

Прикладные задачи анализа данных

Функции ошибки / функционалы качества

Дьяконов А.Г.

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



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!

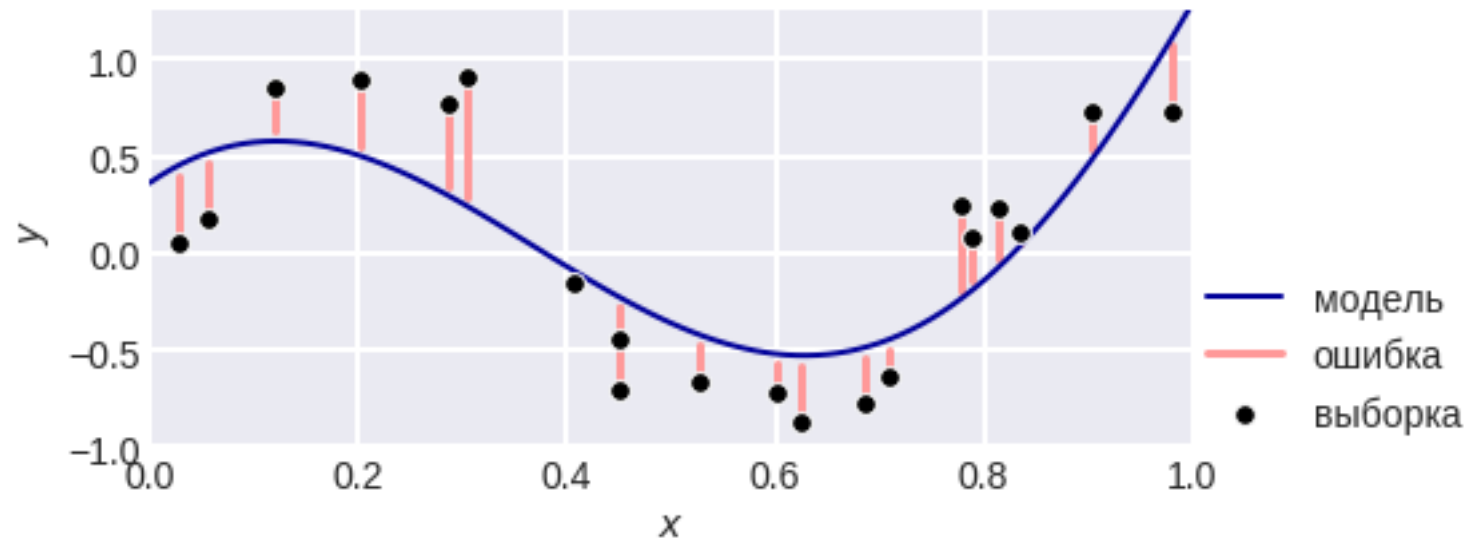
а что такое решение!

В анализе данных:

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

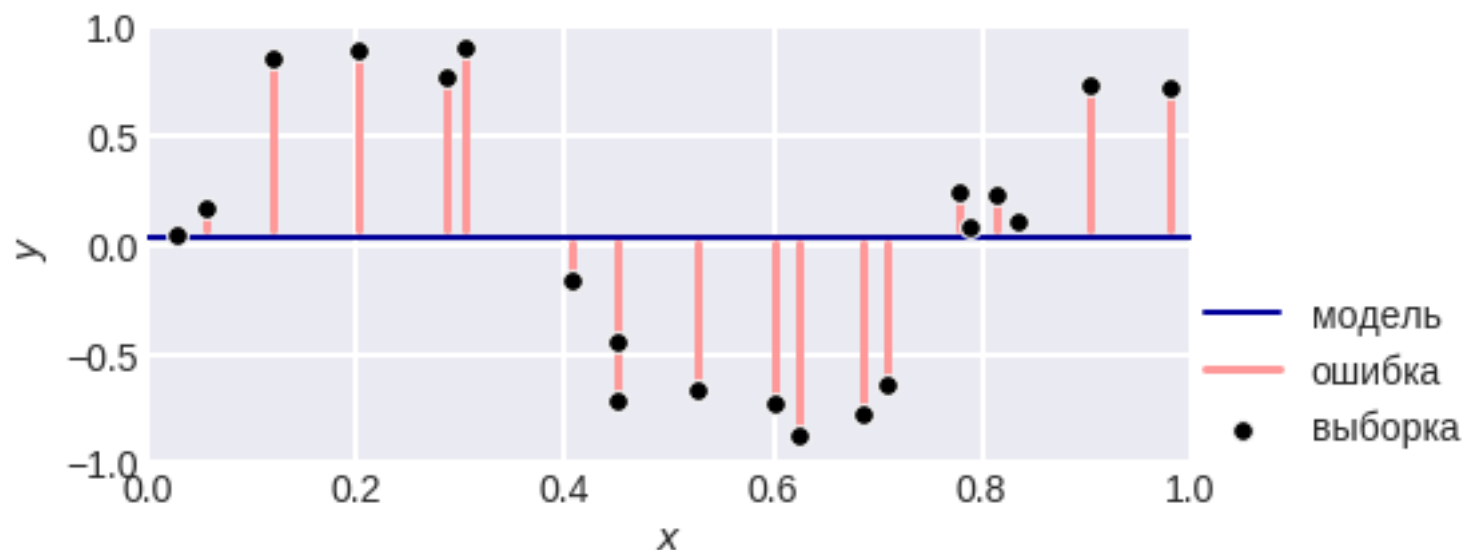
Случай из практики: задача про траектории зрачка
(задача с 3 классами, а не с двумя)

Задача регрессии



Задача регрессии

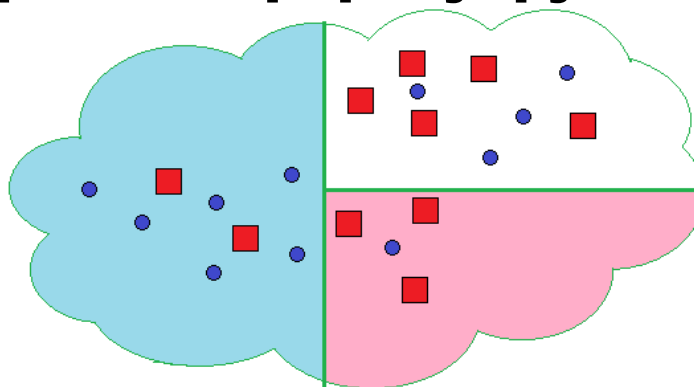
Будем дальше пытаться всё решать в классе констант



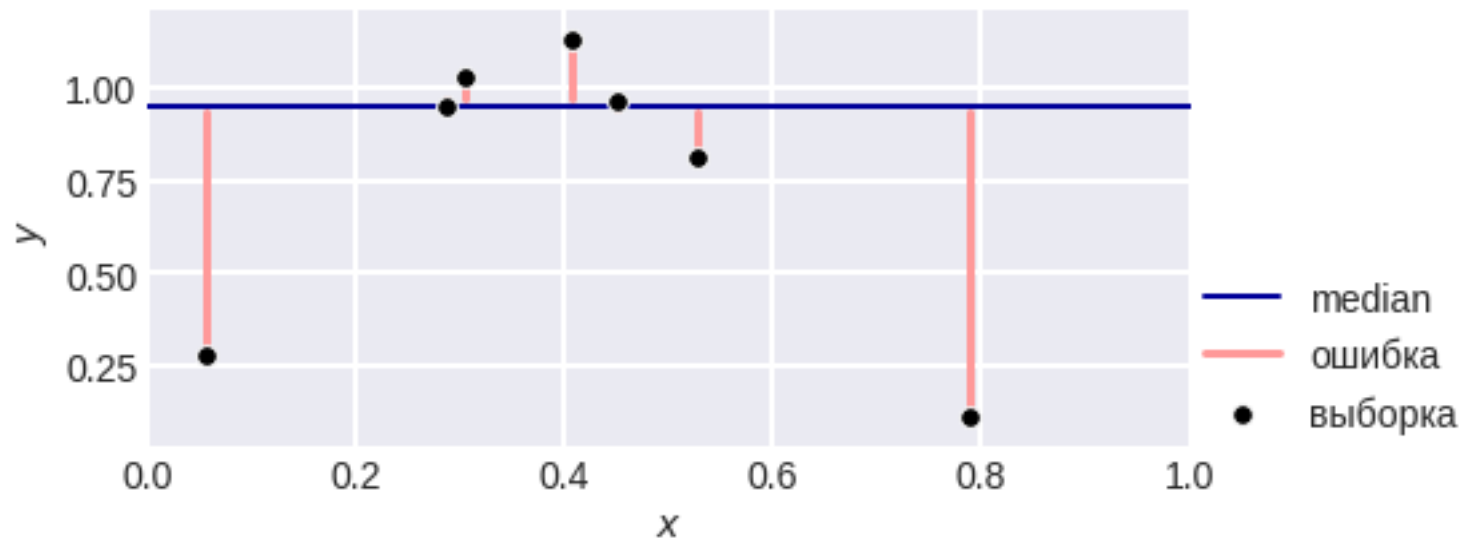
1. Простейшее решение

2. Примерно это и происходит в листьях решающих деревьев

3. Раскрывает природу функционалов



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



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

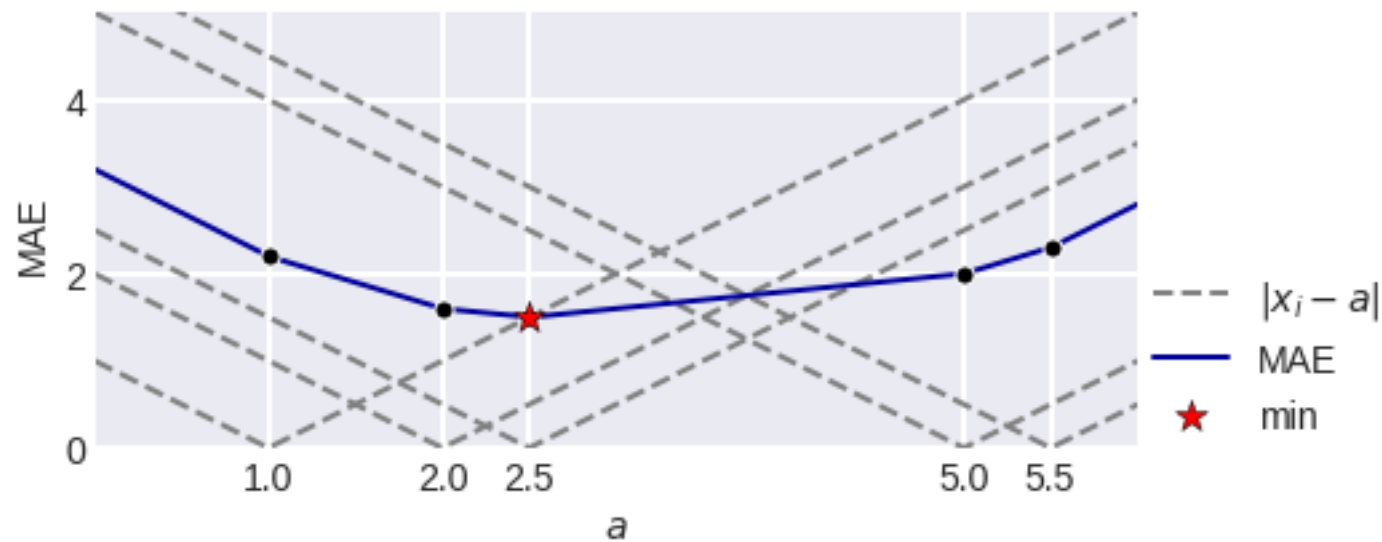
Напоминание:

$$\frac{1}{q} \sum_{i=1}^q |a - y_i| \rightarrow \min$$

$$a = \text{median}(\{y_i\}_{i=1}^q)$$

Это открывает смысл решений!

Средний модуль отклонения



Средний модуль отклонения

Способы использования тайных знаний:

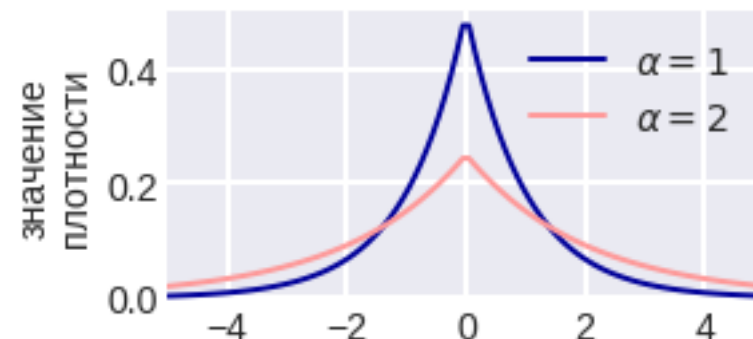
- **медиана, вместо усреднения, в ансамбле**
- **округление ответа (если целевой вектор целочисленный)**

Откуда берётся MAE

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

w – параметры алгоритма $a_w(x)$

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



Для оценки параметров выписываем правдоподобие модели

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

Метод максимального правдоподобия:

$$\begin{aligned} \log L(w) &= \log \prod_{i=1}^m p(y_i | x_i, w) = \\ &= \sum_{i=1}^m \left[\log \frac{\alpha}{2} - \alpha |y_i - a_w(x_i)| \right] \rightarrow \max \end{aligned}$$

Откуда берётся MAE

Получаем

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

т.е. задачу минимизации MAE!

- не зависит от природы модели
- зависит от распределения ошибок
(почему Residual Plots)

Максимизация правдоподобия эквивалентна минимизации MAE!

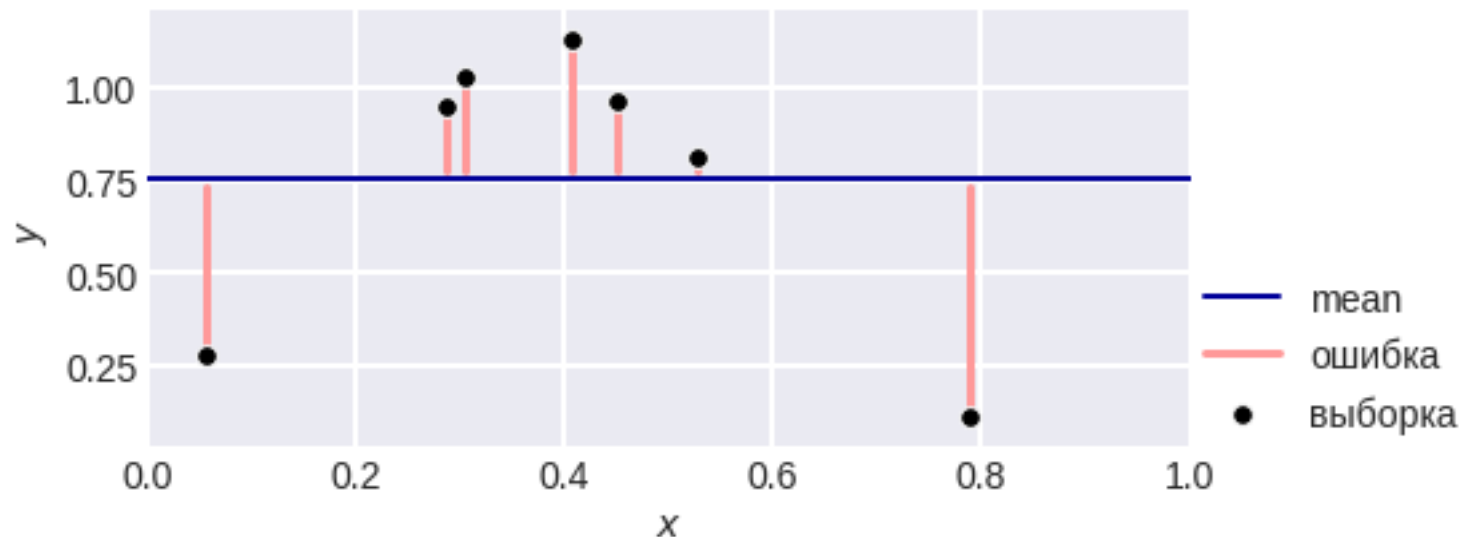
Чему соответствует минимизация весового MAE?

Средний квадрат отклонения ~ Mean Squared Error (MSE)

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

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

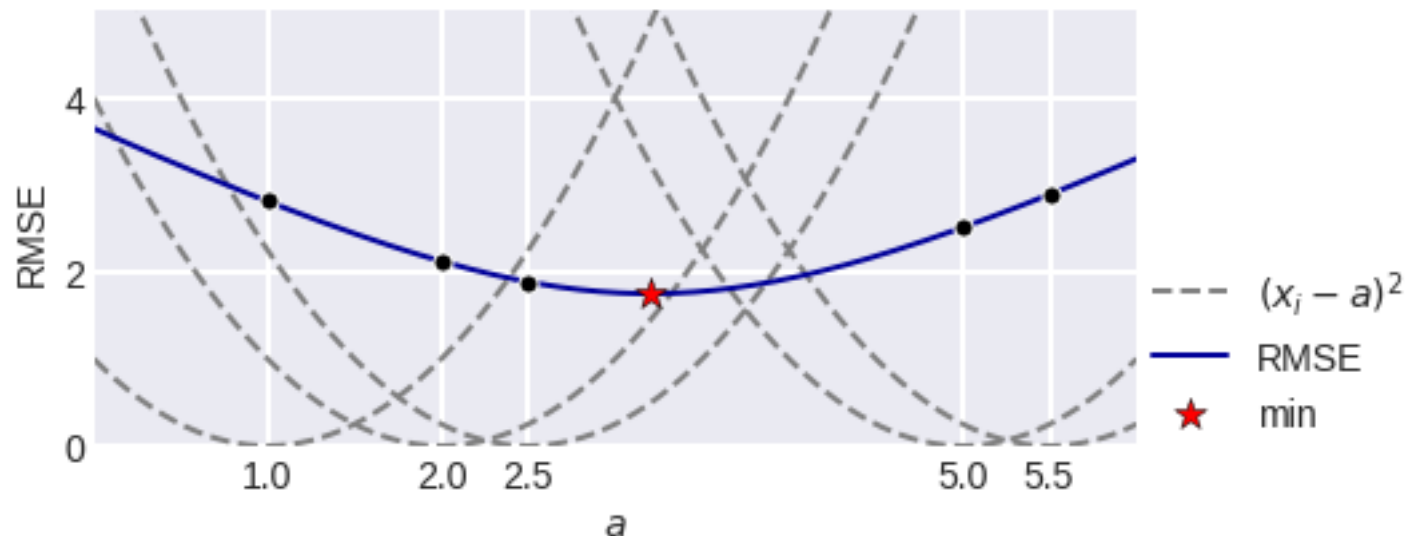
$$a = \frac{1}{q} \sum_{i=1}^q y_i$$



Root Mean Squared Error (RMSE)
или **Root Mean Square Deviation (RMSD)**

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

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)

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

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

В общем случае (в статистике) коэффициент детерминации:

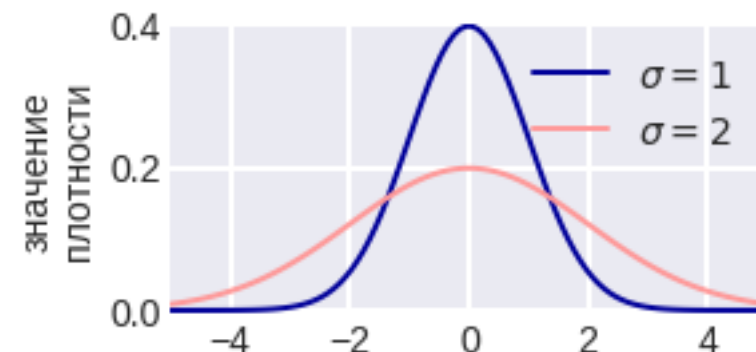
$$R^2 = 1 - \frac{\mathbf{D}(y | x)}{\mathbf{D}(y)}$$

Откуда берётся (R)MSE

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

w – параметры алгоритма $a_w(x)$

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



Для оценки параметров выписываем правдоподобие модели

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

Метод максимального правдоподобия:

$$\begin{aligned} \log L(w) &= \log \prod_{i=1}^m p(y_i | x_i, w) = \\ &= \sum_{i=1}^m \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - a_w(x_i))^2}{2\sigma^2} \right] \rightarrow \max \end{aligned}$$

Откуда берётся (R)MSE

Получаем

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

т.е. задачу минимизации MSE!

- не зависит от природы модели
- зависит от распределения ошибок
(почему Residual Plots)

**Максимизация правдоподобия эквивалентна минимизации
среднеквадратичной ошибки!**

Д3 Каким ещё распределениям какие ошибки соответствуют?

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

Пусть функция ошибки $l(y, a) = g(y - a)$

Что логично?

1. $g(0) = 0$

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

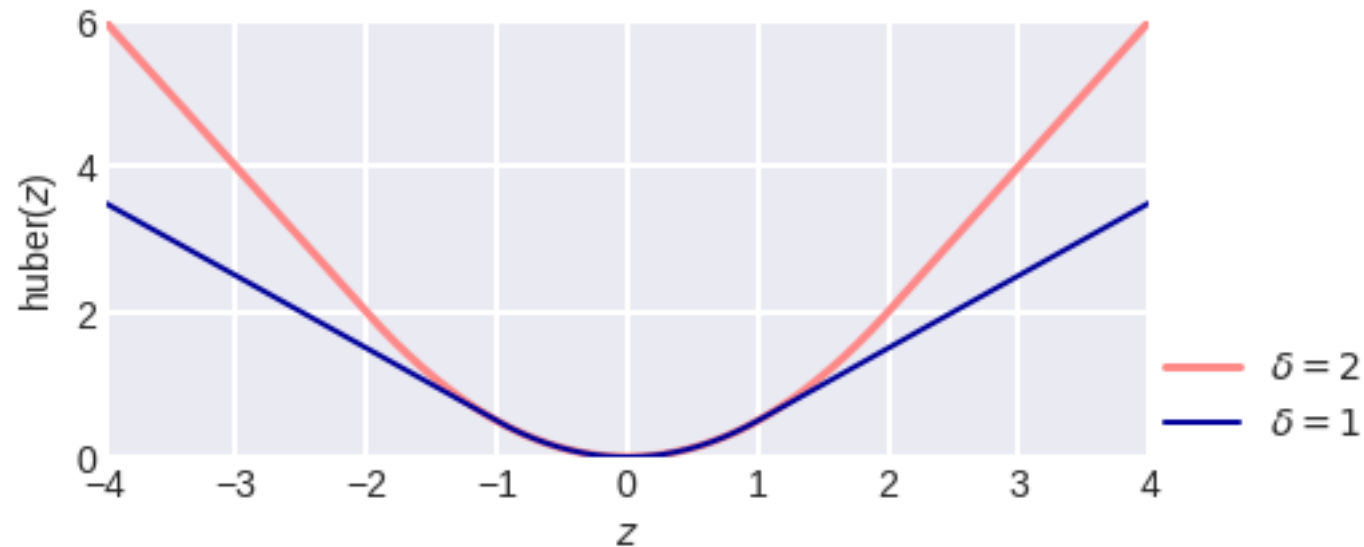
3. достаточно гладкая...

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

но тогда

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

Функция Хьюбера



$$\text{huber}(z) = \begin{cases} \frac{1}{2}z^2, & |z| \leq \delta, \\ \delta\left(|z| - \frac{1}{2}\delta\right), & |z| > \delta. \end{cases}$$

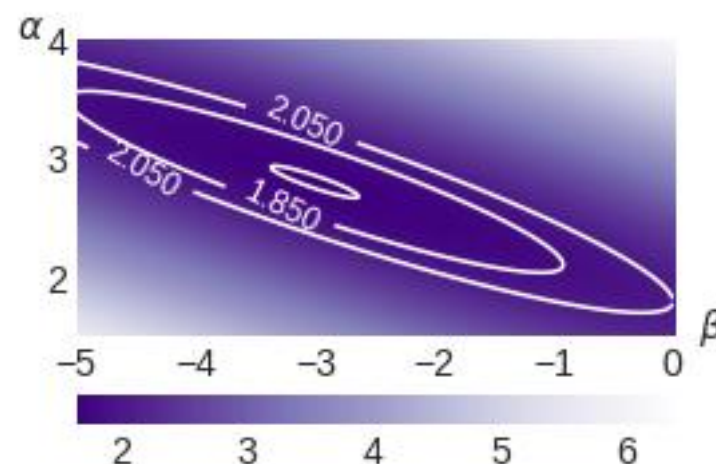
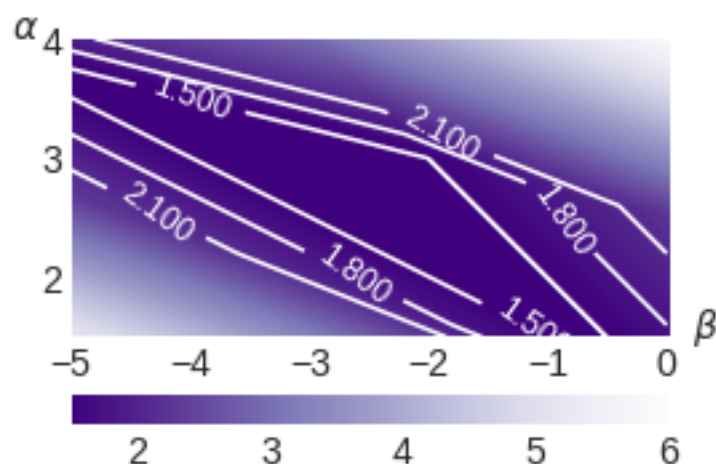
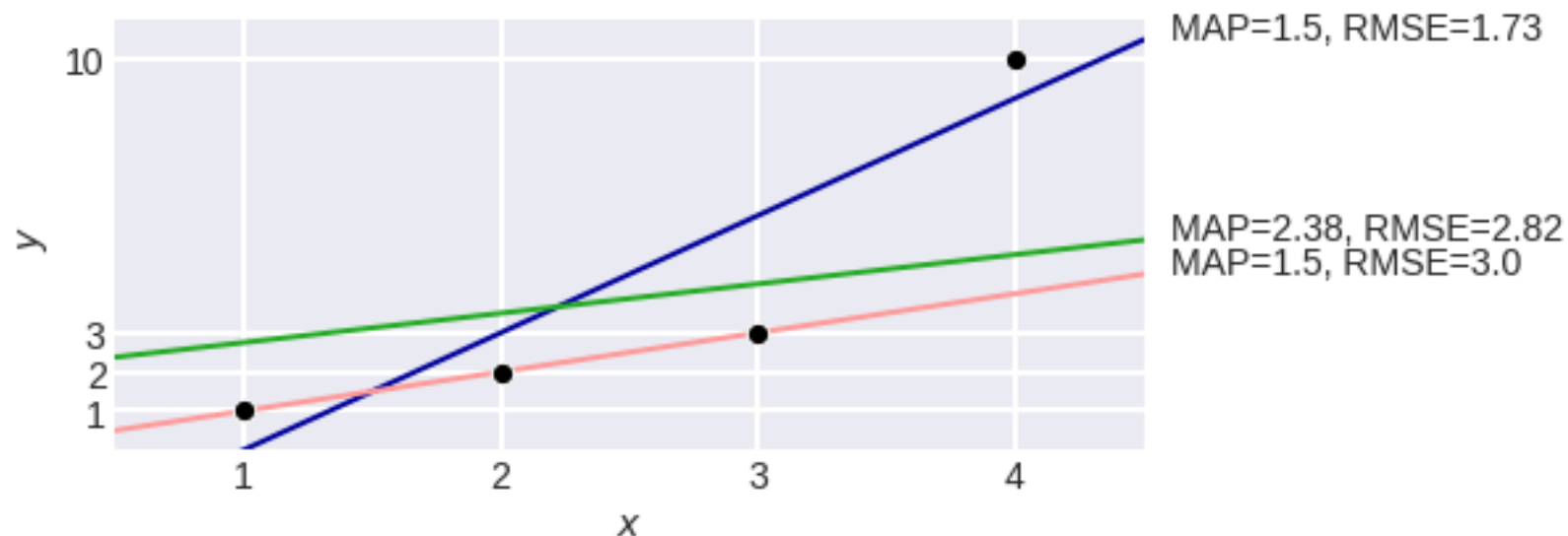
Как только что вывели:

когда отклонение мало – ошибка квадратичная

когда велико (в т.ч. выбросы) – линейная

Различия MSE и MAE: посмотрим на неконстантное решение:

$$\sum_{i=1}^m |y_i - a(x_i)|^p \rightarrow \min, a(x) = \alpha x + \beta$$



Различия MSE и MAE

внутри «треугольника» одинаковый $MAP=1.5$

**можно привести примеры, когда MAP меняется слабо,
а $RMSE$ значительно**

Д3 Хороший нетривиальный пример?

Д3 Может ли быть наоборот?

Обобщения

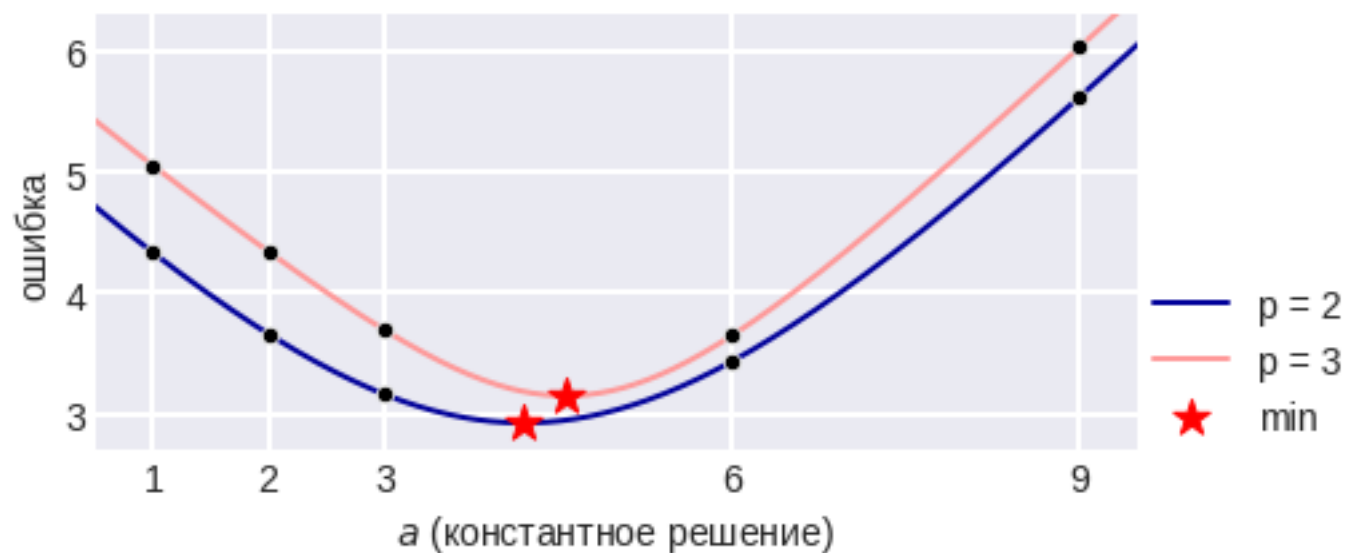
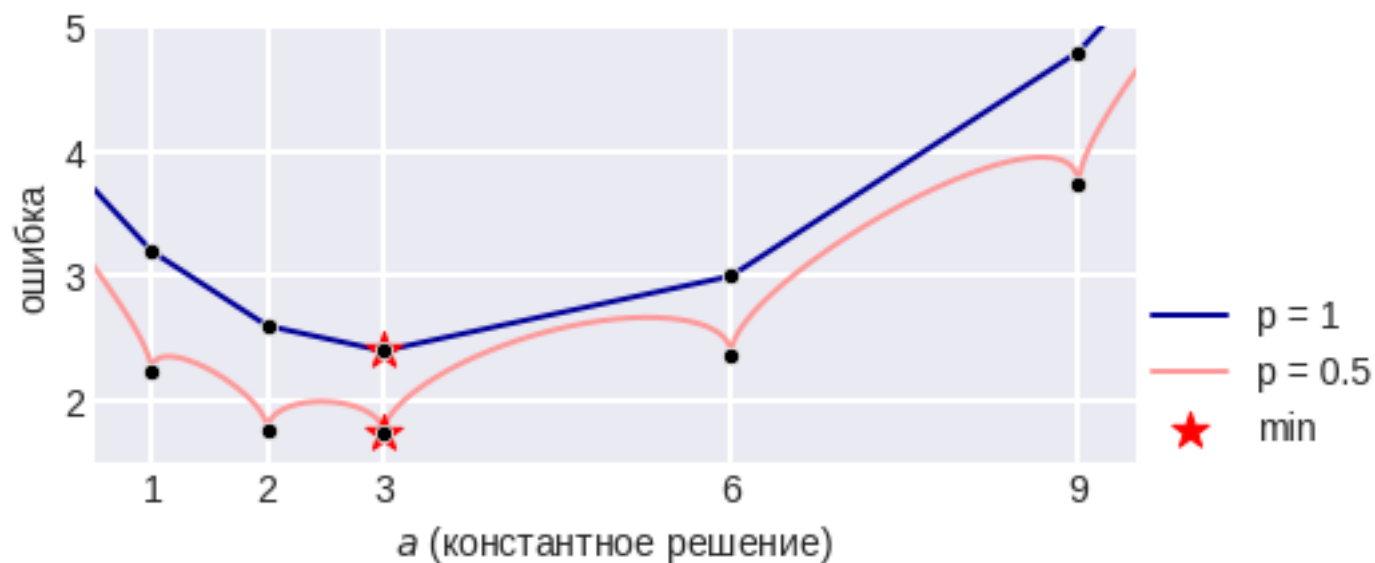
$$\sqrt[p]{\frac{1}{q} \sum_{i=1}^q w_i |\varphi(a_i) - \varphi(y_i)|^p}$$

Рецепты

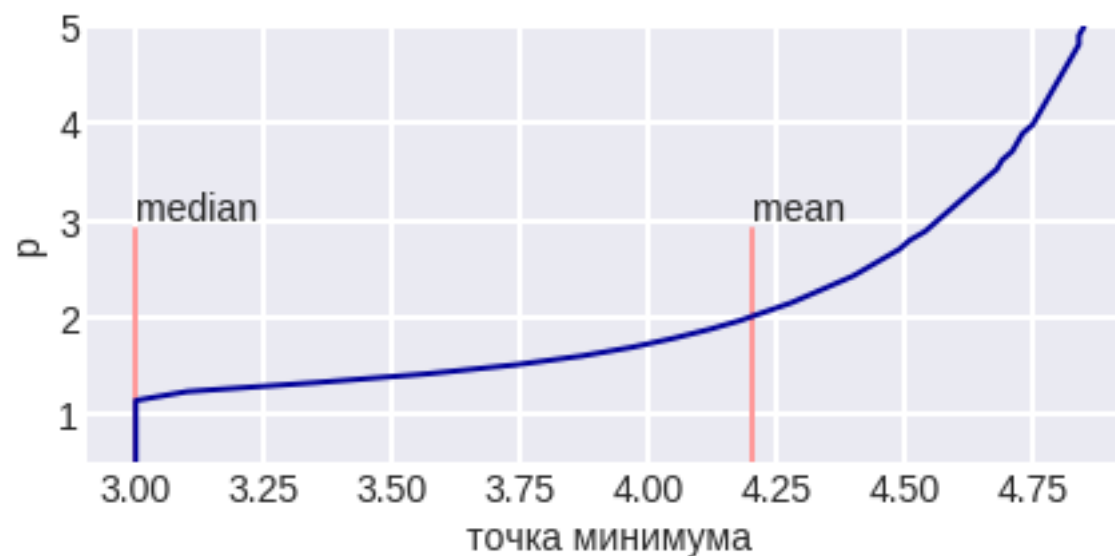
1. Преобразование целевого вектора $\varphi(y)$
2. Веса ~ вероятности появления объектов в сэмплировании
Некоторые модели поддерживают веса объектов
3. В случае нетривиальных p – прямая настройка

Дальше к этому вернёмся...

Про нетривиальные p



Как точка минимума зависит от степени



Symmetric mean absolute percentage error (SMAPE or sMAPE)

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

**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)

$$\text{MAPE} = \frac{1}{q} \sum_{i=1}^q \frac{|y_i - a_i|}{|y_i|}$$

Чем MAPE явно лучше SMAPE на практике?

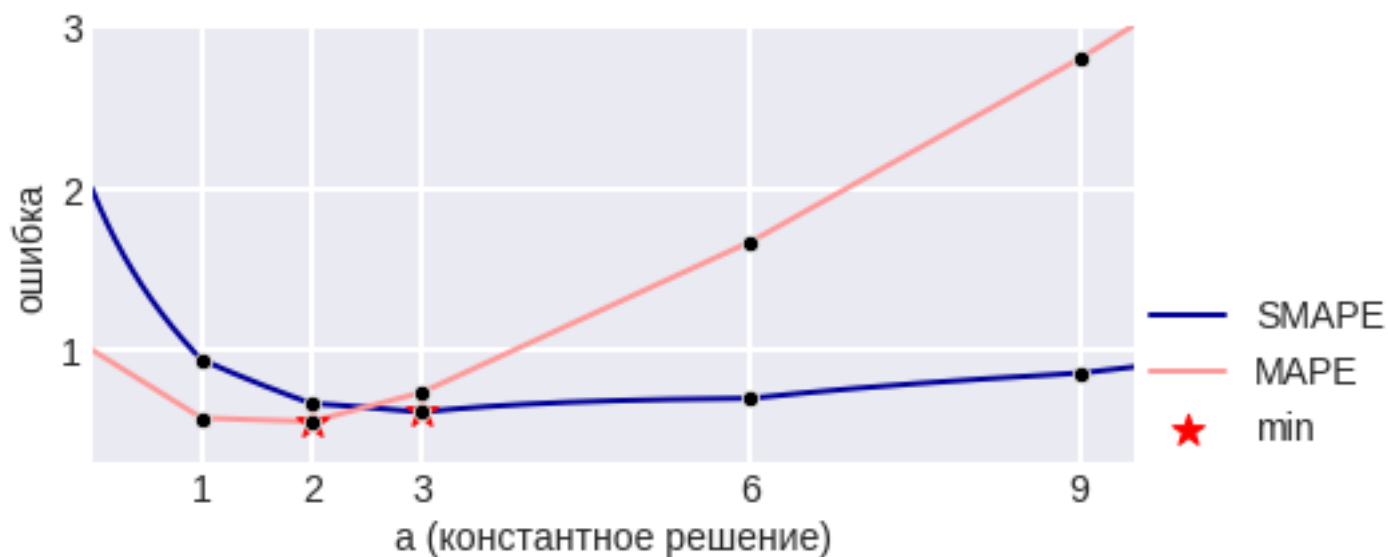
Mean Absolute Percent Error (MAPE)

$$\text{MAPE} = \frac{1}{q} \sum_{i=1}^q w_i |y_i - a_i|$$

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

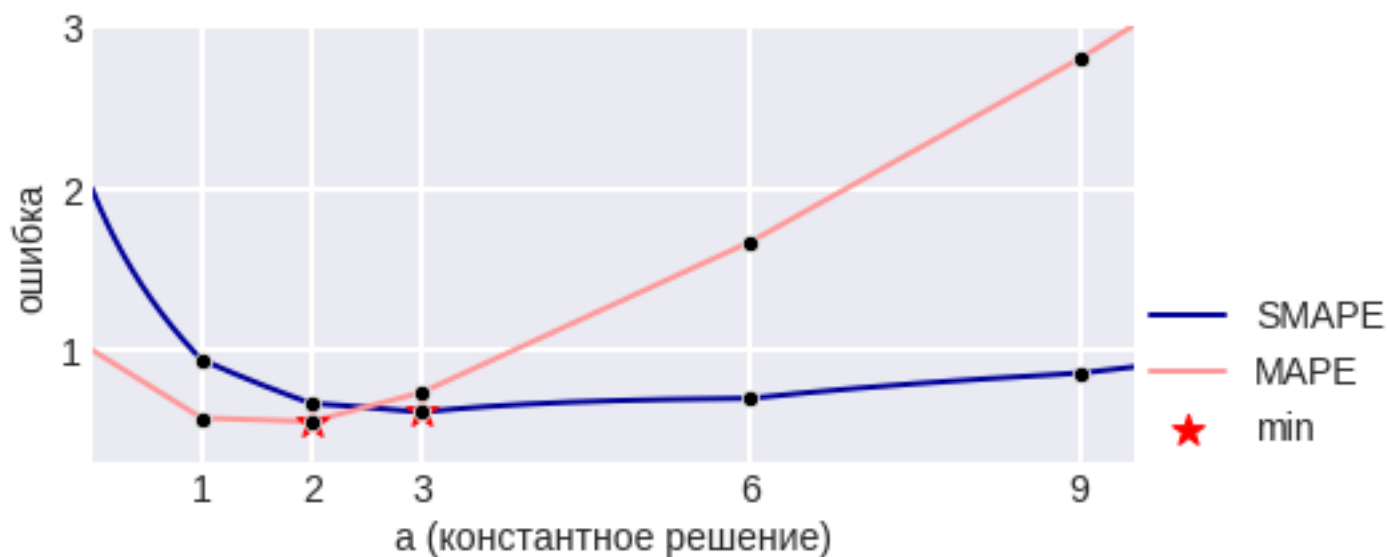
Просто весовой MAE!
как оптимизировать? дальше...

MAPE и SMAPE

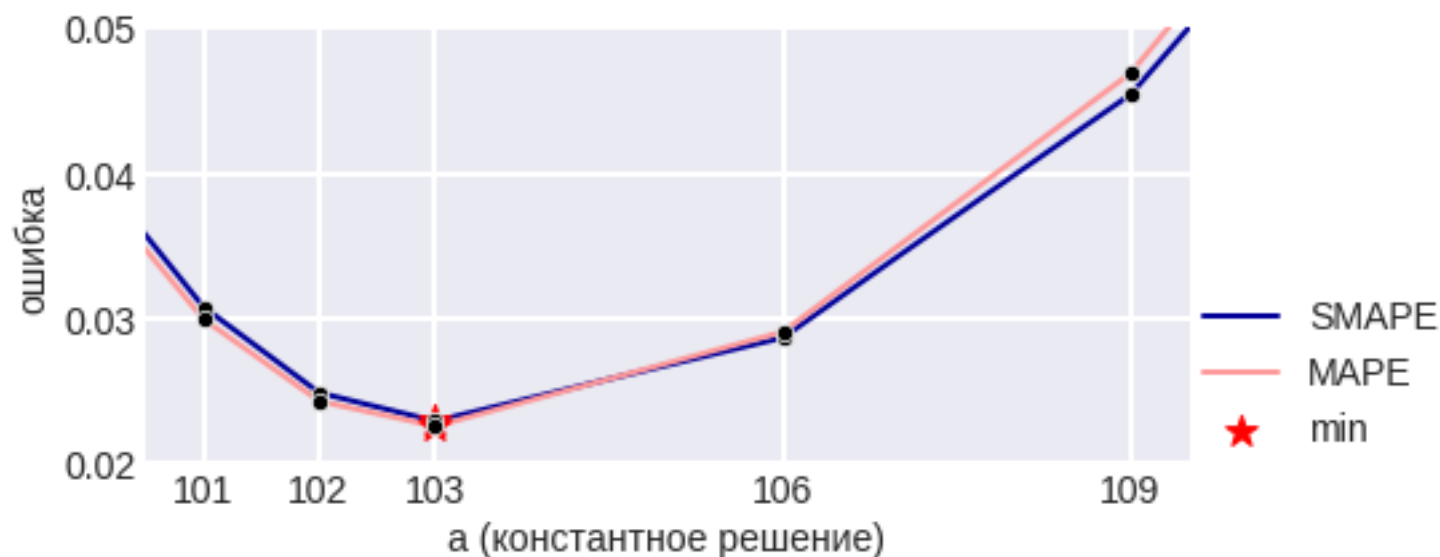


Что настораживает в этом графике?

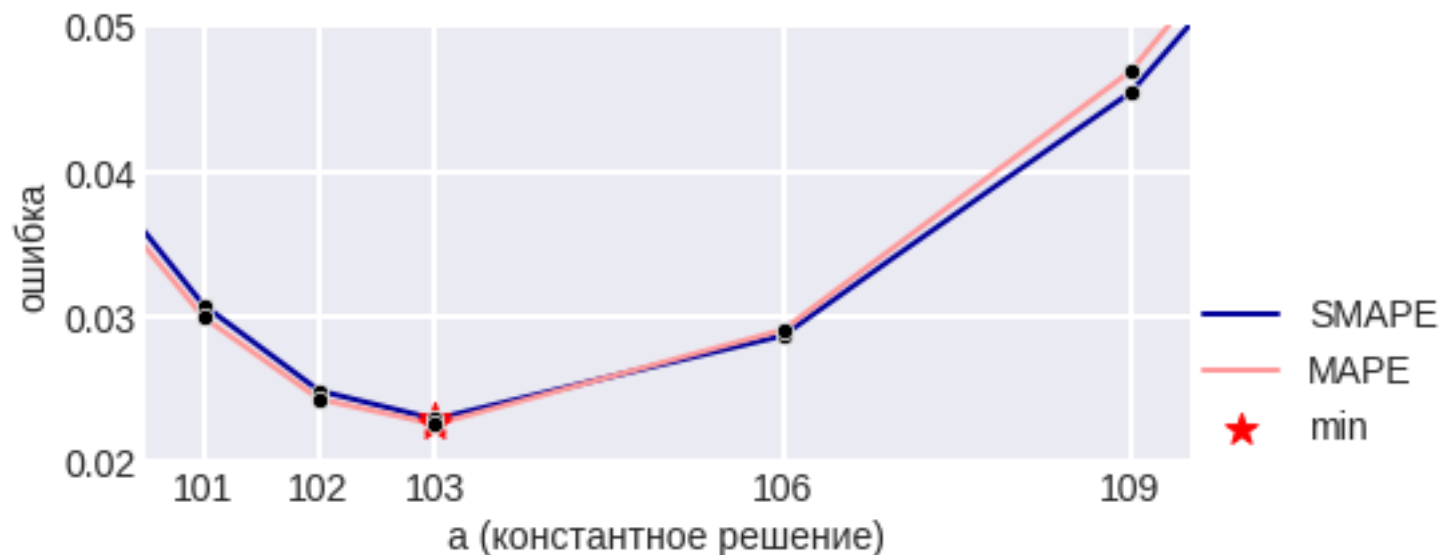
MAPE и SMAPE



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

ДЗ Предложить минимизацию для MAPE и SMAPE

PMAD

Другой способ нормировки ошибки...

$$\text{PMAD} = \frac{\frac{1}{q} \sum_{i=1}^q |y_i - a_i|}{\sum_{i=1}^q |y_i|}$$

эквивалентен MAE

Д3 Как на типичных и специальных выборках соотносятся решения задач минимизации перечисленных функций ошибки?

Меры на сравнении с бенчмарком

Классная идея:

сделать простой алгоритм и смотреть ошибку относительно него

**Mean Relative Absolute Error
(MRAE)**

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

REL_MAE

$$\text{REL_MAE} = \frac{\sum_{i=1}^q |y_i - a_i|}{\sum_{i=1}^q |y_i - a'_i|}$$

Percent Better

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

Меры на сравнении с бенчмарком

Как выбрать бенчмарк в задачах прогнозирования?

Нормированные ошибки

Не зависят от шкалы...

Mean Absolute Scaled Error

$$\text{MASE} = \frac{1}{\frac{q}{q-1} \sum_{i=2}^q |y_{i-1} - y_i|} \sum_{i=1}^q |a_i - y_i|$$

Какие ещё бывают функционалы в регрессии?

С точностью до порога

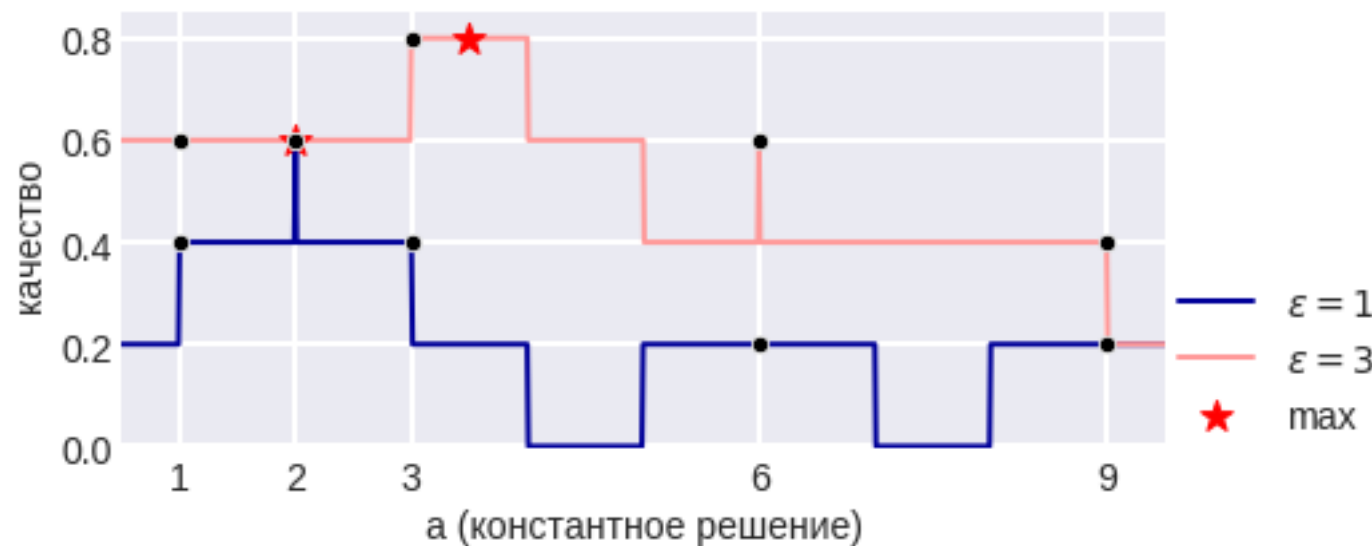
функция ошибки

$$\frac{1}{q} \sum_{i=1}^q I[|y_i - a_i| > \varepsilon]$$

функционал качества

$$\frac{1}{q} \sum_{i=1}^q I[|y_i - a_i| < \varepsilon]$$

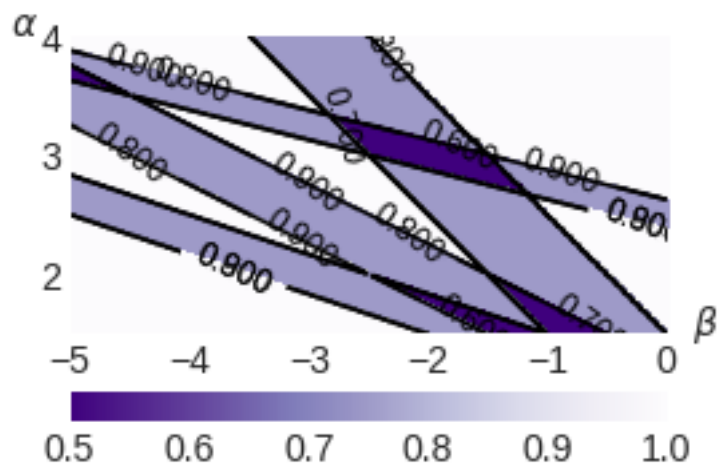
был в задаче **Dunnhumby**



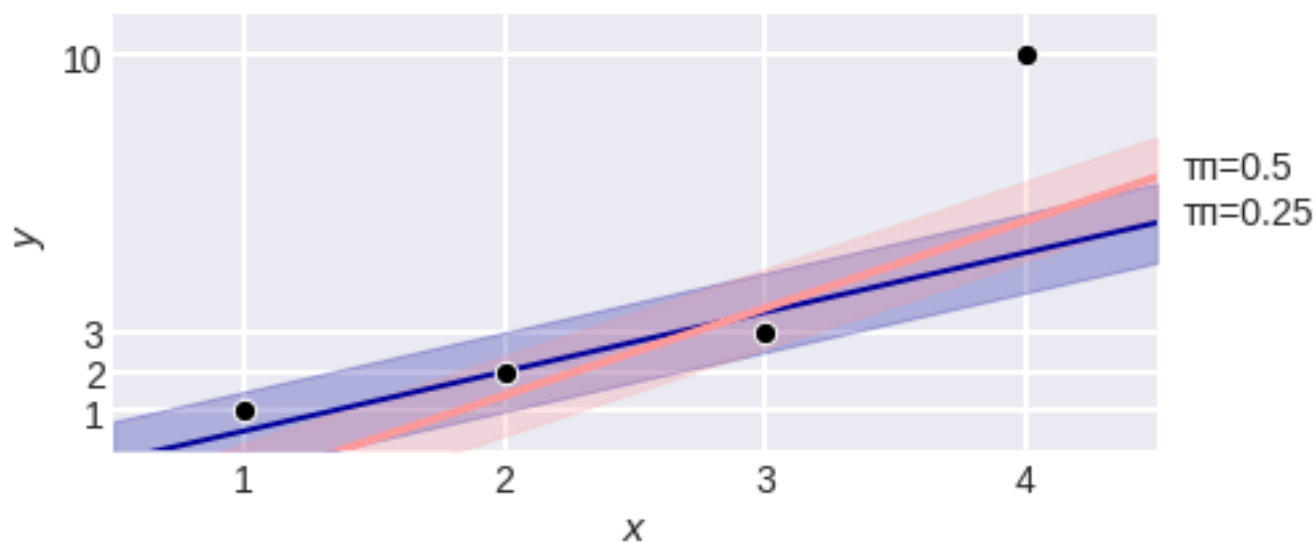
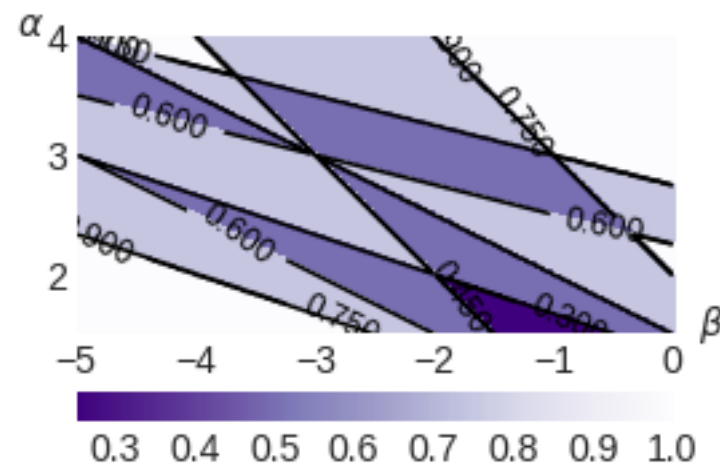
Оптимальное решение – мода парzenовской плотности

С точностью до порога

$$\varepsilon = 0.5$$

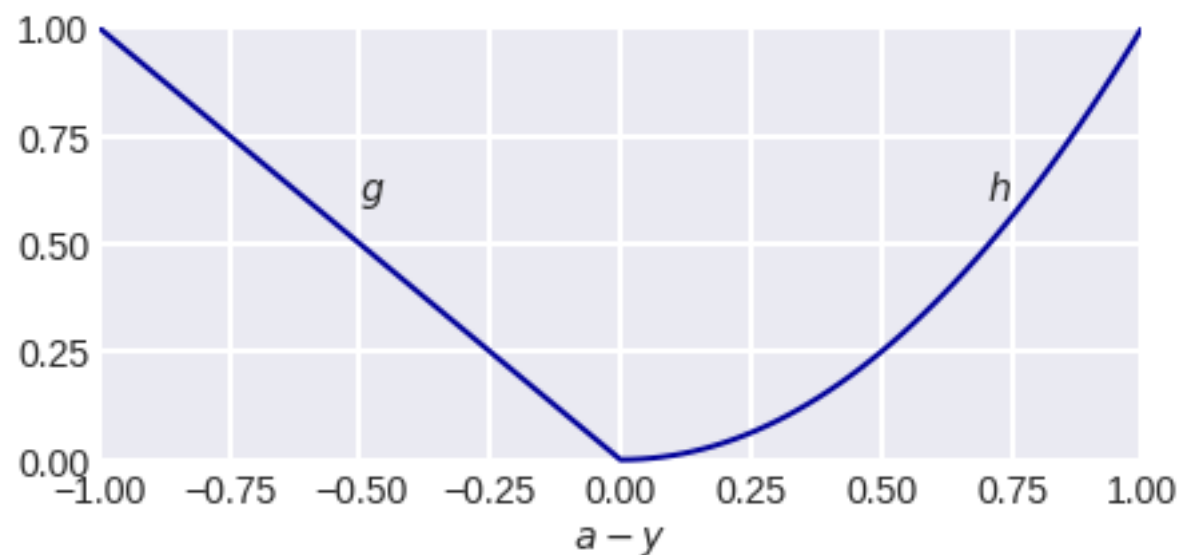


$$\varepsilon = 1.0$$



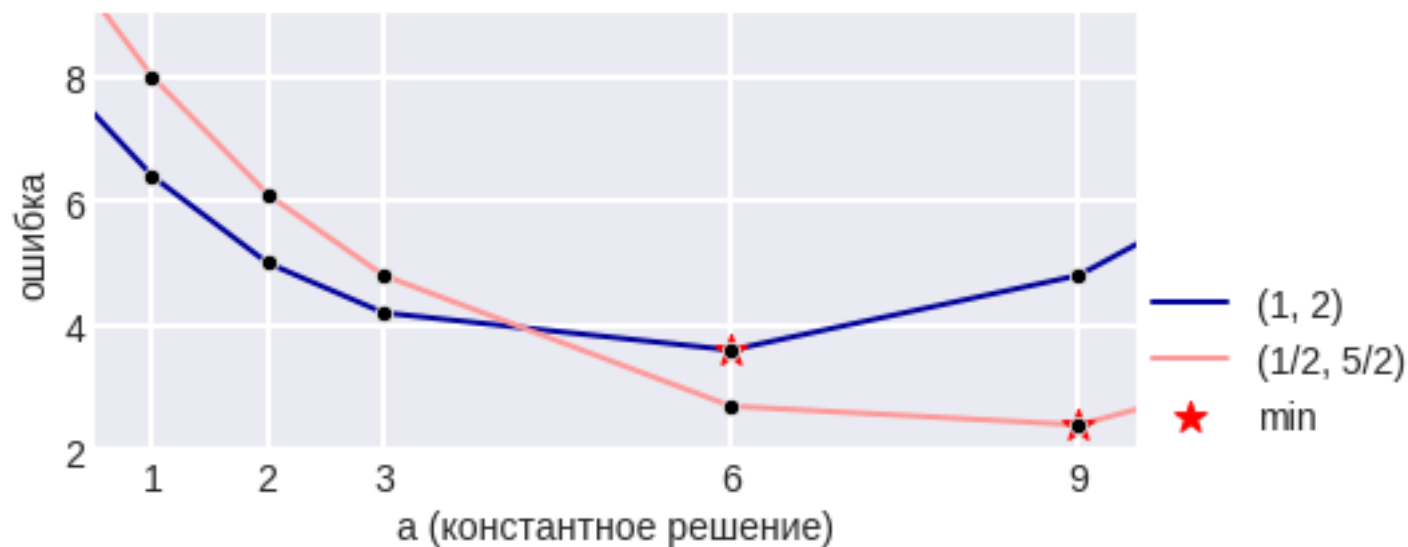
Несимметричные функции потерь

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



Зачем нужны такие функции?

Несимметричные функции потерь



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

Совет

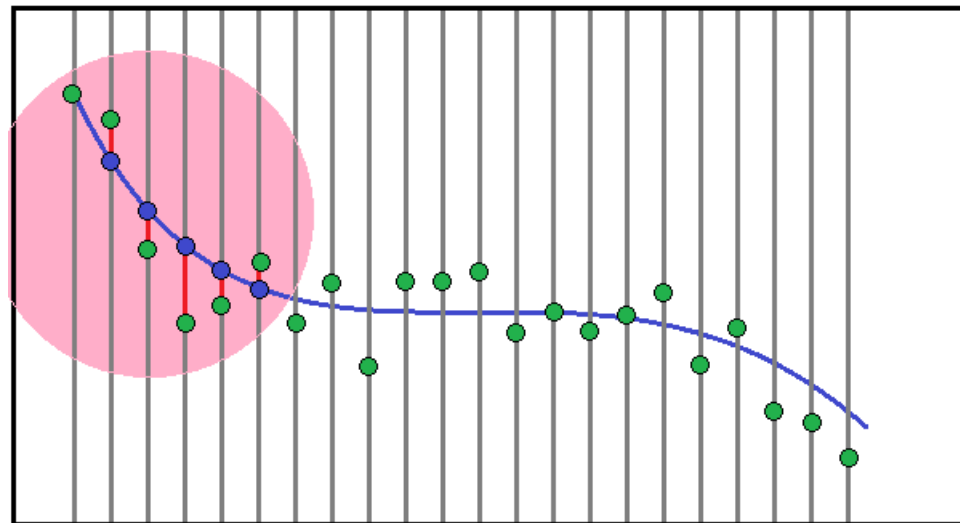
Функции ошибок иногда и классные признаки...

Пример: в Casualty 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:

$$\text{MSE} \rightarrow \min$$

$$\text{RMSE} \rightarrow \min$$

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

$$\sqrt{\frac{1}{q} \sum_{i=1}^q |a_i - y_i|^2} \rightarrow \min$$

Решения на практике могут отличаться...

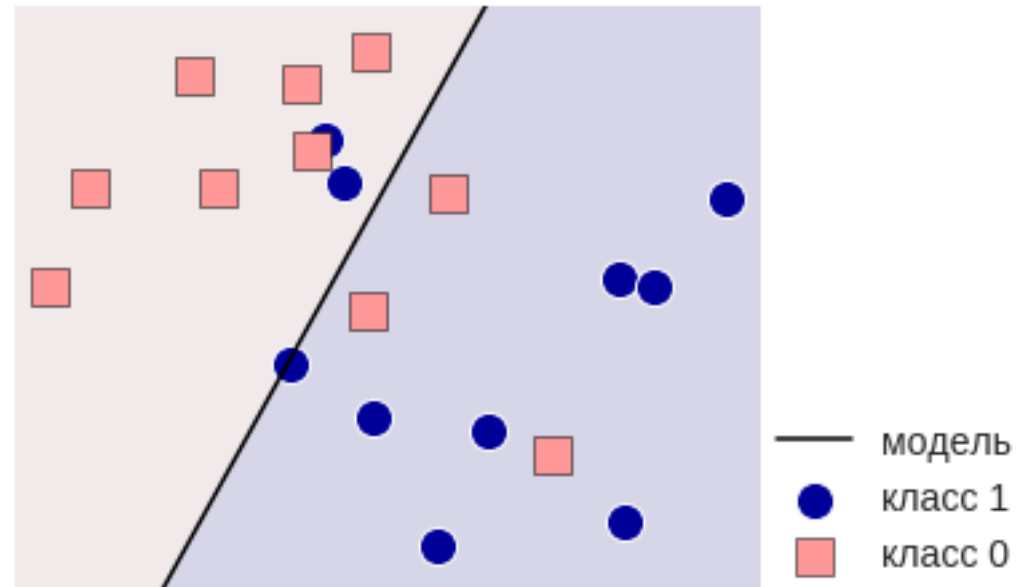
В методе градиентного спуска разные производные

$$\frac{\partial \text{MSE}}{\partial a} = \frac{2}{q} \sum_{i=1}^q (a - y_i)$$

$$\frac{\partial \text{RMSE}}{\partial a} = \frac{1}{q \text{RMSE}} \sum_{i=1}^q (a_i - y_i)$$

Д3 На что это влияет на практике? что лучше минимизировать?

Задача классификации



Задача классификации: матрица ошибок / несоответствий «Confusion Matrix»

ОТВЕТЫ

у	а
0	1 1
1	1 1
2	1 2
3	2 1
4	2 3
5	3 2
6	3 3
7	3 3
8	1 2
9	2 2

матрица ошибок

y	a	1	2	3
1	2	2	0	
2	1	1	1	
3	0	1	2	

Для классов $\{1, 2, \dots, l\}$

$$N = \| n_{ij} \|_{l \times l}$$

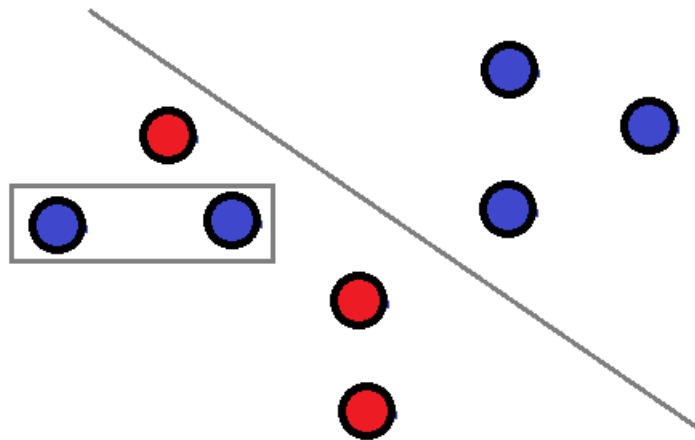
$$n_{ij} = \sum_{t=1}^q I[a_t = i] I[y_t = j]$$

```
from sklearn.metrics import confusion_matrix
n = confusion_matrix(df.y, df.a)
n = pd.crosstab(df.y, df.a)
```

Обычная точность – Accuracy, Mean Consequential Error

$$\text{MCE} = \frac{1}{q} \sum_{i=1}^q I[a_i = y_i] = \frac{\sum_{t=1}^l n_{tt}}{\sum_{t=1}^l \sum_{s=1}^l n_{ts}}$$

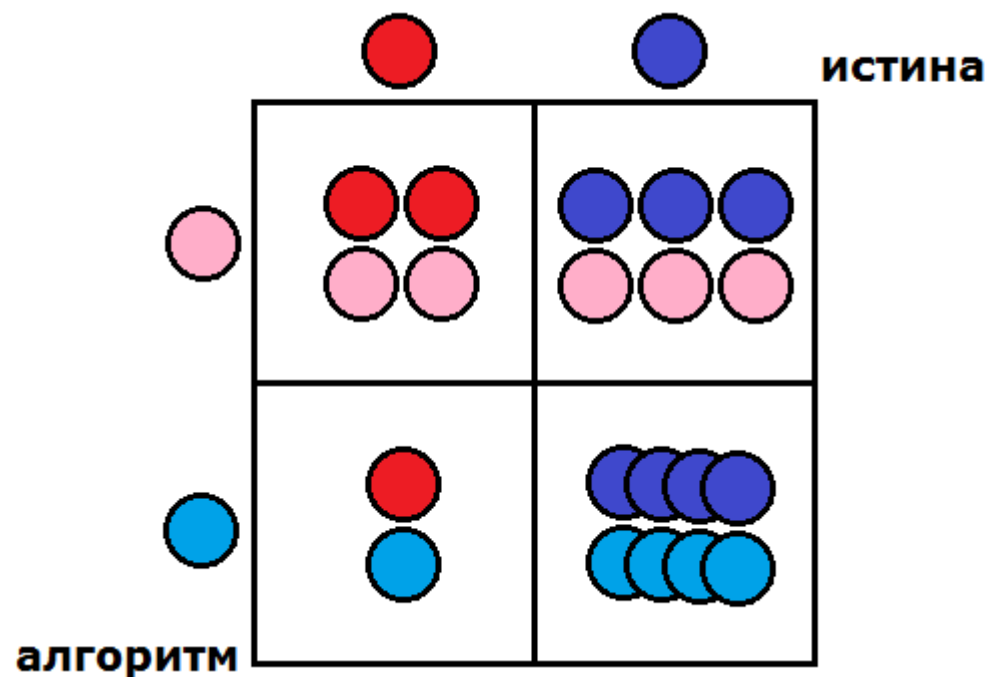
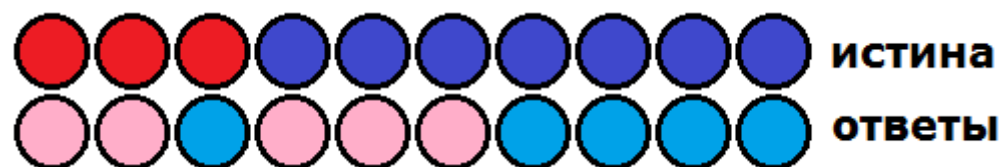
- первое, что приходит в голову
- не учитывает разную мощность классов



$y = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]$

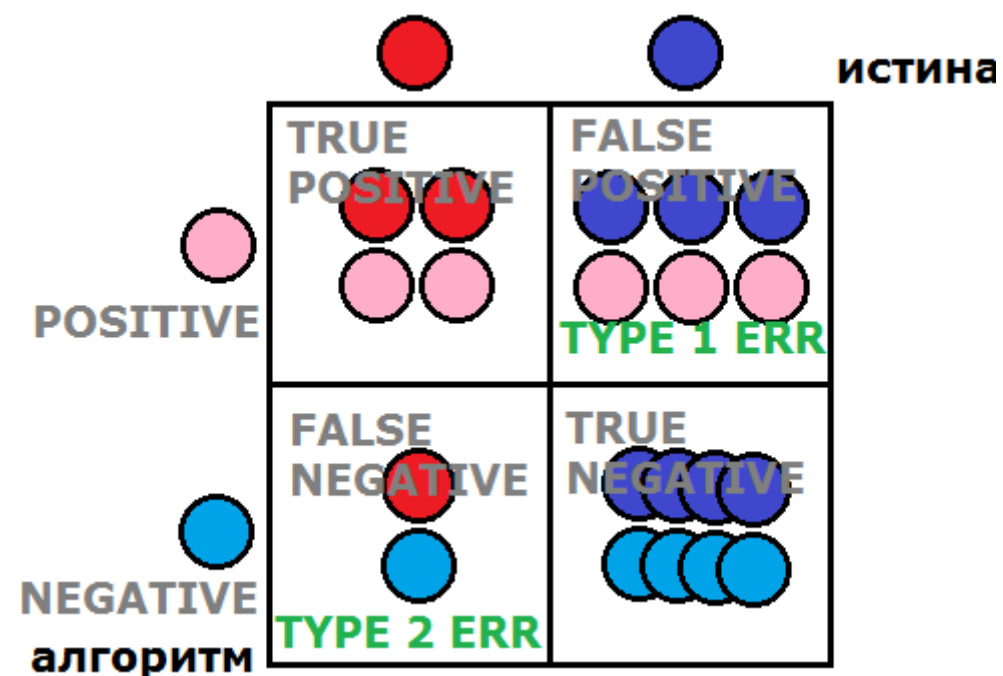
Выгодно выдавать решение – константу 0!

Задача классификации с двумя классами



Confusion Matrix

Задача классификации с двумя классами



Как запомнить названия ошибок

1 рода – **не учил**, но **сдал** (= знает по мнению экзаменатора)

2 рода – **учил**, но **не сдал** (= не знает по мнению экзаменатора)



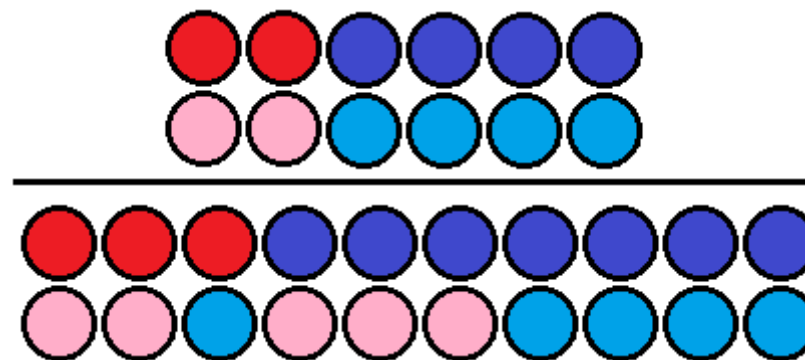
Ошибка 1 рода



Ошибка 2 рода

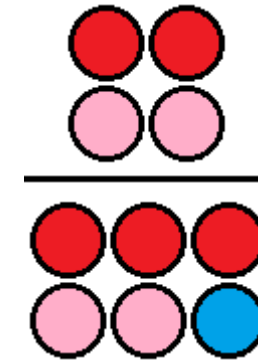
Точность Accuracy

$$ACC = \frac{TP+TN}{ALL}$$



Полнота (Sensitivity, True Positive Rate, Recall, Hit Rate)

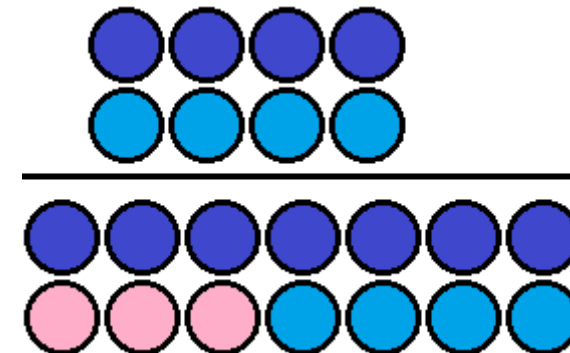
$$TPR = \frac{TP}{TP + FN}$$



TPR = TP / сколько правда 1

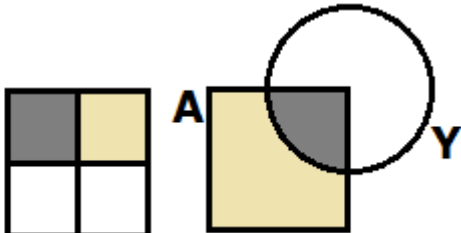
Specificity (True Negative Rate)

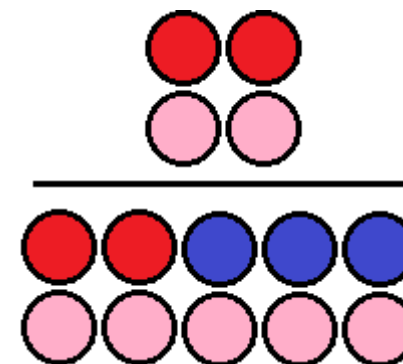
$$SPC = \frac{TN}{FP + TN}$$



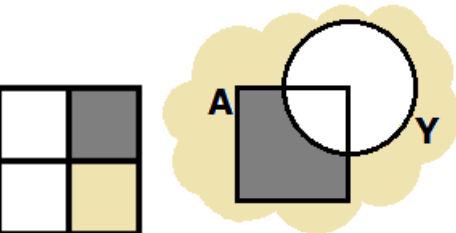
FPR = 1 – Specificity

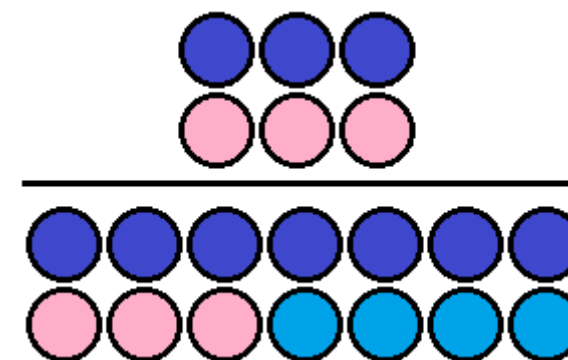
Точность (Precision, Positive Predictive Value)

$$PPV = \frac{TP}{TP + FP}$$


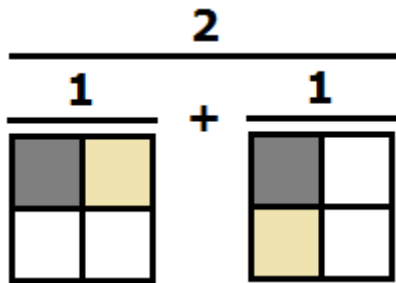


False Positive Rate (FPR, fall-out, false alarm rate)

$$FPR = \frac{FP}{FP + TN}$$




FPR = FP / сколько правда 0

F₁ score

$$\frac{2}{\frac{1}{TP/(TP+FP)} + \frac{1}{TP/(TP+FN)}} = \frac{2TP}{2TP + FP + FN}$$

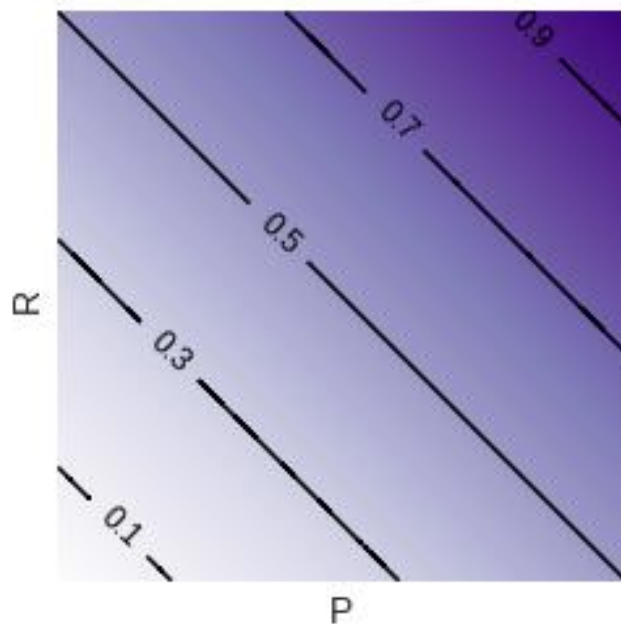
F_β score

$$\frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} = \frac{PR}{\alpha R + (1-\alpha)P} = \frac{1}{\alpha} \frac{PR}{R + \left(\frac{1}{\alpha} - 1\right)P}$$

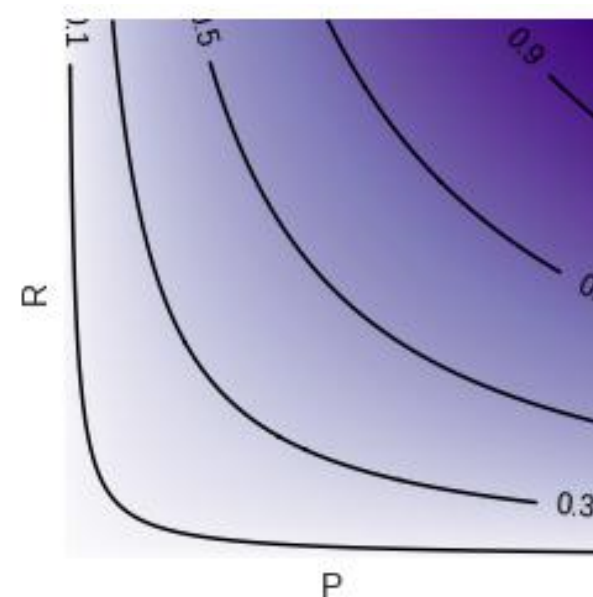
$$\beta^2 = \left(\frac{1}{\alpha} - 1\right)$$

$$F_\beta = (1 + \beta^2) \frac{PR}{R + \beta^2 P}$$

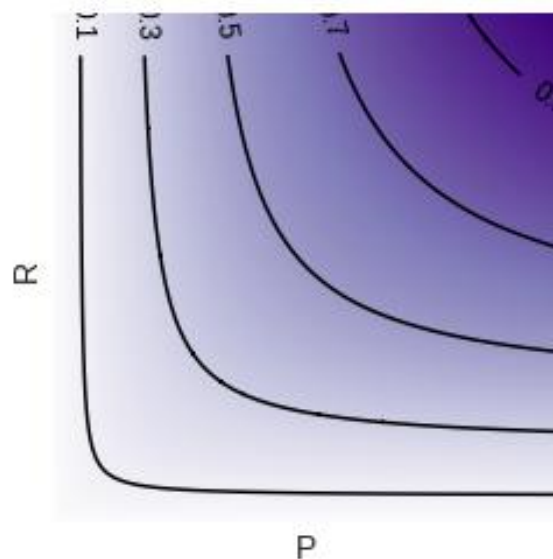
Почему используется F-мера



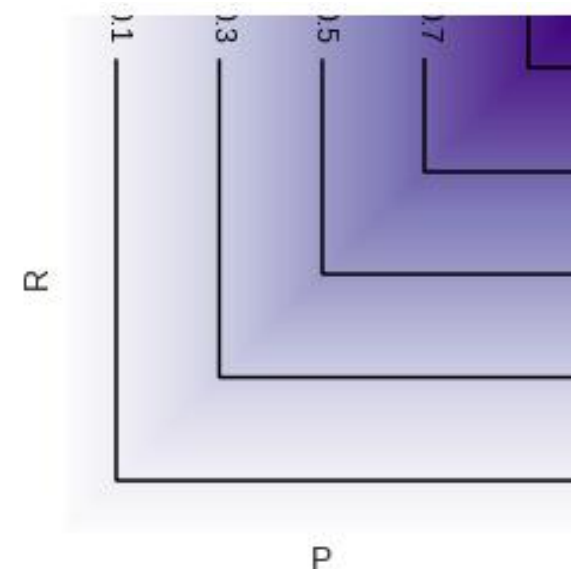
$$(P+R)/2$$



$$\sqrt{P \cdot R}$$



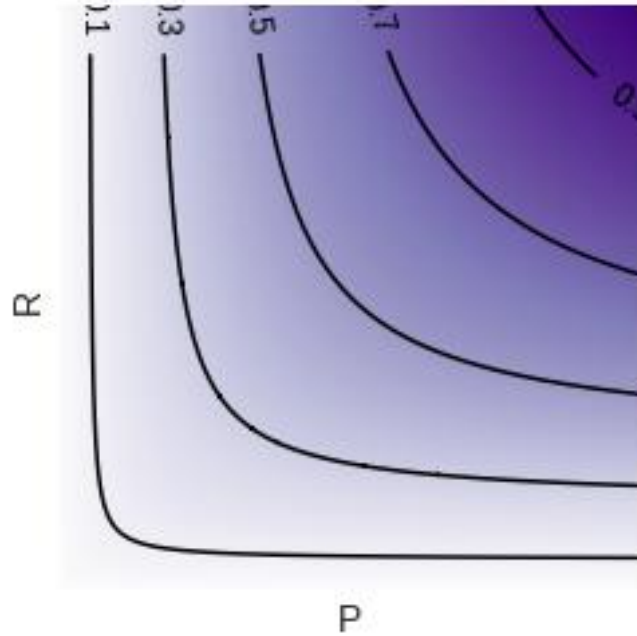
$$2/(1/P + 1/R)$$



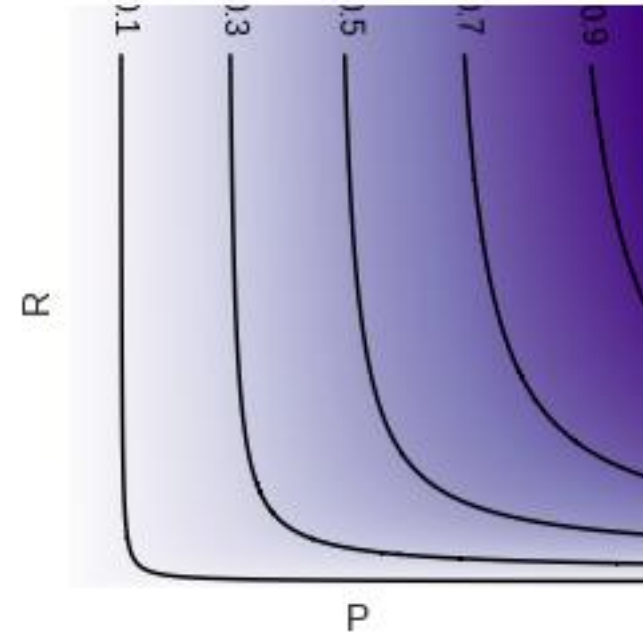
$$\min(P, R)$$

Почему используется F-мера

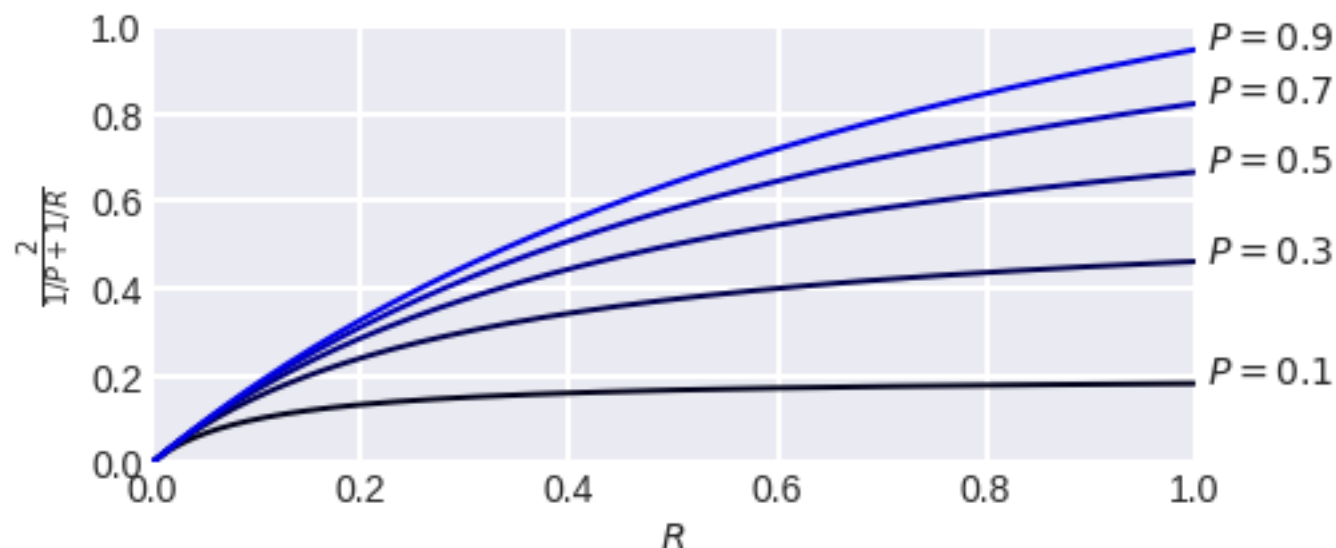
$$2 / (1 / P + 1 / R)$$



$$1 / (0.9 / P + 0.1 / R)$$



Почему используется F-мера



Можно сколь угодно улучшать один из показателей (R), если второй не увеличивается (P), то качество ограничено

Cohen's Kappa в задачах классификации

Chance adjusted index – статистика для измерения согласованности между ответами (p_{observed}) с нормировкой на согласованность по случайности (p_{chance}):

$$r = \frac{p_{\text{observed}} - p_{\text{chance}}}{1 - p_{\text{chance}}}$$

	class 1	class 2
ans 1	n_{11}	n_{12}
ans 2	n_{21}	n_{22}

$$p_{\text{observed}} = \frac{n_{11} + n_{22}}{n}$$

точность – accuracy!

$$p_{\text{chance}} = \frac{n_{11} + n_{12}}{n} \frac{n_{11} + n_{21}}{n} + \frac{n_{21} + n_{22}}{n} \frac{n_{12} + n_{22}}{n}$$

точность по случайности

- – вероятность, что случайно согласован ответ «1»
- – вероятность, что случайно согласован ответ «2»

Cohen's Kappa

смысл: поправка значения точности.

Как раз для решения проблемы дисбаланса классов.

```
from sklearn.metrics import cohen_kappa_score  
cohen_kappa_score(a, y)
```

Cohen's Kappa

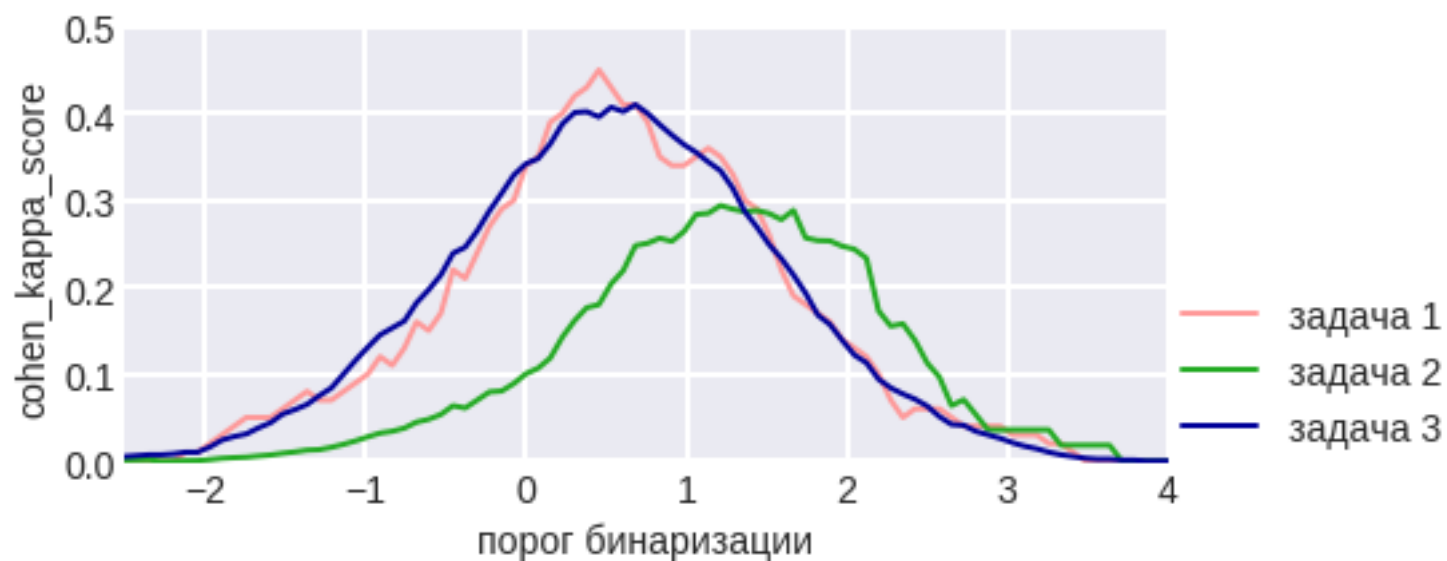
три модельные задачи



Как будет выглядеть график СК от порога бинаризации?
Как меняется ROC AUC?

Cohen's Kappa

график СК от порога бинаризации



ROC AUC: 0.77 во всех задачах!

Weighted kappa

Если есть разумные веса ошибок за конкретные несогласованности

Когда это бывает?

$$K = 1 - \frac{\sum_{i=1}^l \sum_{j=1}^l w_{ij} n_{ij}}{\sum_{i=1}^l \sum_{j=1}^l w_{ij} m_{ij}} \in [-1, +1]$$

**матрица случайных
ответов**

$$m_{ij} = \sum_j n_{ij} \sum_i n_{ij}$$

$$m_{ij} \leftarrow \frac{m_{ij}}{\sum_{ts} m_{ts}} \sum_{ts} n_{ts}$$

квадратичные веса

$$w_{ij} = \frac{(i - j)^2}{(n - 1)^2}$$

м.б. любая весовая схема

Вычисление Quadratic Weighted Кappa

матрица

случайных

ответов

ответы

у а

0	1	1
---	---	---

1	1	1
---	---	---

2	1	2
---	---	---

3	2	1
---	---	---

4	2	3
---	---	---

5	3	2
---	---	---

6	3	3
---	---	---

7	3	3
---	---	---

8	1	2
---	---	---

9	2	2
---	---	---

матрица

ошибок

у а 1 2 3

1	2	2	0
---	---	---	---

2	1	1	1
---	---	---	---

3	0	1	2
---	---	---	---

0 1 2

0	12	16	12
---	----	----	----

1	9	12	9
---	---	----	---

2	9	12	9
---	---	----	---

матрица весов

0 1 2

0	0.00	0.25	1.00
---	------	------	------

1	0.25	0.00	0.25
---	------	------	------

2	1.00	0.25	0.00
---	------	------	------

после нормировки

0 1 2

0	1.2	1.6	1.2
---	-----	-----	-----

1	0.9	1.2	0.9
---	-----	-----	-----

2	0.9	1.2	0.9
---	-----	-----	-----

$$WK = 0.615$$

Вычисление Quadratic Weighted Kappa

```
from sklearn.metrics import cohen_kappa_score  
cohen_kappa_score(df.y, df.a, weights='quadratic')
```

```
n = pd.crosstab(df.y, df.a)  
n = n.values  
m = np.outer(n.sum(axis=1) , n.sum(axis=0))  
m = m / m.sum() * n.sum()  
w = (np.arange(1, 4)[:,np.newaxis] -  
      np.arange(1, 4)) ** 2 / ((3-1)*(3-1))  
1 - (np.sum(n*w) / np.sum(m*w))
```

Quadratic Weighted Kappa

**Применяется в задачах, где классы упорядочены
«ранжирование»**

	y	1.0	0.83	0.83	0.33	0.8	0.0	-1.0
0	0	0	0	0	0	0	0	2
1	0	0	0	0	0	0	1	2
2	0	0	1	0	2	0	2	2
3	1	1	1	1	1	0	0	1
4	1	1	1	1	1	0	1	1
5	1	1	0	2	1	0	2	1
6	2	2	2	2	2	2	0	0
7	2	2	2	2	2	2	1	0
8	2	2	2	1	0	2	2	0

Коэффициент Мэттьюса

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

хорошо для дисбаланса?

Задача бинарной классификации

Теперь выдаём оценку принадлежности к классу 1

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

$$a \in [0, 1]$$

Log Loss

В задаче классификации с двумя непересекающимися классами (0, 1), когда ответ **вероятность** (?) принадлежности к классу 1

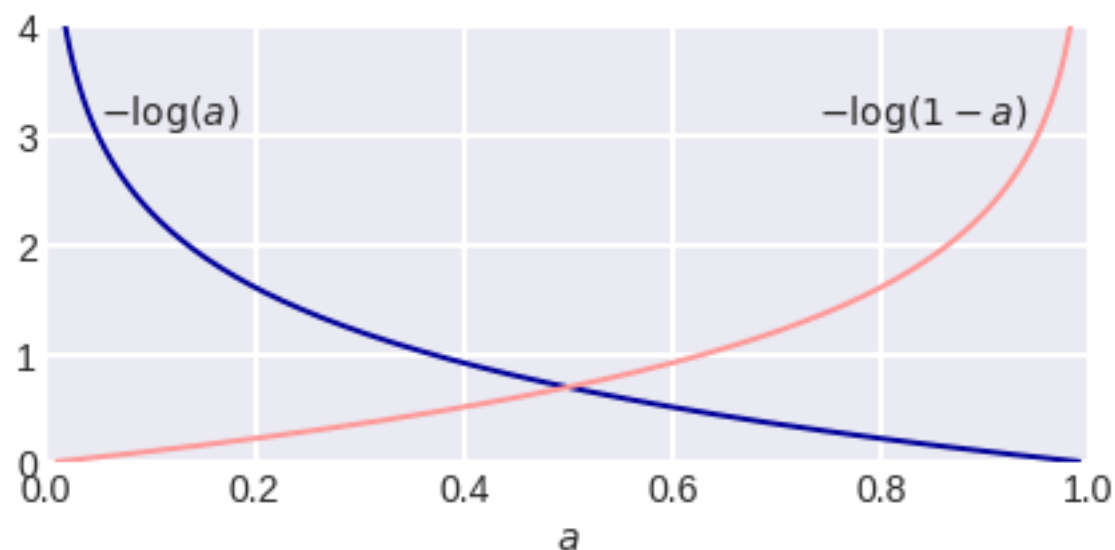
$$\text{LOGLOSS} = -\frac{1}{q} \sum_{i=1}^q (y_i \log a_i + (1 - y_i) \log(1 - a_i))$$

На что похоже?

Так понятнее...

$$-\begin{cases} \log a_i, & y_i = 1, \\ \log(1 - a_i), & y_i = 0. \end{cases}$$

Нельзя ошибаться!



Log Loss

В задаче классификации с двумя непересекающимися классами (0, 1), когда ответ **вероятность** (?) принадлежности к классу 1

$$\text{LOGLOSS} = -\frac{1}{q} \sum_{i=1}^q (y_i \log a_i + (1 - y_i) \log(1 - a_i))$$

На что похоже?

Откуда берётся Log Loss

Обучающая выборка ~ реализация обобщённой схемы Бернулли:

для x_i генерируем

$$y_i = \begin{cases} 1, & p_i, \\ 0, & 1 - p_i. \end{cases}$$

Пусть наша модель генерирует эти вероятности!

$$a_i = a(x_i \mid w)$$

Правдоподобие:

$$p(y \mid X, w) = \prod_i p(y_i \mid x_i, w) = \prod_i a_i^{y_i} (1 - a_i)^{1 - y_i} \rightarrow \max$$

Откуда берётся Log Loss

Максимизация правдоподобия эквивалентна

$$\sum_i (-y_i \log a_i - (1 - y_i) \log(1 - a_i)) \rightarrow \min$$

**Логична ровно настолько, насколько MSE в задаче регрессии
(тоже выводится из ММП)**

Названия

- **логистическая функция ошибки**
 - **«логлосс»**
- **перекрёстная энтропия**
 - **кросс-энтропия**

Log Loss – Оптимальная константа



$$-\frac{1}{q} \sum_{i=1}^q (y_i \log a + (1 - y_i) \log(1 - a)) \rightarrow \min_a$$

$$-\frac{q_1}{q} \log a - \frac{q_0}{q} \log(1 - a) \rightarrow \min_a$$

$$a = \frac{q_1}{q}$$

Интерпретация константного решения

Посчитаем матожидание ошибки –

у нас один (i -й) объект, который с вероятностью p принадлежит классу 1.

$$-p \log(a_i) - (1-p) \log(1-a_i)$$

Минимизируем это выражение:

$$\frac{p}{a_i} - \frac{1-p}{1-a_i} = 0$$

$$a_i = p$$

О чудо!

Но так не всегда...

Вот почему используют `log_loss`

Интерпретация константного решения

Если подставить оптимальное значение $a_i = p$ в

$$-p \log(a_i) - (1-p) \log(1-a_i)$$

получаем энтропию:

$$-p \log(p) - (1-p) \log(1-p)$$

Вот почему используют энтропийный критерий расщепления!

он минимизирует logloss!

Log Loss

В каких пределах варьируется \log_loss ?

Какие недостатки \log_loss ?

Log Loss

В каких пределах варьируется log_loss?

Эффективное изменение в

$$\left[0, -\frac{q_1}{q} \log \frac{q_1}{q} - \frac{q_0}{q} \log \frac{q_0}{q} \right]$$

**Если логарифм по основанию 2,
то на сбалансированной выборке это $[0, 1]$**

Какие недостатки log_loss?

Его значение неинтерпретируемы...

Связь logloss с логистической регрессией

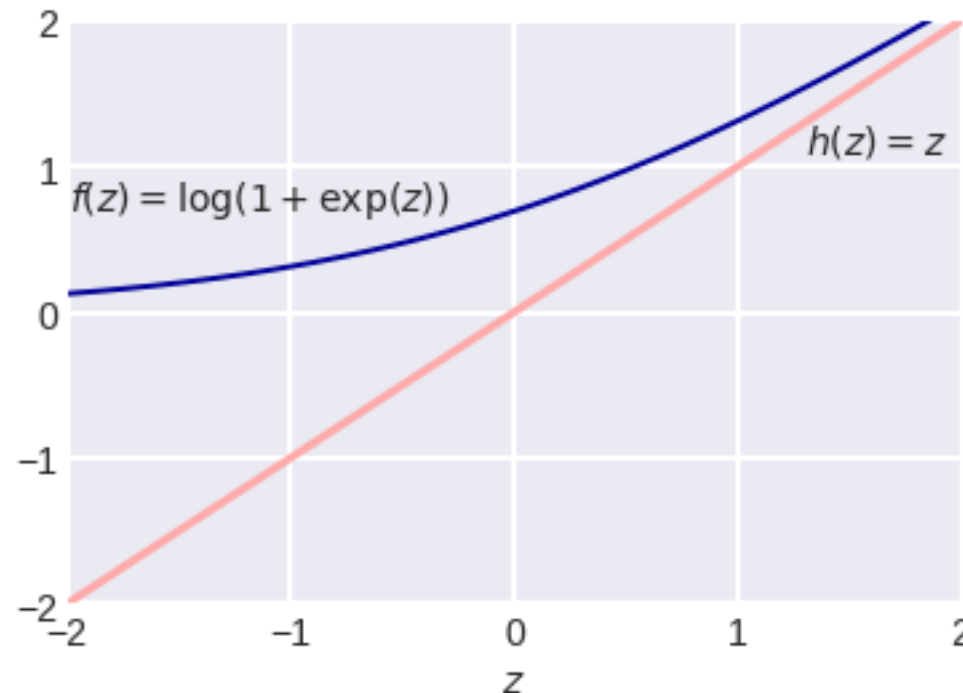
см. лекцию про минимизацию...

Другая форма функционала

Подставим выражение для сигмоиды, сделаем переобозначение:

метки классов теперь -1 и $+1$, тогда

$$\text{logloss}(a, y) = \log(1 + \exp(-y \cdot w^T x))$$



Кстати**SVM**

$$\sum_i \max[1 - y_i w^T x, 0] + \alpha w^T w \rightarrow \min$$

RVM

$$\sum_i \log(1 + \exp(-y_i w^T x)) + w^T \text{diag}(\alpha) w \rightarrow \min$$

Связь logloss с расхождением Кульбака-Лейблера

$$D_{\text{KL}}(P \parallel Q) = \int p(z) \log \frac{p(z)}{q(z)} \partial z$$

$$D_{\text{KL}}(P \parallel Q) = \sum_i P_i \log \frac{P_i}{Q_i}$$

распределение алгоритма: $(1 - a, a)$

истинное: $(1 - y, y)$

расхождение КЛ между ними:

$$(1 - y) \log \frac{(1 - y)}{(1 - a)} + y \log \frac{y}{a} = -(1 - y) \log(1 - a) - y \log a$$

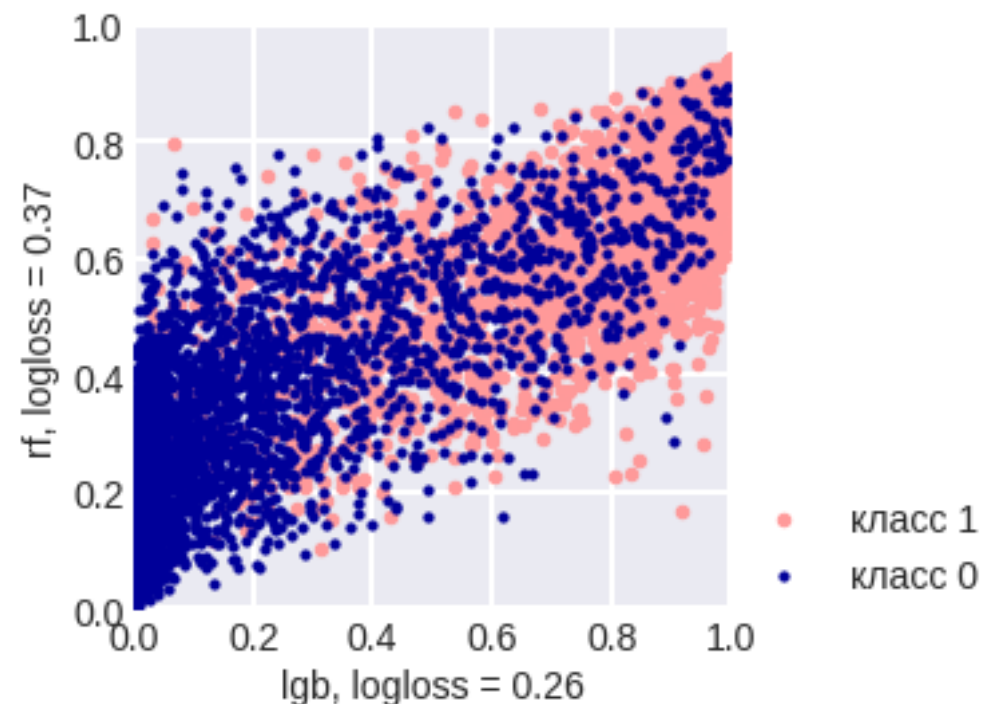
это logloss!!!

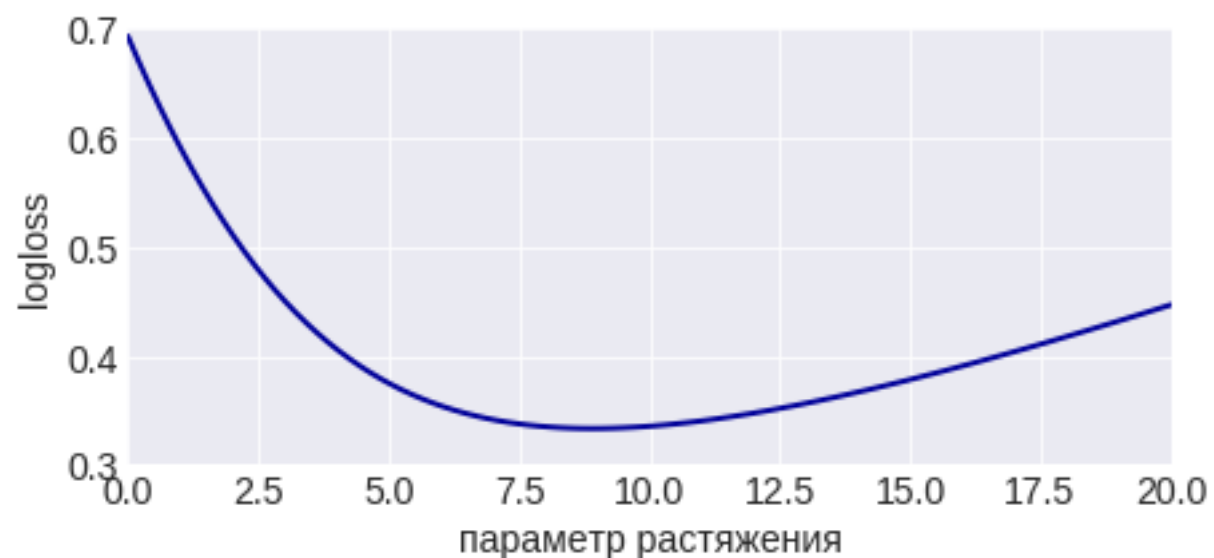
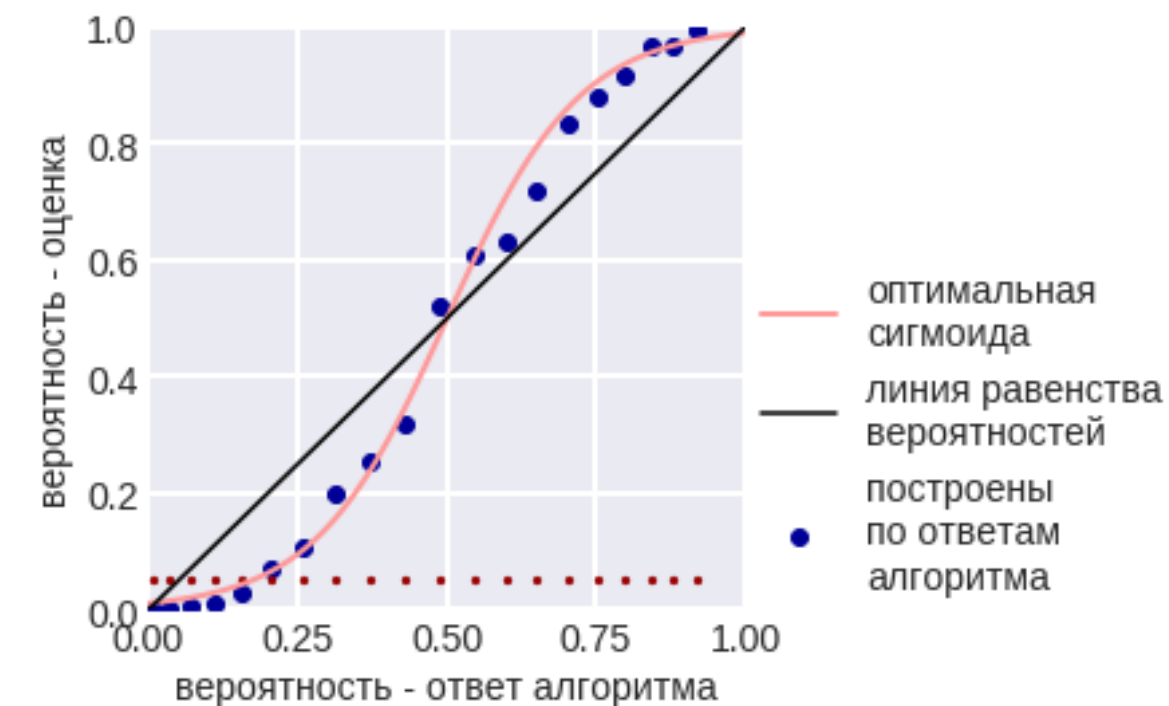
Настройка на Logloss калибровка Платта (Platt calibration)

– для SVM

$$a(x) = \text{sigmoid}(\alpha \cdot r(x) + \beta)$$

ещё есть – монотонная регрессия (Isotonic regression)





Если использовать MSE в задаче классификации

$$L(y, a) = (y - a)^2 = y(1 - a)^2 + (1 - y)a^2$$

**Если объект x с вероятностью p принадлежит классу 1,
то математическое ожидание ошибки**

$$p(1 - a)^2 + (1 - p)a^2$$

**подставляем оптимальный ответ (как делали в logloss,
здесь оптимальный ответ тоже $a = p$):**

$$p(1 - p)^2 + (1 - p)p^2 = p(1 - p)$$

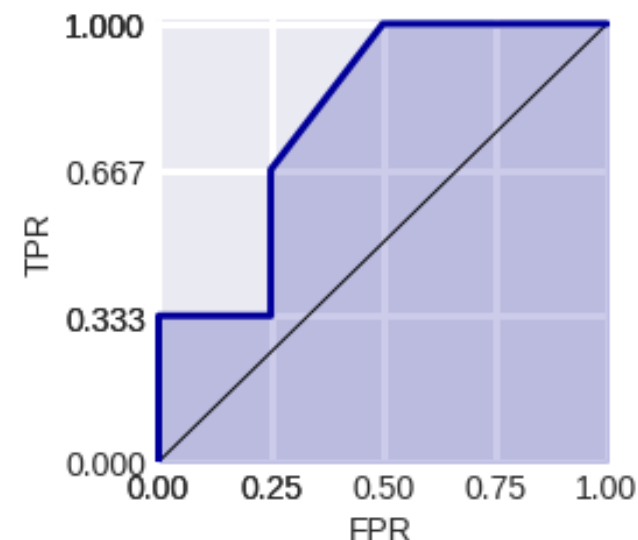
**т.е. критерий расщепления Джини
фактически минимизирует эту функцию ошибки!**

ROC и AUC ROC

Функционал зависит не от конкретных значений, а от их порядка

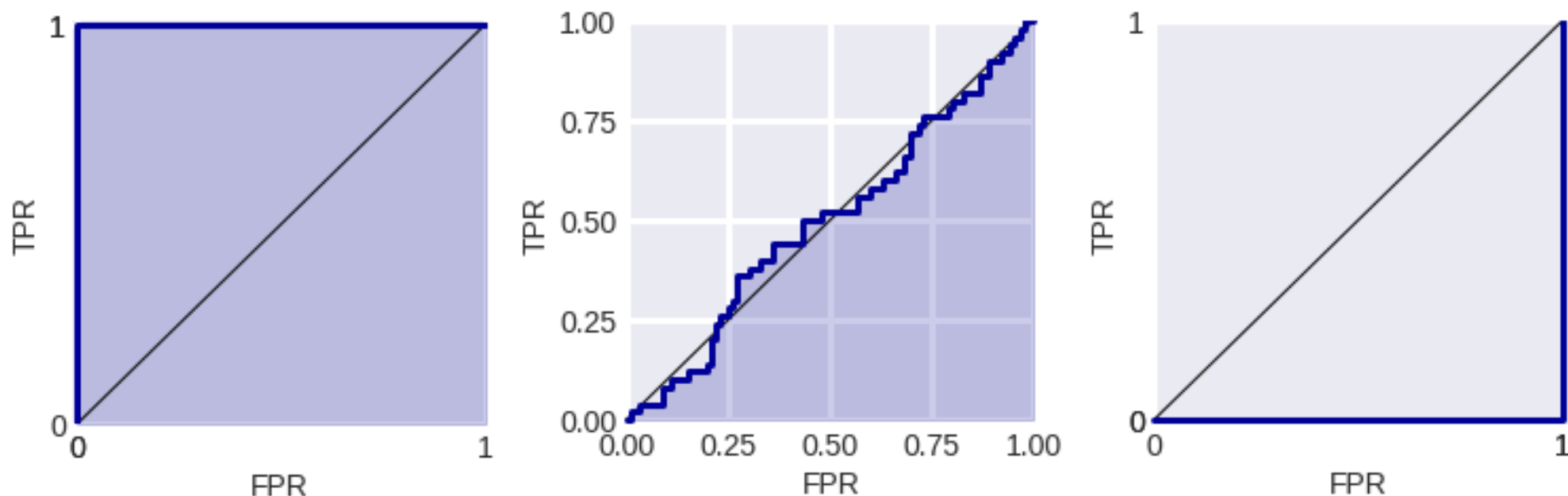
	оценка	класс
0	0.5	0
1	0.1	0
2	0.2	0
3	0.6	1
4	0.2	1
5	0.3	1
6	0.0	0

	оценка	класс	ответ
3	0.6	1	1
0	0.5	0	1
5	0.3	1	1
2	0.2	0	0
4	0.2	1	0
1	0.1	0	0
6	0.0	0	0



```
df['ответ'] = (df['оценка'] > 0.25).astype(int)
df.sort_values('оценка', ascending=False)
```

ROC и AUC ROC



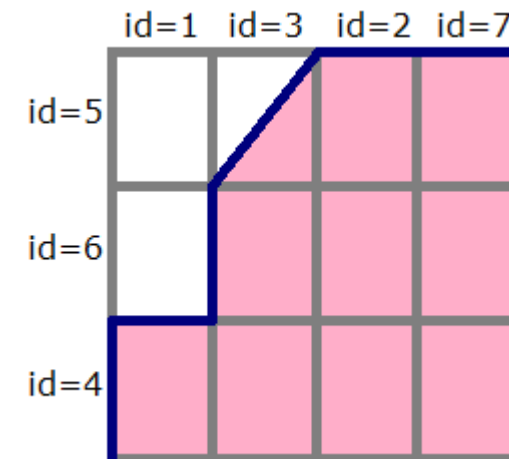
**наилучший (AUC=1), случайный (AUC~0.5) и наихудший (AUC=0)
алгоритма**

Смысл AUC

AUC ~ число правильно отсортированных пар (на рис. «кирпичики»)

Это сложно объяснить заказчику!

$$AUC = \frac{\sum_{i=1}^q \sum_{j=1}^q I[y_i < y_j] I[a_i < a_j]}{\sum_{i=1}^q \sum_{j=1}^q I[y_i < y_j]}$$



Чем хороша эта запись?

Что неправильно (требуется пояснения) в формуле?

Смысл AUC

AUC ~ число правильно отсортированных пар (на рис. «кирпичики»)

Это сложно объяснить заказчику!

$$AUC = \frac{\sum_{i=1}^q \sum_{j=1}^q I[y_i < y_j] I[a_i < a_j]}{\sum_{i=1}^q \sum_{j=1}^q I[y_i < y_j]}$$

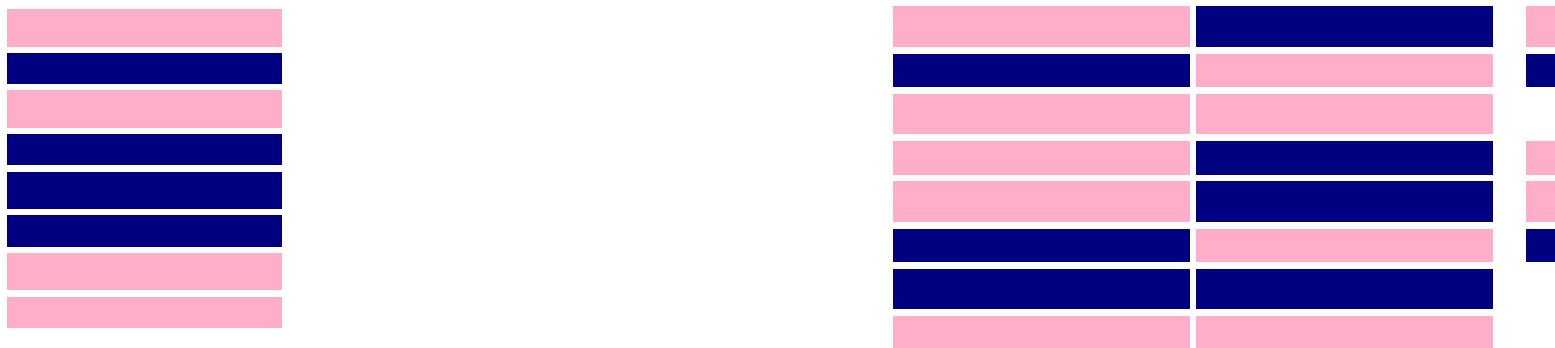
Чем хороша эта запись?

Можно обобщить, например, на регрессию.

$$I[a_i < a_j] = \begin{cases} 1, & a_i < a_j, \\ 1/2, & a_i = a_j, \\ 0, & a_i > a_j. \end{cases}$$

Настройка RF/GBM на AUC ROC

Случай из жизни (Интернет-математика)

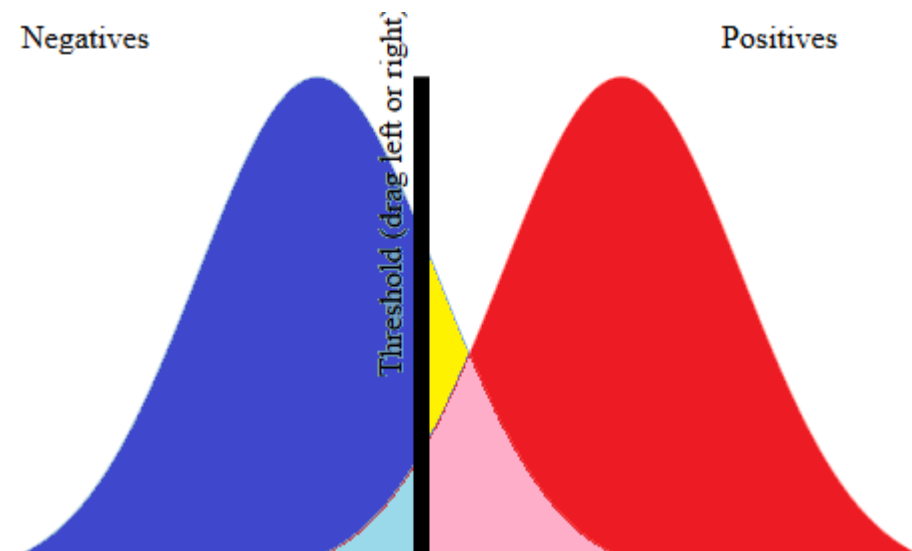


классификация → классификация пар

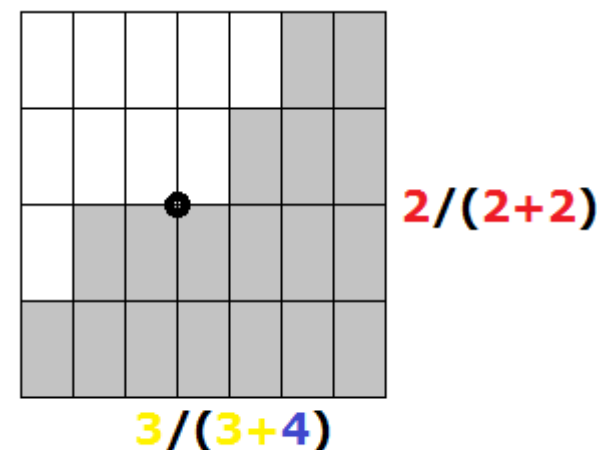
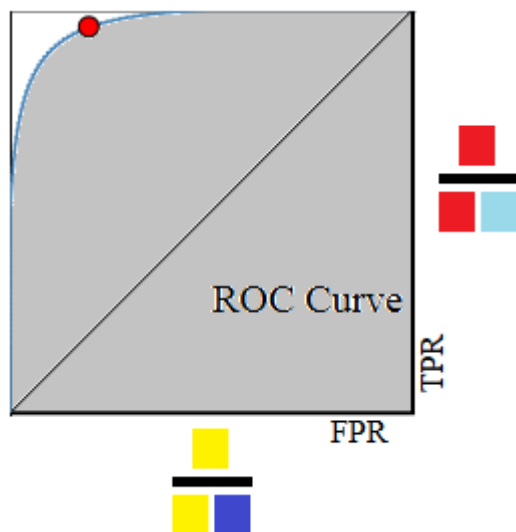
Можно дублировать,

Можно брать разности/отношения.

AUC ROC



	1	0	истина
1			
0			
алгоритм			



0 0 1 0 1 0 0 0 1 0 1

AUC – не всегда ступеньки!

GINI

История... изначально мера расслоения общества относительно какого-нибудь экономического показателя (чаще дохода)

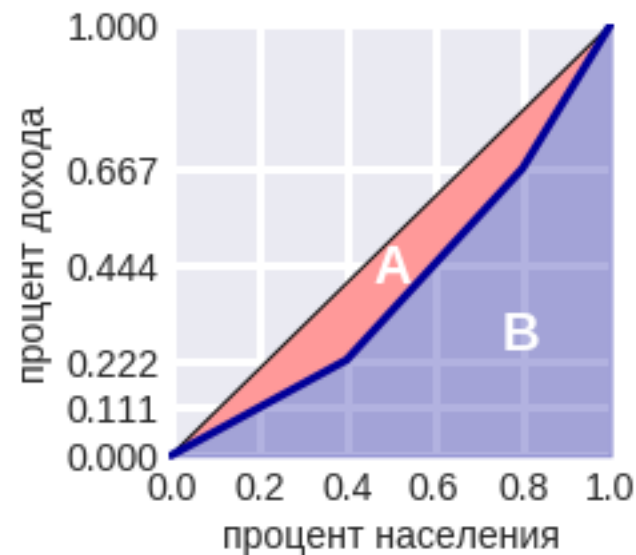


Пример для доходов: 1, 1, 2, 2, 3

40% населения имеют $\frac{2}{9}$ дохода.

GINI

Вычисление



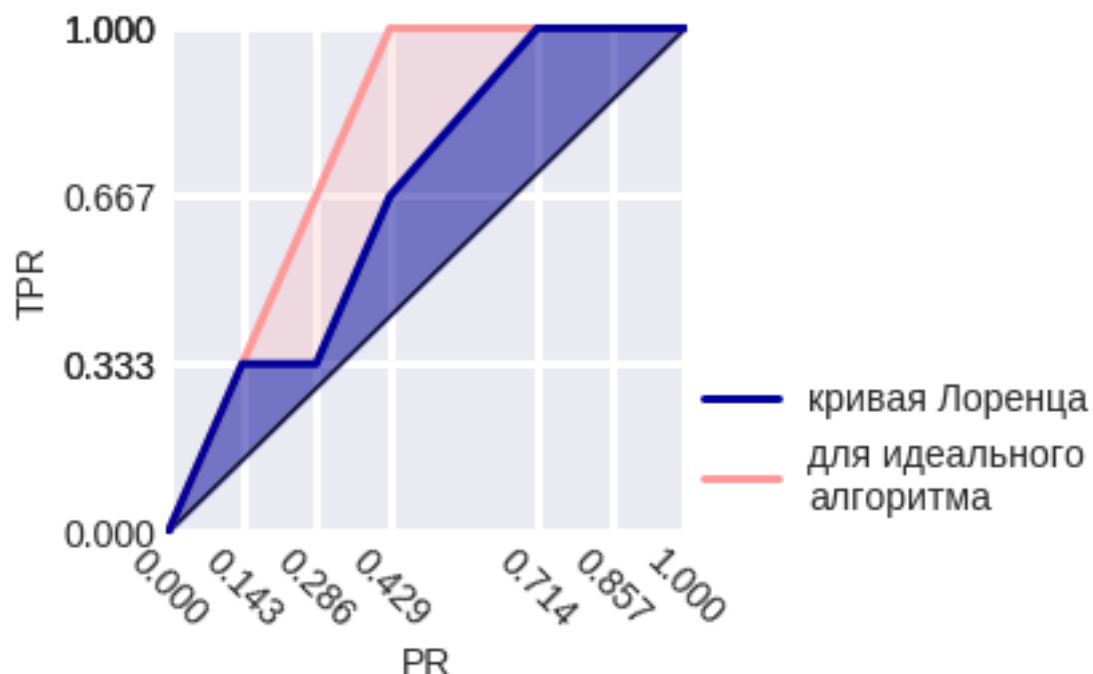
$$\text{gini} = \frac{A}{A + B} = 2A$$

$$\text{gini} = 1 - \sum_{t=1}^m (p_t - p_{t-1})(i_t + i_{t-1}) = 2/9$$

не путать с Gini impurity

GINI в машинном обучении

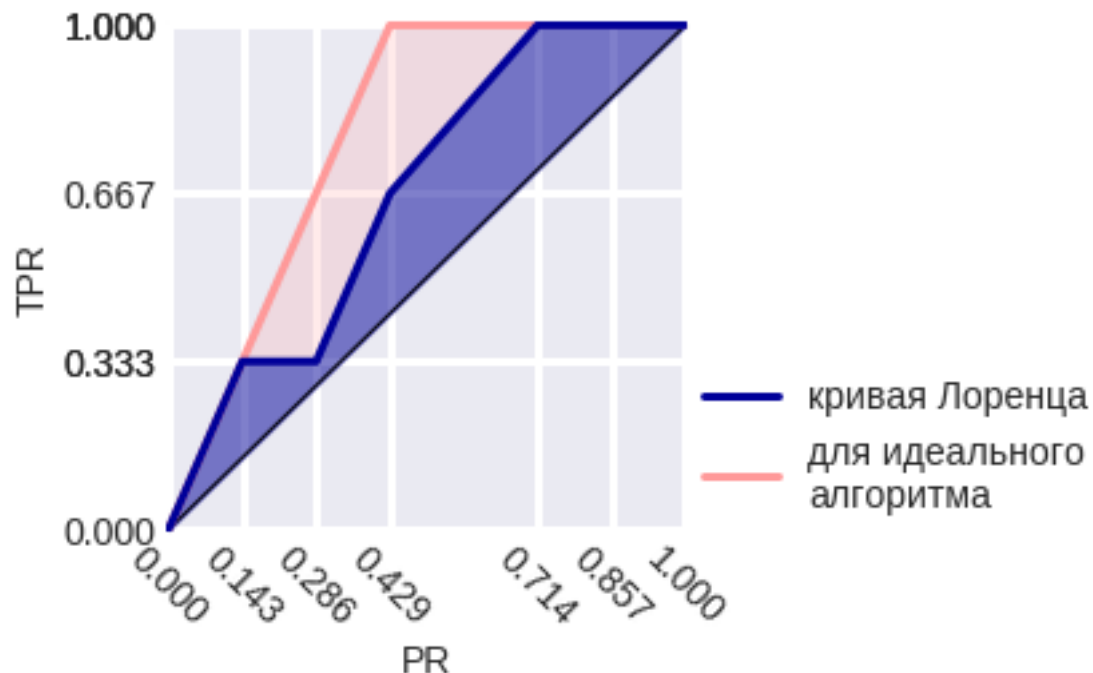
Кривая Лоренца (или CAP – Cumulative Accuracy Profile Curve)



PR = Positive Rate – процент объектов, которые при определённом выборе порога, отнесены к классу 1

Коэффициент Джини – отношение площадей $\frac{\text{blue area}}{\text{blue area} + \text{red area}} = 7/12$

GINI в машинном обучении



$$\text{AUCROC} = \int_0^1 \text{TPR} \partial \text{FPR} = \int_0^1 \frac{\text{TP}}{q_1} \partial \frac{\text{FP}}{q_0} = \frac{1}{q_1 q_0} \int_0^1 \text{TP} \partial \text{FP}$$

$$\text{gini} = \frac{\int_0^1 \text{TPR} \partial \text{PR} - 0.5}{0.5 q_0 / (q_0 + q_1)} = \frac{\int_0^1 \frac{\text{TP}}{q_1} \partial \frac{\text{FP} + \text{TP}}{q_0 + q_1} - 0.5}{0.5 q_0 / (q_0 + q_1)}$$

GINI в машинном обучении

$$\begin{aligned} \text{gini} &= \frac{2}{q_1 q_0} \int_0^1 \text{TP} \partial(\text{FP} + \text{TP}) - \frac{q_0 + q_1}{q_0} = \\ &= 2 \text{AUCROC} + \frac{2}{q_1 q_0} \int_0^1 \text{TP} \partial \text{TP} - \frac{q_1}{q_0} - 1 \end{aligned}$$

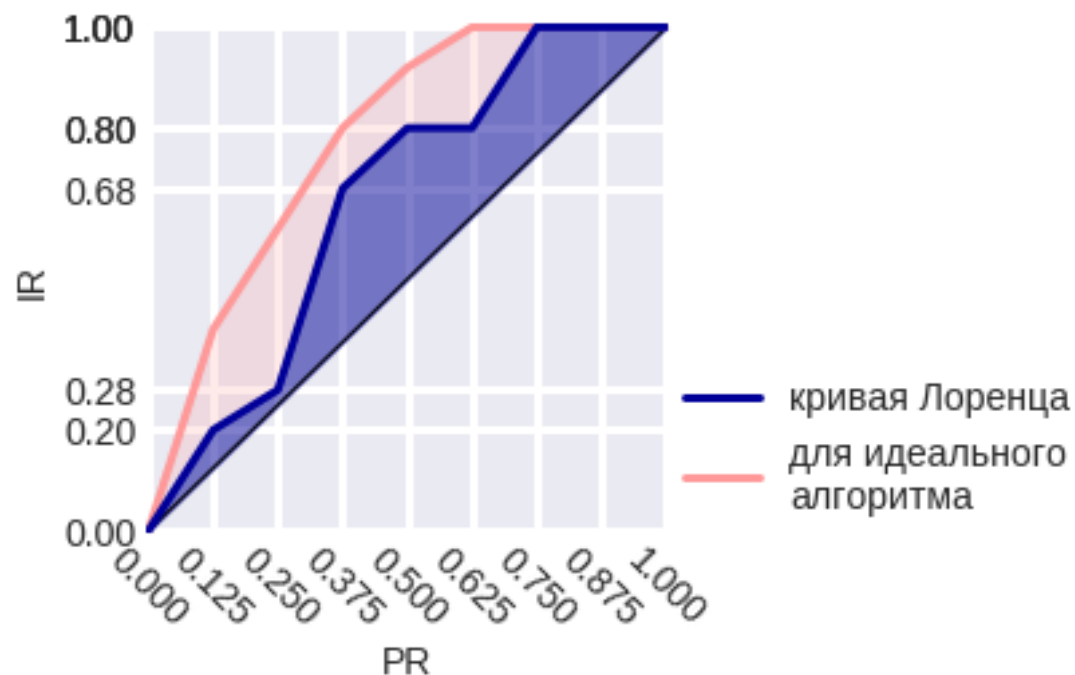
$$\text{gini} = 2 \text{AUCROC} - 1$$

Меняется от -1 до +1 – может сбивать с толку

$$\mathbf{0.9 \text{ AUC} = 0.8 \text{ gini}}$$

GINI в задаче регрессии

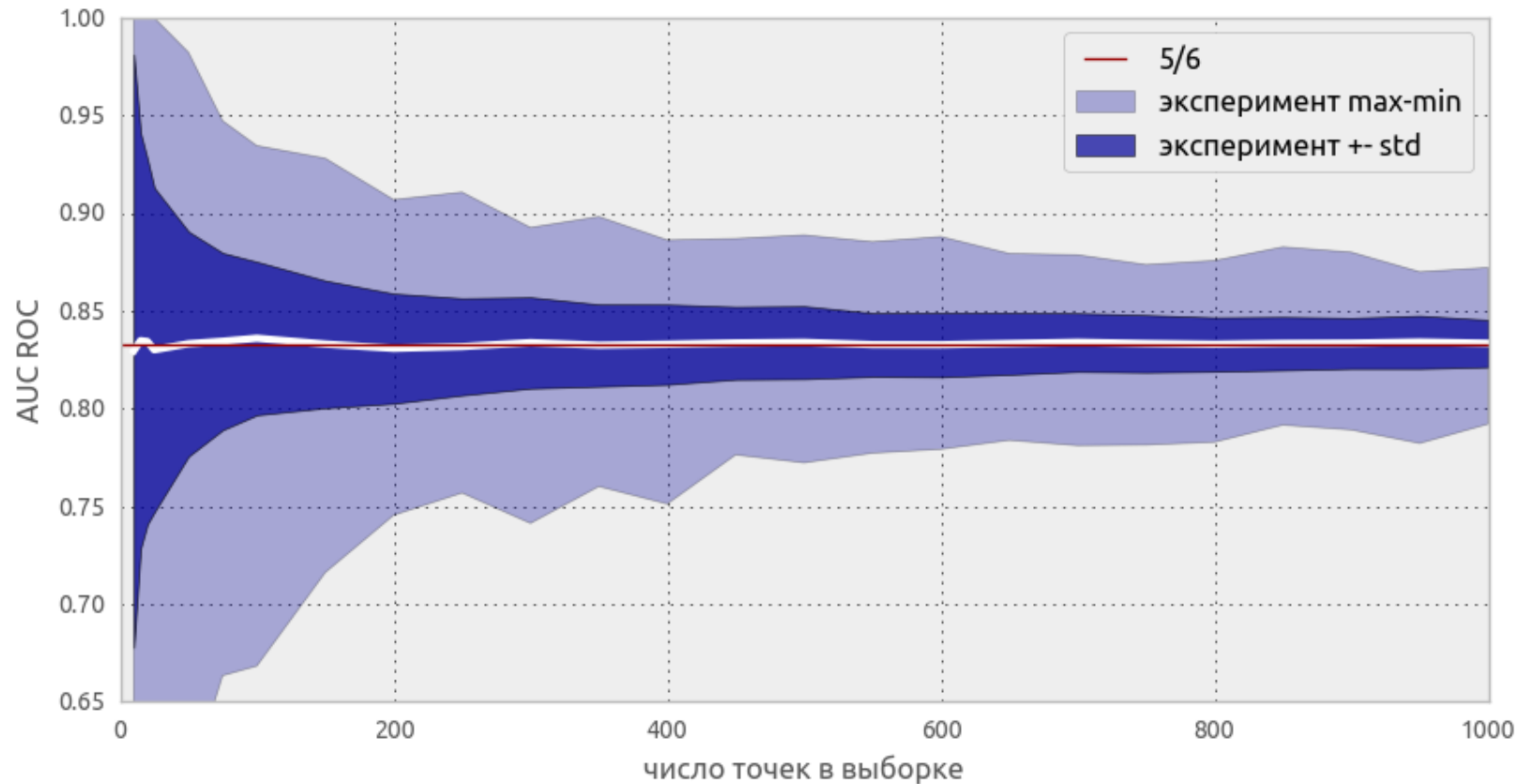
суммы страховых случаев:
0, 0, 5, 0, 3, 10, 2, 5
(так упорядочил алгоритм)



$$\text{gini} \approx 0.57$$

AUC ROC

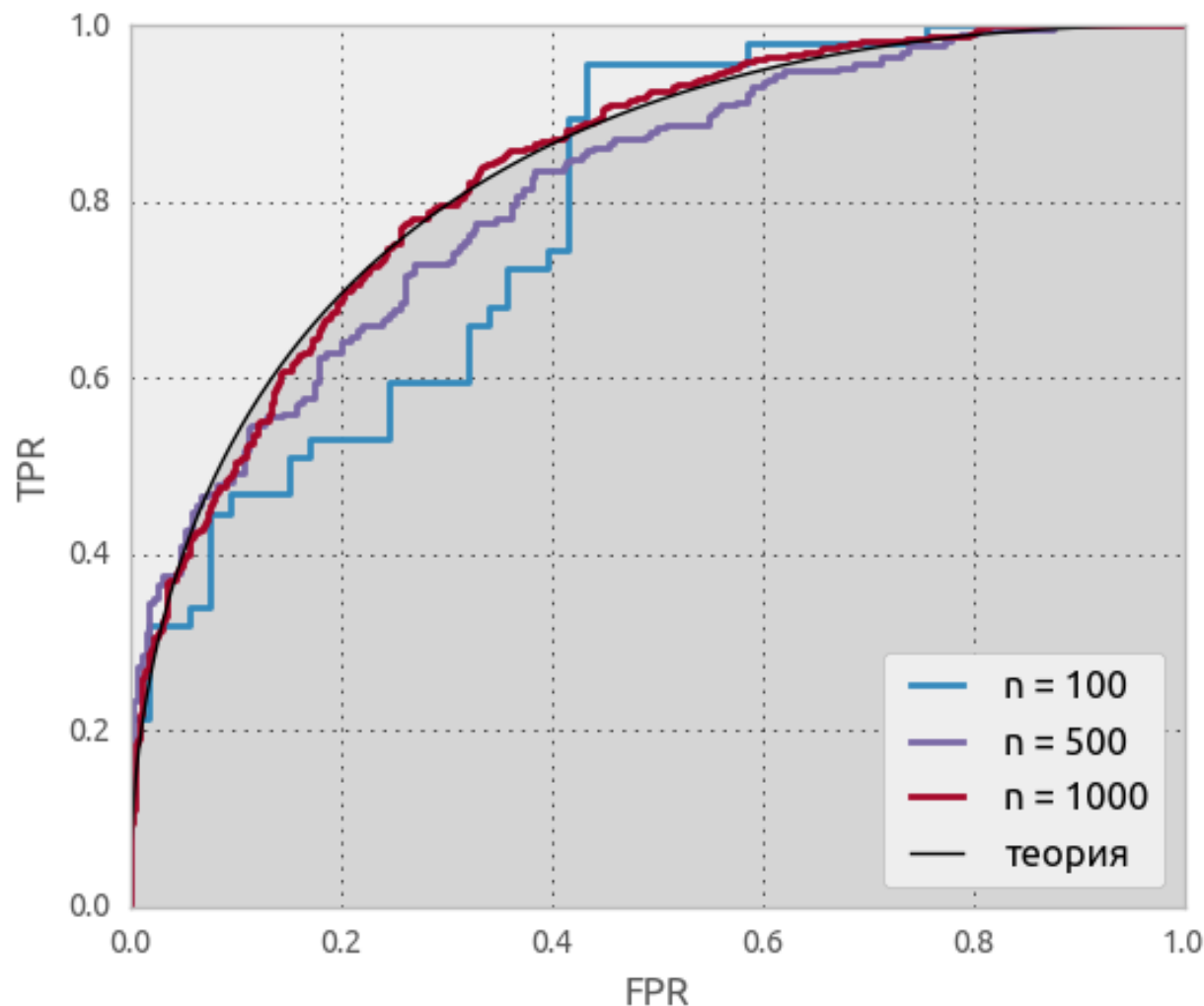
Если задаться распределениями классов (на ответах алгоритма) и получать оценку AUC ROC



Для оценки AUC ROC маленькие выборки не подходят!

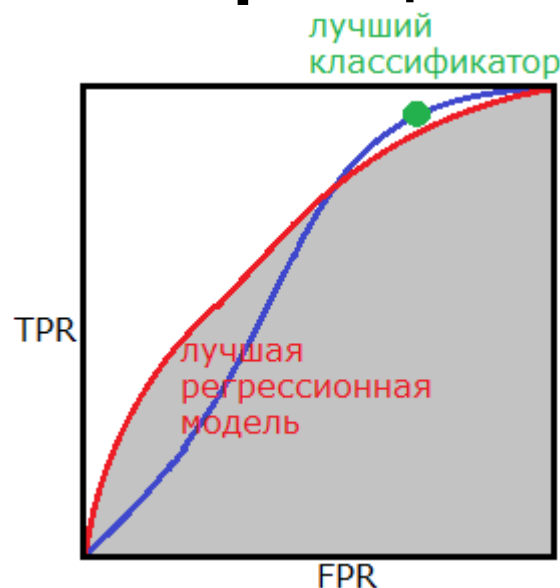
AUC ROC

Если задаться распределениями классов (на ответах алгоритма) и получать оценку AUC ROC

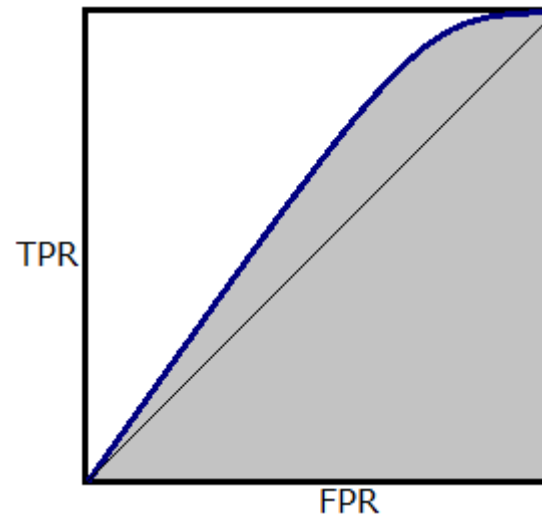


AUC ROC

- + в задачах, где важен порядок
- + учитывает разную мощность классов
- + не важны значения, важен порядок
- + можно использовать для оценки признаков
- «завышает» качество
- оценивает не конкретный классификатор, а регрессию
 - сложно объяснить заказчику
- не путать классификацию и регрессию

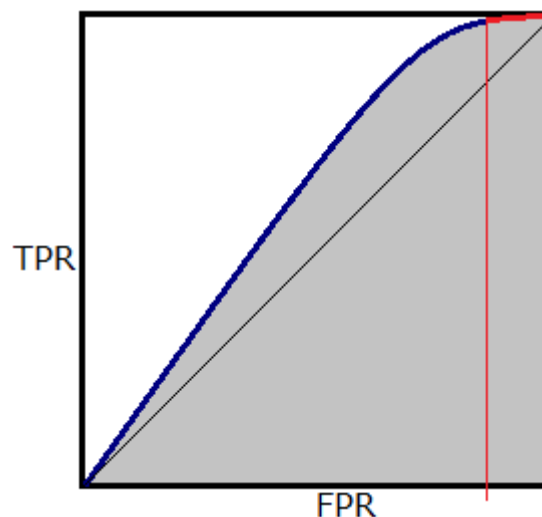


Маленький AUC не всегда плохо...



В каких случаях хороша такая ROC-кривая?

Маленький AUC не всегда плохо...



11010010110...011010100000100000

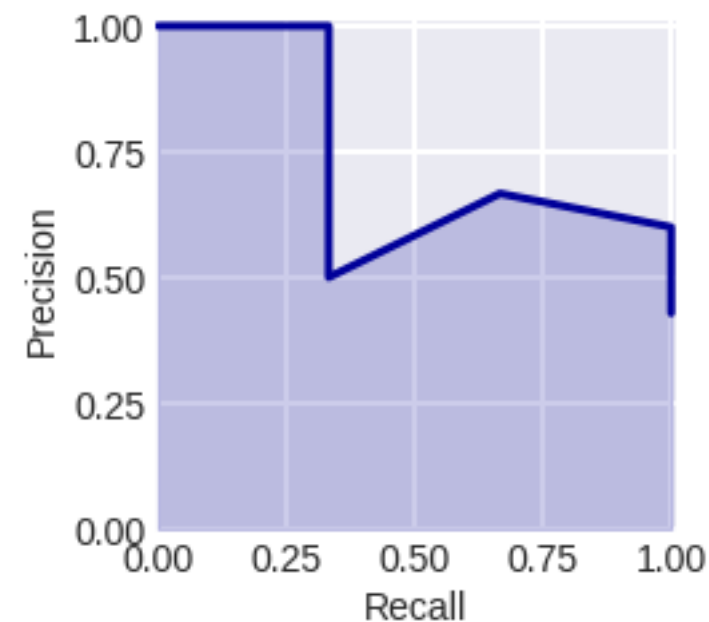
**Мы не можем хорошо решить задачу классификации,
но можем хорошо отделить часть объектов одного класса**

Пример: клиенты, которые не купят билет
(чтобы предложить его им со скидкой)

Ещё примеры кривых... «полнота-точность»

	оценка	класс
0	0.5	0
1	0.1	0
2	0.2	0
3	0.6	1
4	0.2	1
5	0.3	1
6	0.0	0

	оценка	класс	ответ
3	0.6	1	1
0	0.5	0	1
5	0.3	1	1
2	0.2	0	0
4	0.2	1	0
1	0.1	0	0
6	0.0	0	0



Максимизация AUC ROC

- замена индикаторных функций на дифференцируемые
- использование смысла функционала (переход к парам)
 - ансамблирование с ранговой деформацией

Д3 Пройти тест goo.gl/93qkum

Совет

Ищите матожидание!

Пробуйте константные решения.

Многоклассовая задача

Hamming Loss

**Число ошибок в векторе
классификаций**

$$\text{HL}(\tilde{a}, \tilde{y}) = \frac{\|\tilde{a} \oplus \tilde{y}\|}{l}$$

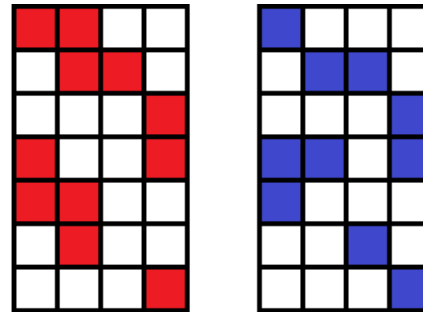
Log Loss

$$\text{LOGLOSS} = -\frac{1}{q} \sum_{i=1}^q \sum_{j=1}^l y_{ij} \log a_{ij}$$

Многоклассовая задача

Полнота и т.п. – всё что придумывается

- по строками матрицы
- по столбцам матрицы



Это множества – и можно усреднять функции сходства множеств

Как использовать на практике (LSHTC)

- Решающее правило с отсечкой:

$$\alpha_{ij} = \begin{cases} 1, & \gamma_{ij} \geq \min(c, \max\{\gamma_{ij}\}_{j=1}^l), \\ 0, & \text{иначе.} \end{cases}$$

- Решать задачу по вертикали / по горизонтали

Функционал в LSHTC

$$\tilde{F} = \frac{2\tilde{P}\tilde{R}}{\tilde{P} + \tilde{R}}$$

Id	Predicted
1,12	35 200
2,54	55
3,11	
4,1	7 101
...	

$$\tilde{P} = \frac{1}{l} \sum_{j=1}^l \frac{TP_j}{TP_j + FP_j}$$

$$\tilde{R} = \frac{1}{l} \sum_{j=1}^l \frac{TP_j}{TP_j + FN_j}$$

Оценка результатов поиска/рекомендаций



Задача с бинарной релевантностью

$$x_1 \prec x_2 \prec \dots \prec x_q$$

$y_i = 1$ – релевантный объект

$y_i = 0$ – нерелевантный объект

Задача ранжирования

Целевой признак может быть бинарным, но это не задача классификации

Precision at n

Точность на первых n элементах

$$p @ n = \frac{y_1 + \dots + y_n}{n}$$

Average Precision at n

Средняя точность на первых n элементах

$$ap @ n = \sum_{k=1}^n \frac{P(k)}{\min(n, m)}$$

**m – мощность множества релевантных объектов
(товаров, документов)**

n – сколько рекомендаций будет учитываться

$$P(k) = \begin{cases} p @ k, & y_k = 1, \\ 0, & y_k = 0, \end{cases}$$

y_i – бинарное значение релевантности

Average Precision at n

Примеры (три релевантных объекта):

$$0 \prec 0 \prec 0$$

$$ap @ 3 = \frac{1}{3} [0 + 0 + 0]$$

$$0 \prec 0 \prec 1$$

$$ap @ 3 = \frac{1}{3} \left[0 + 0 + \frac{1}{3} \right]$$

$$0 \prec 1 \prec 1$$

$$ap @ 3 = \frac{1}{3} \left[0 + \frac{1}{2} + \frac{2}{3} \right]$$

$$1 \prec 0 \prec 0$$

$$ap @ 3 = \frac{1}{3} \left[\frac{1}{1} + 0 + 0 \right]$$

$$0 \prec 0 \prec 1 \prec 1 \prec 1$$

$$ap @ 5 = \frac{1}{3} \left[0 + 0 + \frac{1}{3} + \frac{2}{4} + \frac{3}{5} \right]$$

$$1 \prec 1 \prec 1 \prec 0 \prec 0$$

$$ap @ 5 = \frac{1}{3} \left[\frac{1}{1} + \frac{2}{2} + \frac{3}{3} + 0 + 0 \right]$$

Mean Average Precision

– усреднение $ap@n$ по всем пользователям

Concordant – Discordant ratio

$$\frac{|\{(i, j) \mid y_i > y_j, 1 \leq i < j \leq n\}|}{|\{i \mid y_i = 1\}| \cdot |\{j \mid y_j = 0\}|}$$

Упорядочили: E, D, C, B, A (по убыванию релевантности)

На самом деле: B, E – релевантные

Пары «нерелевантный» – «релевантный»:

BA

EA

BC

EC

BD

ED

Качество упорядочивания: 4 / (2 + 4)

Что ещё может встретиться... в задачах рекомендации

$$\frac{1}{|Z|} \sum_{z \in Z} \frac{|\{x_1, \dots, x_{h(z)}\} \cap \{x'_1, \dots, x'_{h(z)}\}|}{h(z)}$$

x_1, \dots, x_n – **упорядоченный** список ответов

x'_1, \dots, x'_m – **все релевантные**

$$Z \subseteq \{1, 2, \dots, n\}$$

$$Z = \{5, 10, 15, 20, 25, 30\}$$

когда логично применить?

Рекомендации



Рitmix RH-126M, Black Green наушники

Новинка

🔍 Сравнить ❤️ В избранное ➦ Поделиться
















Цвет: черный, зеленый

Тип: Наушники
 Модель: 15119162
 Тип соединения: Проводные
 Вид наушников: Вставные (внутриканальные), Спортивные
 Конструкция наушников: Динамические, С микрофоном

[Перейти к описанию](#)

Ritmix

Рекомендуем также

 от 115 ₽ Ritmix RH-120M наушники Все варианты	 от 282 ₽ Ritmix RH-180M наушники Все варианты	 289 ₽ Ritmix RH-158, Dark Venge наушники В корзину	 от 145 ₽ Ritmix RH-125 наушники Все варианты	 183 ₽ Ritmix RH-115M Luminous наушники В корзину
 от 220 ₽ Ritmix RH-150M наушники Все варианты	 709 ₽ Бестаркий и чувства В корзину	 1 032 ₽ Ritmix RH-567M Gaming игровые наушники В корзину	 934 ₽ Ritmix RH-565M Gaming игровые наушники В корзину	 65 ₽ Ritmix RH-011 наушники Все варианты
 47 ₽ Ritmix RH-004 наушники Все варианты	 65 ₽ Ritmix RH-011, Black наушники В корзину	 45 ₽ Ritmix RH-003 наушники Все варианты	 4 790 ₽ Ritmix RBK-615 электронная книга В корзину	 1 199 ₽ Ritmix RCH-108 WH автомобильный держатель В корзину

Mean Reciprocal Rank (MRR)

– это усреднение Reciprocal rank (RR) по всем ранжированиям, который сделал алгоритм.

$$RR = \frac{1}{\arg \min_i y_i}$$

неправильно

Часто оптимизируют именно его!

Классические функционалы в поиске

Случай небинарной релевантности

Выдали id документов/товаров/..., а их ценность (релевантность):

$$y_1, \dots, y_q$$

Cumulative Gain at n

$$CG@n = y_1 + \dots + y_n$$

Discounted Cumulative Gain at n

$$DCG@n = \sum_{i=1}^n \frac{2^{y_i} - 1}{\log_2(i + 1)}$$

Ещё вариант:

$$DCG@n = y_1 + \sum_{i=2}^n \frac{y_i}{\log_2(i)} = y_1 + y_2 + \frac{y_3}{\log_2 3} + \dots + \frac{y_n}{\log_2 n}$$

Цена ошибок за неправильное ранжирование

$$\frac{1}{\log_2(1+1)} - \frac{1}{\log_2(1+2)} \approx 0.37$$

$$\frac{1}{\log_2(1+10)} - \frac{1}{\log_2(1+11)} \approx 0.01$$

$$\frac{1}{\log_2(1+10)} - \frac{1}{\log_2(1+20)} \approx 0.06$$

Normalized DCG

$$nDCG = \frac{DCG}{IDCG}$$

IDCG = ideal DCG

для того, чтобы не было зависимости от длины выдачи

Ещё подход к сравнению порядков:

Пусть алгоритм выдал

$$x_1 \prec x_2 \prec \dots \prec x_q$$

Правильный порядок

$$x_{i_1} \prec x_{i_2} \prec \dots \prec x_{i_q}$$

Надо сравнить:

$$(1, 2, \dots, q)$$

$$(i_1, i_2, \dots, i_q)$$

Ранговые корреляции...

Ещё подход к оценке ранжирования

Известны вероятности того, что объект является релевантным

$$p_i = p(x_i)$$

~ пользователь выберет ссылку

Expected reciprocal rank (ERR)

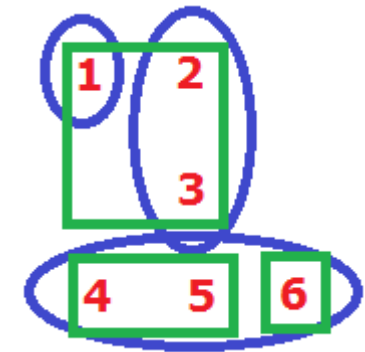
$$ERR @ n = \frac{1}{n} \sum_{k=1}^n \frac{1}{k} p_k \prod_{i < k} (1 - p_i)$$

Как интерпретировать?

Редакторское расстояние

Операции

добавление к кластеру
 создание кластера с одним объектом
 удаление из кластера
 удаление кластера с одним объектом



1 2 3; 4 5; 6

1 2 3; 4 5 [delC]

2 3; 4 5 [del]

2 3; 4 5; 1 [insC]

2 3; 4 5 6; 1 [ins]

	2 3	4 5 6	1
1 2 3	1	6	2
4 5	4	1	3
6	3	2	2

Редакторское расстояние

- Плохо заносить не в тот кластер (целых две операции на перенос)
 - Плохо создавать неправильный кластер

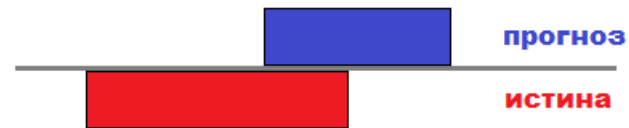
⇒ осторожный алгоритм



- Многое зависит от операций...

Задача с «неклассическим целевым вектором»

**Надо предсказывать не значение,
а интервал $[a, b]$**



Как измерить качество?

Задача с интервальным целевым вектором



Интервал – это множество!

Коэффициент Жаккара (Jaccard)

$$\frac{|A \cap B|}{|A \cup B|}$$

коэффициент Шимкевича-Симпсона (Szymkiewicz, Simpson)

$$\frac{|A \cap B|}{\min(|A|, |B|)}$$

коэффициент Браун-Бланке (Braun-Blanquet)

$$\frac{|A \cap B|}{\max(|A|, |B|)}$$

См. википедию «Коэффициент сходства» для переноса идеи Колмогорова об обобщённом среднем...

Вариации на тему усреднения...

**коэффициент Сёренсена
(Sørensen)**

$$\frac{2|A \cap B|}{|A| + |B|}$$

**коэффициент Кульчинского
(Kulczynsky)**

$$\frac{|A \cap B|}{2} \frac{1}{1/|A| + 1/|B|}$$

коэффициент Отиаи (Ochiai)

$$\frac{|A \cap B|}{\sqrt{|A| \cdot |B|}}$$

Меры включения

$$\frac{\frac{|A \cap B|}{|A|}}{\frac{|A \cap B|}{2|A| - |A \cap B|}}$$

$$\frac{\frac{|A \cap B|}{|B|}}{\frac{|A \cap B|}{2|B| - |A \cap B|}}$$

Как решать задачи с интервалами? Потом вернёмся...

Литература

Tom Fawcett An introduction to ROC analysis // Pattern Recognition Letters Volume 27 Issue 8, 2006, P. 861-874.

<https://ccrma.stanford.edu/workshops/mir2009/references/ROCintro.pdf>

Стрижов В.В. Функция ошибки в задачах восстановления регрессии // Заводская лаборатория, 2013, 79(5): 65-73.

<http://strijov.com/papers/Strijov2012ErrorFn.pdf>

К.Д. Маннинг, П. Рагхаван, Х. Шютце «Введение в информационный поиск» // . — Вильямс, 2011.

Jeffrey M Girard «Inter-observer reliability» //

<https://github.com/jmgirard/mReliability/wiki>