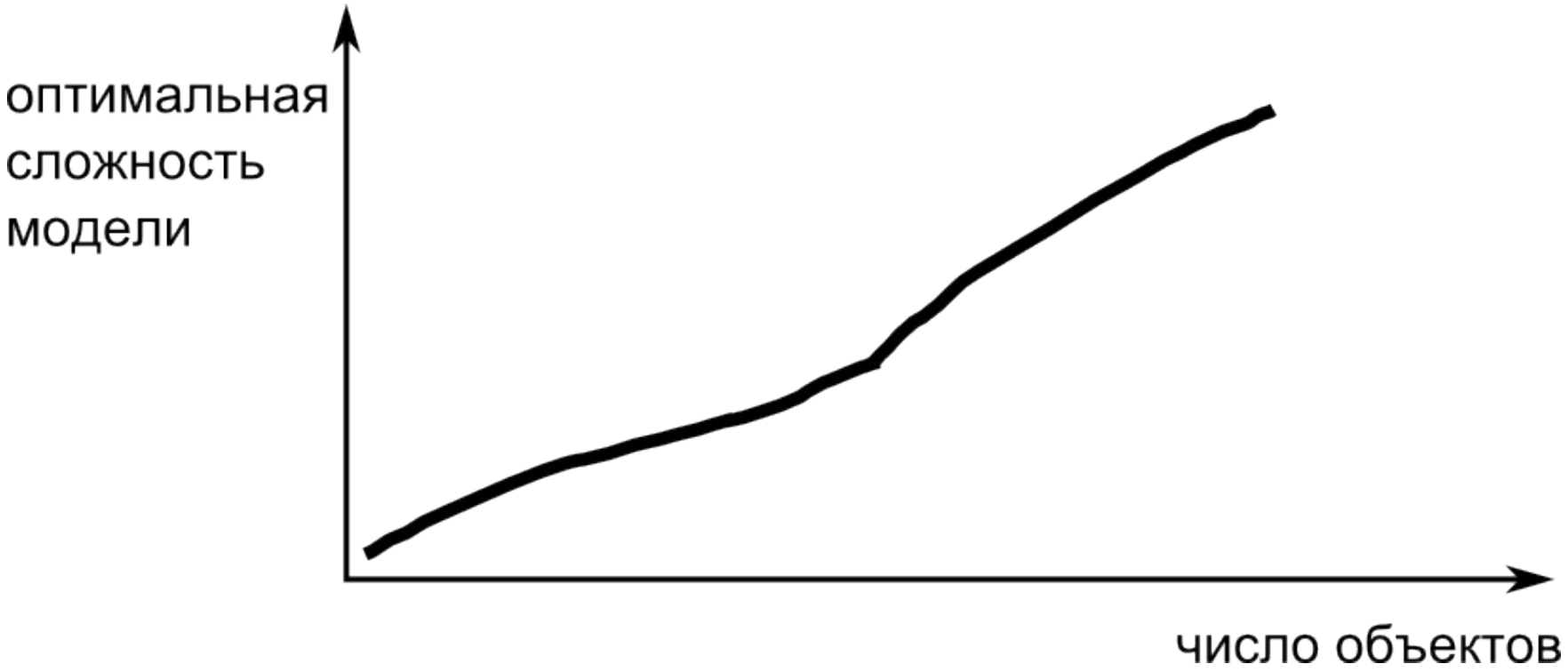
Проблемы семплирования: Переобучение



Проблемы семплирования: Переобучение



Проблемы семплирования: Переобучение



Проблемы семплирования

- Большое количество фич

Проблемы семплирования

- Большое количество фич
- Чем больше фич, тем сложнее модель

Проблемы семплирования

- Большое количество фич
- Чем больше фич, тем сложнее модель
- Чем сложнее модель, тем проще она переобучается

Проблемы семплирования

- Большое количество фич
- Чем больше фич, тем сложнее модель
- Чем сложнее модель, тем проще она переобучается

В LSML приходится иметь дело с задачами, содержащими 10^3 — 10^6 фич

Решение: LSML

- Использование редких событий и признаков

Решение: LSML

- Использование редких событий и признаков
- Использование большого числа фич

Решение: LSML

- Использование редких событий и признаков
- Использование большого числа фич
- Увеличение сложности моделей

Решение: LSML

- Способы разделения алгоритмов
  - Параллельность по данным

Решение: LSML

- Способы разделения алгоритмов
  - Параллельность по данным

    схема master-worker

Решение: LSML

- Способы разделения алгоритмов
  - Параллельность по данным

    схема master-worker

  - Паралельность по итерациям

Решение: LSML

- Способы разделения алгоритмов
  - Параллельность по данным

    схема master-worker

  - Паралельность по итерациям

    DAG

Решение: LSML

- Способы разделения алгоритмов
  - Параллельность по данным

    схема master-worker

  - Паралельность по итерациям

    DAG

- Пример: mapReduce парадигма.
- Map-only задачи — паралельны по данным
- Map + reduce — паралельность по итерациям

Распараллеливание алгоритмов

Как распараллеливаются классические алгоритмы в парадигме mapreduce

Распараллеливание можно аддитивные статистики

Пример: Naive Bayes

Признаки

$$\vec{x} = \begin{pmatrix} x^{(1)} \\ \dots \\ x^{(n)} \end{pmatrix}$$

Класс

$$y \in \{0, 1\}$$

Формула Наивного байейса:

$$P(y|x^{(1)}, \dots, x^{(n)}) = P(y)\frac{\prod_i P(x^{(i)}|y)}{C}$$

Пример: Naive Bayes

- Выражение $$P(y) = \frac{\nu(y)}{\nu(all)} = \frac{\sum_j 1(y_j = y)}{\sum_j 1}$$

Пример: Naive Bayes

- Выражение $P(y) = \frac{\nu(y)}{\nu(all)} = \frac{\sum_j 1(y_j = y)}{\sum_j 1}$

- **Map**: (y_j, Σ_{sub}) $(total, \Sigma_{sub})$

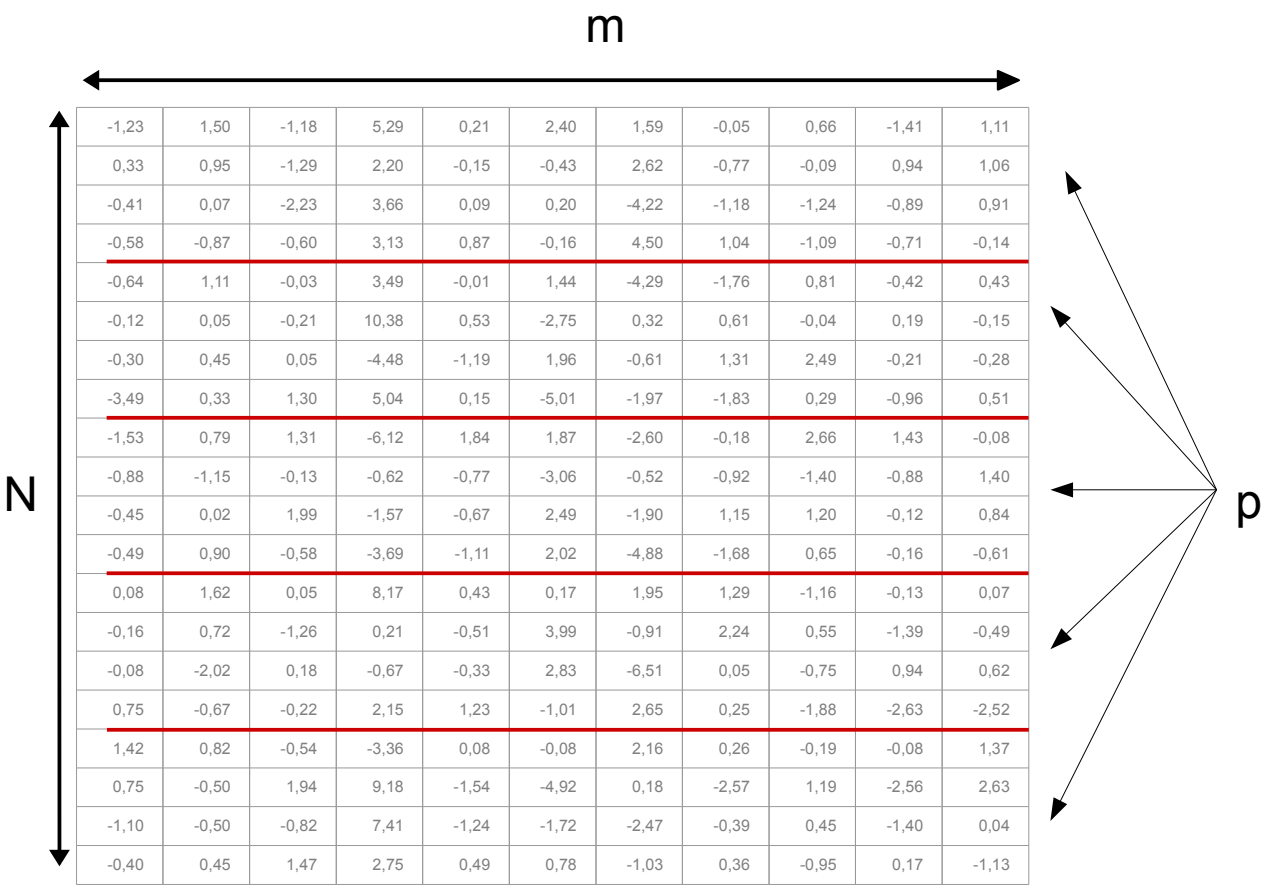
Пример: Naive Bayes

- Выражение $P(y) = \frac{\nu(y)}{\nu(all)} = \frac{\sum_j 1(y_j = y)}{\sum_j 1}$
- **Map**: (y_j, Σ_{sub}) $(total, \Sigma_{sub})$
- **Reduce**: (y_j, Σ) $(total, \Sigma)$

Пример: Naive Bayes

- Выражение $P(x^{(i)}|y) = \frac{\nu(x^{(i)}, y)}{\nu(y)}$
- **Map**: $((x^i, y_j); \Sigma_{sub})$
- **Reduce**: $((x^i, y_j); \Sigma)$

Пример: Naive Bayes. Асимптотика



-1,23	1,50	-1,18	5,29	0,21	2,40	1,59	-0,05	0,66	-1,41	1,11
0,33	0,95	-1,29	2,20	-0,15	-0,43	2,62	-0,77	-0,09	0,94	1,06
-0,41	0,07	-2,23	3,66	0,09	0,20	-4,22	-1,18	-1,24	-0,89	0,91
-0,58	-0,87	-0,60	3,13	0,87	-0,16	4,50	1,04	-1,09	-0,71	-0,14
-0,64	1,11	-0,03	3,49	-0,01	1,44	-4,29	-1,76	0,81	-0,42	0,43
-0,12	0,05	-0,21	10,38	0,53	-2,75	0,32	0,61	-0,04	0,19	-0,15
-0,30	0,45	0,05	-4,48	-1,19	1,96	-0,61	1,31	2,49	-0,21	-0,28
-3,49	0,33	1,30	5,04	0,15	-5,01	-1,97	-1,83	0,29	-0,96	0,51
-1,53	0,79	1,31	-6,12	1,84	1,87	-2,60	-0,18	2,66	1,43	-0,08
-0,88	-1,15	-0,13	-0,62	-0,77	-3,06	-0,52	-0,92	-1,40	-0,88	1,40
-0,45	0,02	1,99	-1,57	-0,67	2,49	-1,90	1,15	1,20	-0,12	0,84
-0,49	0,90	-0,58	-3,69	-1,11	2,02	-4,88	-1,68	0,65	-0,16	-0,61
0,08	1,62	0,05	8,17	0,43	0,17	1,95	1,29	-1,16	-0,13	0,07
-0,16	0,72	-1,26	0,21	-0,51	3,99	-0,91	2,24	0,55	-1,39	-0,49
-0,08	-2,02	0,18	-0,67	-0,33	2,83	-6,51	0,05	-0,75	0,94	0,62
0,75	-0,67	-0,22	2,15	1,23	-1,01	2,65	0,25	-1,88	-2,63	-2,52
1,42	0,82	-0,54	-3,36	0,08	-0,08	2,16	0,26	-0,19	-0,08	1,37
0,75	-0,50	1,94	9,18	-1,54	-4,92	0,18	-2,57	1,19	-2,56	2,63
-1,10	-0,50	-0,82	7,41	-1,24	-1,72	-2,47	-0,39	0,45	-1,40	0,04
-0,40	0,45	1,47	2,75	0,49	0,78	-1,03	0,36	-0,95	0,17	-1,13

Классический алгоритм

$$O(Nm+mc)$$

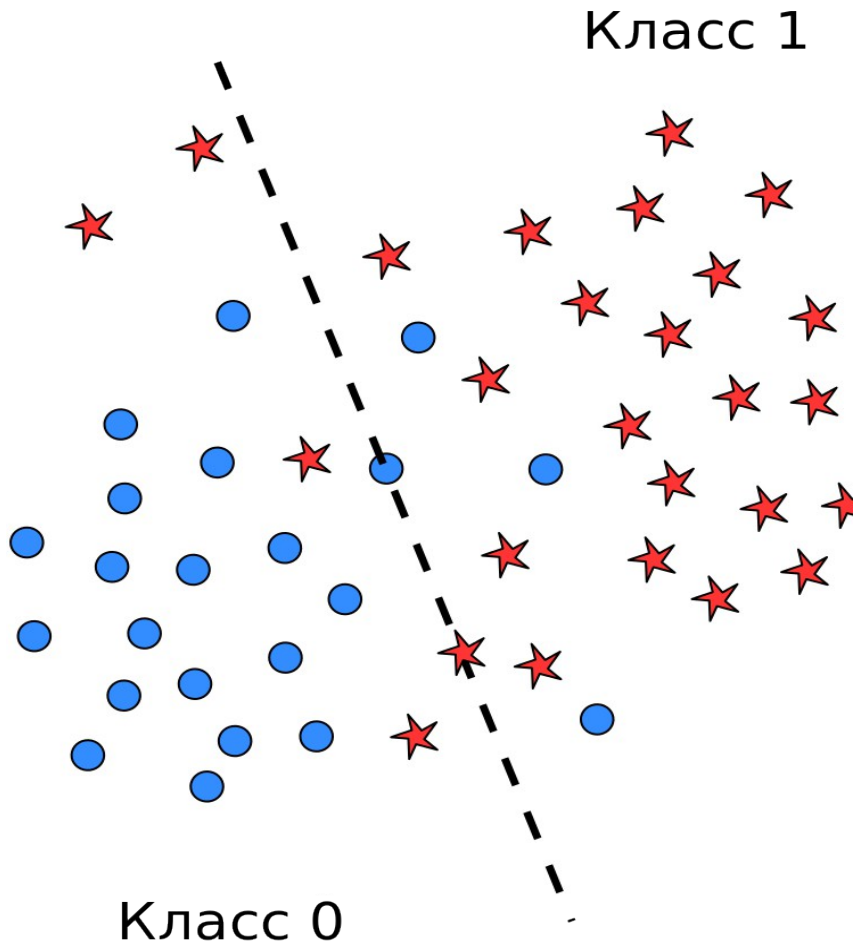
MapReduce алгоритм

$$O\left(\frac{Nm}{P}+mc\log P\right)$$

Пример: Логистическая регрессия

$$w^T \cdot x = 0$$

$$P_w(x) = \frac{1}{1+\exp(-w \cdot x)}$$



$$l(w) = \sum_{i=1}^{n} y_i \log p(x_i) + (1-y_i) \log(1-p(x_i))$$

Пример: Логистическая регрессия

$$w = argmin_w l(w)$$

Метод Ньютона-Рафсона $$w = w - H^{-1} \nabla_w l(w)$$

$$\nabla_w l(w) = \begin{pmatrix} \frac{\partial l(w)}{\partial w_1} \\ \dots \\ \frac{\partial l(w)}{\partial w_n} \end{pmatrix} \qquad H = \begin{pmatrix} \frac{\partial^2 l(w)}{\partial w_1 \partial w_1} & \dots & \frac{\partial^2 l(w)}{\partial w_1 \partial w_n} \\ \dots & \dots & \dots \\ \frac{\partial^2 l(w)}{\partial w_n \partial w_1} & \dots & \frac{\partial^2 l(w)}{\partial w_n \partial w_n} \end{pmatrix}$$

Пример: Логистическая регрессия

- Выражение $\frac{\partial l(w)}{\partial w_k} = \sum_{i=1}^{m} (y_i - p_w(x_i)) x_i^{(k)}$

- **Map**: $grad_{sub}[k] = \sum_i (y - p_w(x_i)) x^{(k_i)}$

  $(k, grad_{sub}[k])$

- **Reduce**: $grad[k] = \sum grad_{sub}[k]$

  $(k, grad[k])$

Пример: Логистическая регрессия

- Выражение

$$\frac{\partial^2 l(w)}{\partial w_k \partial w_j} = \sum_{i=1}^{m} p_w(x_i)\left(p_w(x_i) - 1\right) x_i^{(j)} x_i^{(k)}$$

- **Map**:

$$H_{sub}[k,j] = \sum_{sub} \ldots.$$

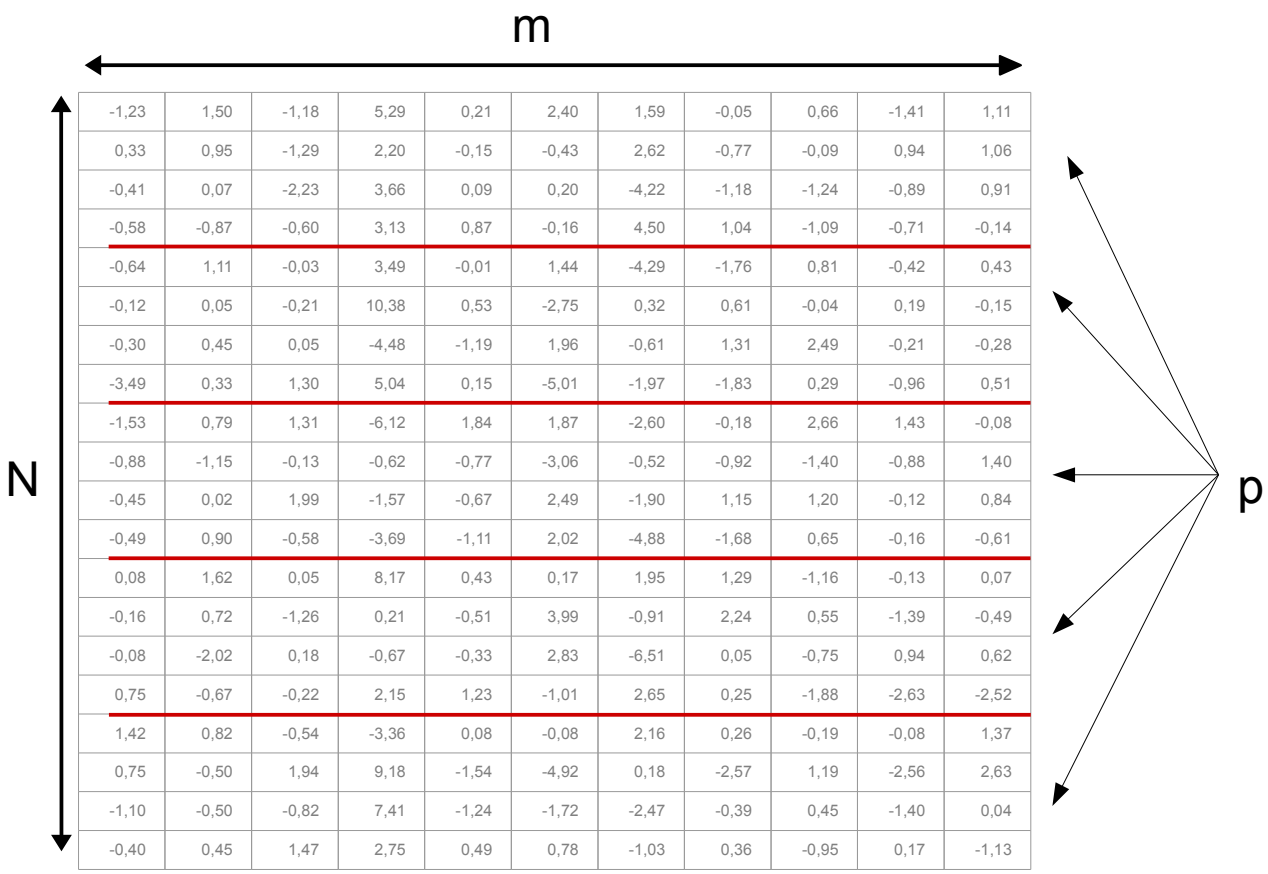
$$\left((k,j); H_{sub}[k,j]\right)$$

- **Reduce**:

$$H[k,j] = H_{sub}[k,j]$$

$$\left((k,j); H[k,j]\right)$$

Логистическая регрессия. Асимптотика



-1,23	1,50	-1,18	5,29	0,21	2,40	1,59	-0,05	0,66	-1,41	1,11
0,33	0,95	-1,29	2,20	-0,15	-0,43	2,62	-0,77	-0,09	0,94	1,06
-0,41	0,07	-2,23	3,66	0,09	0,20	-4,22	-1,18	-1,24	-0,89	0,91
-0,58	-0,87	-0,60	3,13	0,87	-0,16	4,50	1,04	-1,09	-0,71	-0,14
-0,64	1,11	-0,03	3,49	-0,01	1,44	-4,29	-1,76	0,81	-0,42	0,43
-0,12	0,05	-0,21	10,38	0,53	-2,75	0,32	0,61	-0,04	0,19	-0,15
-0,30	0,45	0,05	-4,48	-1,19	1,96	-0,61	1,31	2,49	-0,21	-0,28
-3,49	0,33	1,30	5,04	0,15	-5,01	-1,97	-1,83	0,29	-0,96	0,51
-1,53	0,79	1,31	-6,12	1,84	1,87	-2,60	-0,18	2,66	1,43	-0,08
-0,88	-1,15	-0,13	-0,62	-0,77	-3,06	-0,52	-0,92	-1,40	-0,88	1,40
-0,45	0,02	1,99	-1,57	-0,67	2,49	-1,90	1,15	1,20	-0,12	0,84
-0,49	0,90	-0,58	-3,69	-1,11	2,02	-4,88	-1,68	0,65	-0,16	-0,61
0,08	1,62	0,05	8,17	0,43	0,17	1,95	1,29	-1,16	-0,13	0,07
-0,16	0,72	-1,26	0,21	-0,51	3,99	-0,91	2,24	0,55	-1,39	-0,49
-0,08	-2,02	0,18	-0,67	-0,33	2,83	-6,51	0,05	-0,75	0,94	0,62
0,75	-0,67	-0,22	2,15	1,23	-1,01	2,65	0,25	-1,88	-2,63	-2,52
1,42	0,82	-0,54	-3,36	0,08	-0,08	2,16	0,26	-0,19	-0,08	1,37
0,75	-0,50	1,94	9,18	-1,54	-4,92	0,18	-2,57	1,19	-2,56	2,63
-1,10	-0,50	-0,82	7,41	-1,24	-1,72	-2,47	-0,39	0,45	-1,40	0,04
-0,40	0,45	1,47	2,75	0,49	0,78	-1,03	0,36	-0,95	0,17	-1,13

Классический алгоритм

$$O\left(Nm^2+m^3\right)$$

MapReduce алгоритм

$$O\left(\frac{Nm^2}{P}+\frac{m^3}{P}+m^2\log P\right)$$

Асимптотика

Алгоритм	**Классическая сложность**	**Сложность MapReduce**
Наивный Байес	$O(Nm+mc)$	$O(\frac{Nm}{P}+mc\log P)$
K-средних	$O(Nmc)$	$O(\frac{Nmc}{P}+Nm\log P)$
Логистическая регрессия	$O(Nm^2+m^3)$	$O(\frac{N m^2}{P}+\frac{m^3}{P}+m^2\log P)$
GDA	$O(Nm^2+m^3)$	$O(\frac{N m^2}{P}+\frac{m^3}{P}+m^2\log P)$
SVM	$O(N^2 m)$	$O(\frac{N^2 m}{P}+m\log P)$

Выводы

- LSML может дать профит!

Выводы

- LSML может дать профит!
- N и m велико - требуется модификация алгоритмов