курс «Машинное обучение»

Функции ошибки / функционалы качества Часть 1: Задачи регрессии

Александр Дьяконов

16 февраля 2021 года

План на несколько лекций

задача регрессии

задача бинарной классификации
- чёткая классификация
- скоринговые функции

задача классификации с несколькими классами

задачи ранжирования

задачи кластеризации

Задача – ДНК

Дано

Найти

Критерий

Построить алгоритм легко!

Чтобы улучшить… надо уметь оценивать.

~~Метрики~~

- **функции ошибки**
- **функционалы качества**

Функции ошибки / функционалы качества

Пожалуй, самое главное при решении задачи…
иногда важнее данных!

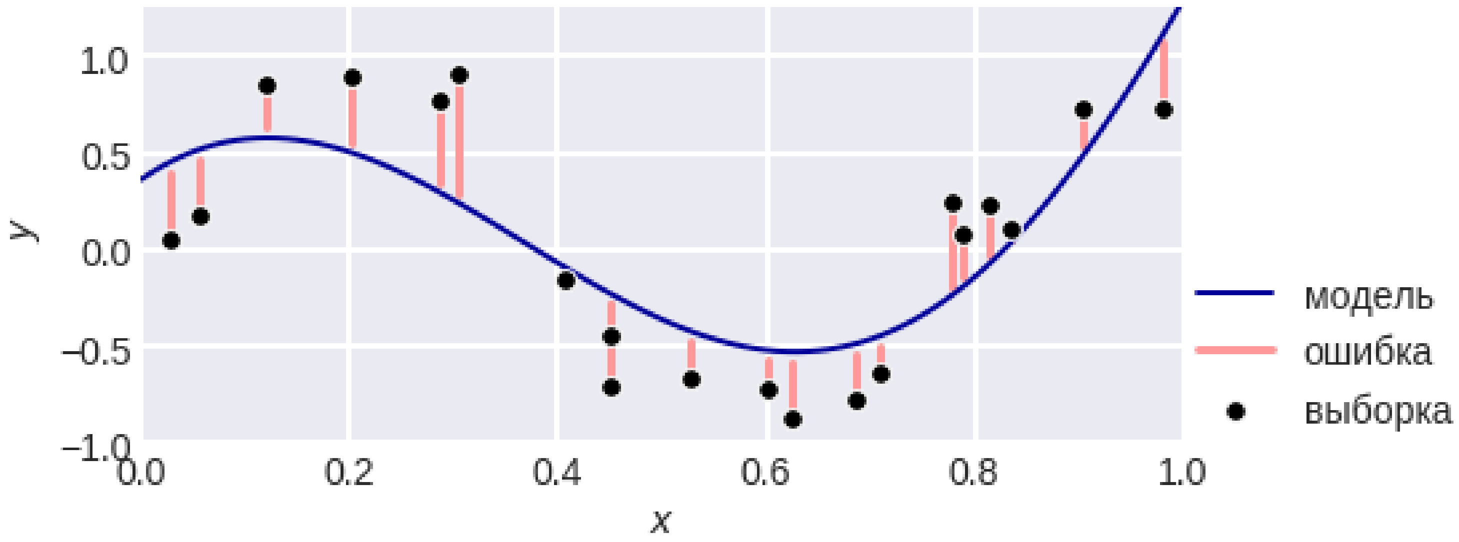
а что такое решение!

В анализе данных:

- **формализация ответа (формат)**
- **как ответ оценивается (критерий качества)**

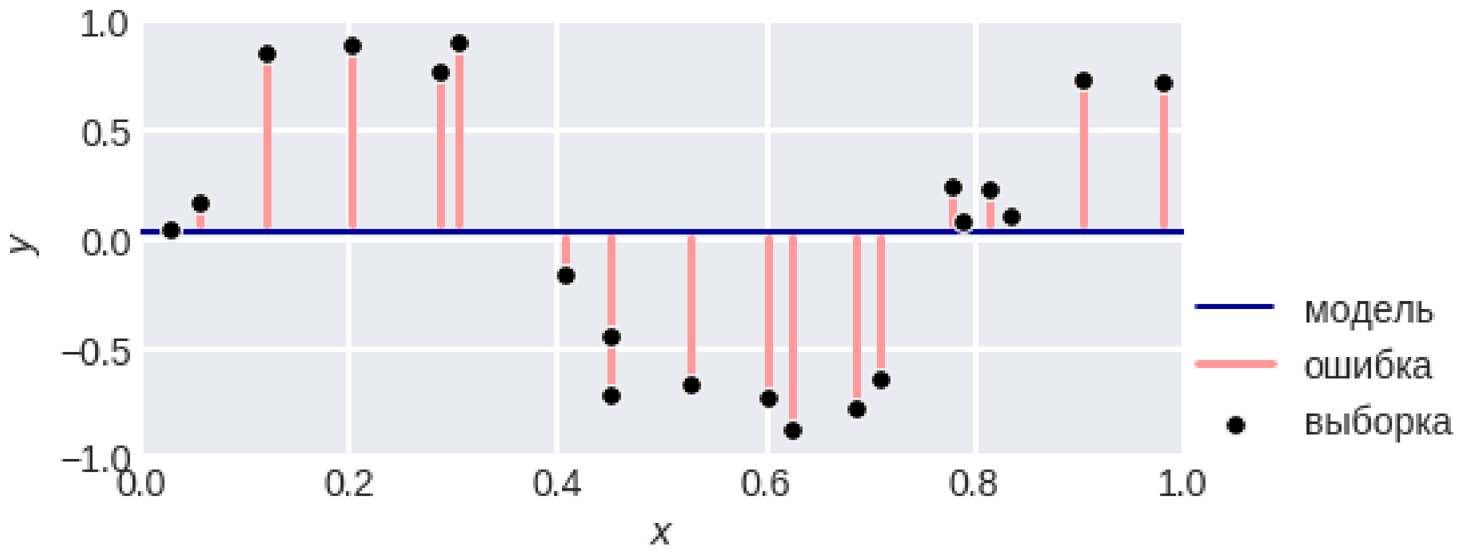
Случай из практики: задача про траектории зрачка
(задача с 3 классами, а не с двумя)

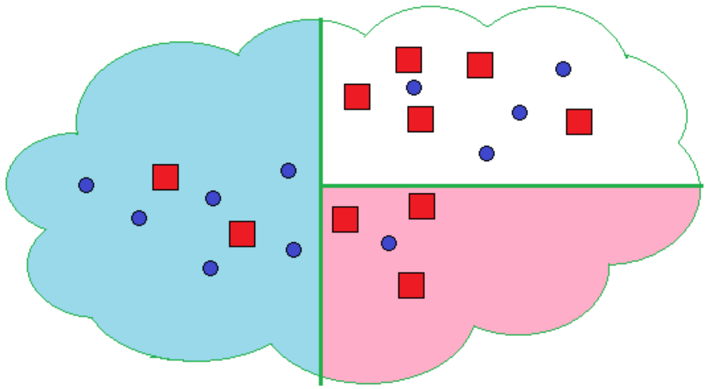
Задача регрессии



Задача регрессии

Будем дальше пытаться всё решать в классе констант



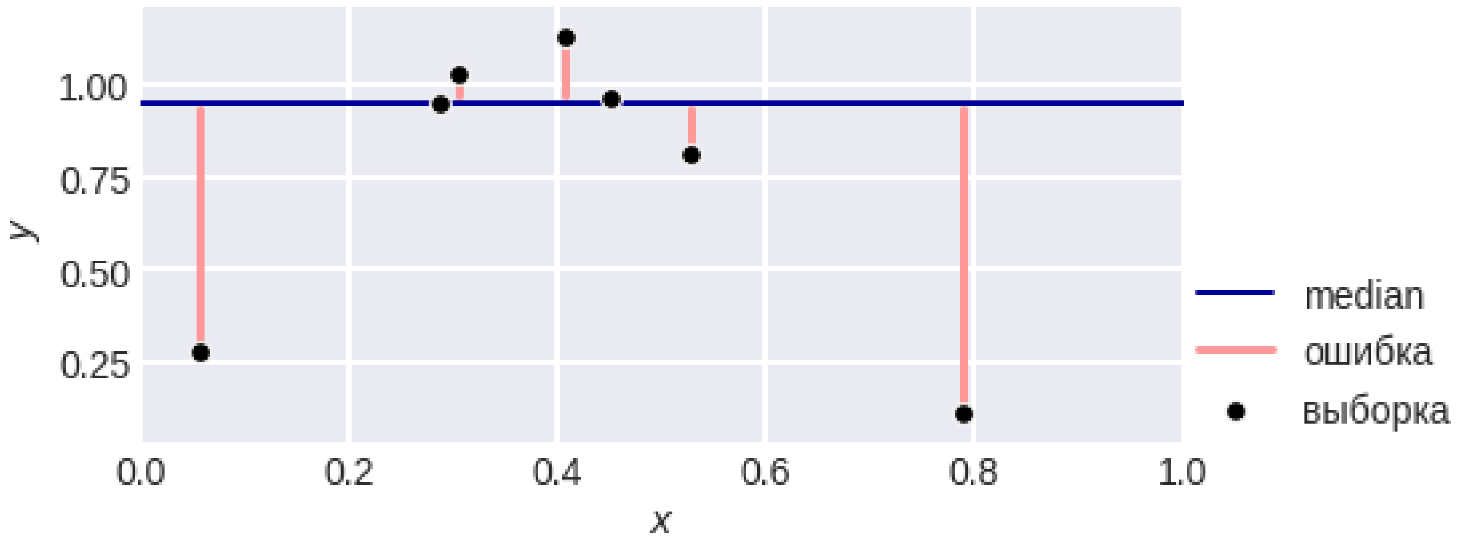


1. Простейшее решение

2. Примерно это и происходит в листьях решающих деревьев

3. Раскрывает природу функционалов

Средний модуль отклонения – Mean Absolute Error (MAE), Mean Absolute Deviation (MAD)



$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m} | a_i - y_i |$$

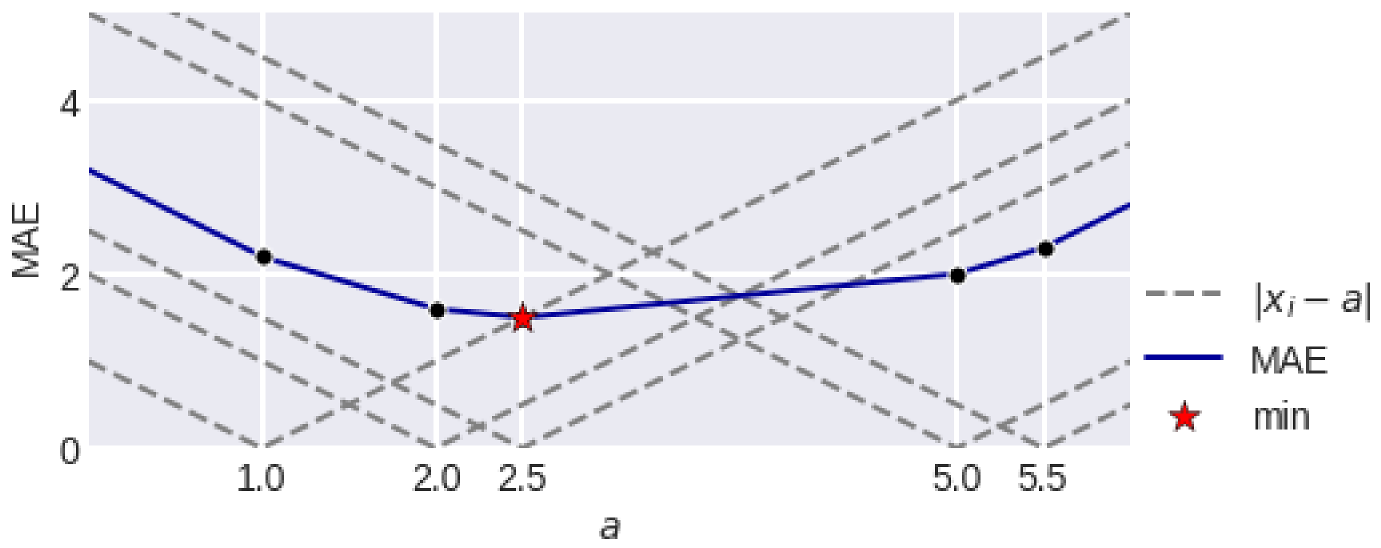
Напоминание:

$$\frac{1}{m}\sum_{i=1}^{m} | a - y_i | \to \min$$

$$a = \mathrm{median}(\{y_i\}_{i=1}^{m})$$

Это открывает смысл решений!

Средний модуль отклонения



Средний модуль отклонения

Способы использования тайных знаний:

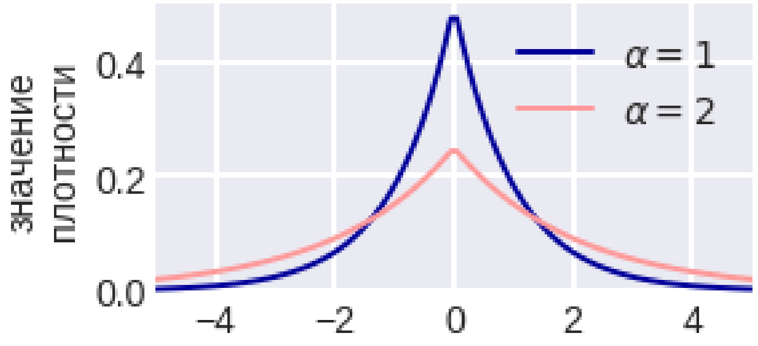
- медиана, вместо усреднения, в ансамбле
- округление ответа (если целевой вектор целочисленный)

Откуда берётся MAE

$$y = a_w(x) + \varepsilon$$

w – **параметры алгоритма** $a_w(x)$

$$\varepsilon \sim \mathrm{laplace}(0, \alpha)$$



Для оценки параметров выписываем правдоподобие модели

$$p(y \mid x, w) = \frac{\alpha}{2}\exp\left[-\alpha \mid y - a_w(x) \mid\right]$$

Метод максимального правдоподобия:

$$\log L(w) = \log \prod_{i=1}^{m} p(y_i \mid x_i, w) =$$

$$= \sum_{i=1}^{m}\left[\log\frac{\alpha}{2} - \alpha \mid y_i - a_w(x_i) \mid\right] \to \max$$

Откуда берётся MAE

Получаем

$$\alpha \sum_{i=1}^{m} | y_i - a_w(x_i) | \to \min$$

т.е. задачу минимизации MAE!

- **не зависит от природы модели**
- **зависит от распределения ошибок**

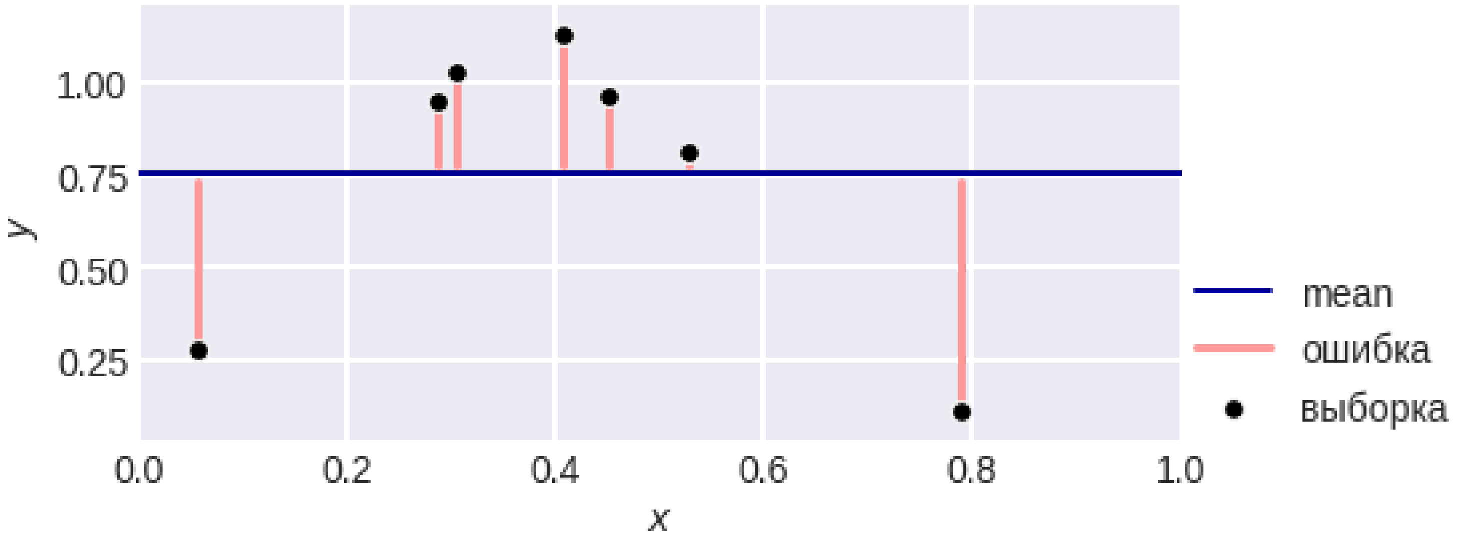
(почему Residual Plots)

Максимизация правдоподобия эквивалентна минимизации MAE!

Чему соответствует минимизация весового MAE?

Средний квадрат отклонения ~ Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m} | a_i - y_i |^2$$



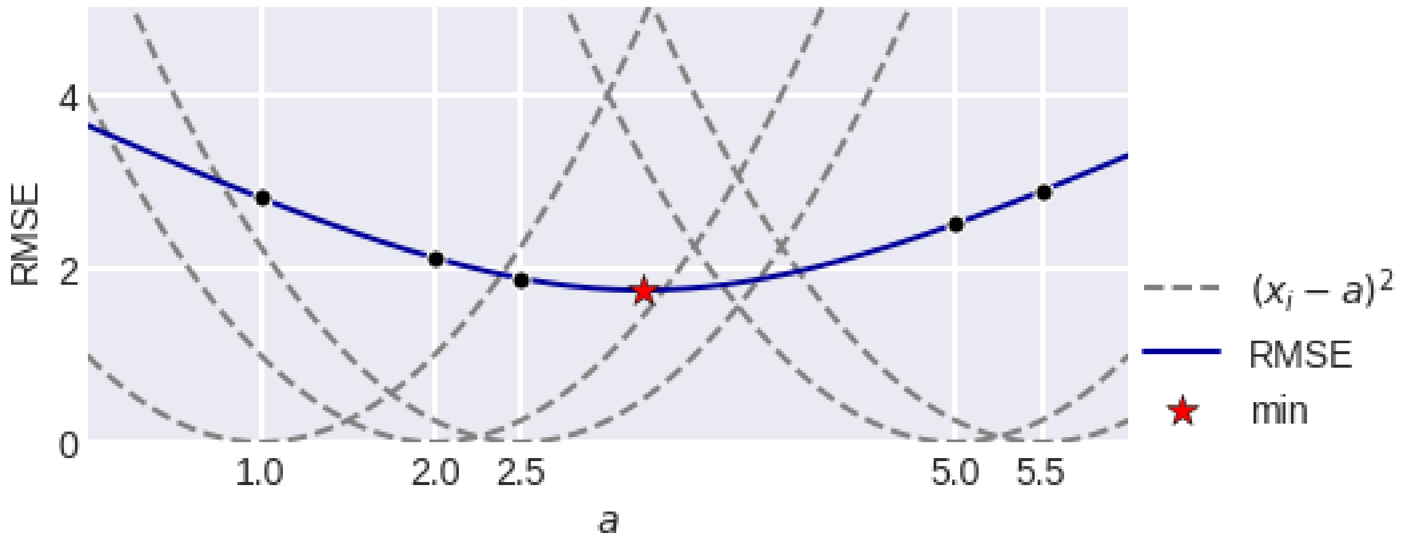
$$\frac{1}{m}\sum_{i=1}^{m} | a - y_i |^2 \to \min$$

$$a = \frac{1}{m}\sum_{i=1}^{m} y_i$$

Root Mean Squared Error (RMSE) / Root Mean Square Deviation (RMSD)

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} | a_i - y_i |^2}$$

Средний квадрат отклонения ~ Mean Squared Error (MSE)



Способы использования тайных знаний

- ничего не делать (в RF, GBM и т.д. всё равно усредняют)
- метод НСКО – классическая регрессия!

Нормированная версия: коэффициент детерминации R^2 (Coefficient of Determination)

$$R^2 = 1 - \frac{\sum_{i=1}^{m} |a_i - y_i|^2}{\sum_{i=1}^{m} |\bar{y} - y_i|^2}$$

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

В общем случае (в статистике) коэффициент детерминации:

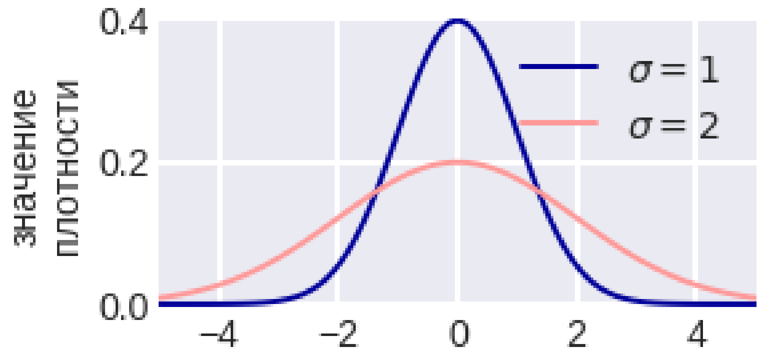
$$R^2 = 1 - \frac{\mathbf{D}(y\,|\,x)}{\mathbf{D}(y)}$$

Откуда берётся (R)MSE

$$y = a_w(x) + \varepsilon$$

w – **параметры алгоритма** $a_w(x)$

$$\varepsilon \sim \mathrm{norm}(0, \sigma^2)$$



Для оценки параметров выписываем правдоподобие модели

$$p(y \mid x, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y - a_w(x))^2}{2\sigma^2}\right]$$

Метод максимального правдоподобия:

$$\log L(w) = \log \prod_{i=1}^{m} p(y_i \mid x_i, w) =$$

$$= \sum_{i=1}^{m} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - a_w(x_i))^2}{2\sigma^2}\right] \to \max$$

Откуда берётся (R)MSE

Получаем

$$\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - a_w(x_i))^2 \to \min$$

т.е. задачу минимизации MSE!

- **не зависит от природы модели**
- **зависит от распределения ошибок**

(почему Residual Plots)

Максимизация правдоподобия эквивалентна минимизации среднеквадратичной ошибки!

Откуда берётся (R)MSE: ещё одно «оправдание»

Пусть функция ошибки $l(y,a)=g(y-a)$

Что логично?

1. $g(0)=0$

2. $|z_1|\le|z_2|\Rightarrow g(z_1)\le g(z_2)$

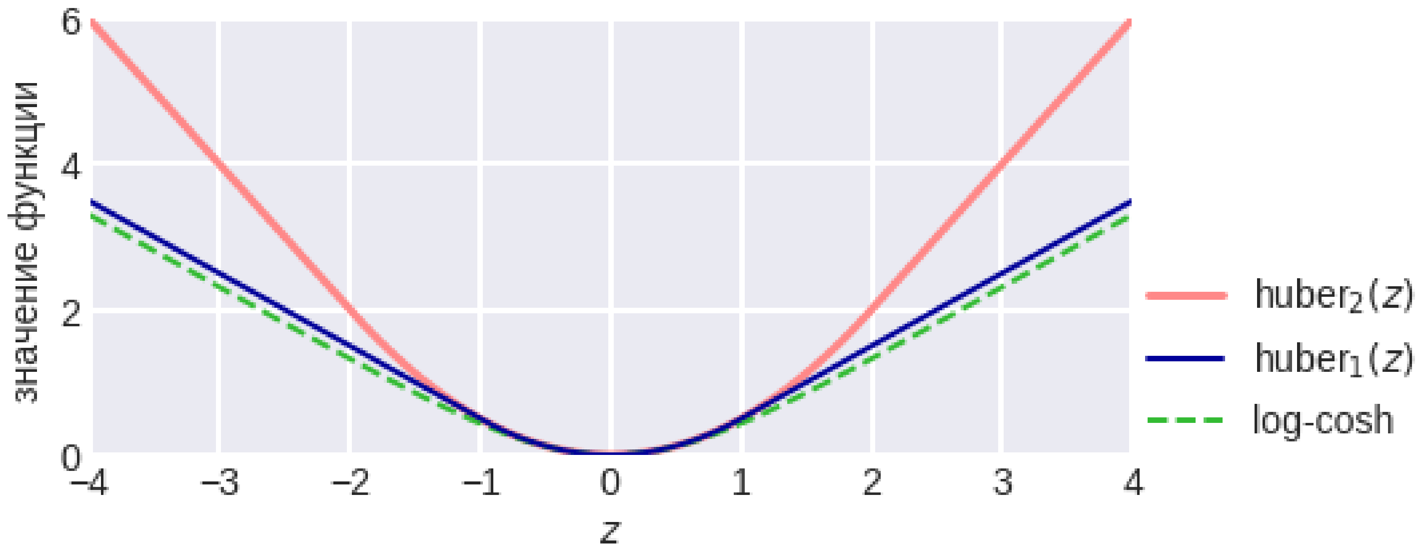
3. достаточно гладкая...

$$g(z)=g(0)+g'(0)z+\frac{g''(0)}{2}z^2+o(z^2)$$

но тогда

$$l(y,a)=g(y-a)\approx \underset{=0\,(1)}{g(0)}+\underbrace{g'(0)(y-a)}_{=0\,(2)}+\frac{g''(0)}{2}(y-a)^2=\underset{>0}{C}(y-a)^2$$

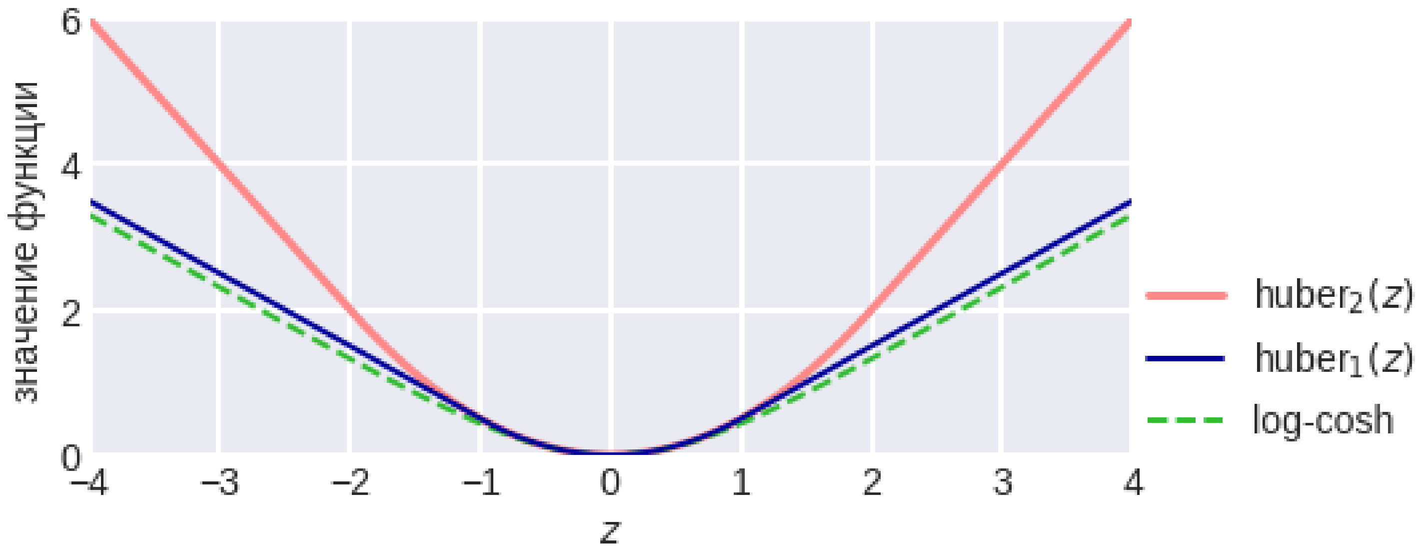
Функция Хьюбера и logcosh



$$\mathrm{huber}_\delta(z) = \begin{cases} \dfrac{1}{2}z^2, & |z| \le \delta, \\ \delta\left(|z| - \dfrac{1}{2}\delta^2\right), & |z| > \delta. \end{cases}$$

Как только что вывели:
когда отклонение мало – ошибка квадратичная
когда велико (в т.ч. выбросы) – линейная

Функция Хьюбера и logcosh

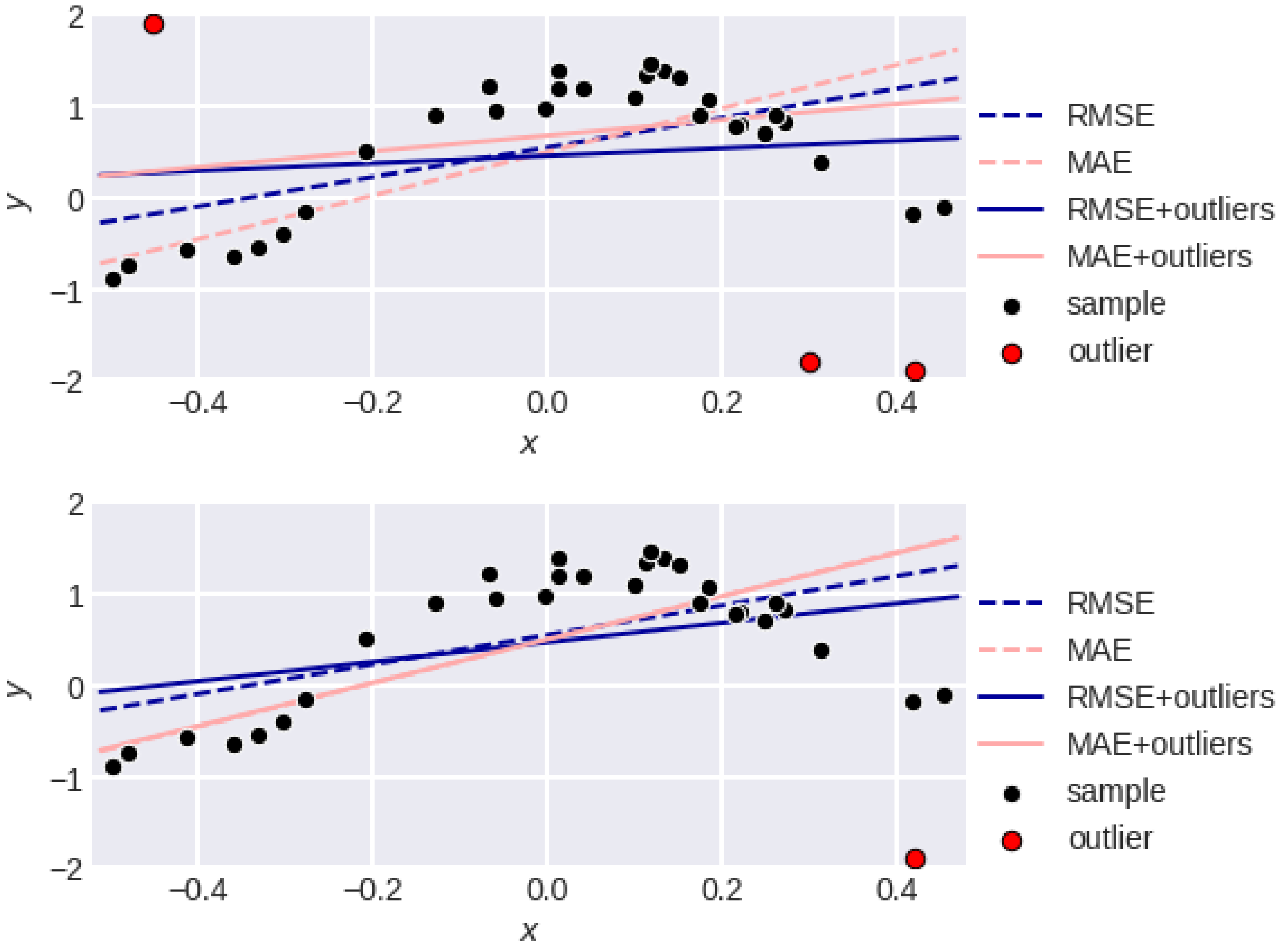


$$\text{logcosh} = \log\left(\frac{\exp(z)+\exp(-z)}{2}\right)$$

непараметрическая,
но используется редко.

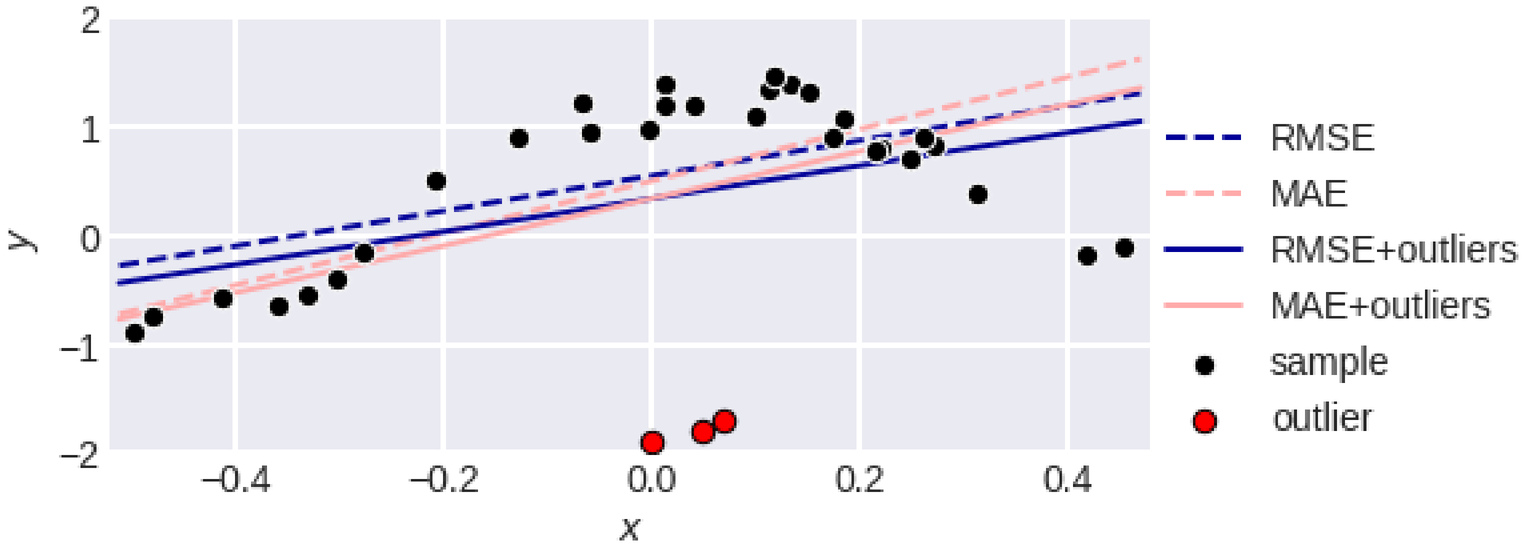
Различия MSE и MAE

Устойчивость к выбросам…



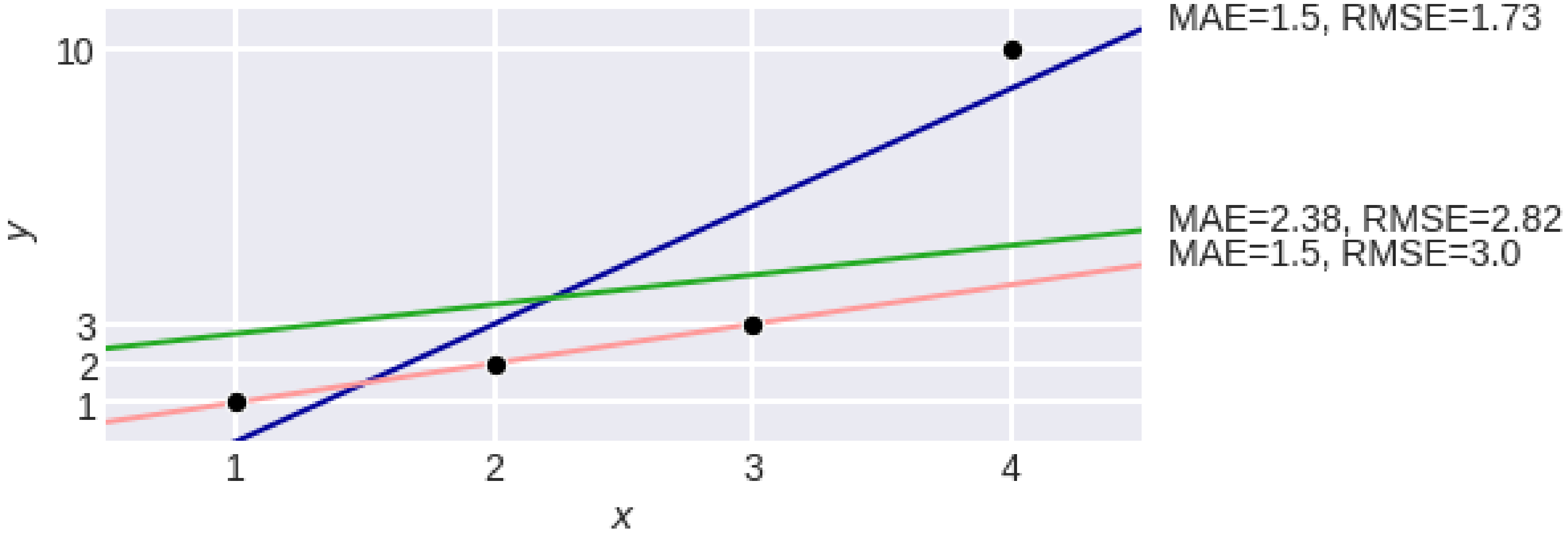
Различия MSE и MAE

Устойчивость к выбросам…



Считается, что MAE устойчивее к выбросам…

Различия MSE и MAE

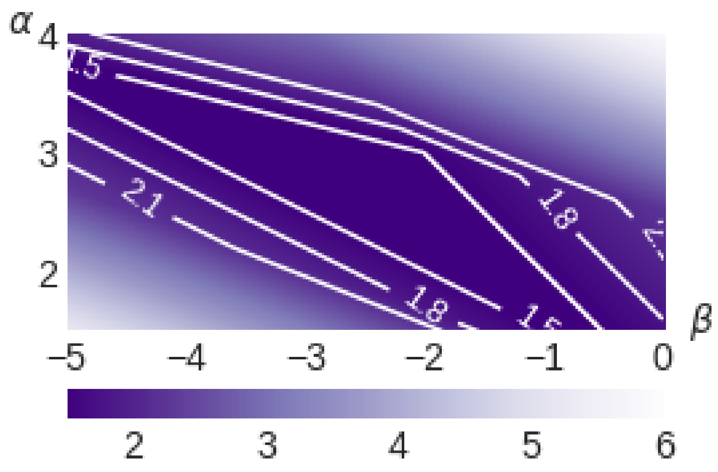


посмотрим на неконстантное решение:

$$\sum_{i=1}^{m} | y_i - a(x_i) |^p \to \min,$$

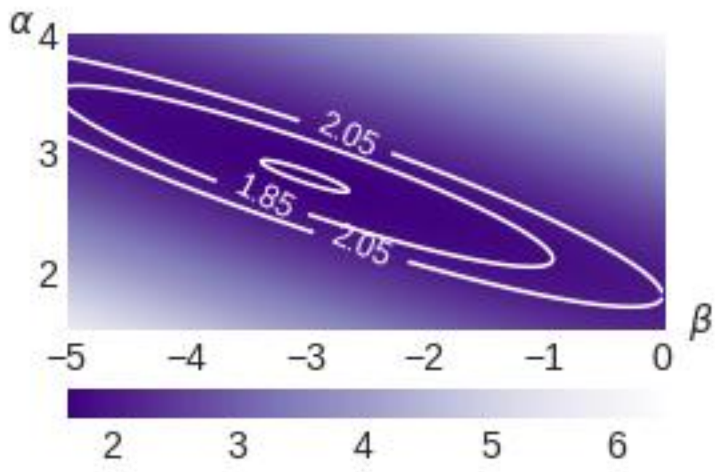
$$a(x) = \alpha x + \beta$$

MAE

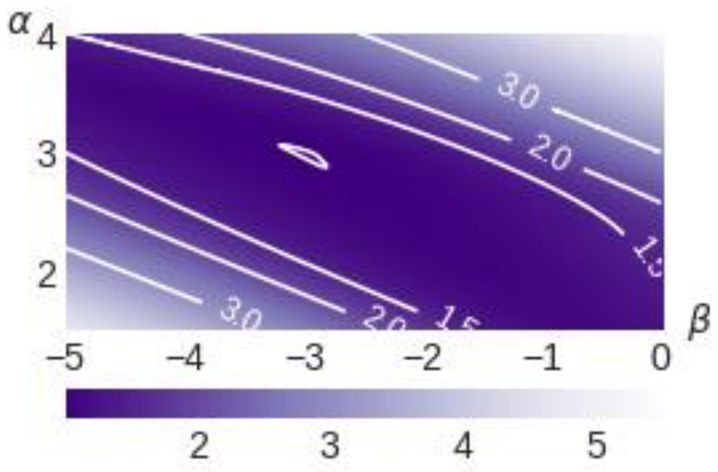


внутри «треугольника» одинаковый MAE=1.5

RMSE



Huber $(\delta = 1)$



Различия MSE и MAE

внутри «треугольника» одинаковый MAE=1.5

можно привести примеры, когда MAE меняется слабо, а RMSE значительно

Обобщения

$$\sqrt[p]{\frac{1}{m}\sum_{i=1}^{m} w_i \,|\, \varphi(a_i) - \varphi(y_i) \,|^p} \qquad \sum_{i=1}^{m} w_i = 1, \quad w_1, \ldots, w_m \geq 0$$

Рецепты

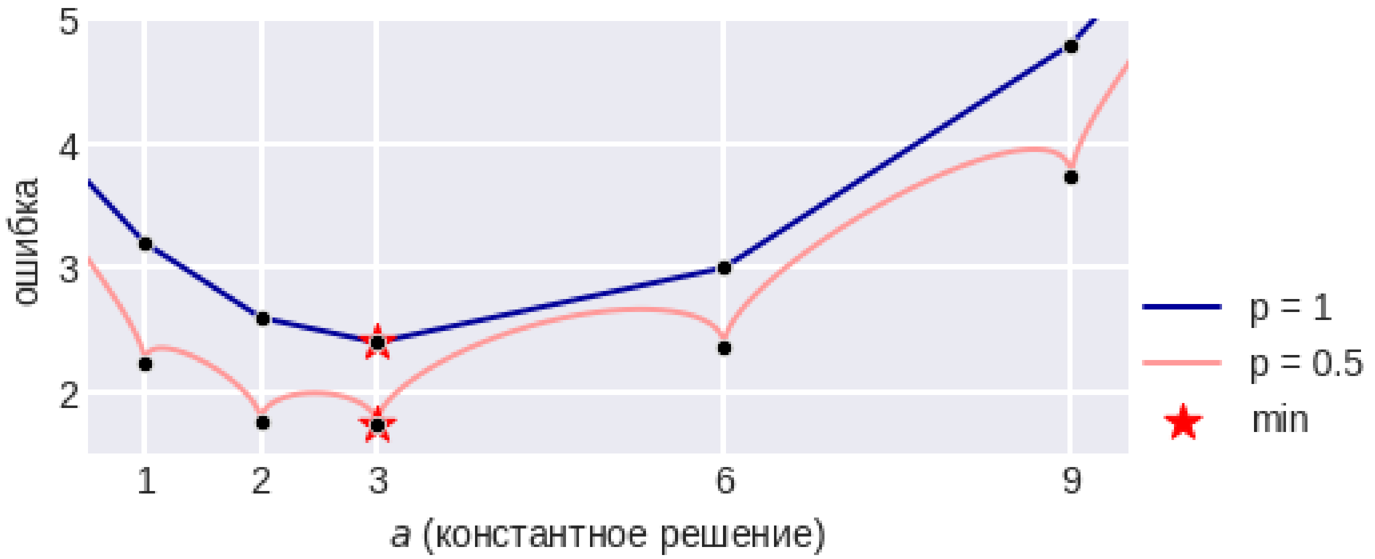
1. Преобразование целевого вектора $\varphi(y)$

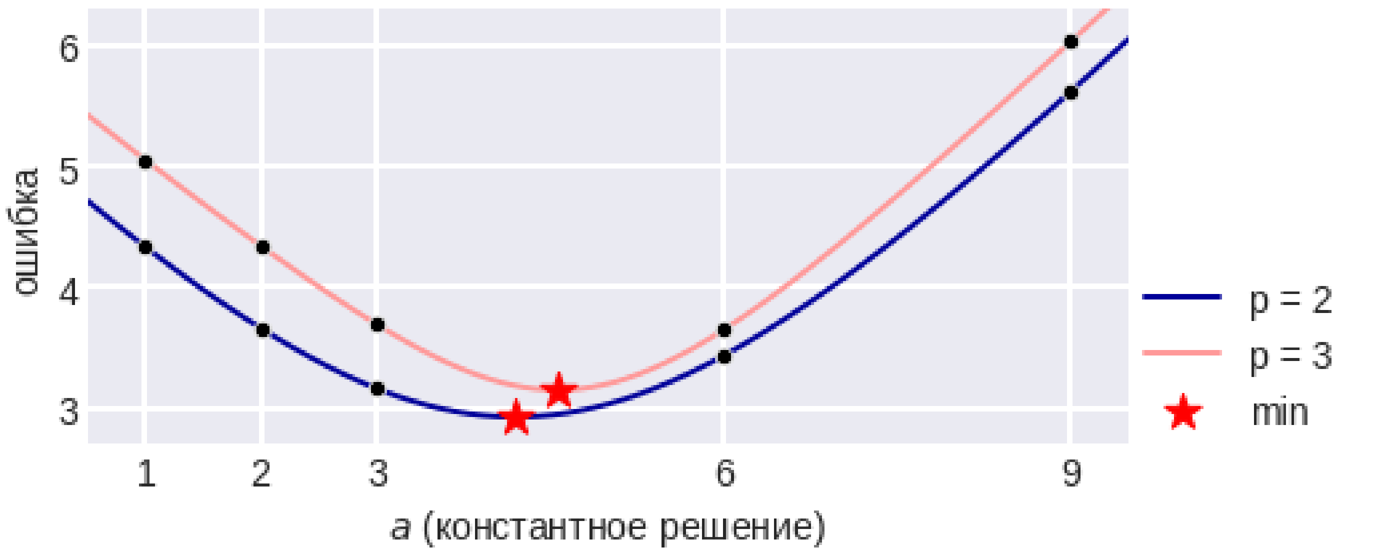
2. Веса ~ вероятности появления объектов в сэмплировании
Некоторые модели поддерживают веса объектов

3. В случае нетривиальных p – прямая настройка

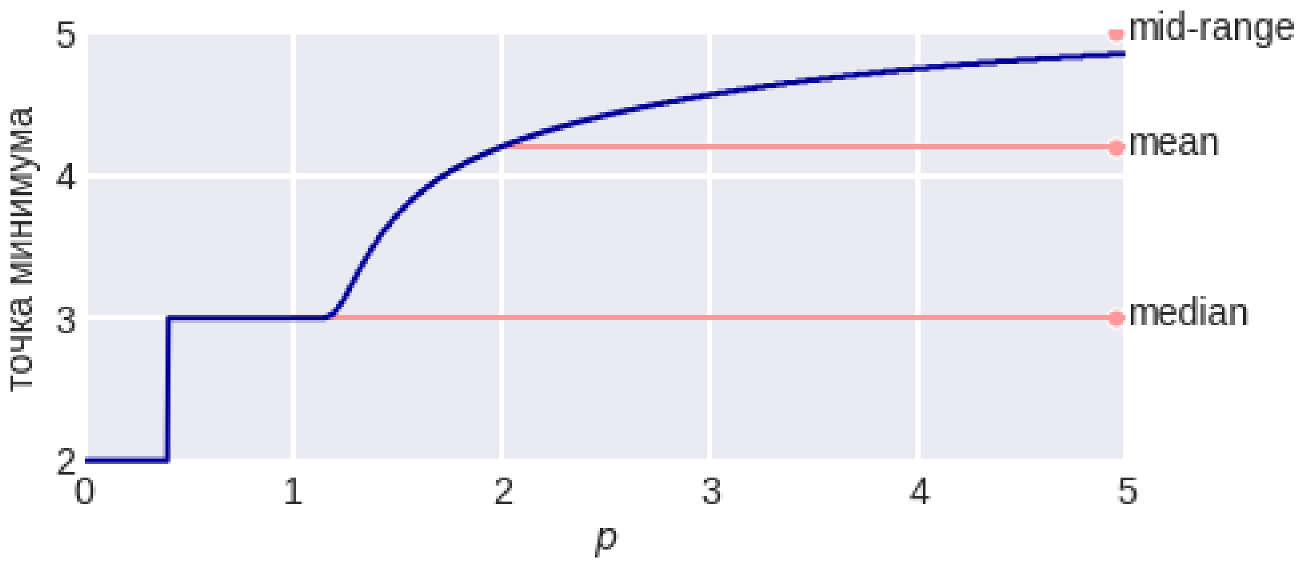
Дальше к этому вернёмся...

Про нетривиальные p





Как точка минимума зависит от степени



Разные постановки задачи регрессии

$$\| Xw - y \|_p \to \min$$

$p < 1$	**Это не норма, NP-сложная задача**
$p = 1$	**Линейное программирование**
$1 < p < 2$	**Нет стандартных методов**
$p = 2$	**Аналитическое решение (линейная алгебра)**
$p > 2$	**Градиентные методы**
$p = \infty$	**Линейное программирование**

Symmetric mean absolute percentage error (SMAPE or sMAPE)

$$\mathrm{SMAPE} = \frac{2}{m}\sum_{i=1}^{m}\frac{|y_i - a_i|}{y_i + a_i} = 100\% \cdot \frac{1}{m}\sum_{i=1}^{m}\frac{|y_i - a_i|}{(y_i + a_i)/2}$$

Когда надо интерпретировать погрешность как проценты
– плохо, если есть нули (и отрицательные значения)

1 – 2
SMAPE = 67%

100 – 101
SMAPE = 1%

0 – 1
SMAPE = 200%

Начальники не знают, что такое проценты…

Применение SMAPE – прогноз временных рядов

Mean Absolute Percent Error (MAPE)

$$\mathrm{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\frac{|y_i - a_i|}{|y_i|}$$

Чем MAPE явно лучше SMAPE на практике?

Mean Absolute Percent Error (MAPE)

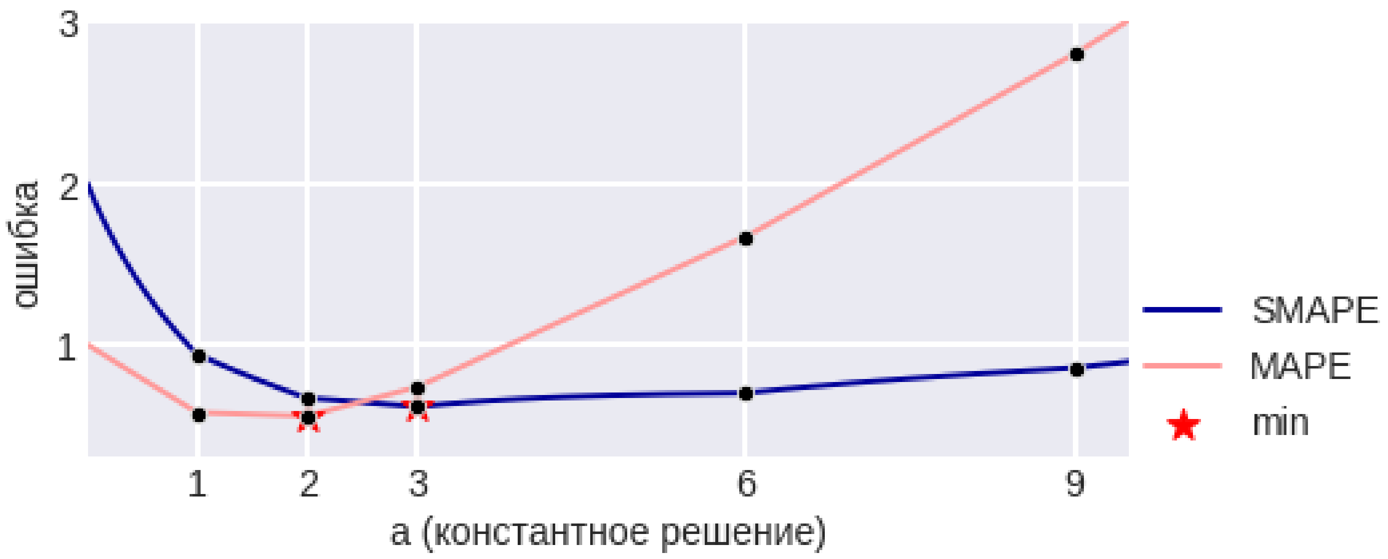
$$\mathrm{MAPE} = \frac{1}{m}\sum_{i=1}^{m} w_i \, | y_i - a_i |$$

$$w_i = \frac{1}{| y_i |}$$

Просто весовой MAE!

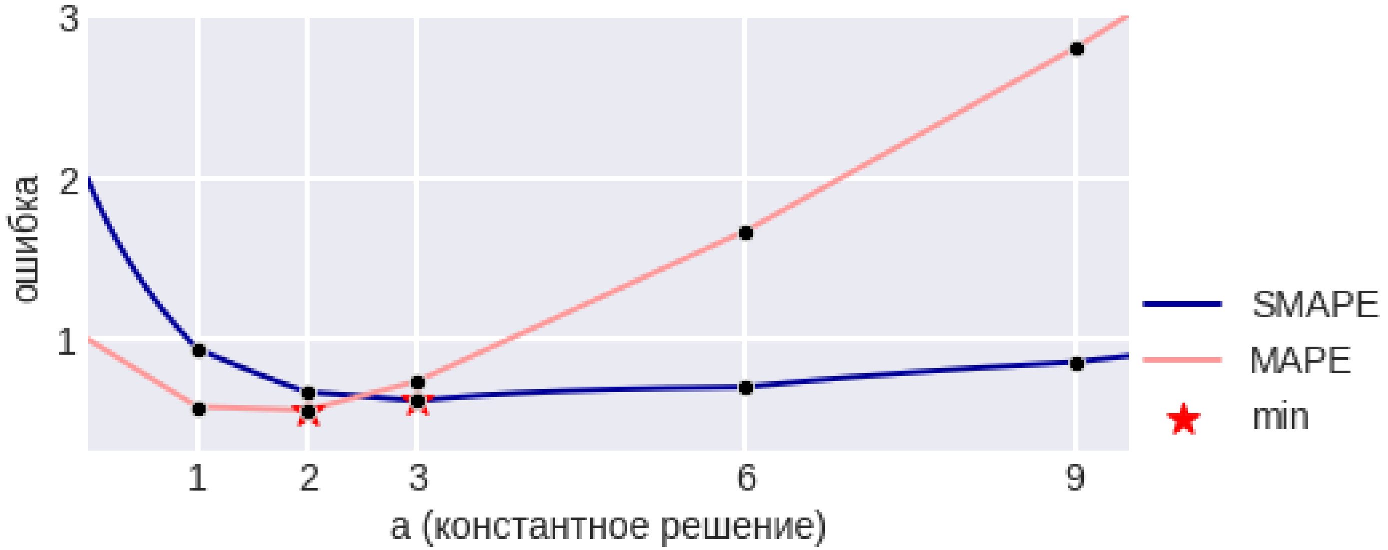
как оптимизировать? дальше...

MAPE и SMAPE

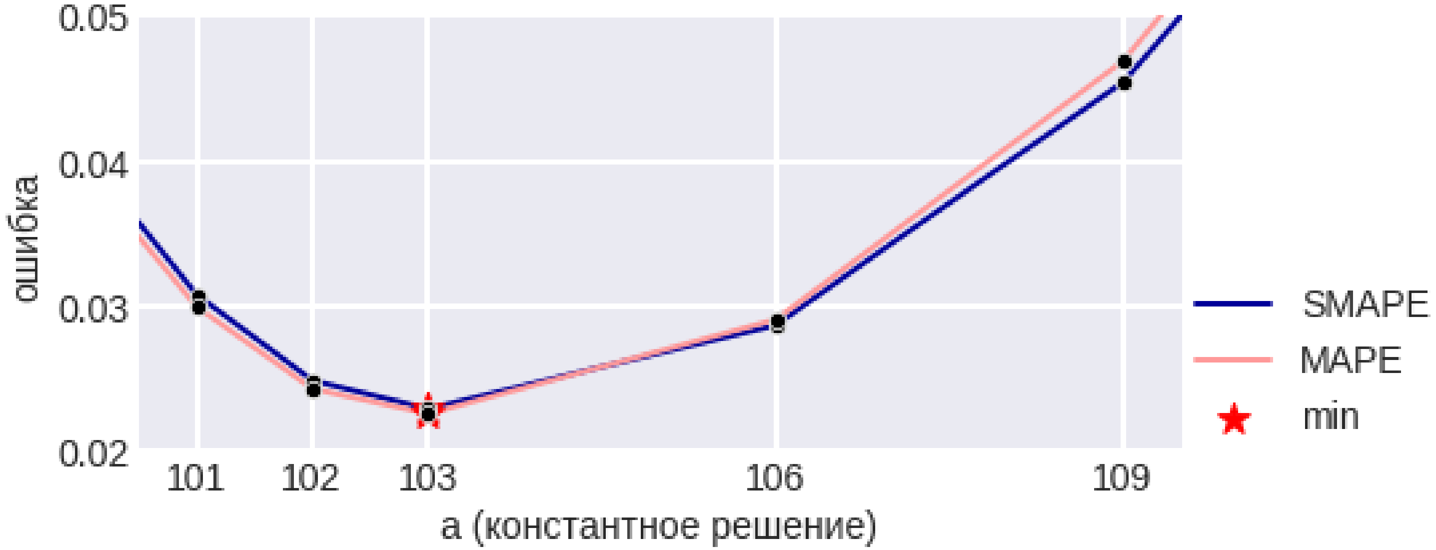


Что настораживает в этом графике?

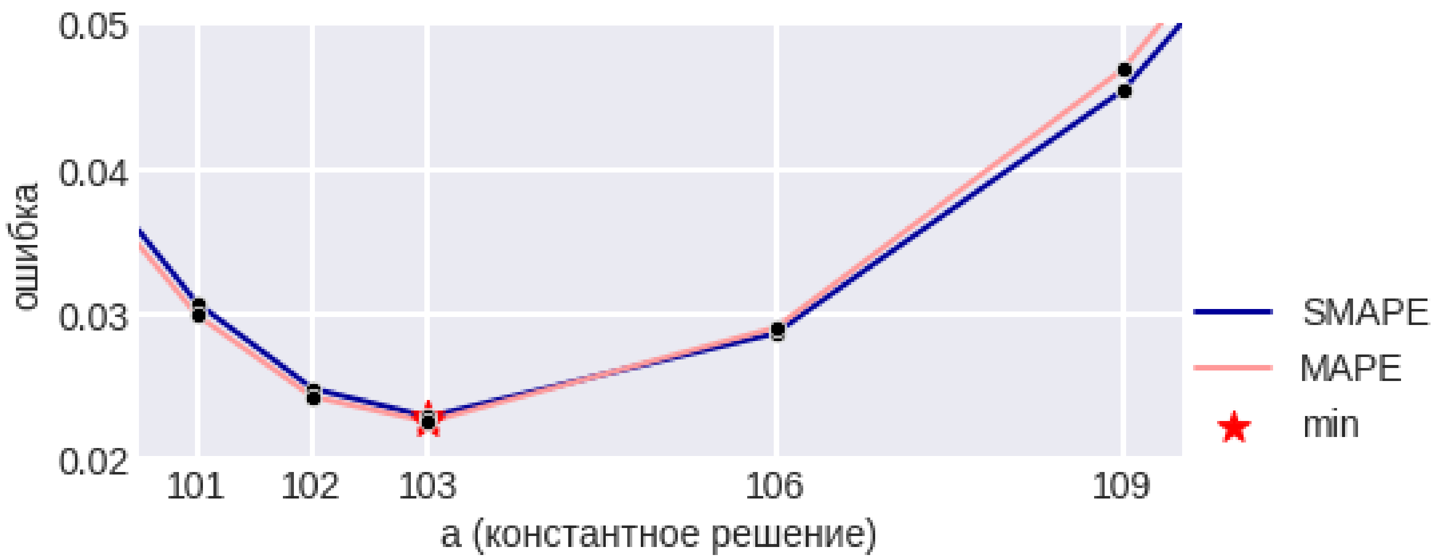
MAPE и SMAPE



Масштаб! Типичная ошибка (и во многих курсах)



MAPE и SMAPE



Например, MAPE – весовой MAE,
но на практике веса не сильно отличаются!

Поэтому решение около медианы

PMAD

Другой способ нормировки ошибки…

$$\mathrm{PMAD} = \frac{\sum_{i=1}^{m} | y_i - a_i |}{\sum_{i=1}^{m} | y_i |}$$

эквивалентен MAE

Меры на сравнении с бенчмарком

Классная идея:
сделать простой алгоритм и смотреть ошибку относительно него

Mean Relative Absolute Error (MRAE)

$$\mathrm{MRAE} = \frac{1}{m}\sum_{i=1}^{m}\frac{|y_i - a_i|}{|y_i - a'_i|}$$

REL_MAE

$$\mathrm{REL\_MAE} = \frac{\sum_{i=1}^{m}|y_i - a_i|}{\sum_{i=1}^{m}|y_i - a'_i|}$$

Percent Better

$$\mathrm{PB(MAE)} = \frac{1}{m}\sum_{i=1}^{m} I[|y_i - a_i| < |y_i - a'_i|]$$

Меры на сравнении с бенчмарком

Как выбрать бенчмарк в задачах прогнозирования?

Нормированные ошибки

Не зависят от шкалы…

Mean Absolute Scaled Error

$$\mathrm{MASE} = \frac{1}{\frac{m}{m-1}\sum_{i=2}^{m} | y_i - y_{i-1} |} \sum_{t=1}^{m} | y_i - a_i |$$

Какие ещё бывают функционалы в регрессии?

С точностью до порога

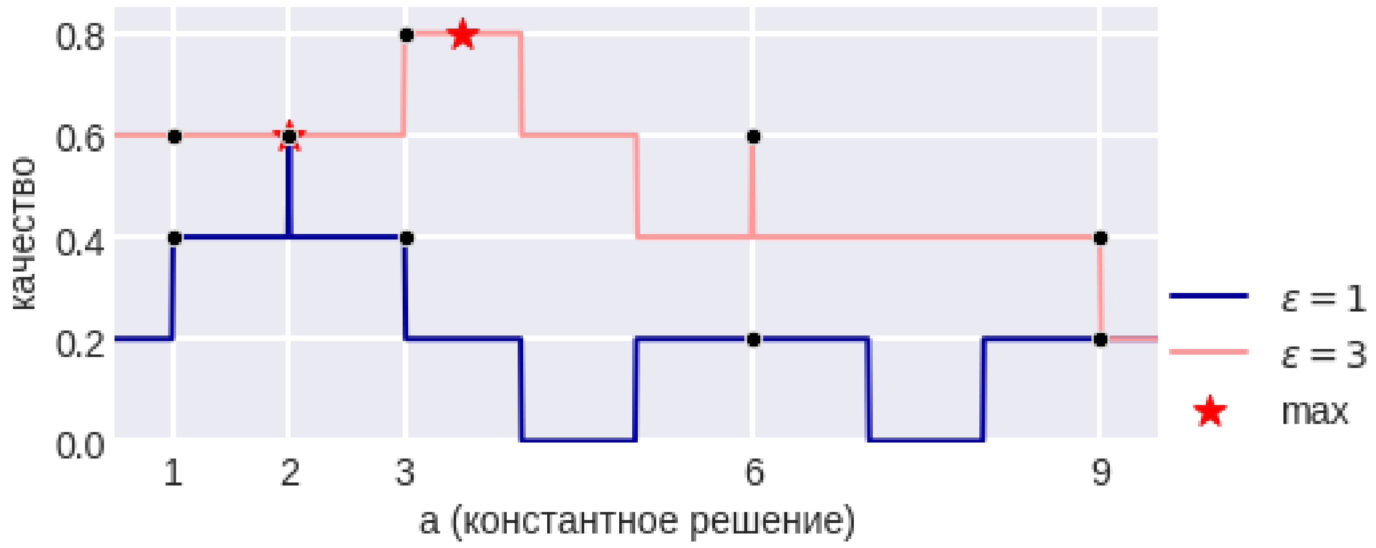
функция ошибки

$$\mathrm{eB} = \frac{1}{m}\sum_{i=1}^{m} I[|\,y_i - a_i\,| > \varepsilon]$$

функционал качества

$$\mathrm{eB} = \frac{1}{m}\sum_{i=1}^{m} I[|\,y_i - a_i\,| < \varepsilon]$$

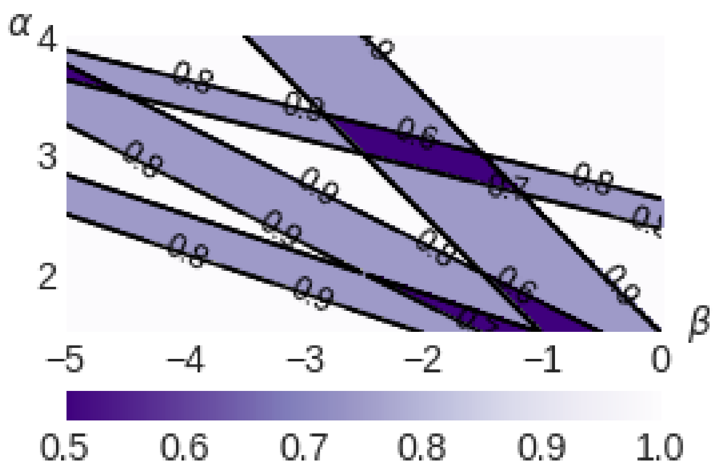
был в задаче Dunnhumby



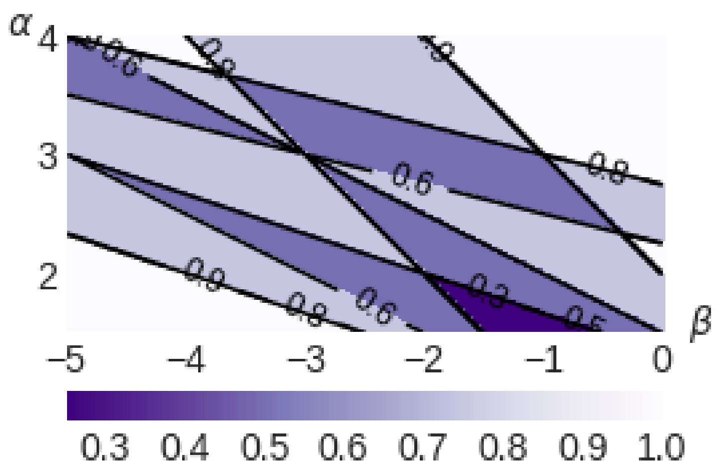
Оптимальное решение – мода парзеновской плотности

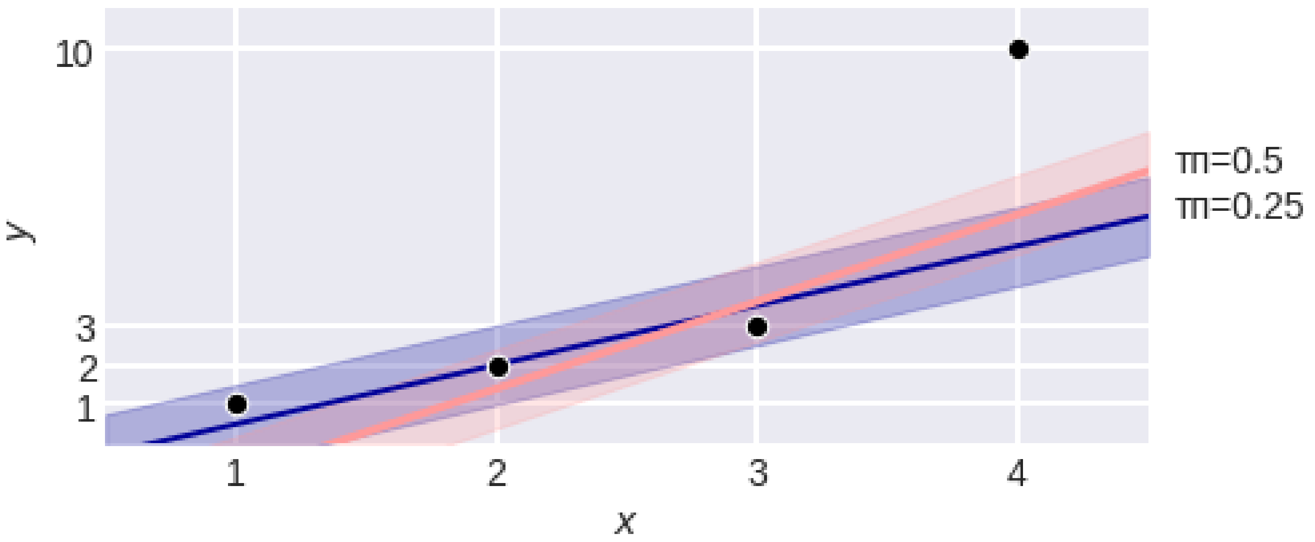
С точностью до порога

$\varepsilon = 0.5$



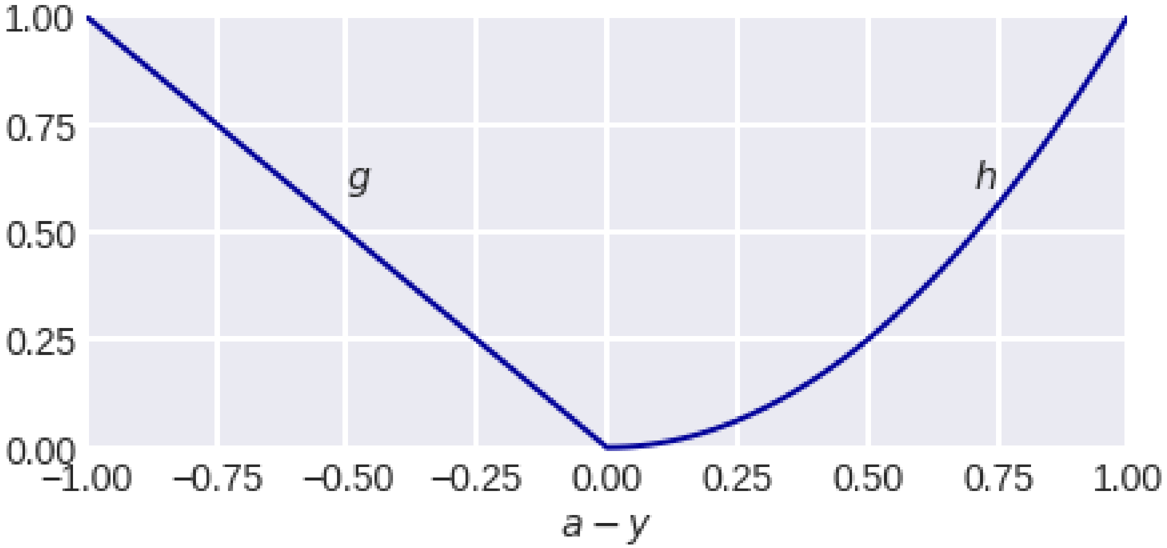
$\varepsilon = 1.0$





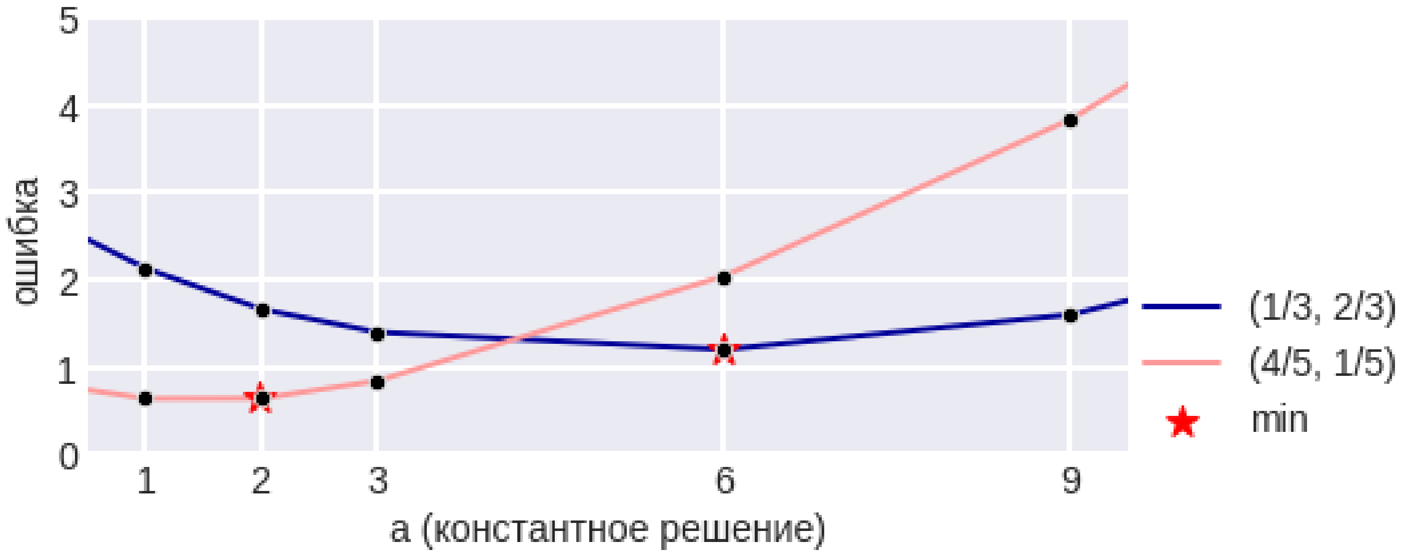
Несимметричные функции потерь

$$\frac{1}{m}\sum_{i=1}^{m}\begin{cases} g(|y_i - a_i|), & y_i < a_i, \\ h(|y_i - a_i|), & y_i \geq a_i, \end{cases}$$



Зачем нужны такие функции?

Несимметричные функции потерь



$$\frac{1}{m}\sum_{i=1}^{m}\begin{cases} k_1 \,|\, y_i - a_i \,|, & y_i < a_i, \\ k_2 \,|\, y_i - a_i \,|, & y_i \geq a_i, \end{cases}$$

Совет

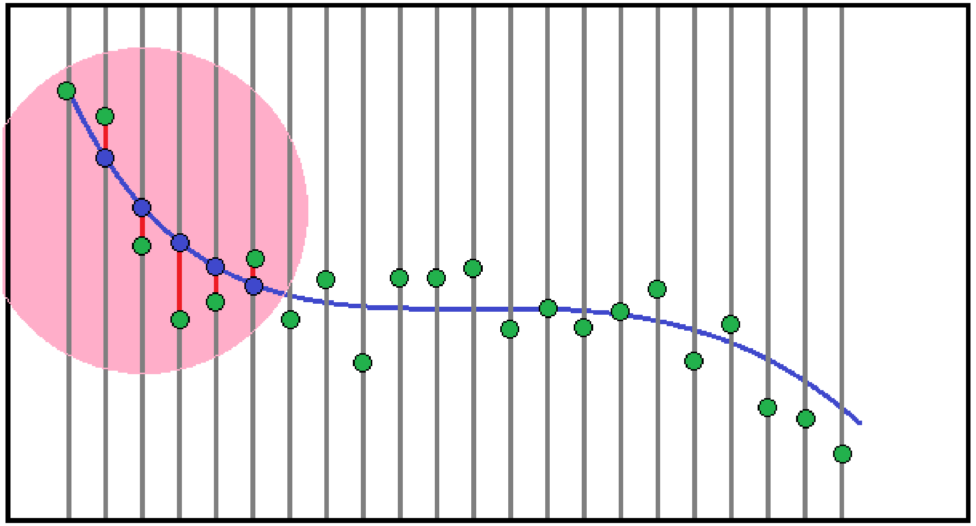
Функции ошибок иногда и классные признаки…

Пример: **в Casuality придумываем бенчмарки (восстановление одной переменной по другой), признаки – их относительные ошибки, т.к. абсолютные брать нельзя**

Почему?

Совет

Аналогично во многих задачах с сигналами…



Признак – не только коэффициенты в приближении, но и ошибка приближения!

~ отклонение от типичного поведения

Монотонное изменение функции ошибки

Формально задачи эквивалентные:

$$\mathrm{MSE} \to \min \qquad \mathrm{RMSE} \to \min$$

$$\frac{1}{m}\sum_{i=1}^{m} | a - y_i |^2 \to \min \qquad \sqrt{\frac{1}{m}\sum_{i=1}^{m} | a_i - y_i |^2} \to \min$$

Решения на практике могут отличаться…

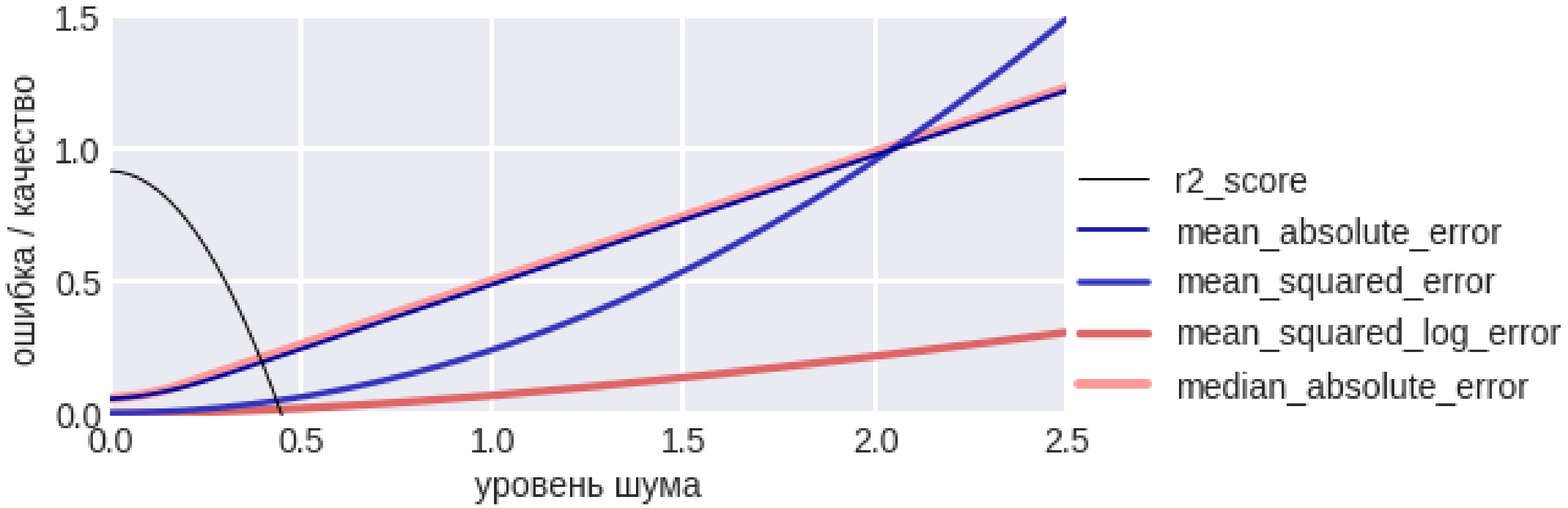
В методе градиентного спуска разные производные

$$\frac{\partial \mathrm{MSE}}{\partial a} = \frac{2}{m}\sum_{i=1}^{m}(a - y_i) \qquad \frac{\partial \mathrm{RMSE}}{\partial a} = \frac{1}{m\,\mathrm{RMSE}}\sum_{i=1}^{m}(a_i - y_i)$$

Метрики в регрессии: минутка кода

```
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_squared_log_error
from sklearn.metrics import median_absolute_error
from sklearn.metrics import explained_variance_score
# R^2
print (r2_score(y, a),
       1 - np.mean((y - a) ** 2) / np.mean((y - np.mean(y)) ** 2))
# MAE
print (mean_absolute_error(y, a),
       np.mean(np.abs(y - a)))
# MSE
print (mean_squared_error(y, a),
       np.mean((y - a) ** 2))
# MSLp1E
print (mean_squared_log_error(y, a),
       np.mean((np.log1p(y) - np.log1p(a)) ** 2))
# MedAE
print (median_absolute_error(y, a),
       np.median(np.abs(y - a)))
```

Сравнение метрик в одном эксперименте



Итоги

Функции ошибки имеют вероятностное обоснование
(через правдоподобие)

средний модуль отклонения MAE (MAD)

средний квадрат отклонения MSE
+ RMSE, коэффициент детерминации R2, функция Хьюбера, Logcosh
Можно невероятностно обосновать для малых отклонений

Иногда попадаются обобщения MAE и RMSE

Итоги

Процентные функции ошибок
(SMAPE, MAPE, PMAD)

Основанные на сравнении с бенчмарком
(MRAE, REL_MAE, PB)

Нормированные ошибки
(MASE)

Несимметричные ошибки

Ошибки с точностью до порога

Есть нетрадиционные применения функций ошибок
для генерации признаков

Литература

Стрижов В.В. Функция ошибки в задачах восстановления регрессии // Заводская лаборатория, 2013, 79(5): 65-73.
http://strijov.com/papers/Strijov2012ErrorFn.pdf

«How to Win a Data Science Competition: Learn from Top Kagglers»
https://ru.coursera.org/learn/competitive-data-science

конспект этих лекций
https://dyakonov.org/2018/10/23/функции-ошибок-в-задачах-регрессии/