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!

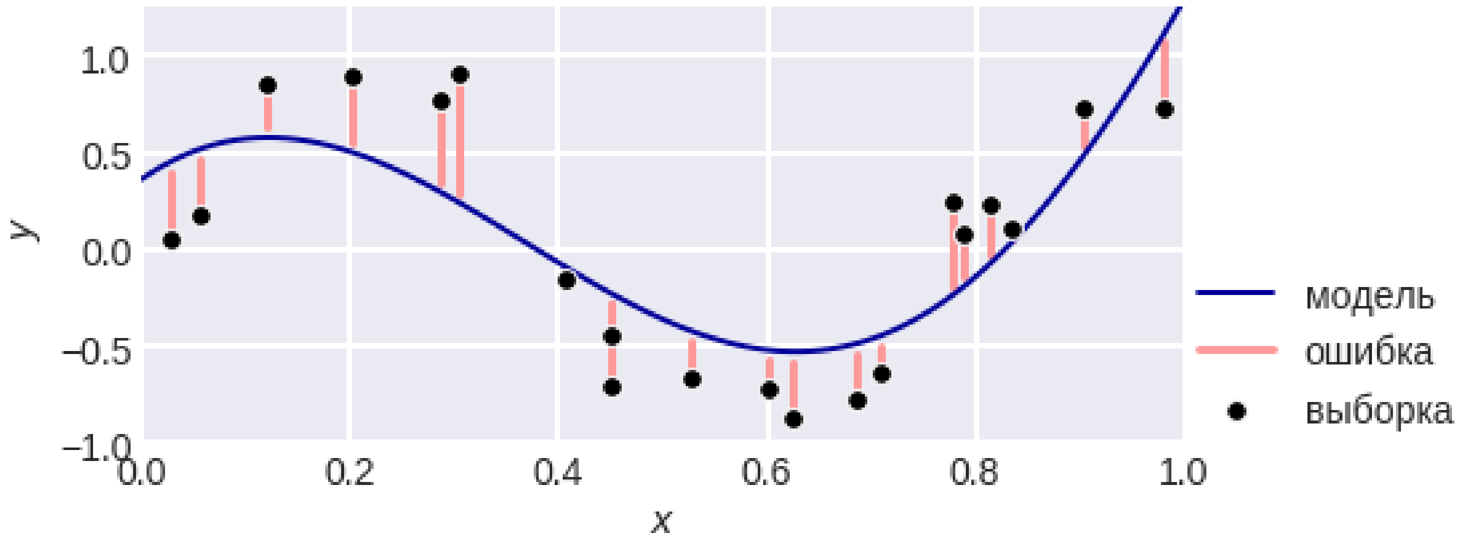
а что такое решение!

В анализе данных:

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

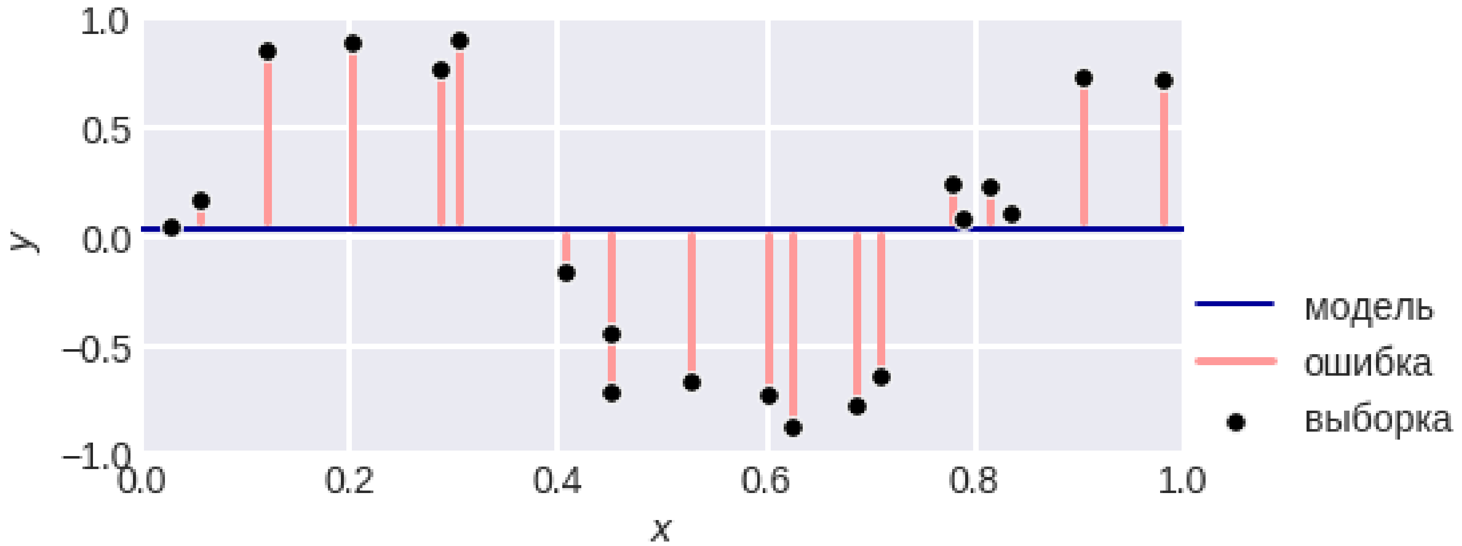
Случай из практики: задача про траектории зрачка
(задача с 3 классами, а не с двумя)

Задача регрессии



Задача регрессии

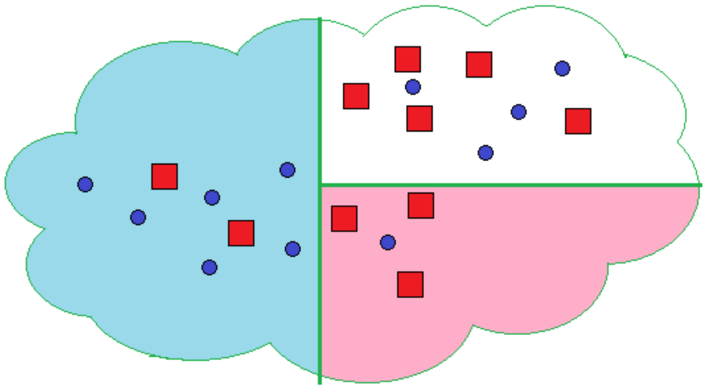
Будем дальше пытаться всё решать в классе констант



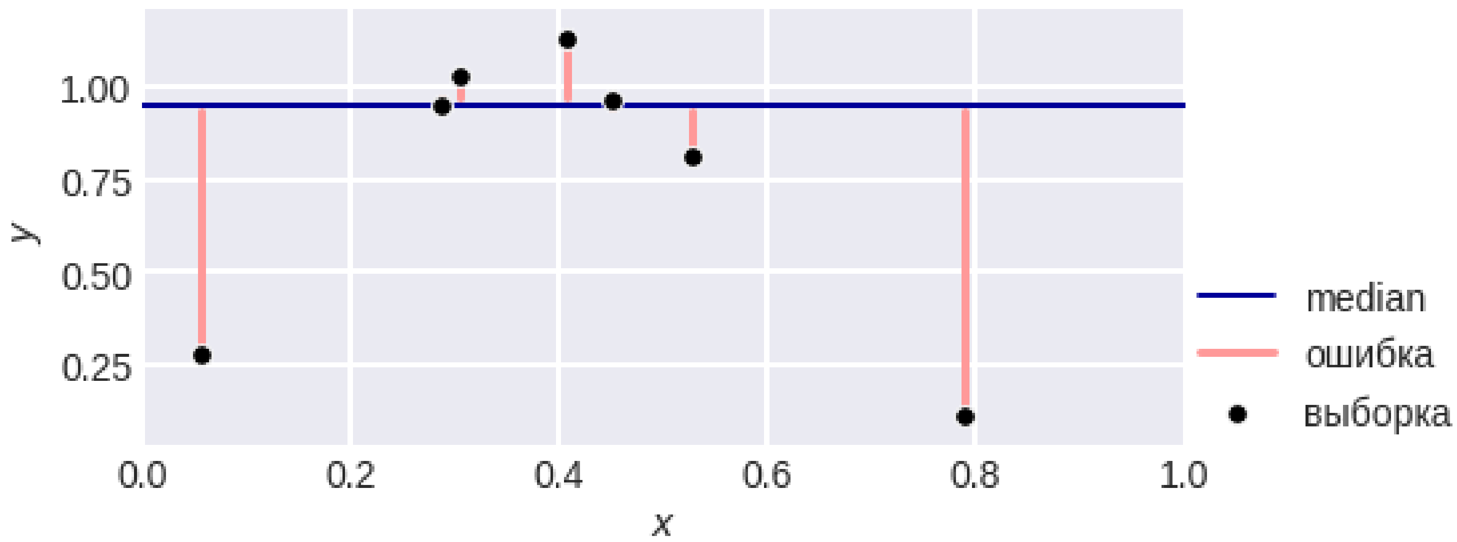
1. Простейшее решение

2. Примерно это и происходит в листьях решающих деревьев

3. Раскрывает природу функционалов



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



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

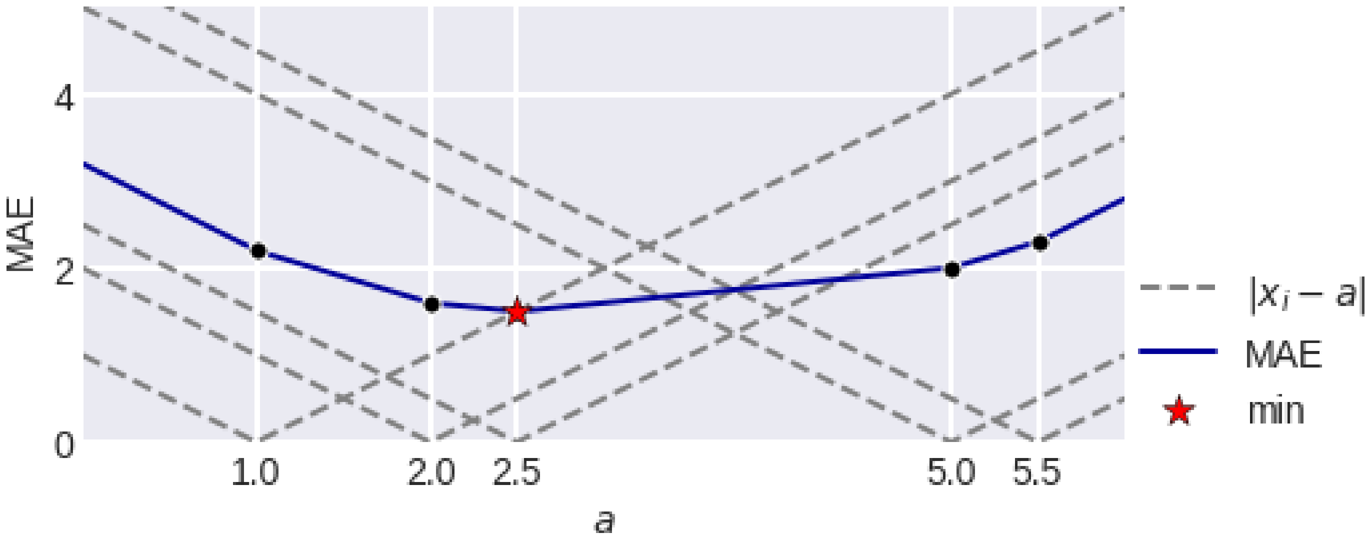
Напоминание:

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

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

Это открывает смысл решений!

Средний модуль отклонения



Средний модуль отклонения

Способы использования тайных знаний:

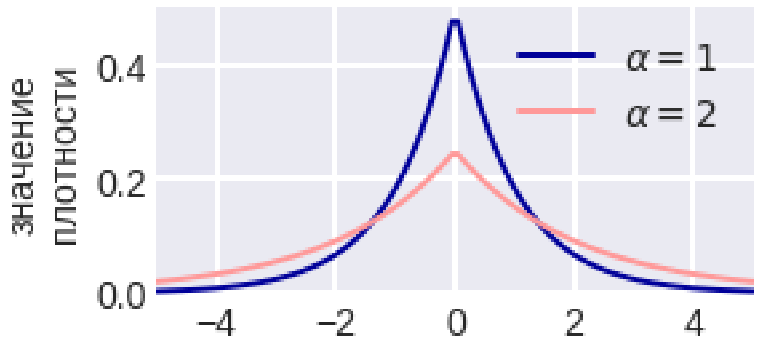
- **медиана, вместо усреднения, в ансамбле**
- **округление ответа (если целевой вектор целочисленный)**

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!

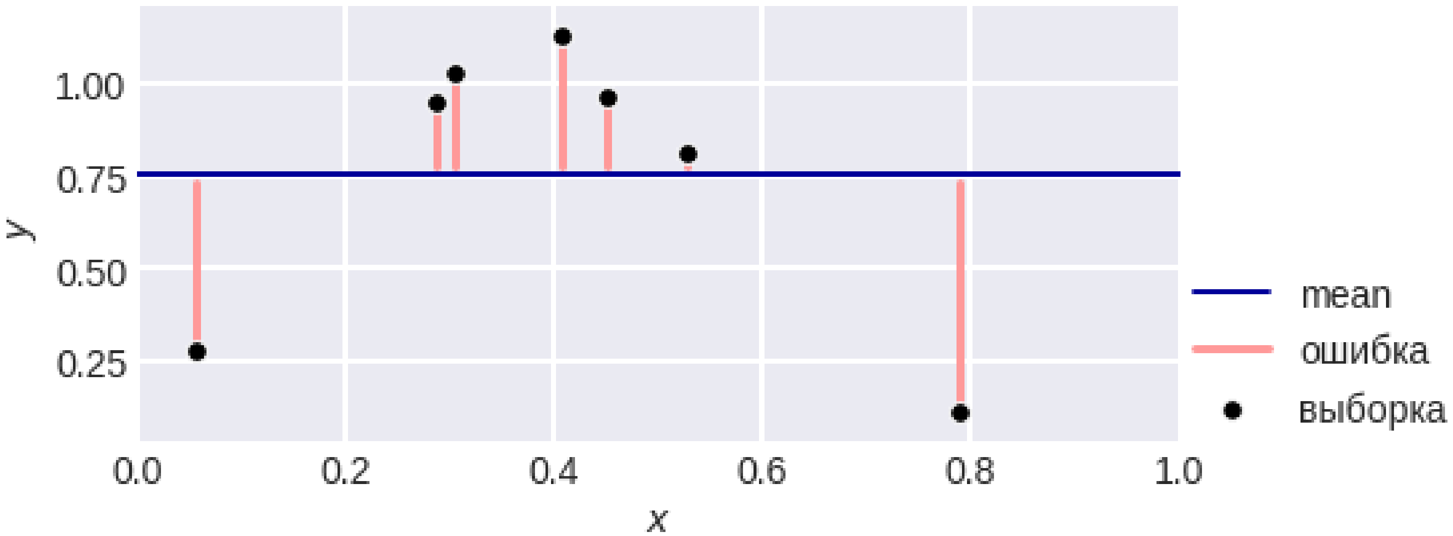
Чему соответствует минимизация весового MAE?

Средний квадрат отклонения ~ Mean Squared Error (MSE)

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

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

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

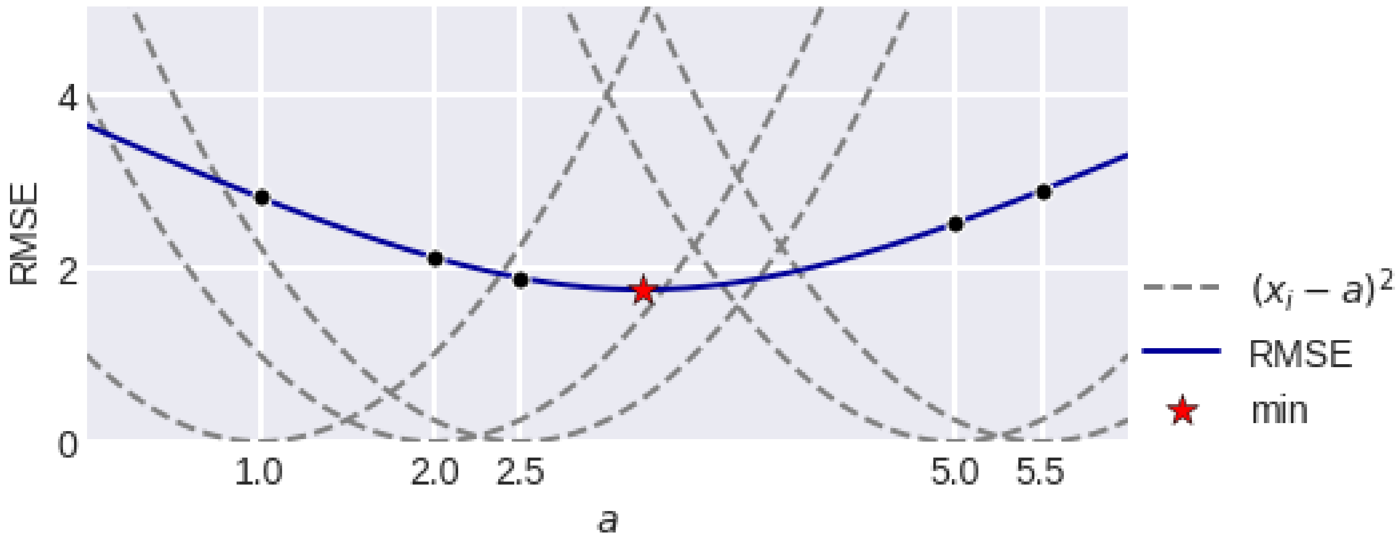


Root Mean Squared Error (RMSE)

или 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)

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

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

В общем случае (в статистике) коэффициент детерминации:

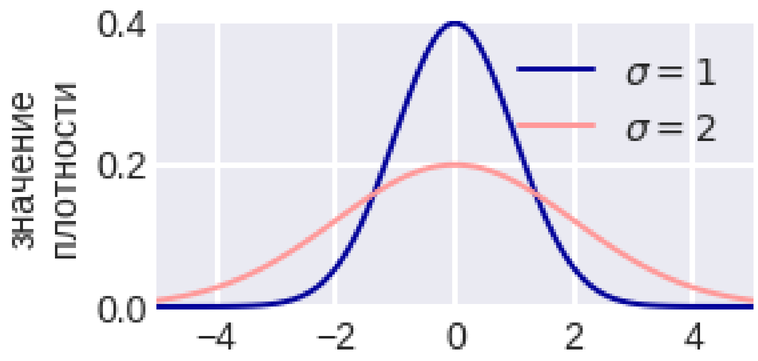
$$R^2 = 1 - \frac{\mathbf{D}(y \mid x)}{\mathbf{D}(y)}$$

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!

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

3. достаточно гладкая…

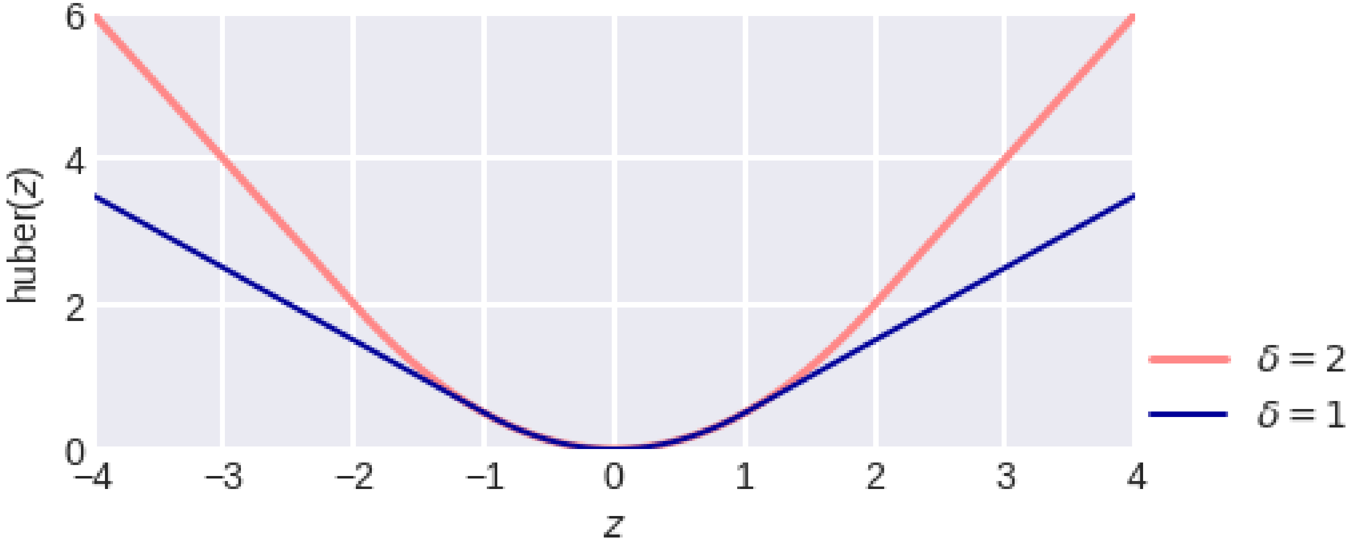
$$g(z)=g(0)+g'(0)z+\frac{g''(0)}{2}z^2+o(z^2)$$

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



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

Как только что вывели:

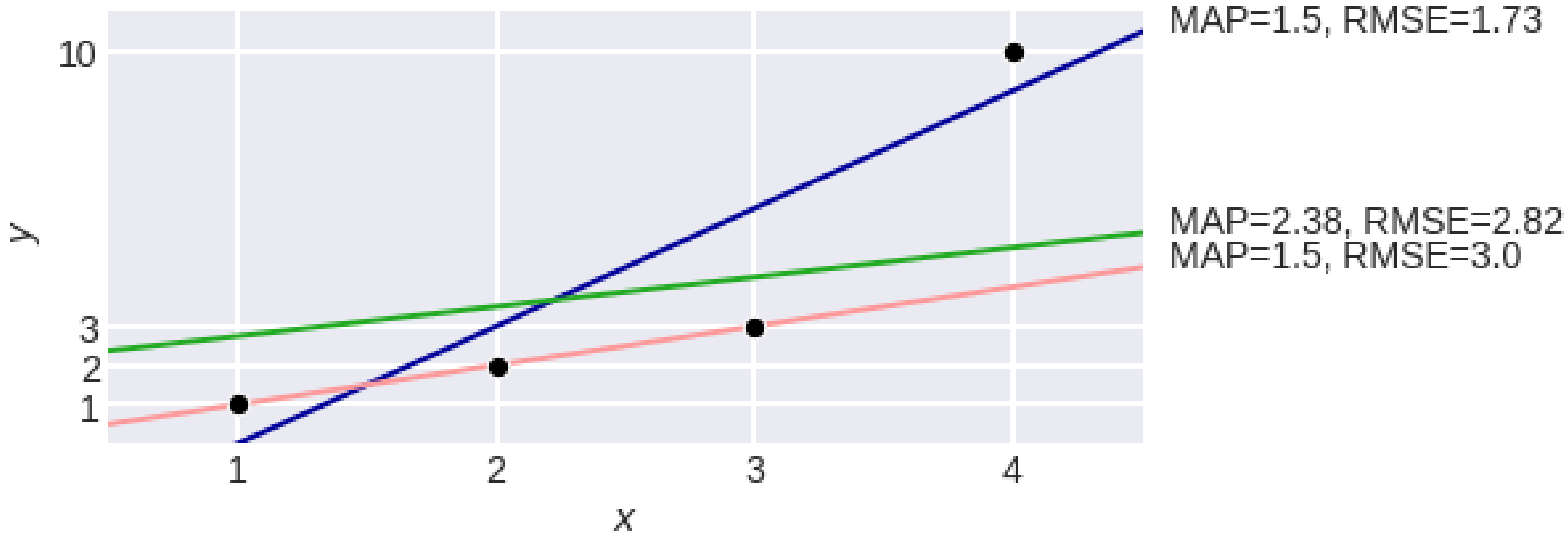
когда отклонение мало – ошибка квадратичная

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

Различия MSE и MAE

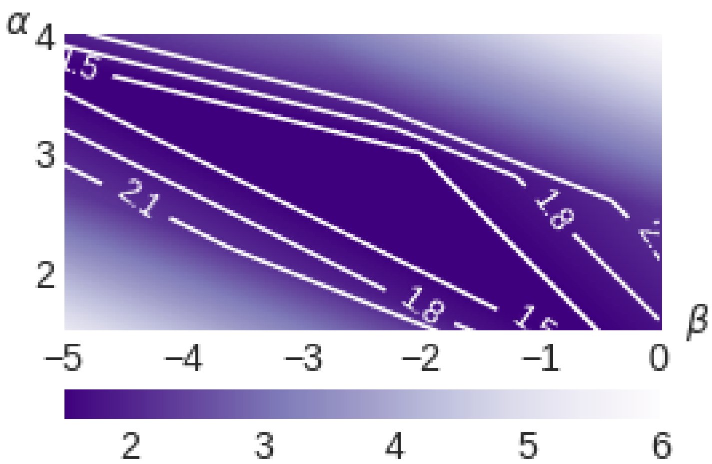
посмотрим на неконстантное решение:

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

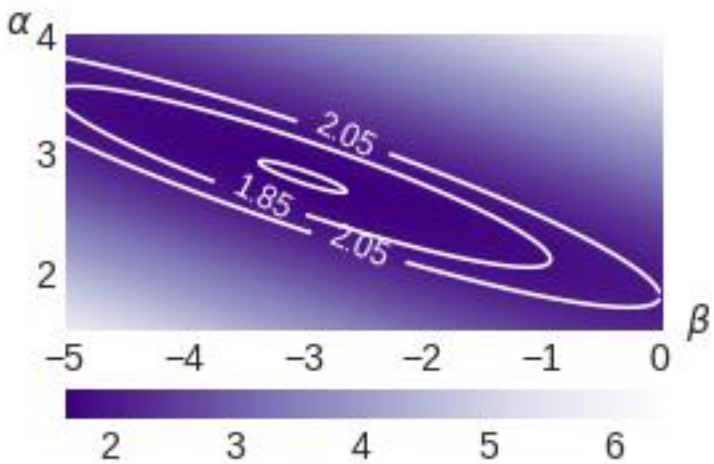


Различия MSE и MAE

MAE

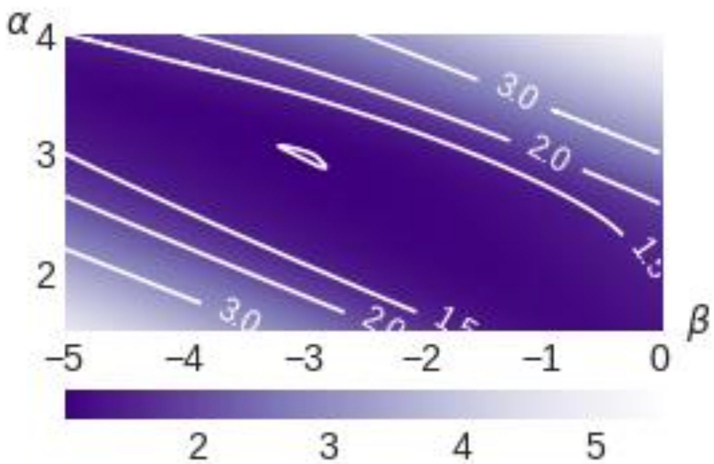


RMSE



внутри «треугольника» одинаковый MAP=1.5

Huber ($\delta = 1$)



Различия MSE и MAE

внутри «треугольника» одинаковый MAP=1.5

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

**ДЗ Хороший нетривиальный пример?
ДЗ Может ли быть наоборот?**

Обобщения

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

Рецепты

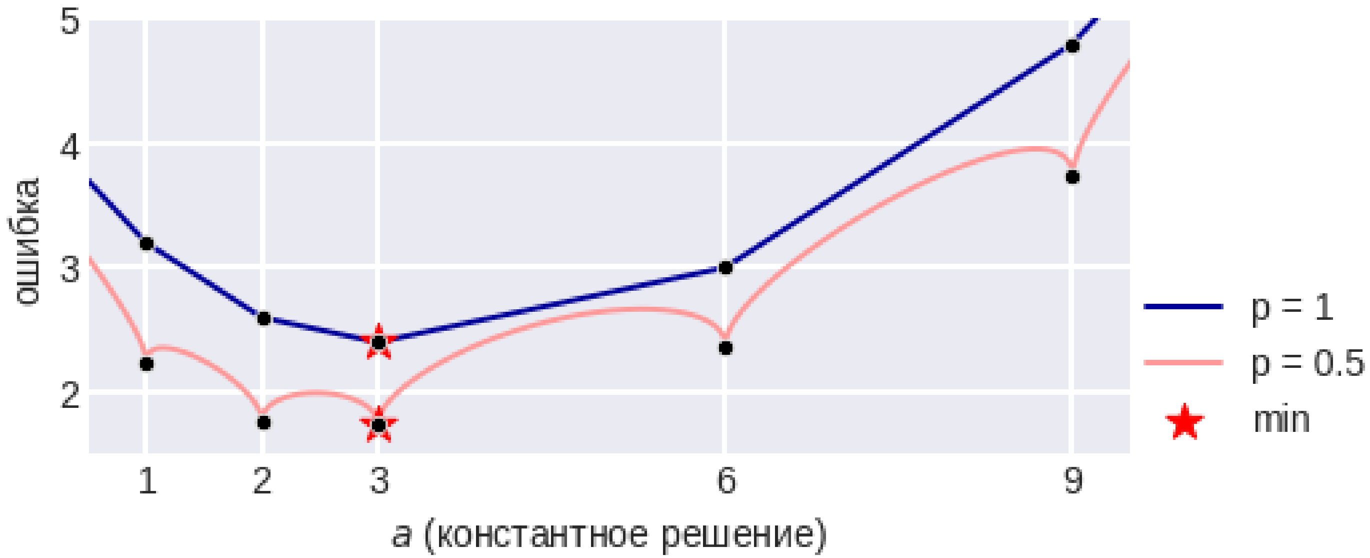
1. Преобразование целевого вектора $\varphi(y)$

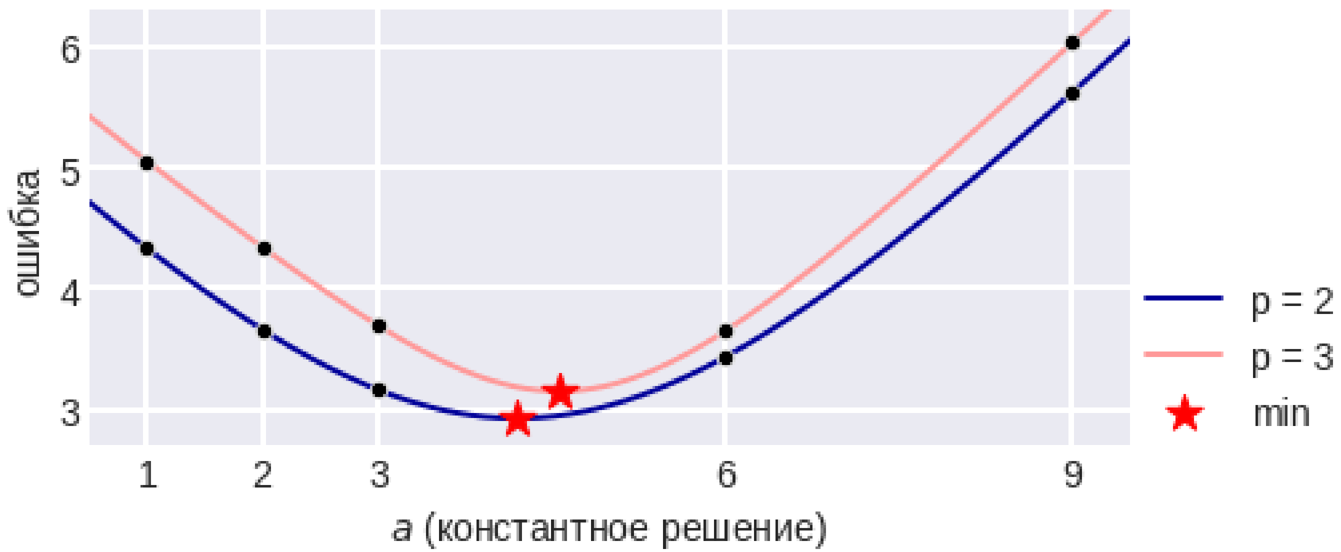
2. Веса ~ вероятности появления объектов в сэмплировании
Некоторые модели поддерживают веса объектов

3. В случае нетривиальных p **– прямая настройка**

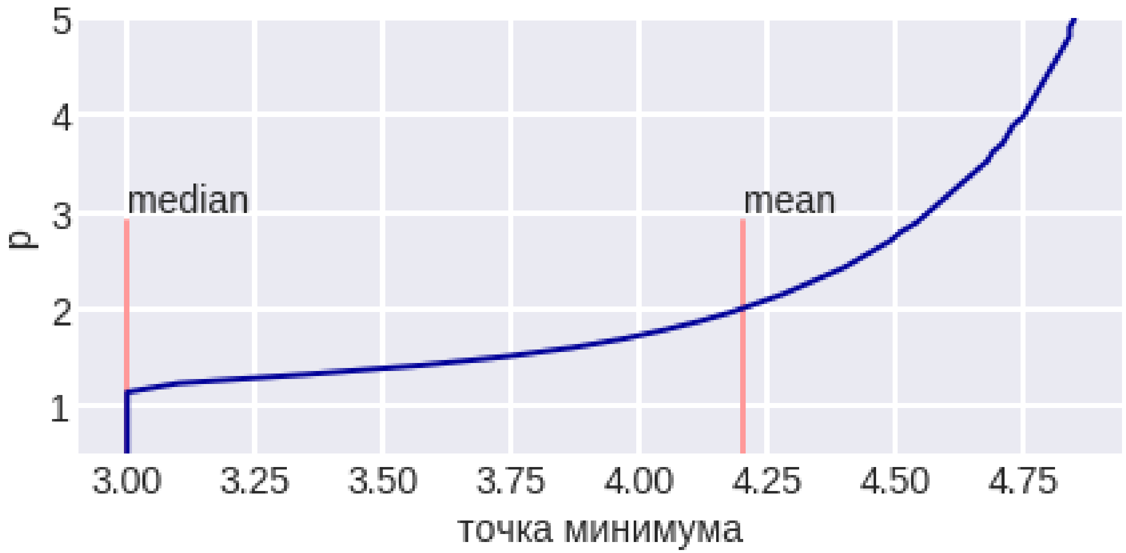
Дальше к этому вернёмся...

Про нетривиальные p





Как точка минимума зависит от степени



Symmetric mean absolute percentage error (SMAPE or sMAPE)

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

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

1 – 2	100 – 101	0 – 1
SMAPE = 67%	SMAPE = 1%	SMAPE = 200%

Начальники не знают, что такое проценты...

Применение SMAPE – прогноз временных рядов

Mean Absolute Percent Error (MAPE)

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

Чем MAPE явно лучше SMAPE на практике?

Mean Absolute Percent Error (MAPE)

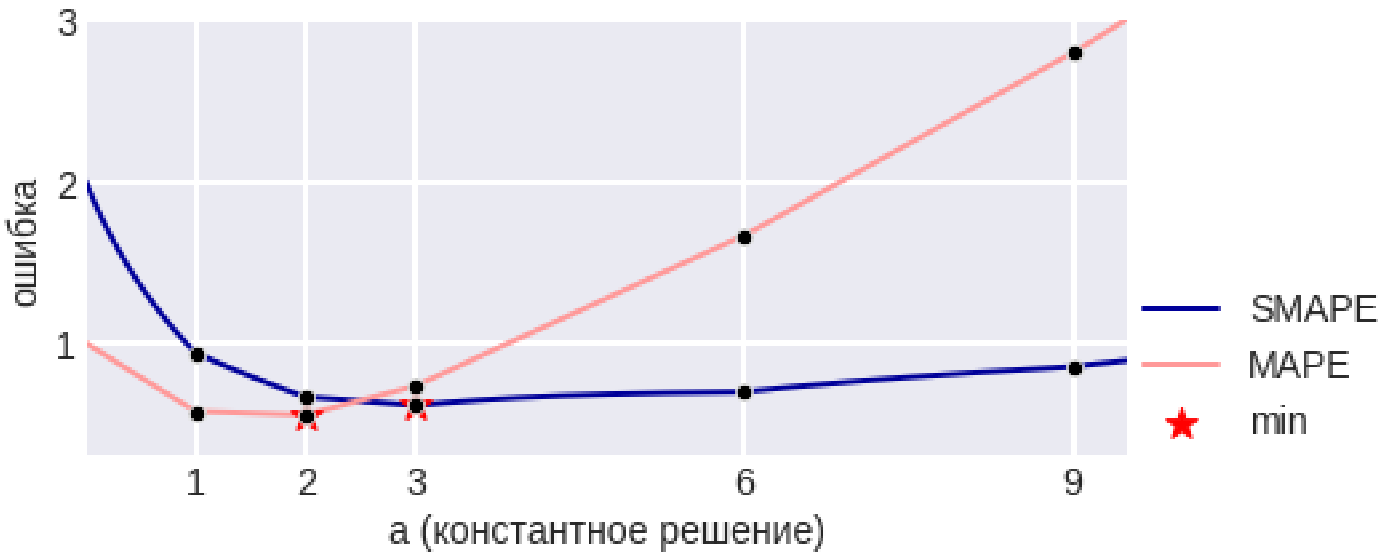
$$\mathrm{MAPE} = \frac{1}{q}\sum_{i=1}^{q} w_i \,|\, y_i - a_i \,|$$

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

Просто весовой MAE!

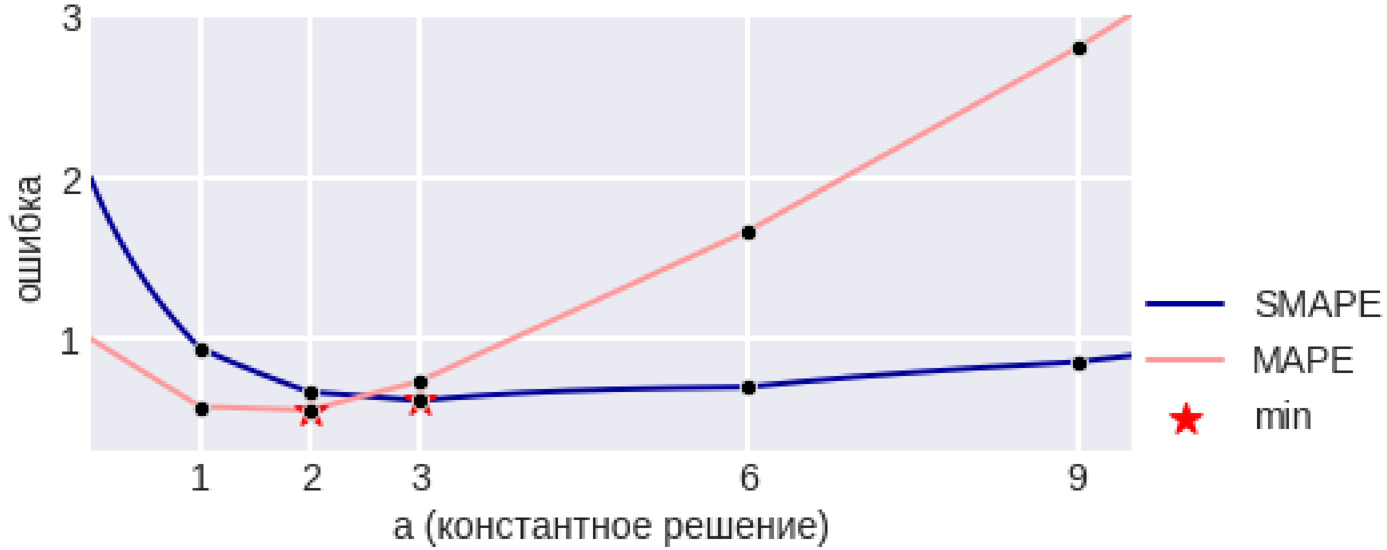
как оптимизировать? дальше...

MAPE и SMAPE

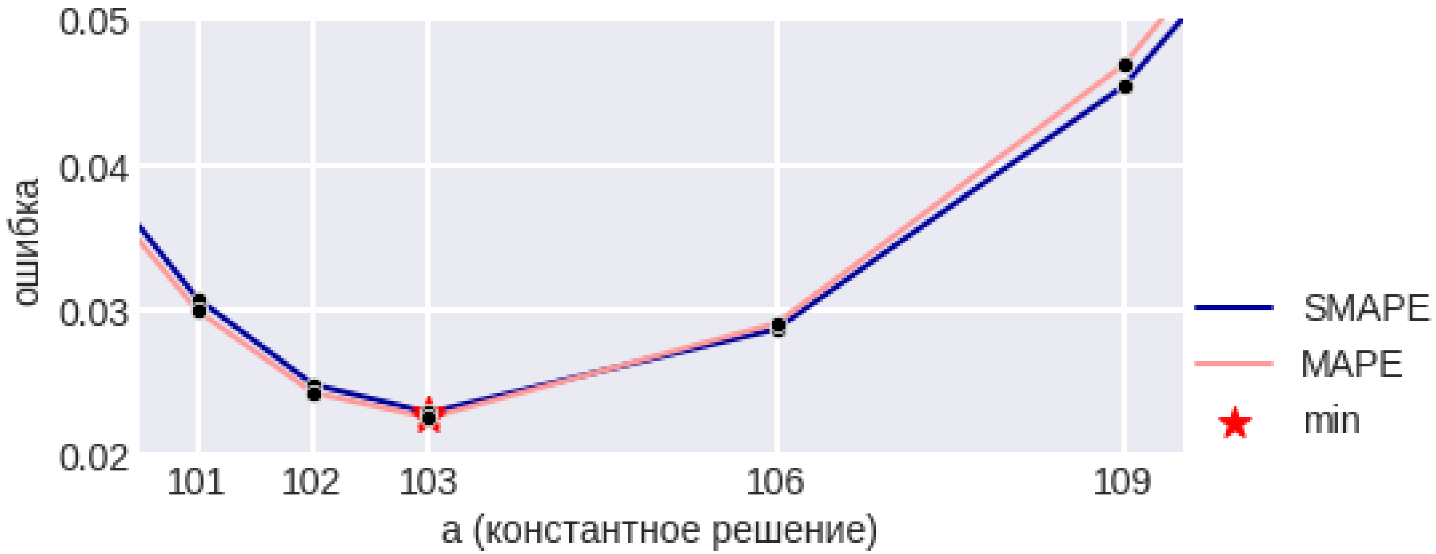


Что настораживает в этом графике?

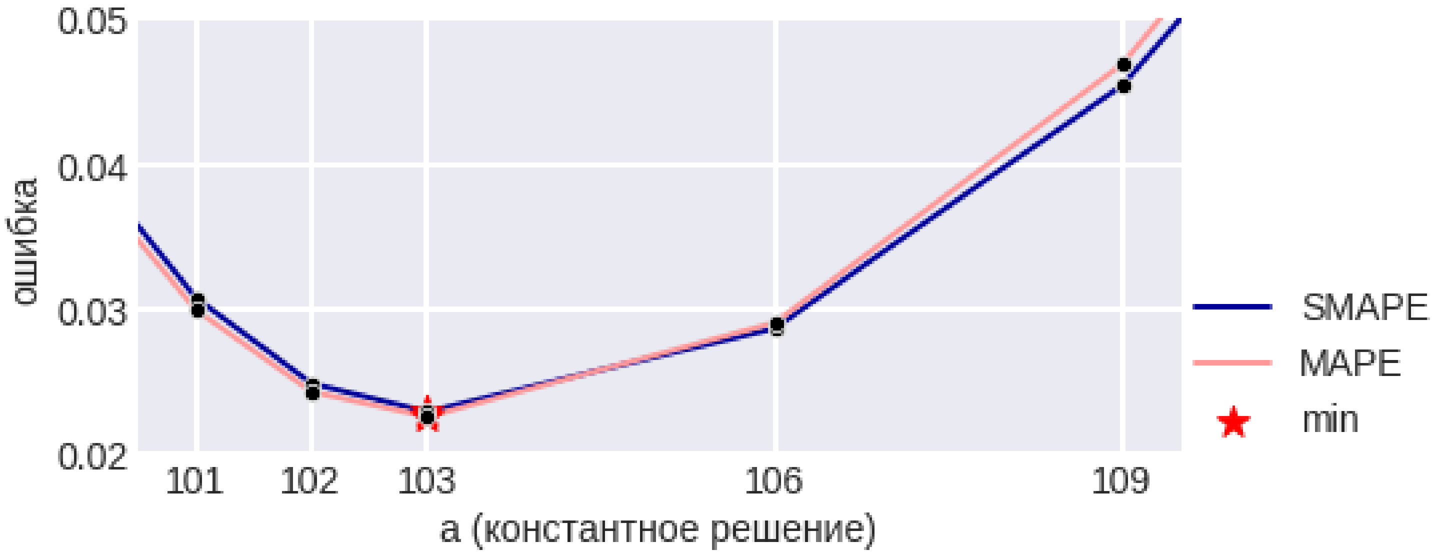
MAPE и SMAPE



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…

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

эквивалентен MAE

ДЗ 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)

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

REL_MAE

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

Percent Better

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

Меры на сравнении с бенчмарком

Как выбрать бенчмарк в задачах прогнозирования?

Нормированные ошибки

Не зависят от шкалы…

Mean Absolute Scaled Error

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

Какие ещё бывают функционалы в регрессии?

С точностью до порога

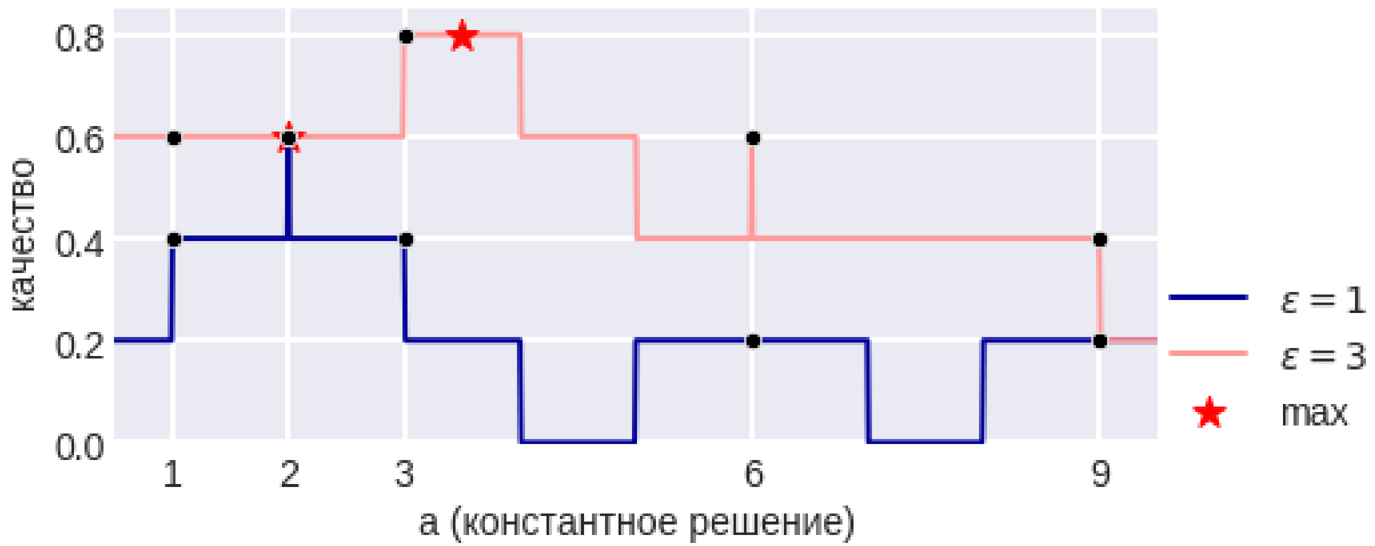
функция ошибки

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

функционал качества

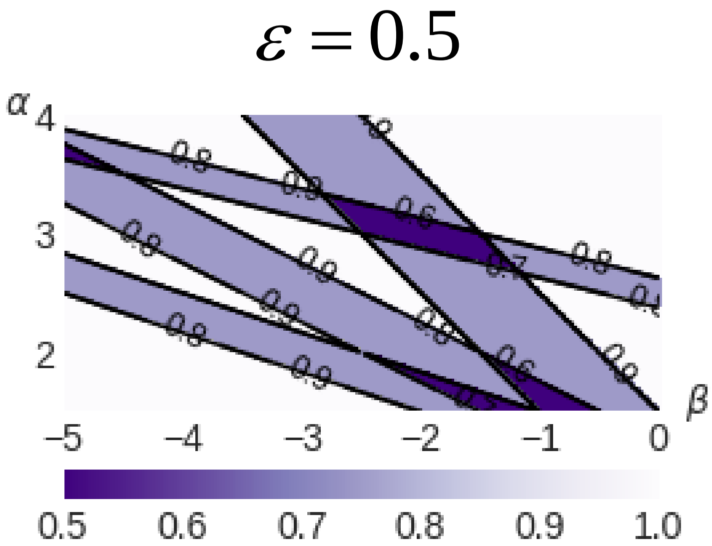
$$\frac{1}{q}\sum_{i=1}^{q} I[|y_i - a_i| < \varepsilon]$$

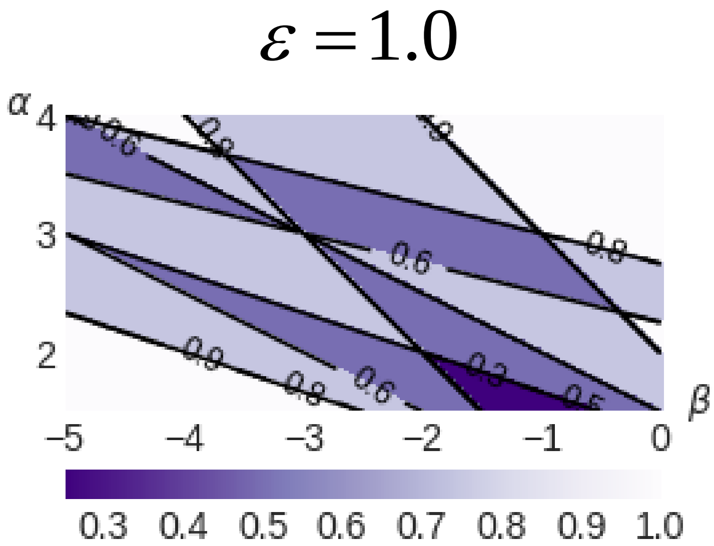
был в задаче Dunnhumby

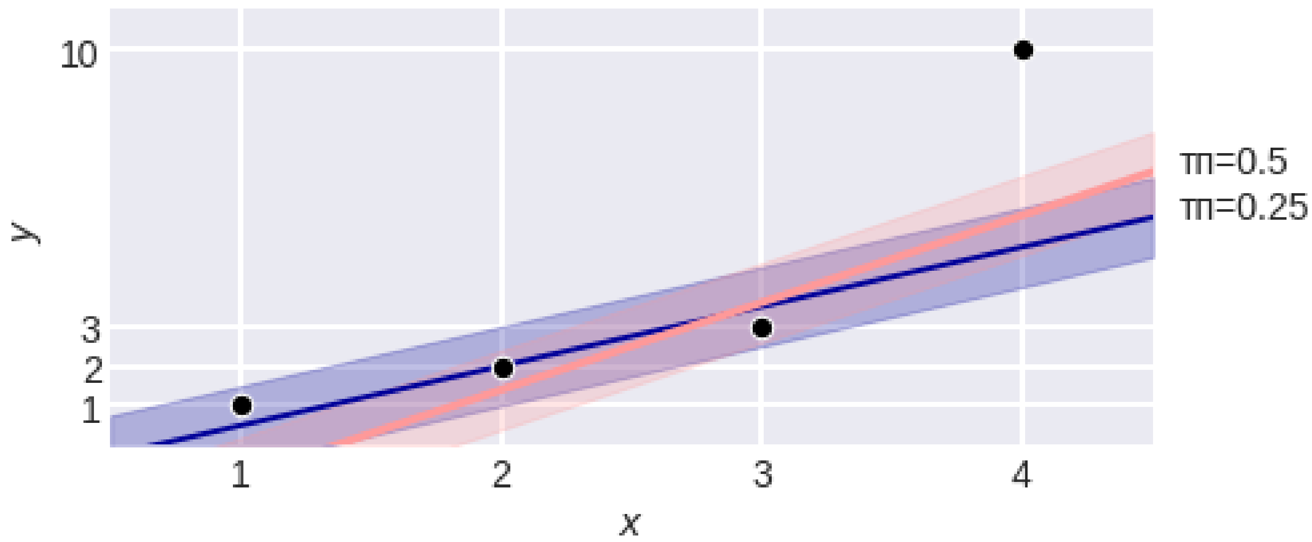


Оптимальное решение – мода парзеновской плотности

С точностью до порога

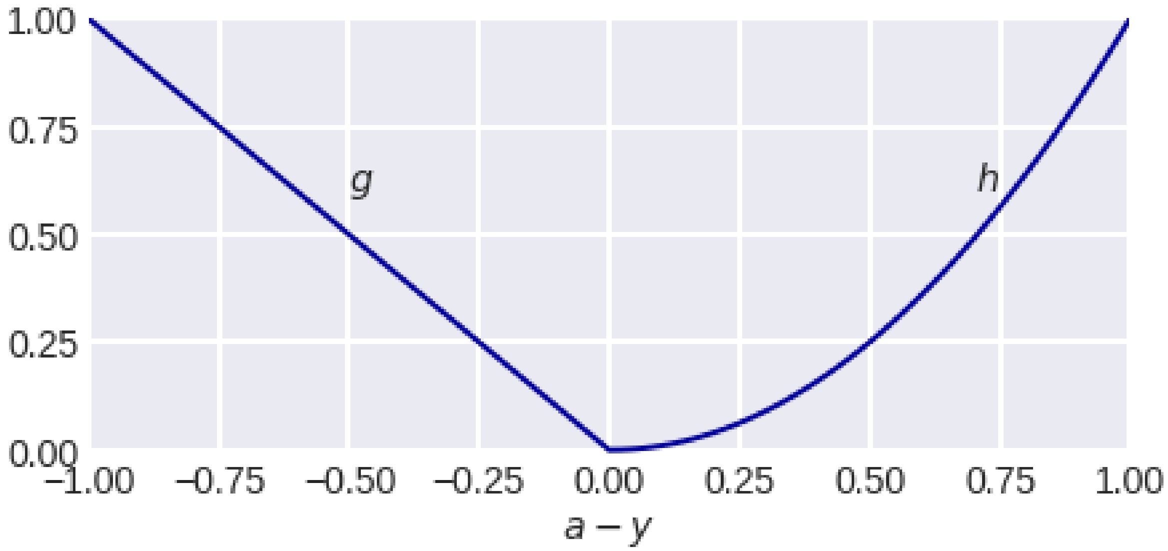






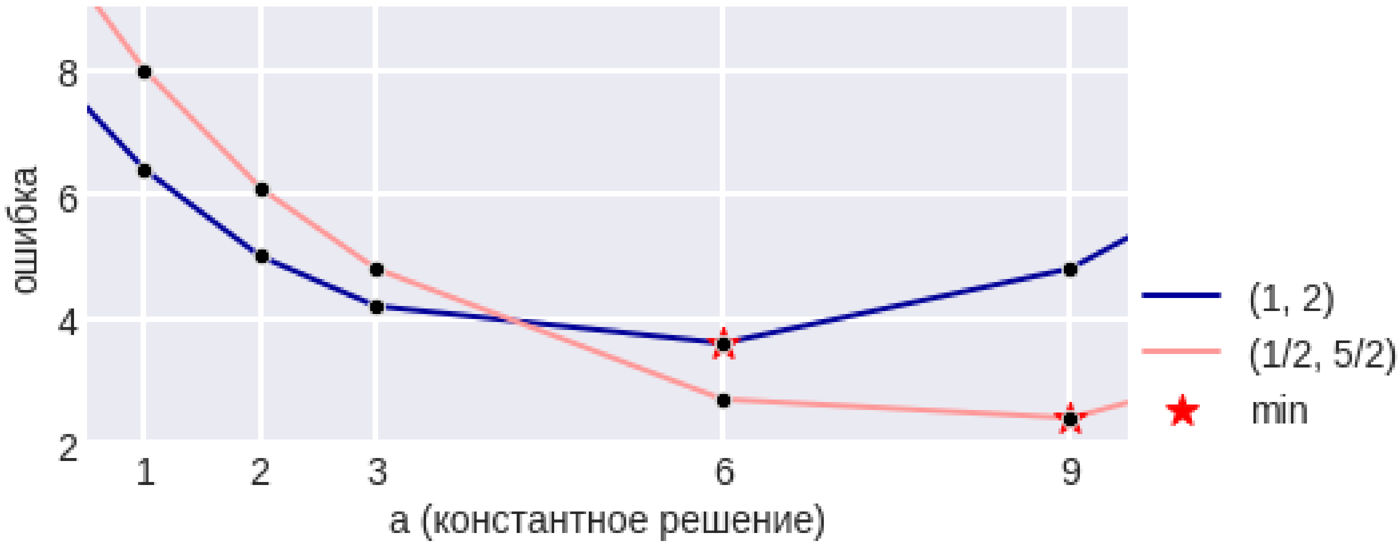
Несимметричные функции потерь

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



Зачем нужны такие функции?

Несимметричные функции потерь



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

Совет

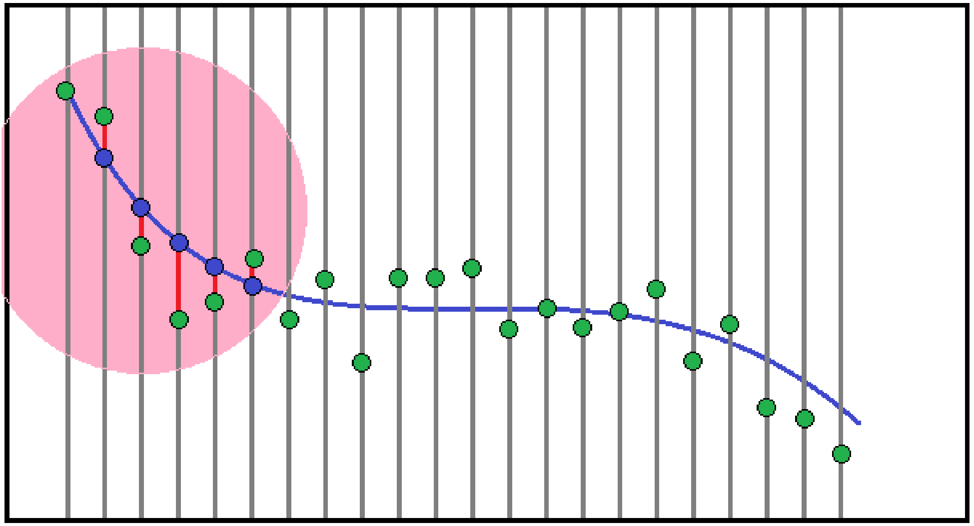
Функции ошибок иногда и классные признаки…

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:

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

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

Решения на практике могут отличаться...

В методе градиентного спуска разные производные

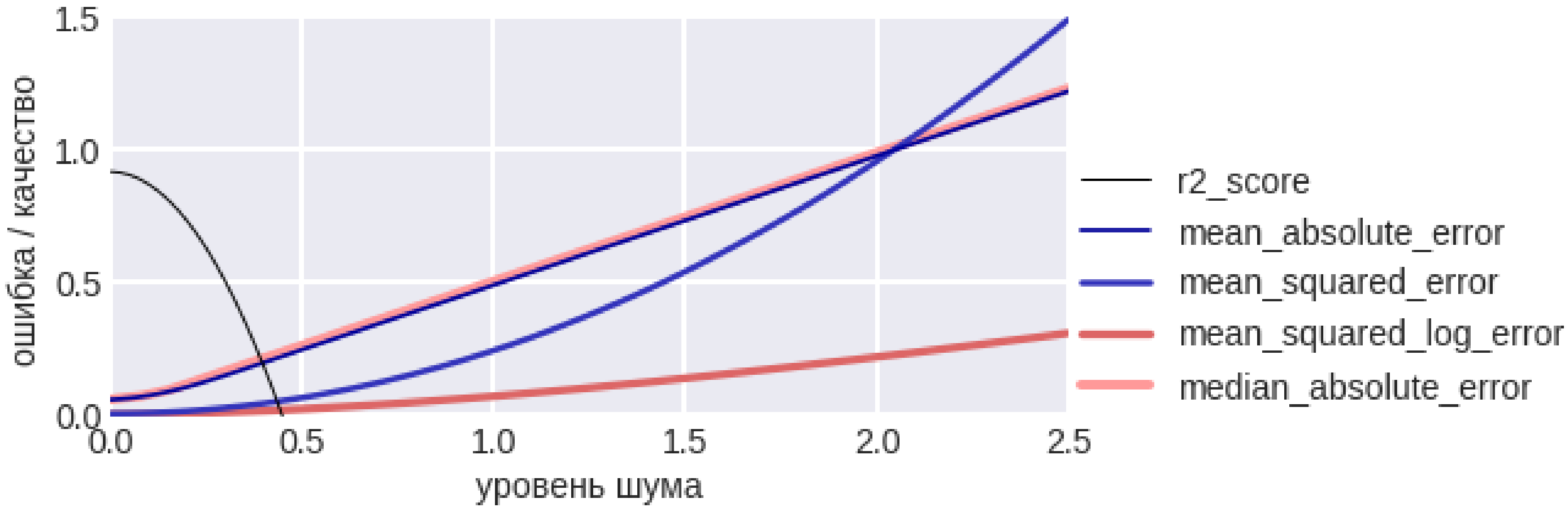
$$\frac{\partial\,\mathrm{MSE}}{\partial a} = \frac{2}{q}\sum_{i=1}^{q}(a - y_i) \qquad \frac{\partial\,\mathrm{RMSE}}{\partial a} = \frac{1}{q\,\mathrm{RMSE}}\sum_{i=1}^{q}(a_i - y_i)$$

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

Метрики в регрессии: минутка кода

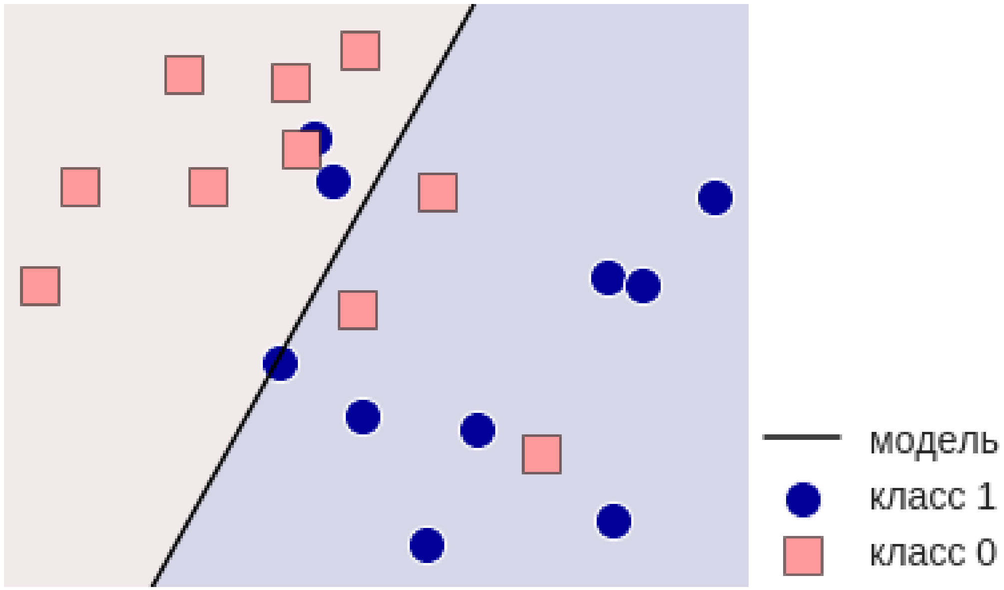
```
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_squared_log_error
from sklearn.metrics import median_absolute_error
from sklearn.metrics import explained_variance_score
# R^2
print (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



NEW: Что за эксперимент? Почему ошибки ведут себя так?

Задача классификации



Задача классификации: матрица ошибок / несоответствий «Confusion Matrix»

ответы

	y	a
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 $\{1, 2, \ldots, l\}$

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

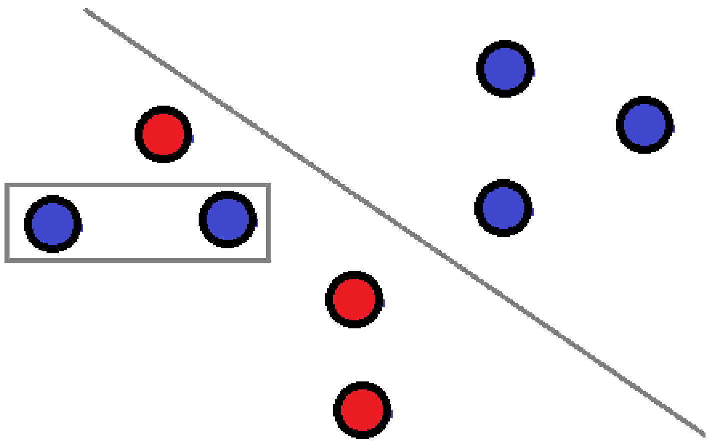
$$n_{ij} = \sum_{t=1}^{q} I[a_t = i] I[y_t = j]$$

```
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

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

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

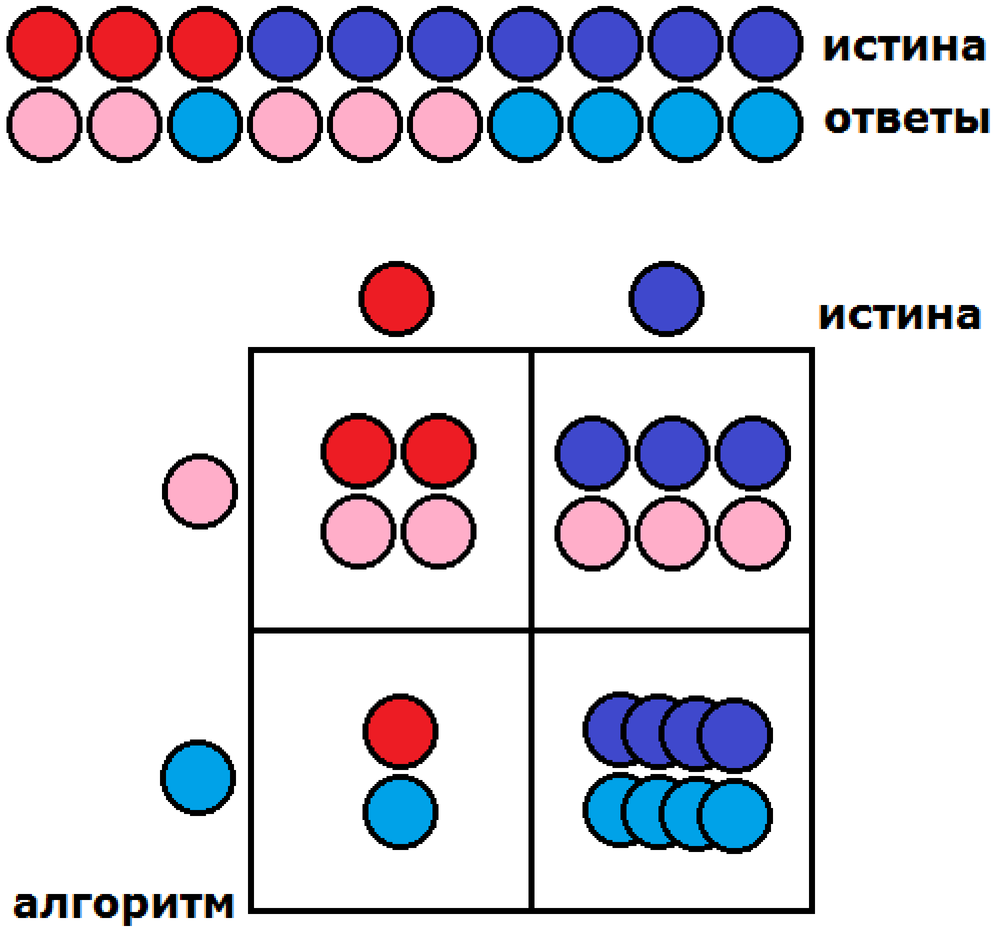


```
y = [0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0]
```

Выгодно выдавать решение – константу 0!

Задача классификации с двумя классами

Confusion Matrix

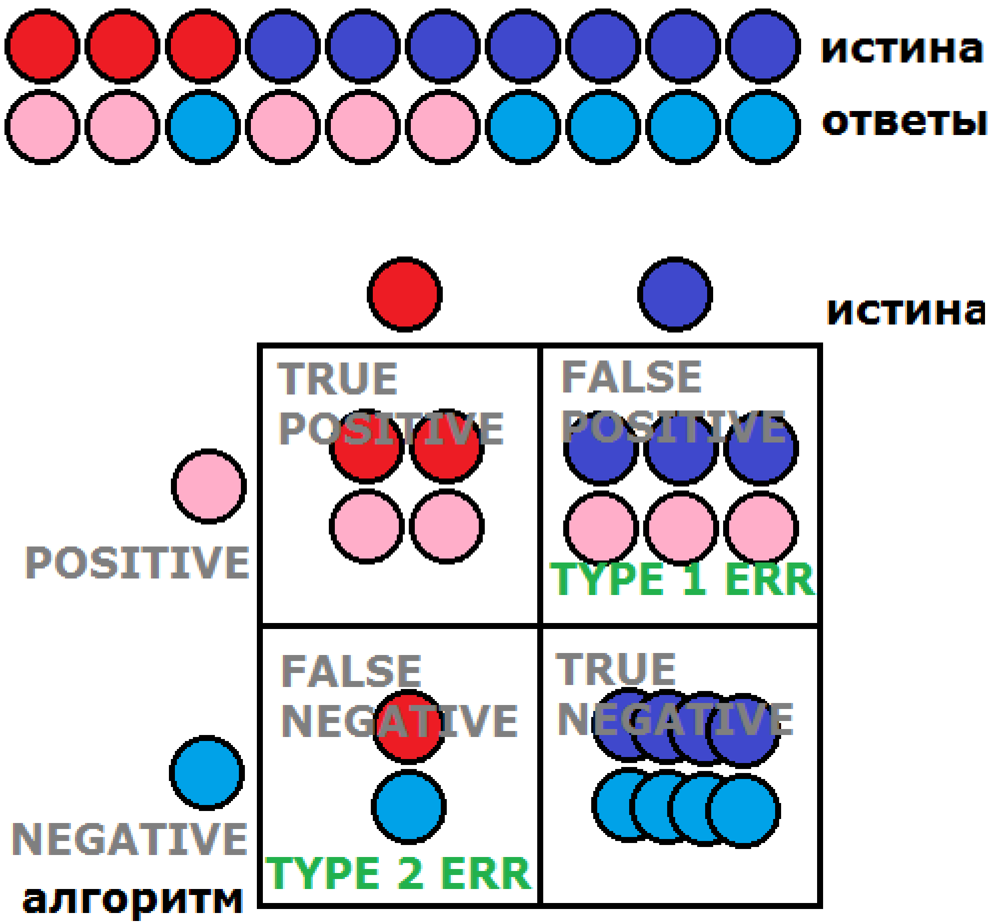


	$a=0$	$a=1$
$y=0$	13599	2600
$y=1$	898	903

в scikit-learn-е другая ориентация!

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, a_test)
```

Задача классификации с двумя классами




```
tn, fp, fn, tp = confusion_matrix(y, a).ravel()
```

Как запомнить названия ошибок

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

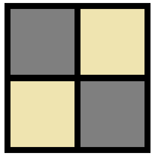
2 рода – учил, но не сдал (= не знает по мнению экзаменатора)

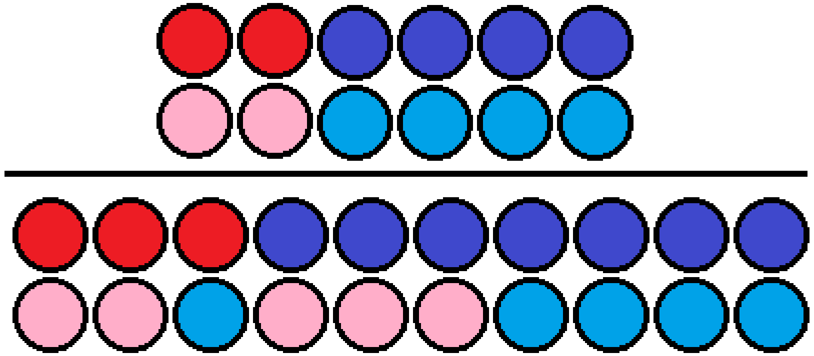
Ошибка 1 рода

Ошибка 2 рода

Точность Accuracy

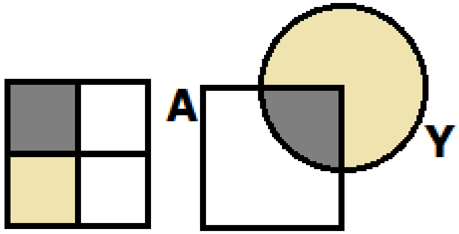
$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{ALL}}$$

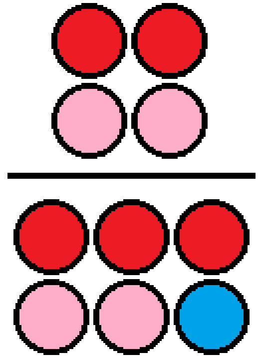




Полнота (Sensitivity, True Positive Rate, Recall, Hit Rate)

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

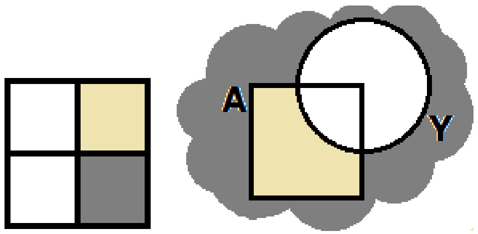


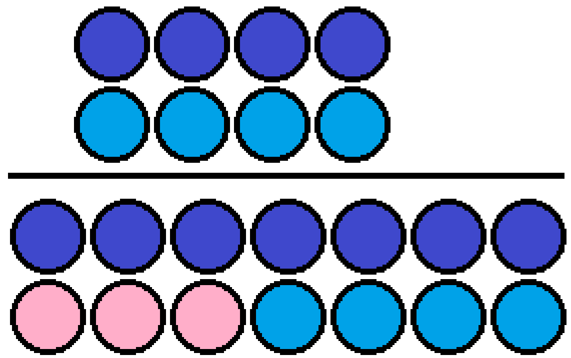


TPR = TP / сколько правда 1

Specificity (True Negative Rate)

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

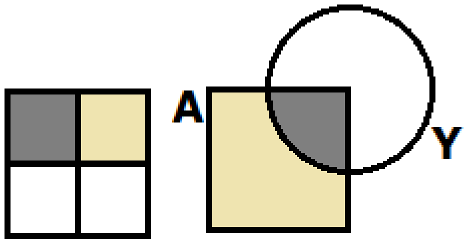


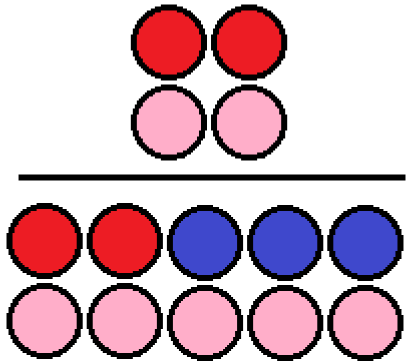


FPR = 1 – Specificity

Точность (Precision, Positive Predictive Value)

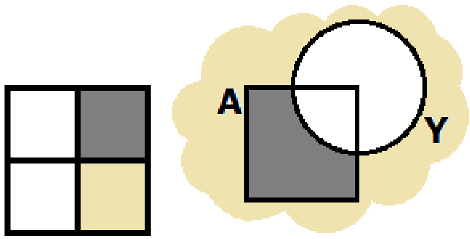
$$PPV = \frac{TP}{TP+FP}$$

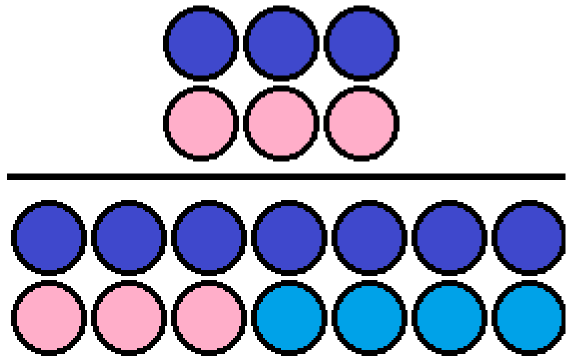




False Positive Rate (FPR, fall-out, false alarm rate)

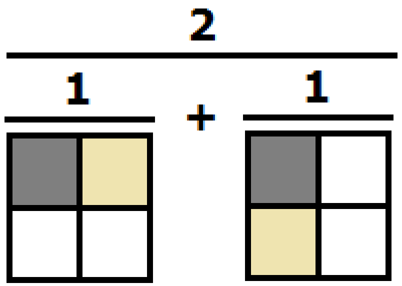
$$FPR = \frac{FP}{FP+TN}$$





FPR = FP / сколько правда 0

F_1 score



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

$$=\frac{2TP}{2TP+FP+FN}$$

F_β score

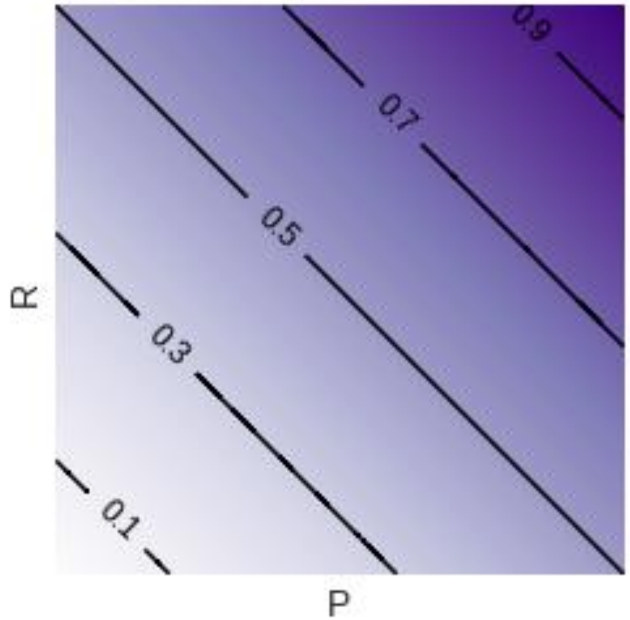
$$\frac{1}{\dfrac{\alpha}{P}+\dfrac{1-\alpha}{R}}=\frac{PR}{\alpha R+(1-\alpha)P}=$$

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

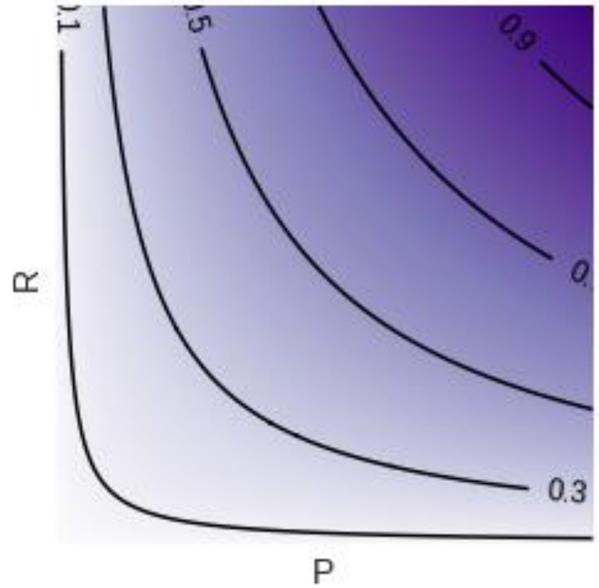
$$\beta^2=\left(\frac{1}{\alpha}-1\right)$$

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

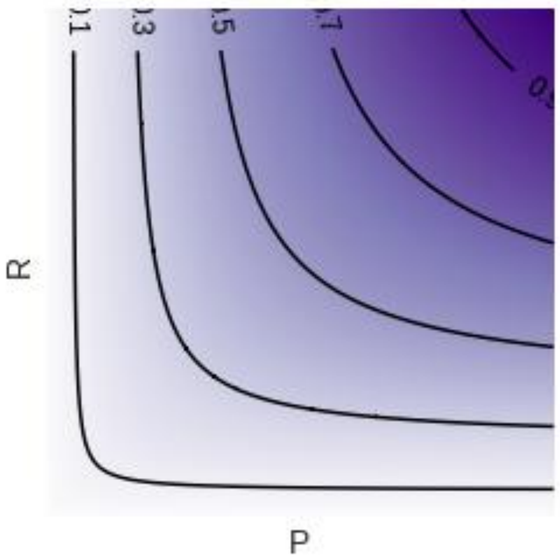
Почему используется F-мера



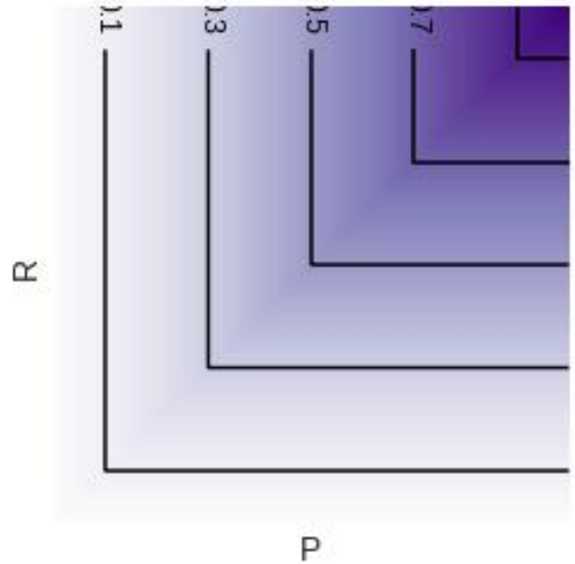
$(P+R)/2$



$\sqrt{P\cdot R}$



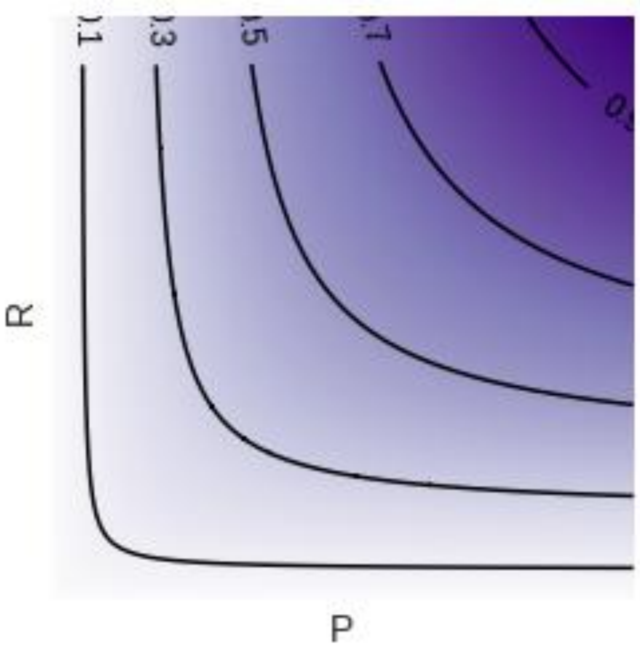
$2/(1/P+1/R)$



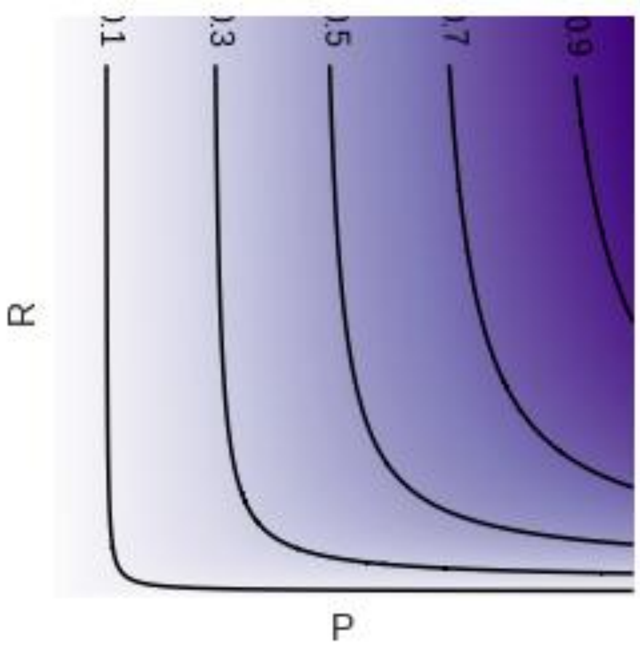
$\min(P,R)$

Почему используется F-мера

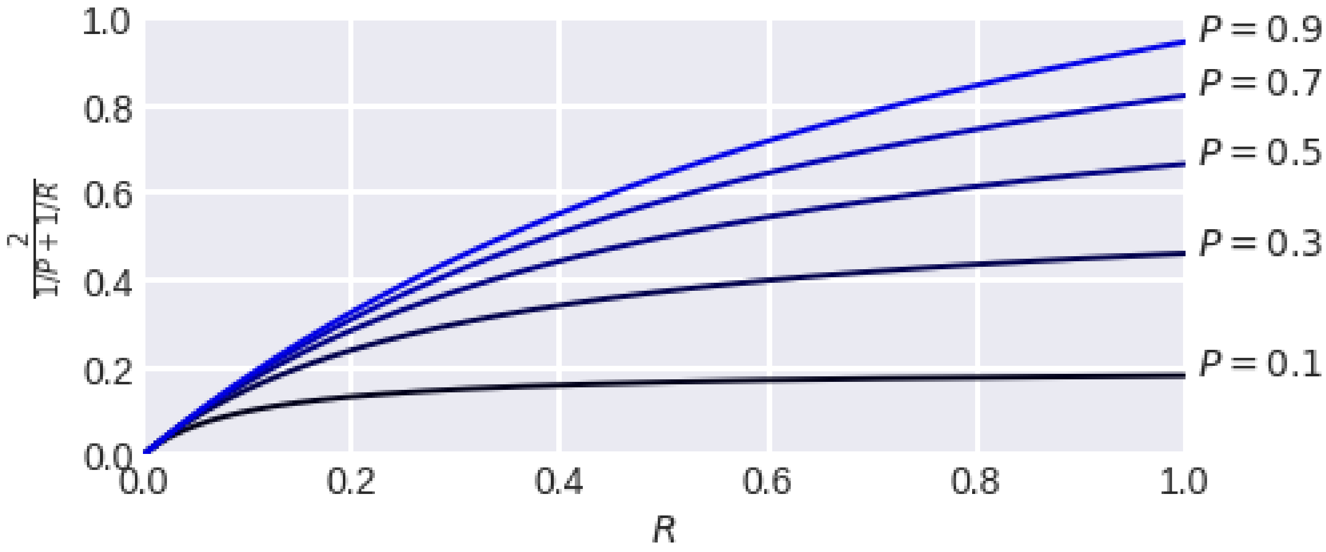
$$2/(1/P+1/R)$$



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



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

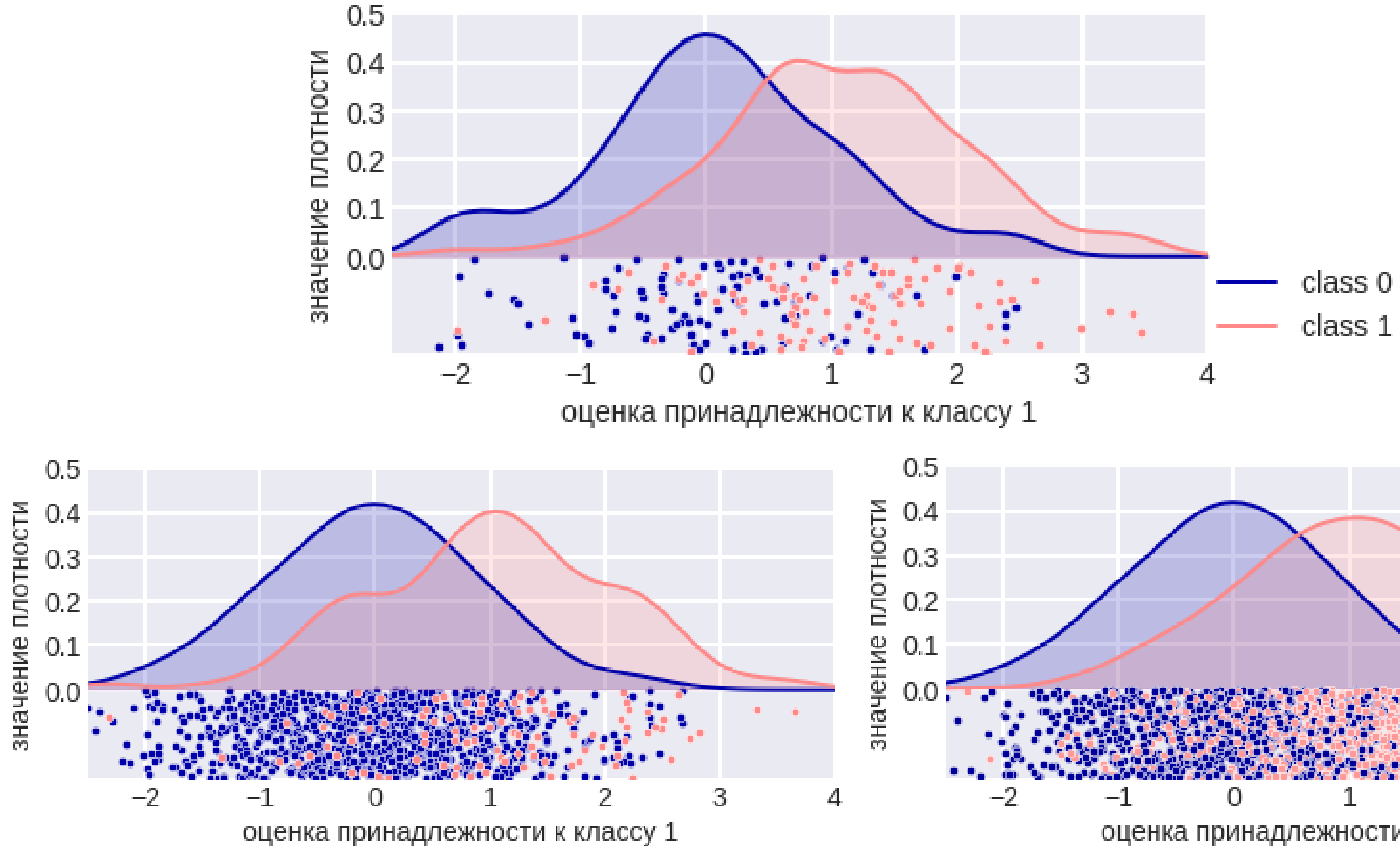
смысл: поправка значения точности.

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

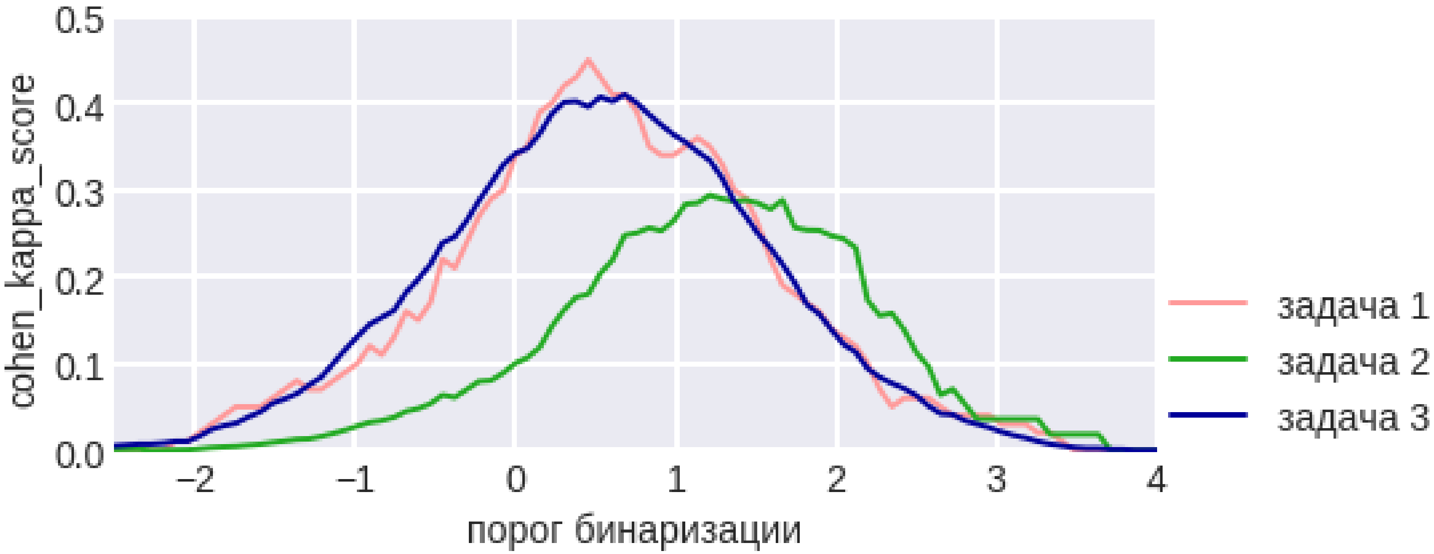


Как будет выглядеть график CK от порога бинаризации?

Как меняется ROC AUC?

Cohen's Kappa

график CK от порога бинаризации



ROC AUC: 0.77 во всех задачах!

Weighted kappa

Если есть разумные веса ошибок за конкретные несогласованности

Когда это бывает?

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

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 Kappa

ответы

	y	a
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

матрица случайных ответов

	0	1	2
0	12	16	12
1	9	12	9
2	9	12	9

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

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

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
Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \in [-1,+1]$$

Подходит для несбалансированных выборок

ДЗ Показать преимущество перед другими функционалами

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

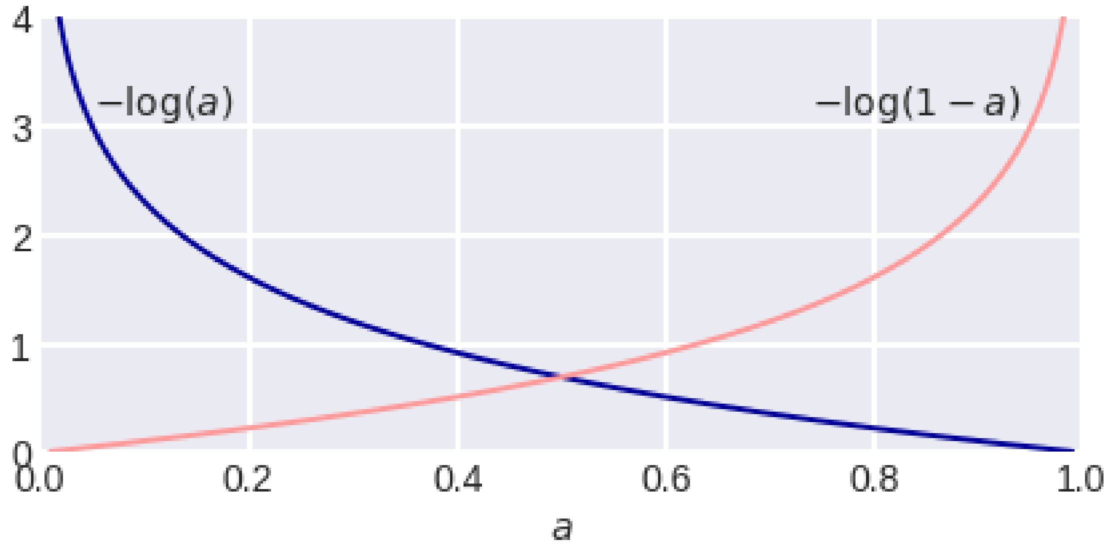
$$\mathrm{LOGLOSS} = -\frac{1}{q}\sum_{i=1}^{q}(y_i \log a_i + (1-y_i)\log(1-a_i))$$

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

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

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:

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

Откуда берётся Log Loss

Максимизация правдоподобия эквивалентна

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

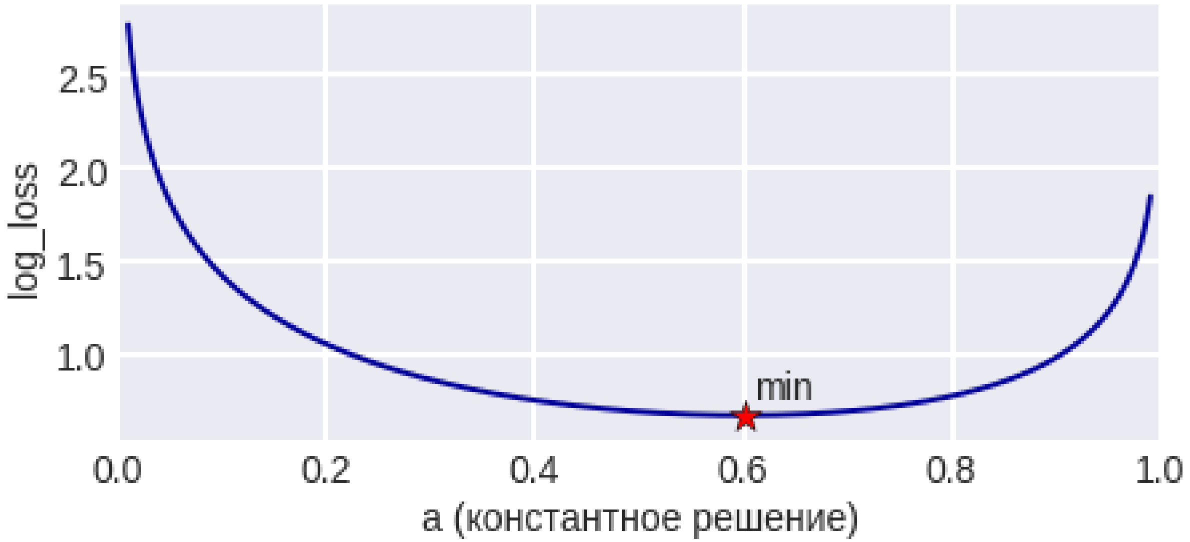
Логична ровно настолько, насколько MSE в задаче регрессии

(тоже выводится из ММП)

Названия

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

Log Loss – Оптимальная константа



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

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

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

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…

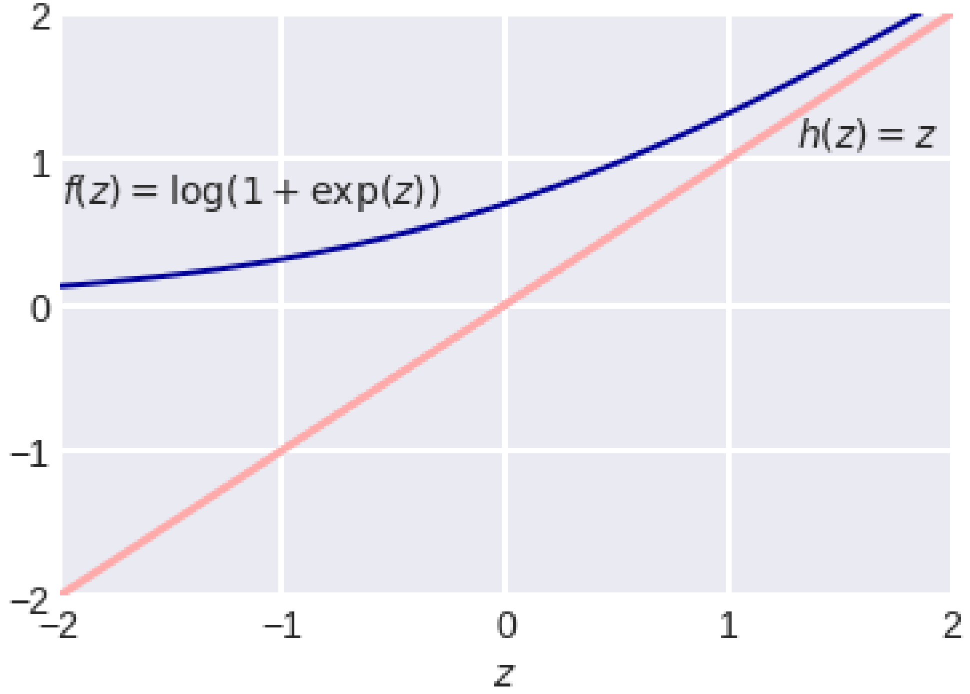
Связь logloss с логистической регрессией

см. лекцию про минимизацию...

Другая форма функционала

Подставим выражение для сигмоиды, сделаем переобозначение: метки классов теперь –1 и +1, тогда

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



Кстати

SVM

$$\sum_i \max[1 - y_i w^{\mathrm{T}} x, 0] + \alpha w^{\mathrm{T}} w \to \min$$

RVM

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

Связь logloss с расхождением Кульбака-Лейблера

$$D_{\mathrm{KL}}(P \,\|\, Q) = \int p(z) \log \frac{p(z)}{q(z)} \partial z$$

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_i P_i \log \frac{P_i}{Q_i}$$

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!!!

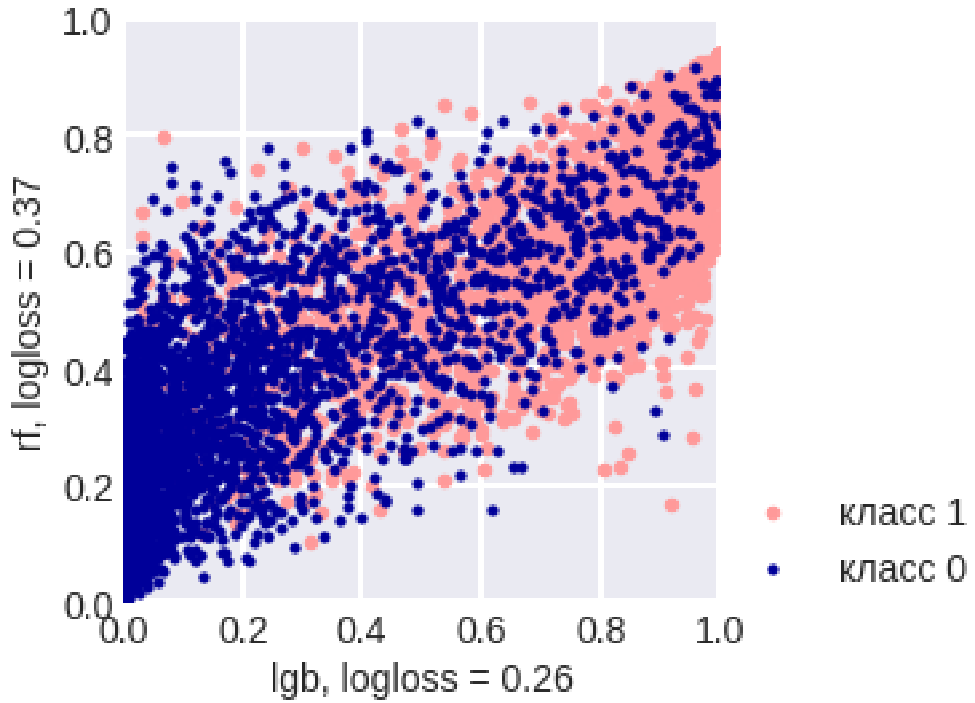
Настройка на Logloss

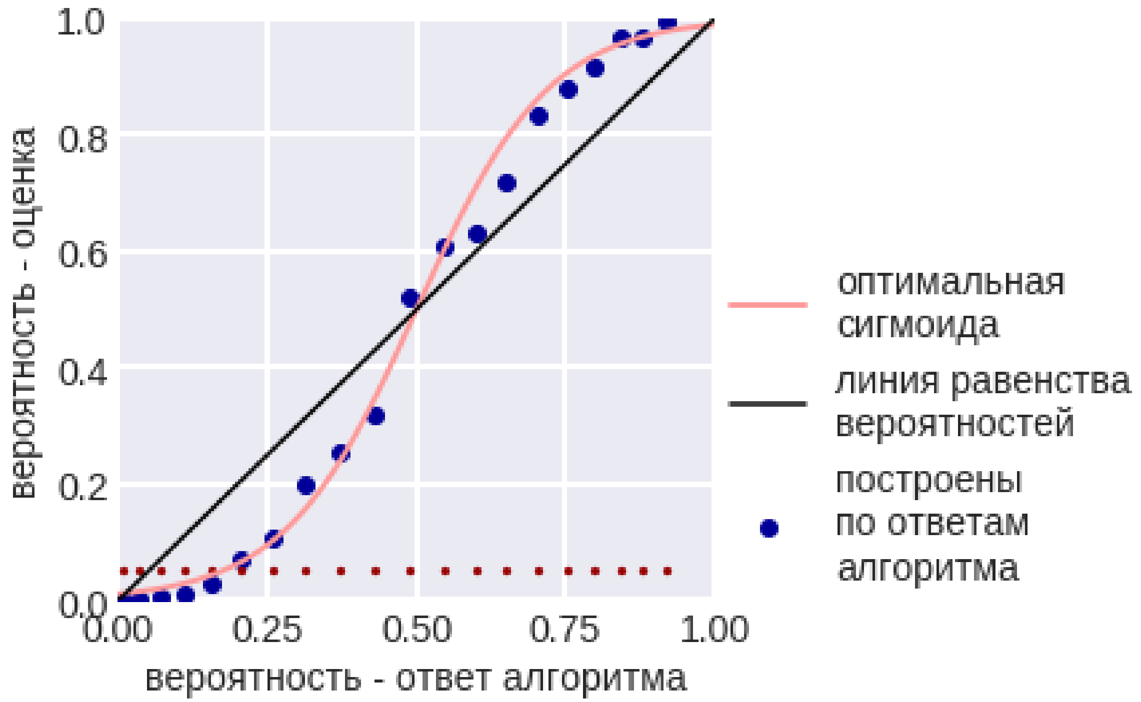
калибровка Платта (Platt calibration)

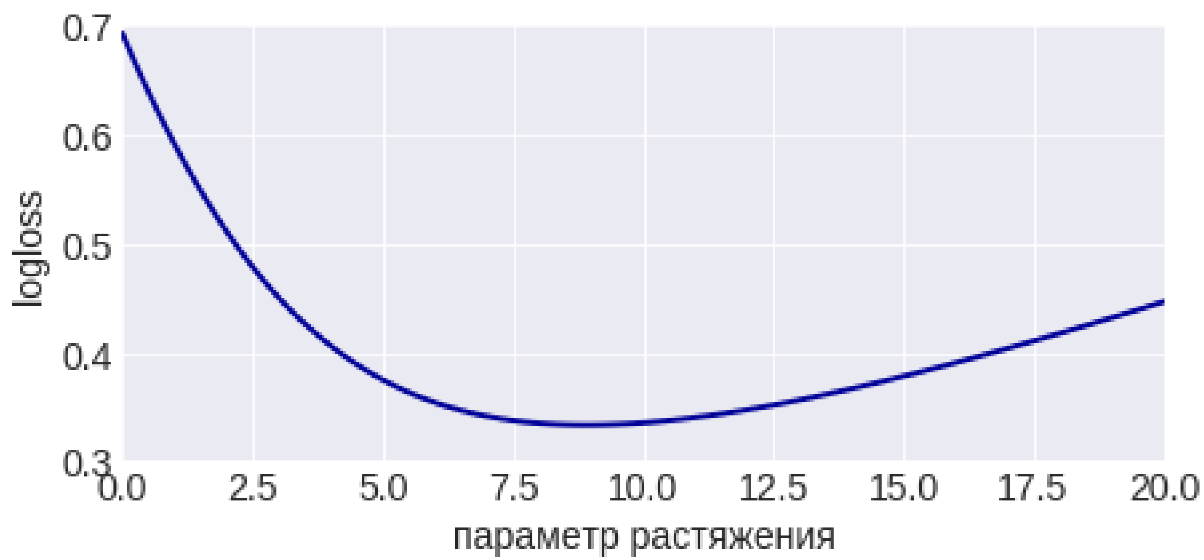
– для SVM

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

ещё есть – монотонная регрессия (Isotonic regression)




1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
вероятность - оценка
вероятность - ответ алгоритма
оптимальная сигмоида
линия равенства вероятностей
построены по ответам алгоритма


0.7
0.6
0.5
0.4
0.3
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
logloss
параметр растяжения

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 матожидание 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!

это называется «Brier score»

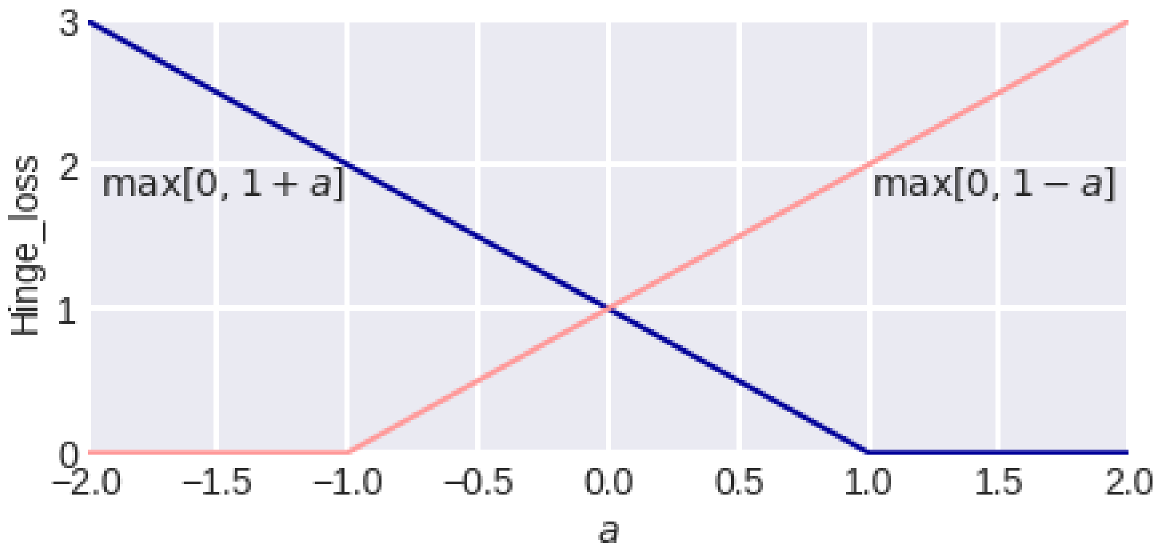
```
from sklearn.metrics import brier_score_loss
brier_score_loss(y_true, y_prob)
```

Hinge loss

$$y_i \in \{\pm 1\},\ a_i \in \mathbb{R}$$

когда вещественная оценка, вспоминаем SVM…

$$\frac{1}{m}\sum_{i=1}^{m}\max[1-y_i a_i, 0]$$



Часто делают сглаживание аналогичное Huber Loss.

Скоринговые ошибки

– ошибки в задаче бинарной классификации, для которых оптимальный ответ на каждом объекте – вероятность его принадлежности к классу 1.

$L(y,a)$**:**

$$p = \arg\min_a \mathrm{E}_y L(y,a)$$ **для** $$y \sim \mathrm{Bernoulli}(p).$$

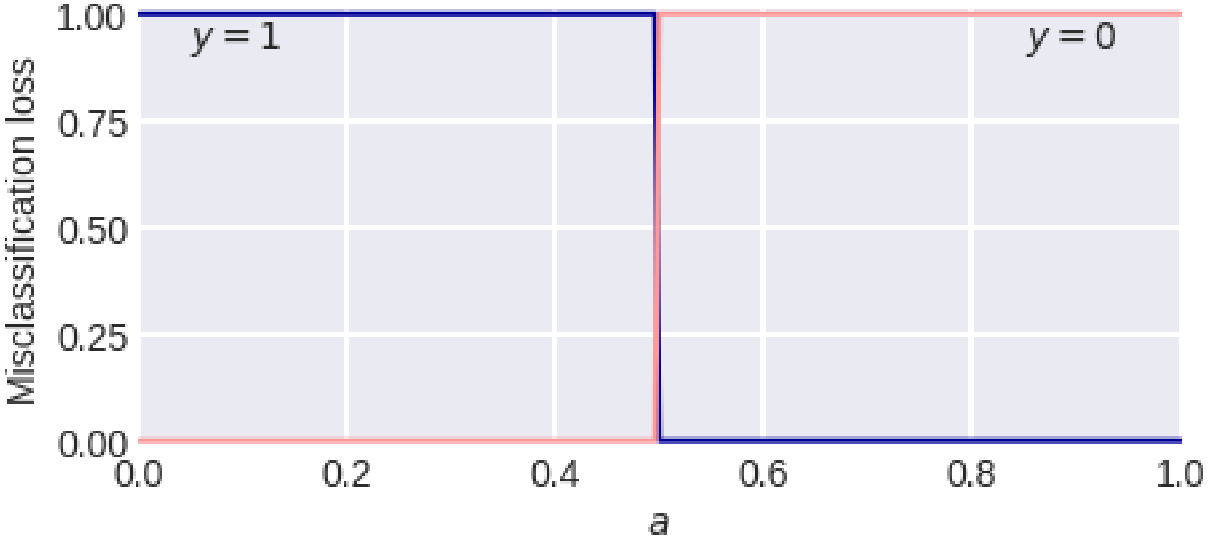
- **Log Loss**
- **MSE**
- **Exploss**
- **Misclassification Loss**

но не все…

~~**MAE**~~

Misclassification Loss

$$\mathrm{ME} = y\,I[a \le 0.5] + (1-y)\,I[a > 0.5]$$



немного искусственная функция

почему?

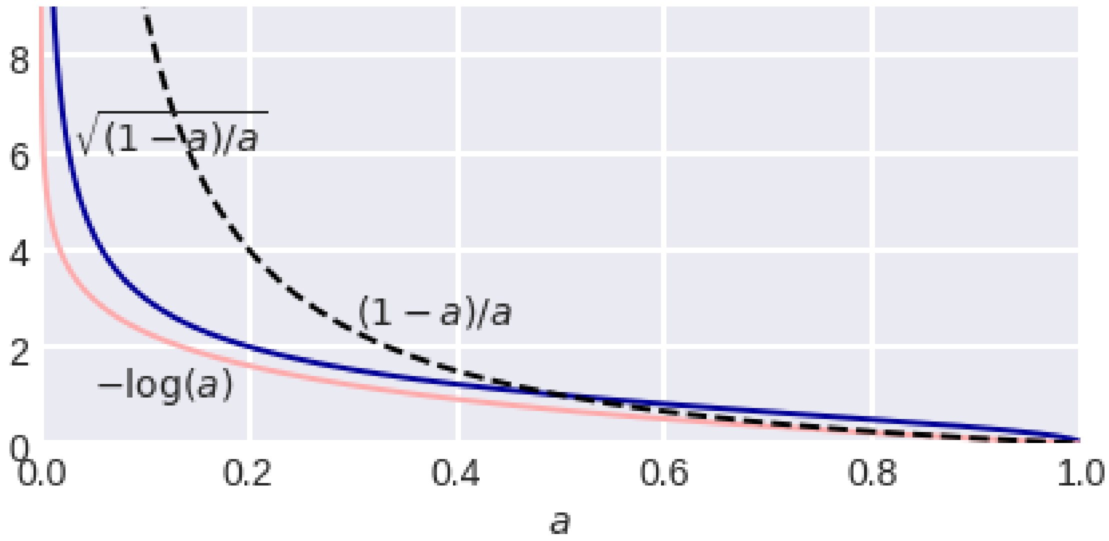
$$\mathbf{E}_y\,\mathrm{ME} = p\,I[a \le 0.5] + (1-p)\,I[a > 0.5],\ y \sim \mathrm{Bernoully}(p),$$

нет единственного решения:

$$\arg\min \mathbf{E}_y\,\mathrm{ME} \in \begin{cases} [0, 0.5], & a \le 0.5, \\ [0.5, 1], & a > 0.5, \end{cases}$$

$$\min \mathbf{E}_y\,\mathrm{ME} = \min(p, 1-p)$$

Exploss: попытка немного изменить LogLoss

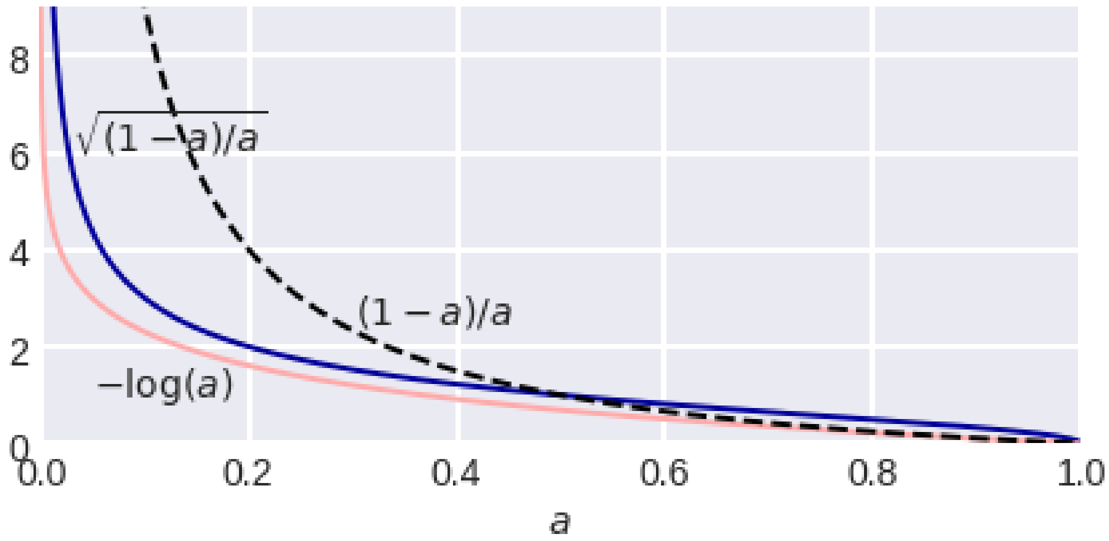


$$\text{exploss} = y\sqrt{\frac{1-a}{a}} + (1-y)\sqrt{\frac{a}{1-a}}$$

$$p\sqrt{\frac{1-a}{a}} + (1-p)\sqrt{\frac{a}{1-a}} \to \min$$

$$p = a$$

Exploss: попытка немного изменить LogLoss



$$p\sqrt{\frac{1-p}{p}}+(1-p)\sqrt{\frac{p}{1-p}}=2\sqrt{p(1-p)}$$

Что это?

ДЗ А другие похожие на Log Loss функции?

Exploss: почему логичная функция

задача классификации на два класса $\{\pm 1\}$

алгоритм выдаёт оценки принадлежности к классу 1

$$a(x) \in (-\infty, +\infty)$$

Естественна функция ошибки:

$$\exp(-y a)$$

(изначально использовалась в бустинге)

Матожидание на объекте

$$p \exp(-a) + (1 - p) \exp(+a)$$

если взять производную и приравнять к нулю, то получим

$$a = \ln \sqrt{\frac{p}{1 - p}}$$

А это как вероятность превратить в оценку на $(-\infty, +\infty)$

Exploss: почему логичная функция

Подставляем…

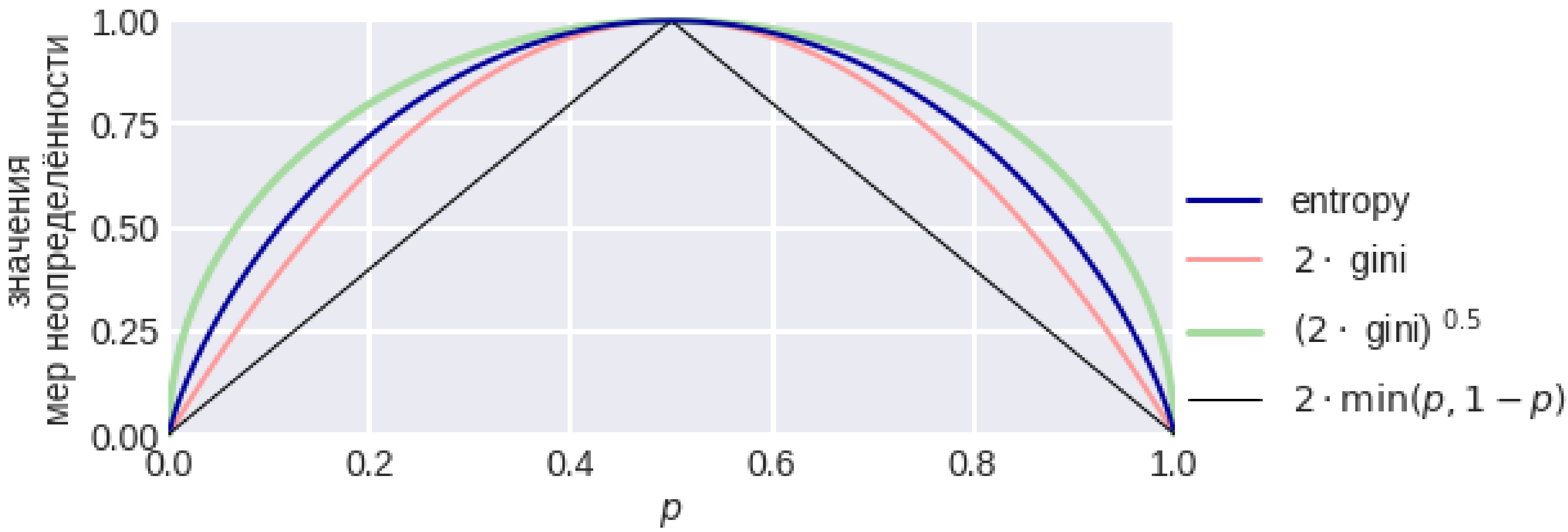
$$\exp(-y\,a) = \exp\left(-y\ln\sqrt{\frac{p}{1-p}}\right) =$$

$$= \left(\sqrt{\frac{p}{1-p}}\right)^{-y}$$

– выражение exploss
(вместо ответов алгоритма там стоит вероятность)

Таким образом, это «естественная поправка» экспоненты, если мы хотим ответы нашего алгоритма интерпретировать как вероятности.

Критерии расщепления



Любая скоринговая функция порождает информационную меру, которая может быть использована в критерии расщепления

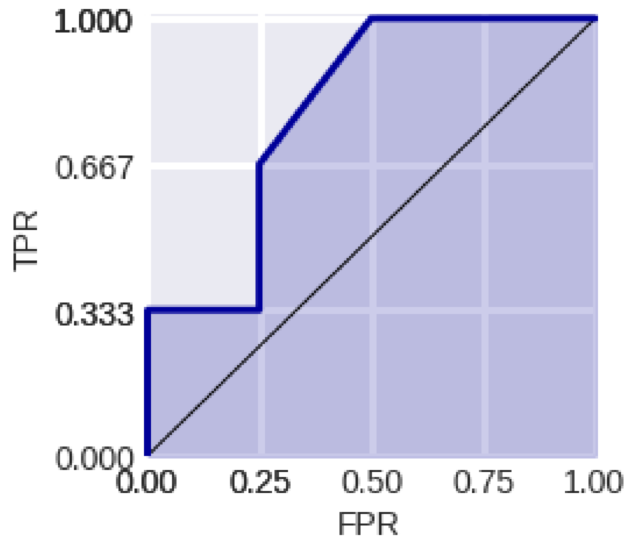
Функция ошибки	Минимальное матожидание
LogLoss	$-p\log(p)-(1-p)\log(1-p)$
MSE	$1-p^2-(1-p)^2=2p(1-p)$
ExpLoss	$2\sqrt{p(1-p)}$
ME	$\min(p,1-p)$ **ДЗ где ещё?**

ROC и AUC ROC

Функционал зависит не от конкретных значений, а от их порядка

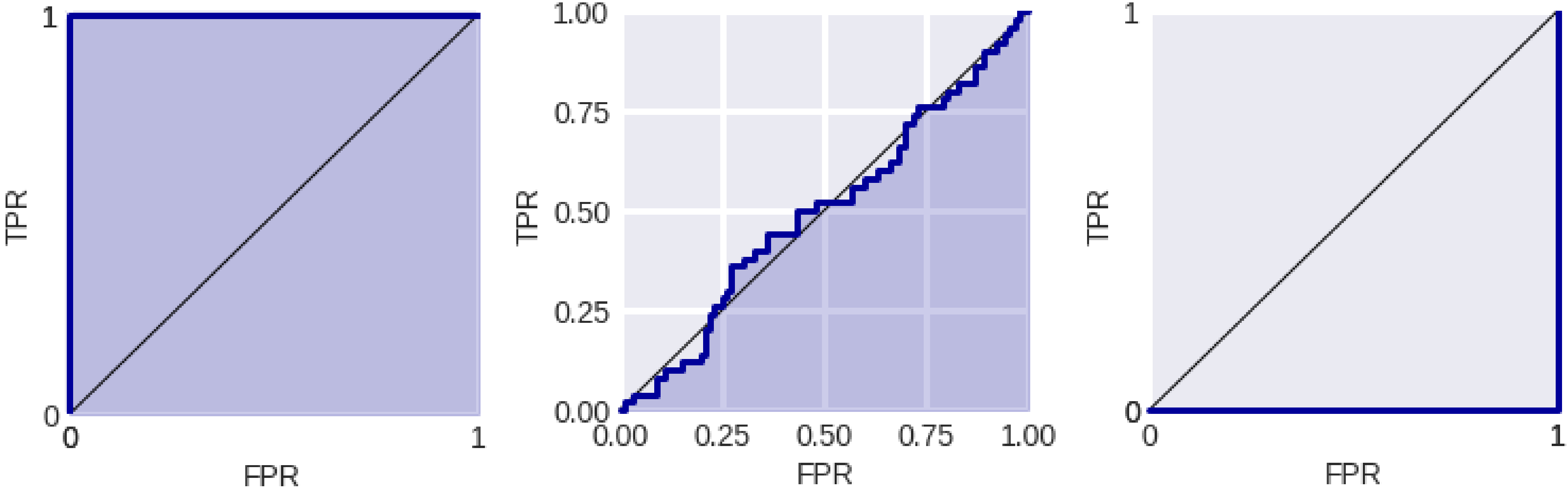
	оценка	класс
0	0.5	0
1	0.1	0
2	0.2	0
3	0.6	1
4	0.2	1
5	0.3	1
6	0.0	0

	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

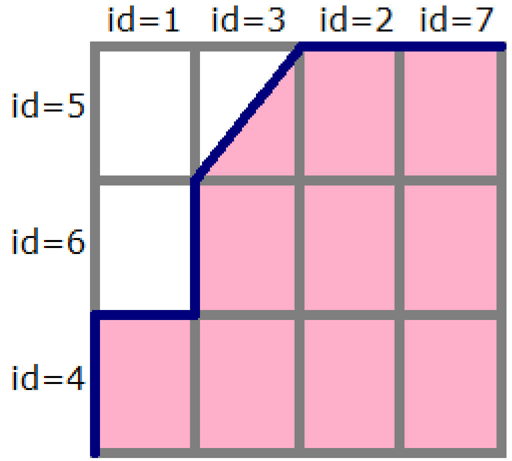
```
from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_test, a)
plt.plot(fpr, tpr, lw=3, c='#000099')
```

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 (требует пояснения) в формуле?

Смысл AUC

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

Это сложно объяснить заказчику!

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

Чем хороша эта запись?

Можно обобщить, например, на регрессию.

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

Обобщения AUC

Иногда используют «естественные обобщения»:

$$\mathrm{AUC} = \frac{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} I[y_i < y_j]\cdot \max(a_j - a_i, 0)}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} I[y_i < y_j]}$$

Если есть веса объектов…

как обобщить AUC?

Обобщения AUC

Напишем, что есть FPR, TPR

$$\mathrm{TPR} = \frac{\sum_{i=1}^{m} I[a_i = 1] I[y_i = 1]}{\sum_{i=1}^{m} I[y_i = 1]} \qquad \mathrm{FPR} = \frac{\sum_{i=1}^{m} I[a_i = 1] I[y_i = 0]}{\sum_{i=1}^{m} I[a_i = 1]}$$

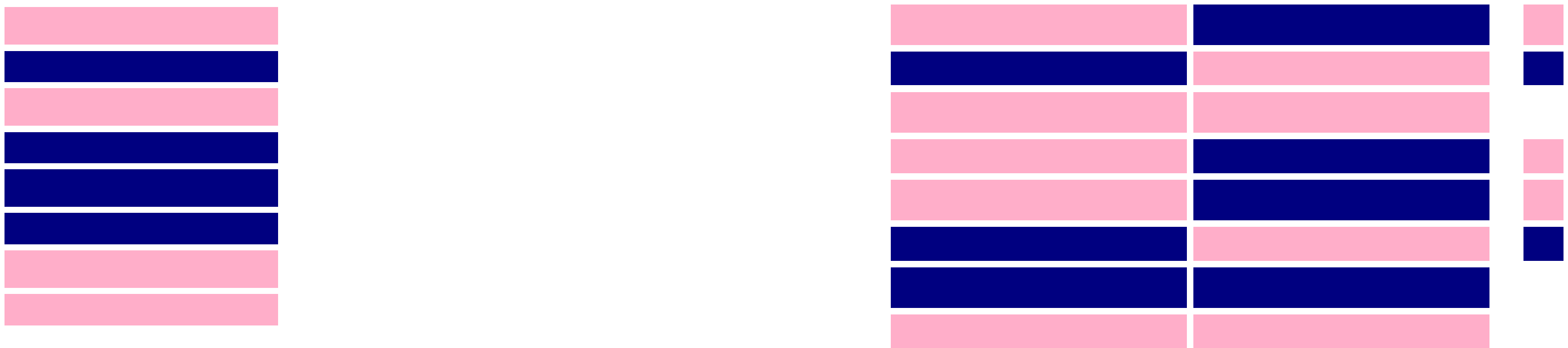
Теперь всё ясно…

$$\mathrm{TP} = \frac{\sum_{i=1}^{m} w_i I[a_i = 1] I[y_i = 1]}{\sum_{i=1}^{m} w_i I[y_i = 1]}$$

ДЗ Исследовать взвешенный AUC ROC.

Настройка RF/GBM на AUC ROC

Случай из жизни (Интернет-математика)

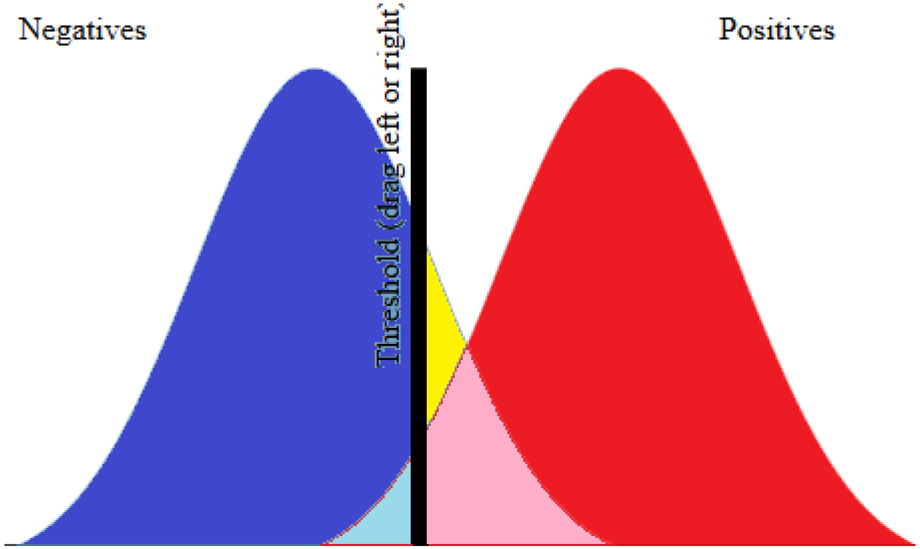


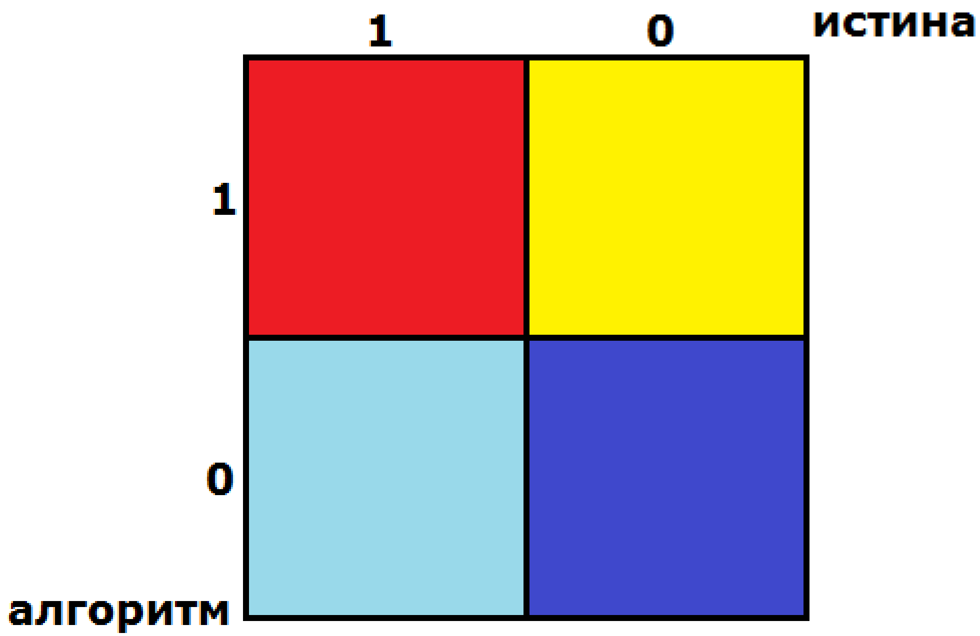
классификация → классификация пар

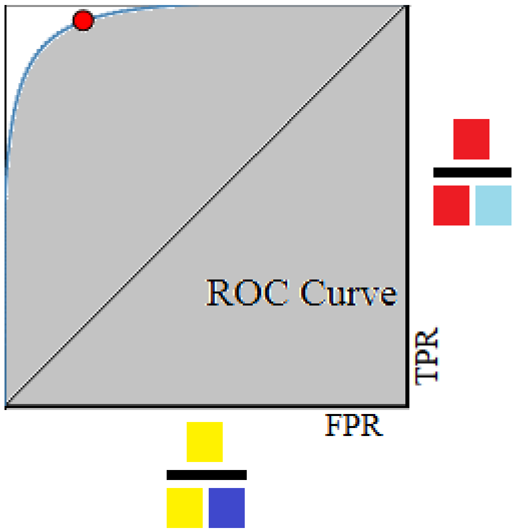
Можно дублировать,

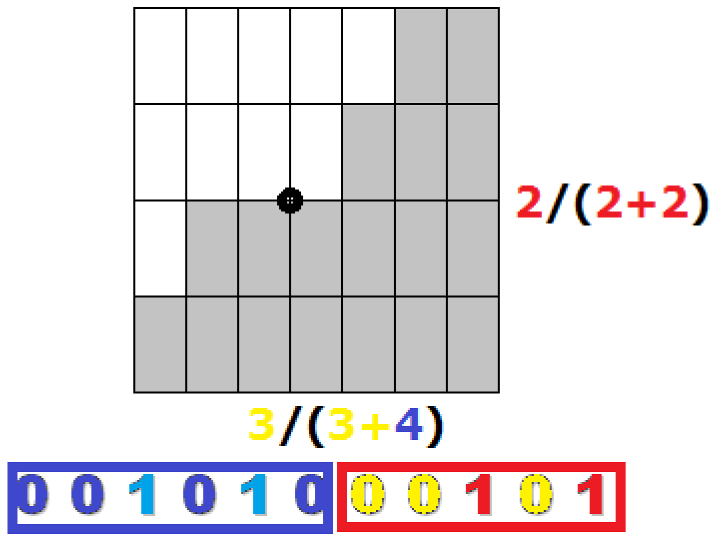
Можно брать разности/отношения.

AUC ROC







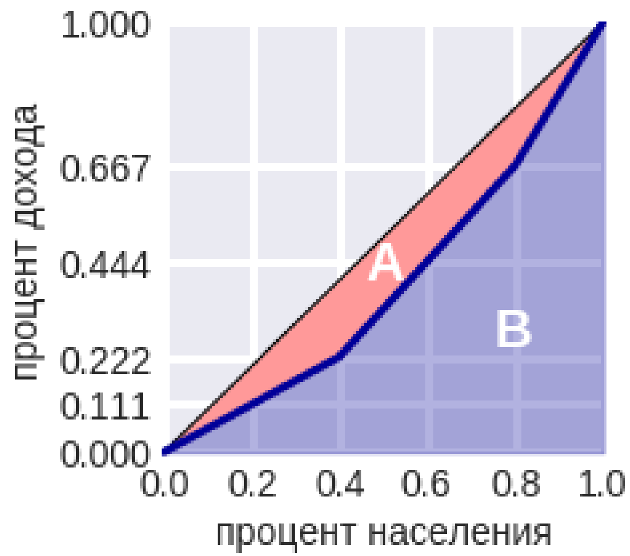


AUC – не всегда ступеньки!

GINI

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

Кривая Лоренца

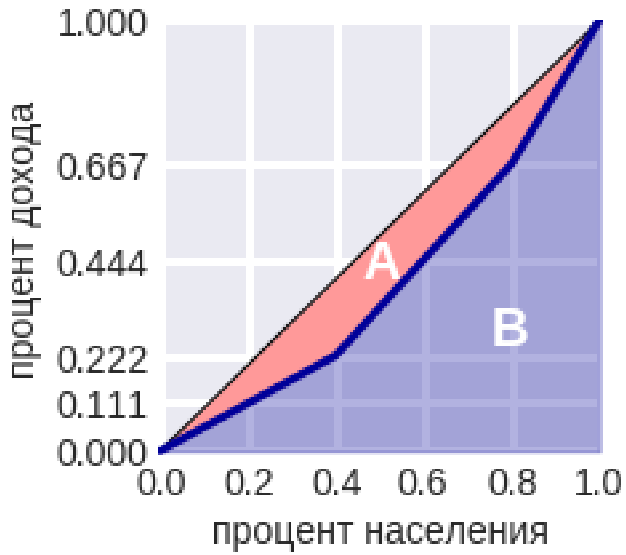


Пример для доходов: 1, 1, 2, 2, 3

40% населения имеют 2/9 дохода.

GINI

Вычисление



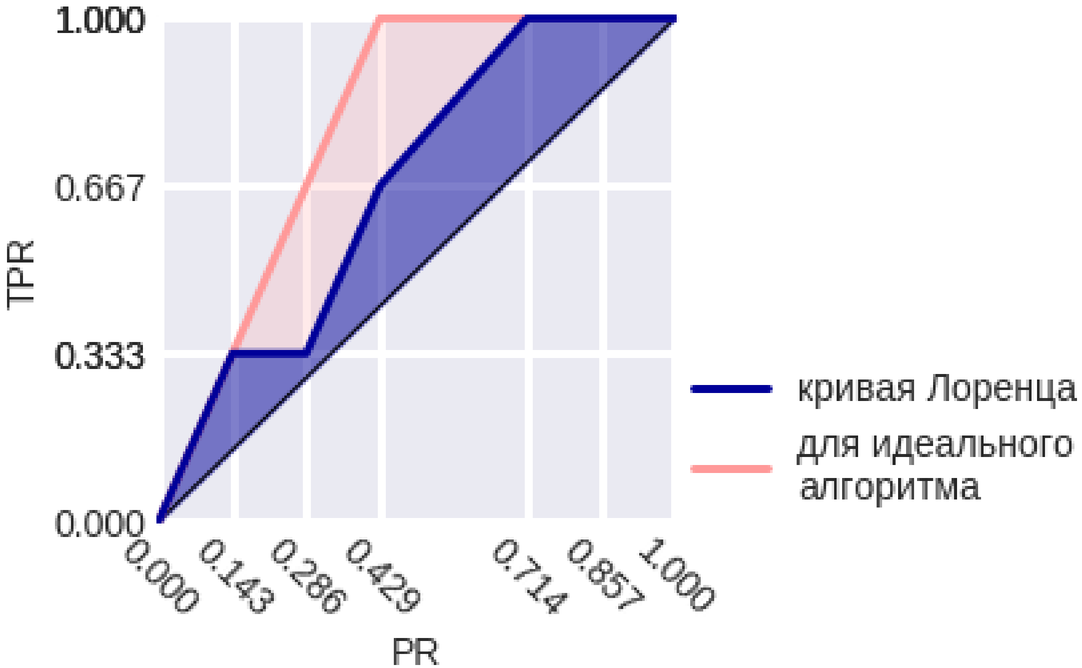
$$\text{gini} = \frac{A}{A+B} = 2A$$

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

не путать с Gini impurity

GINI в машинном обучении

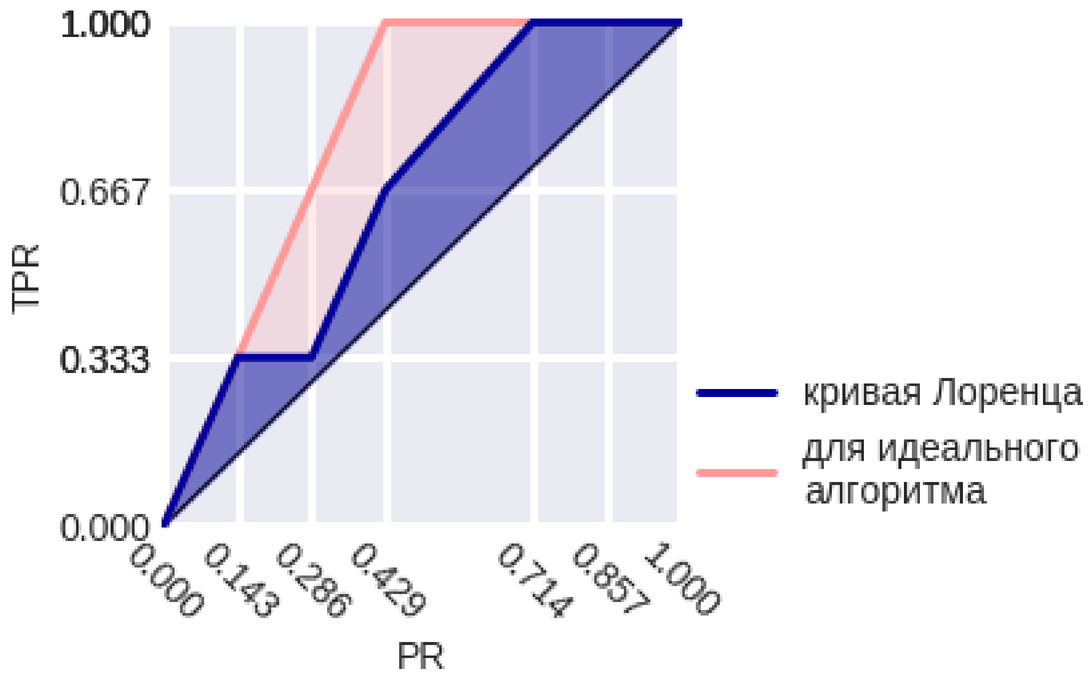
Кривая Лоренца (или CAP – Cumulative Accuracy Profile Curve)



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

Коэффициент Джини – отношение площадей / (+) =7/12

GINI в машинном обучении



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

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

GINI в машинном обучении

$$\mathrm{gini} = \frac{2}{q_1 q_0}\int_0^1 \mathrm{TP}\,\partial(\mathrm{FP}+\mathrm{TP}) - \frac{q_0+q_1}{q_0} =$$

$$= 2\,\mathrm{AUCROC} + \frac{2}{q_1 q_0}\int_0^1 \mathrm{TP}\,\partial\,\mathrm{TP} - \frac{q_1}{q_0} - 1$$

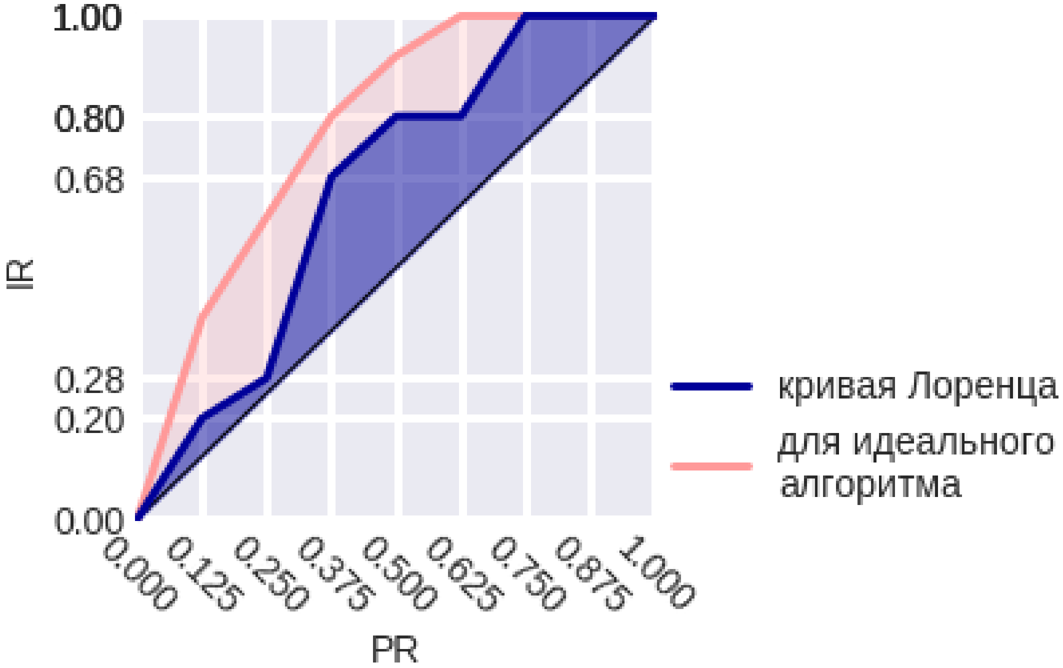
$$\mathrm{gini} = 2\,\mathrm{AUCROC} - 1$$

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

0.9 AUC = 0.8 gini

GINI в задаче регрессии

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

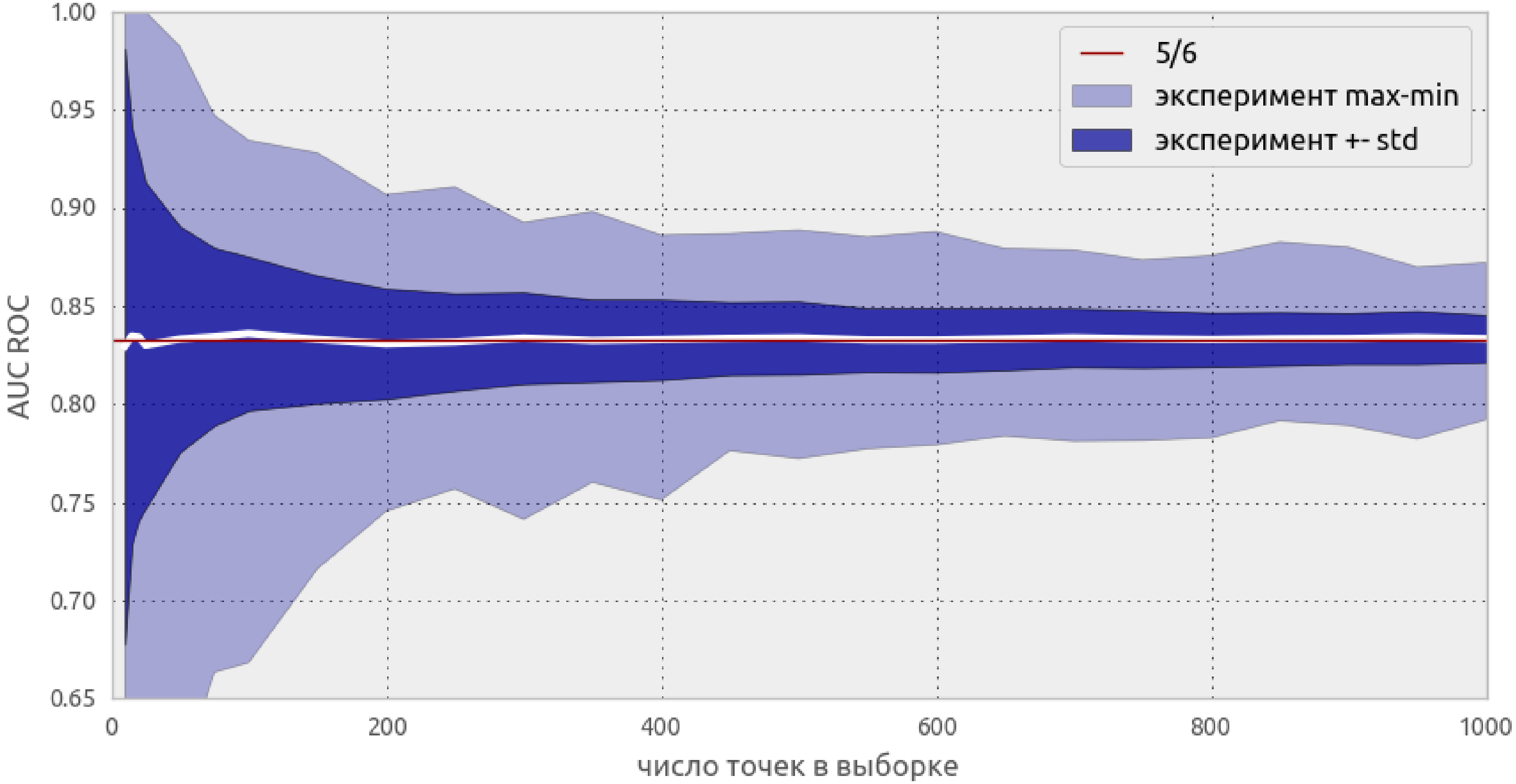


Идеальный алгоритм: 10, 5, 5, 3, 2, 0, 0, 0

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

AUC ROC

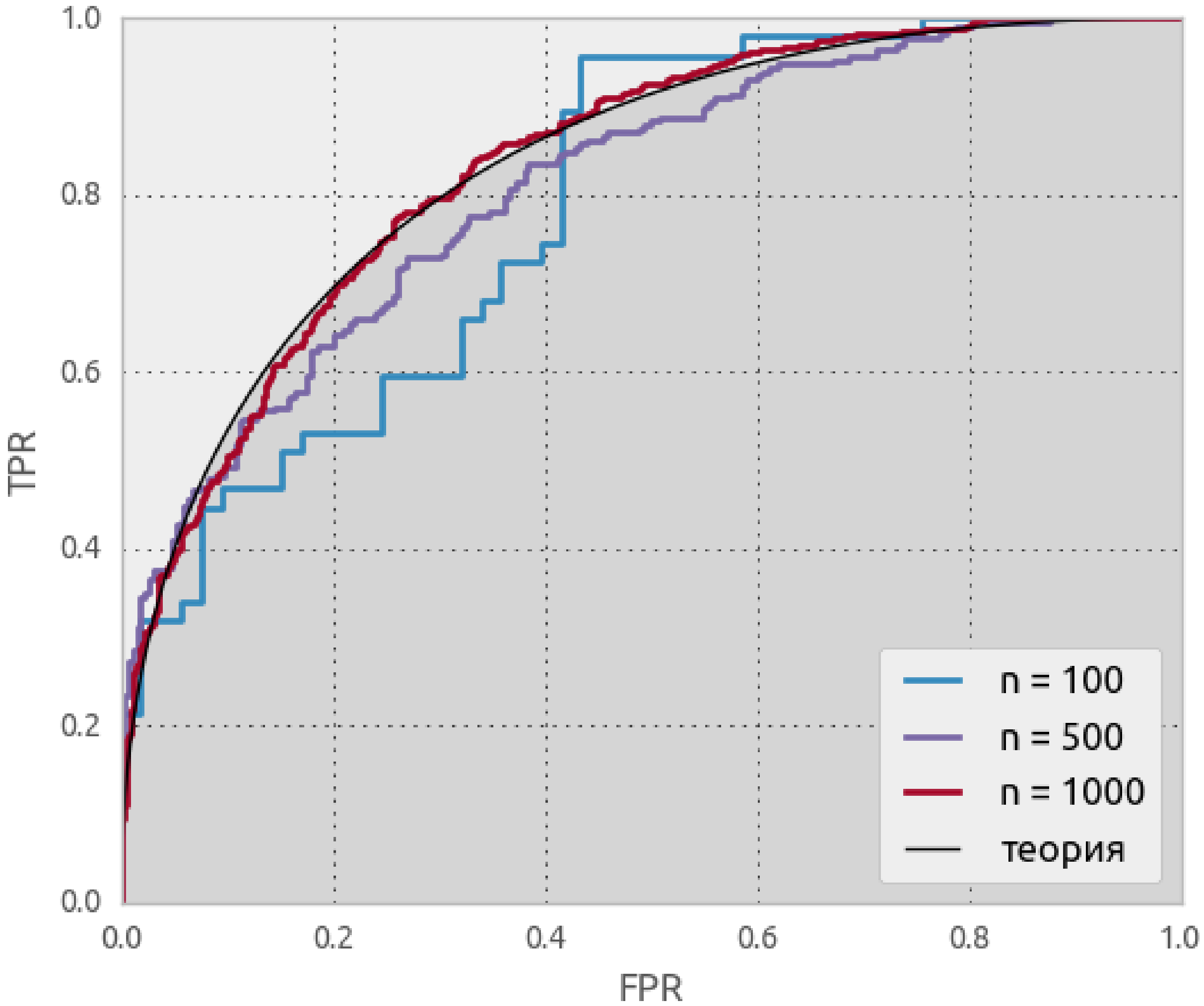
Если задаться распределениями классов (на ответах алгоритма) и получать оценку AUC ROC



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

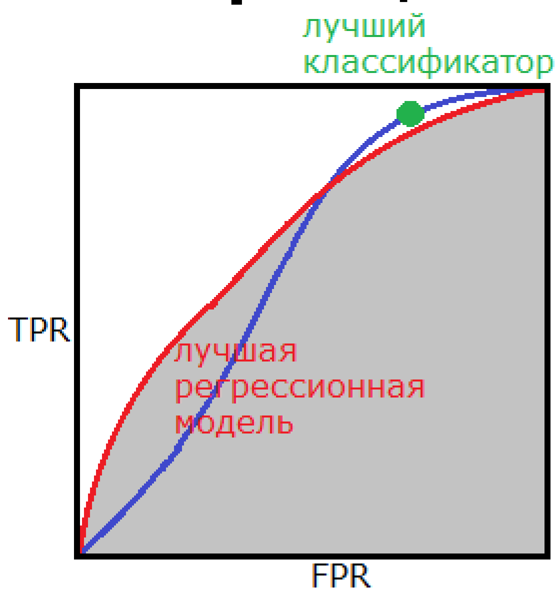
+ можно использовать для оценки признаков

- «завышает» качество

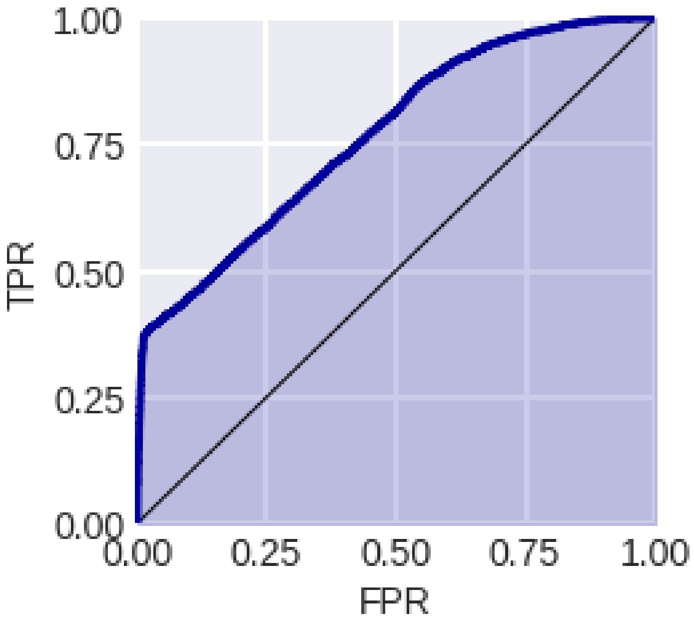
- оценивает не конкретный классификатор, а регрессию

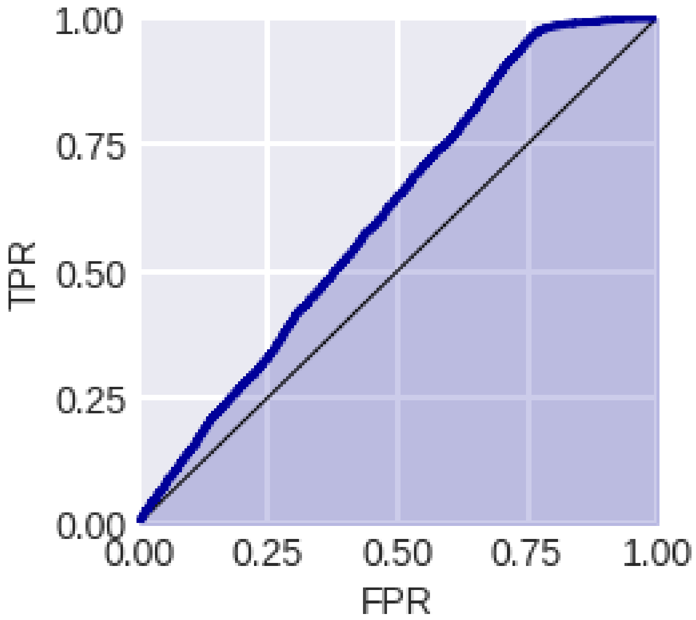
- сложно объяснить заказчику

- не путать классификацию и регрессию



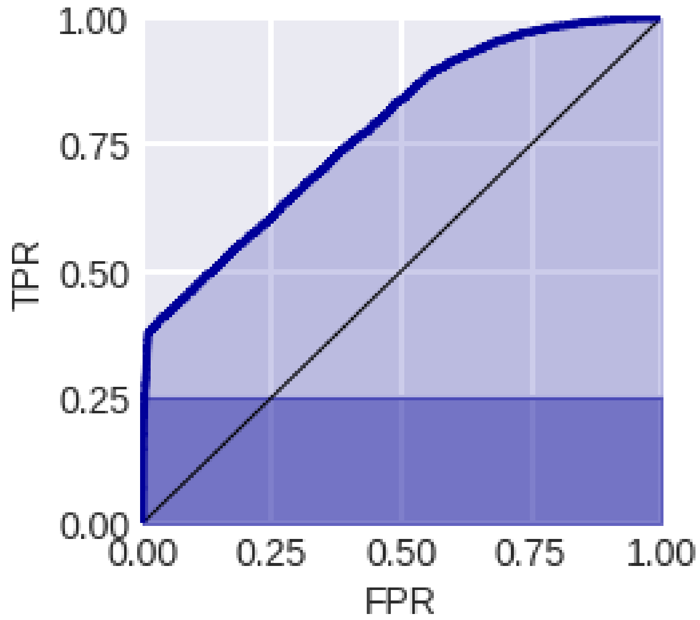
Маленький AUC не всегда плохо

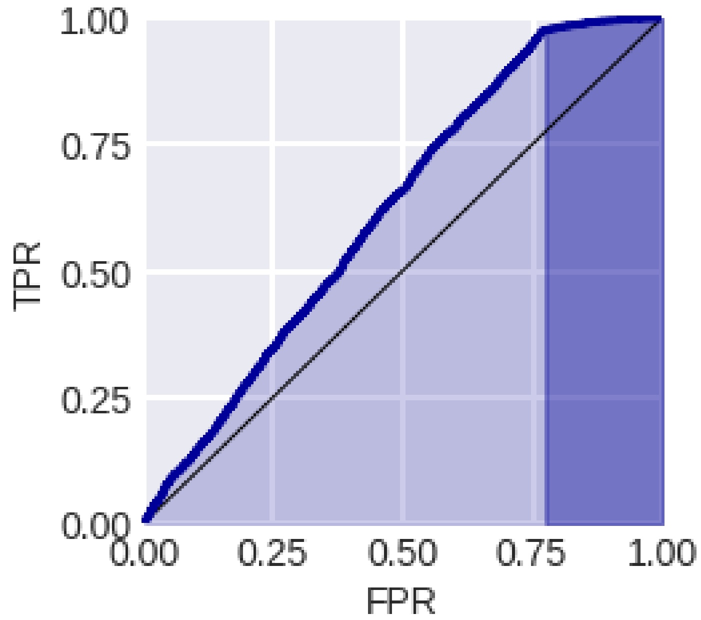




Чем хороши эти ROC-кривые?

Маленький AUC не всегда плохо





y = [... 1, 1, 1, 1]
если оценка большая почти всегда это правда «класс 1»

y = [0, 0, 0, 0, ...]
если оценка маленькая почти всегда это правда «класс 0»

Можем хорошо отделить часть объектов одного класса

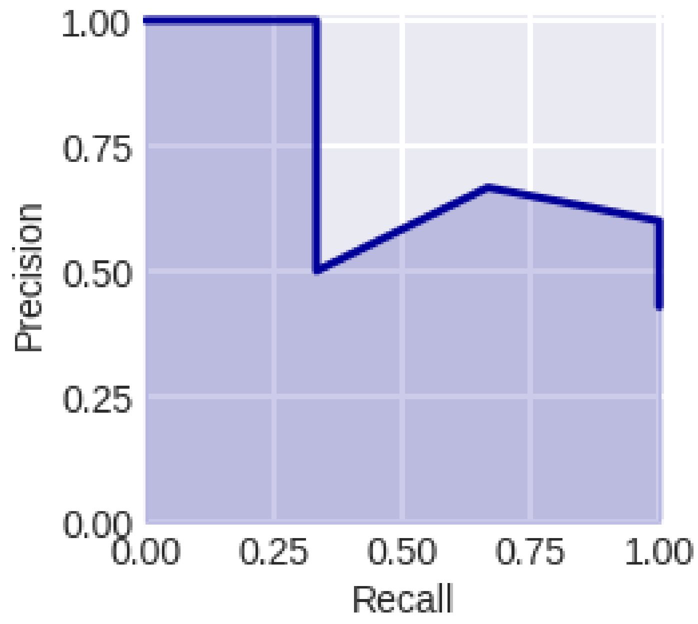
Пример: клиенты, которые точно не купят билет (чтобы предложить его им со скидкой)

Ещё примеры кривых... «полнота-точность»

Площадь под кривой.. «Average Precision» (есть и другой смысл)

	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




```
from sklearn.metrics import precision_recall_curve
precision, recall, thresholds = precision_recall_curve(y_test, a)
plt.plot(recall, precision)
# вычисление площади методом трапеций
from sklearn.metrics import auc
auc(recall, precision)
# или готовую функцию использовать
from sklearn.metrics import average_precision_score
```

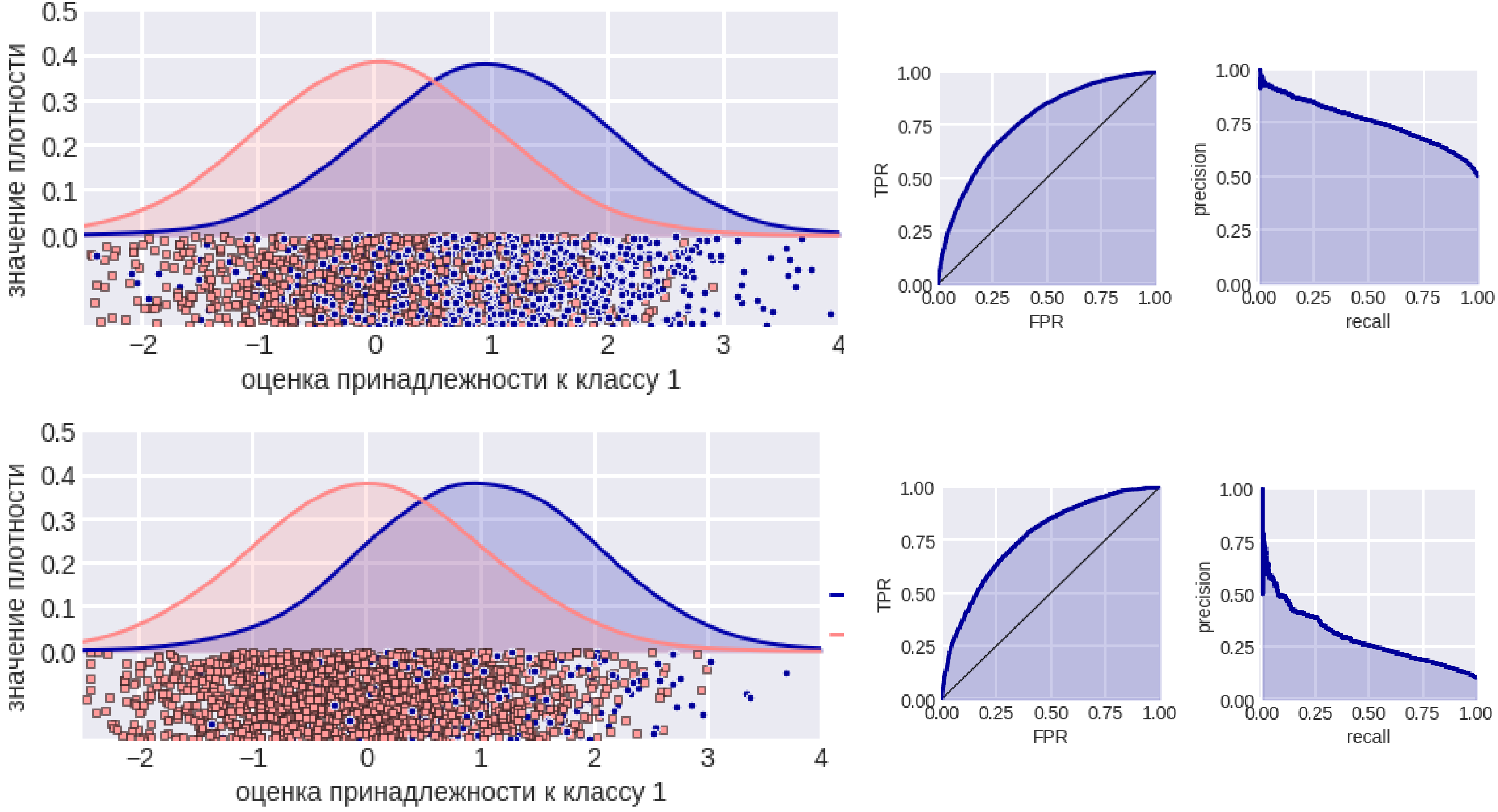
Максимизация AUC ROC

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

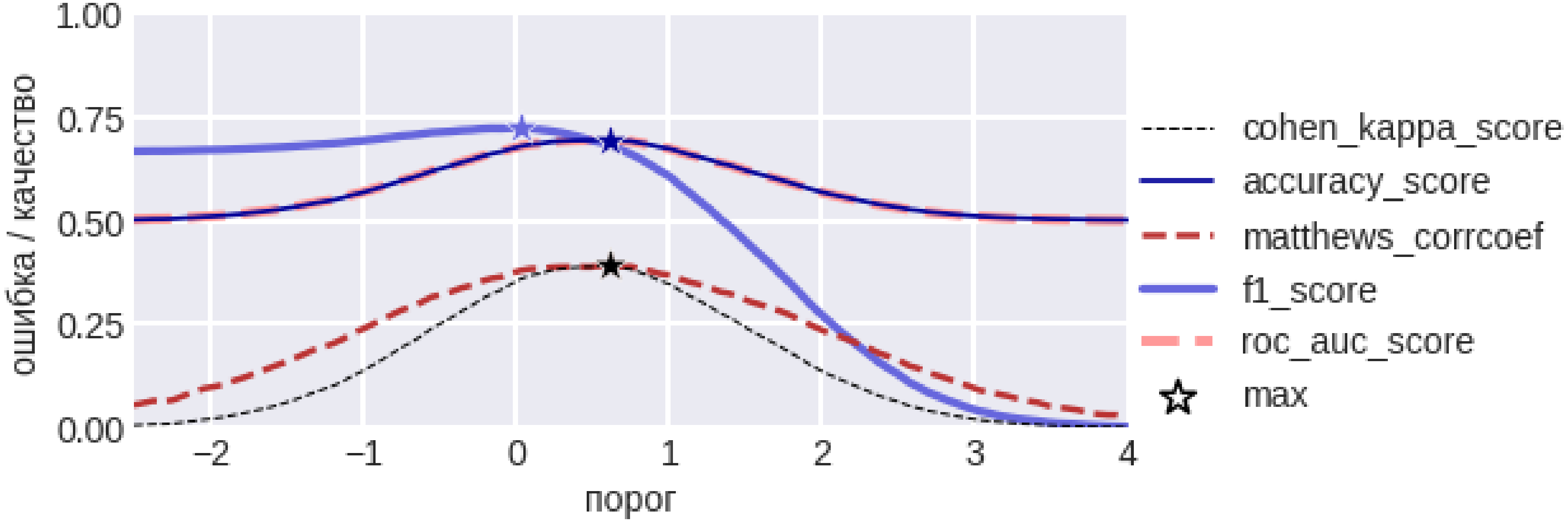
ДЗ Пройти тест goo.gl/93qkum

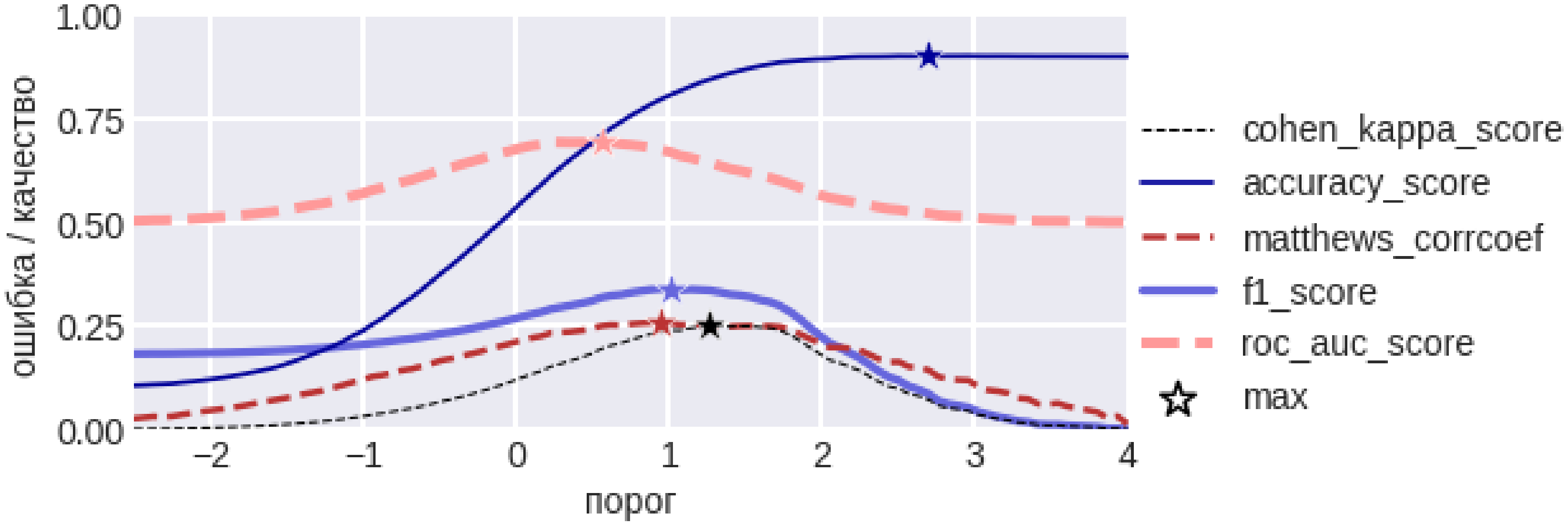
Сравнение метрик в задачах классификации

Модельные задачи



Сравнение метрик в задачах классификации

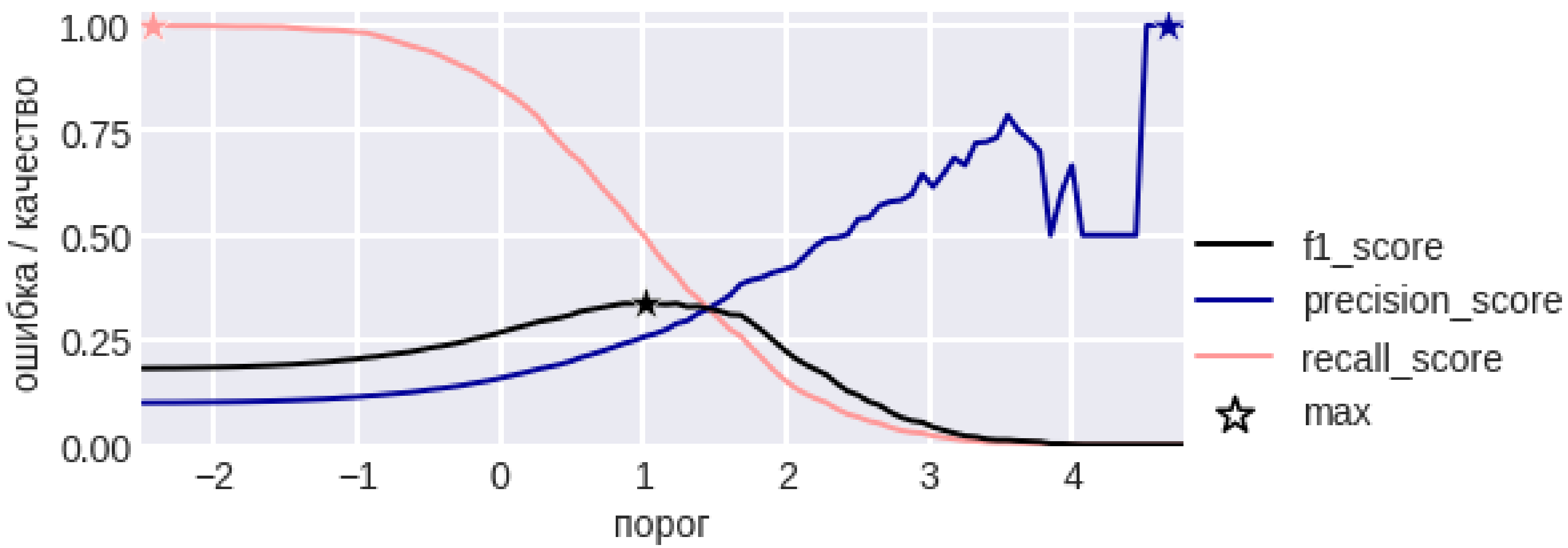




Сравнение метрик в задачах классификации

ДЗ Совпадение максимумов – случайность?
Обосновать!

ROC AUC у бинарного ответа (потом обсудим)
но это же совпадает с balanced_accuracy_score (ниже)



Почему прыгает точность?

Минутка кода

```
from sklearn.metrics import classification_report
print (classification_report(y_test, a_test)) # нужен print

              precision    recall  f1-score   support

         0.0       0.94      0.84      0.89     16199
         1.0       0.26      0.50      0.34      1801

   micro avg       0.81      0.81      0.81     18000
   macro avg       0.60      0.67      0.61     18000
weighted avg       0.87      0.81      0.83     18000
```

```
from sklearn.metrics import cohen_kappa_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import matthews_corrcoef
from sklearn.metrics import f1_score
from sklearn.metrics import roc_auc_score
from sklearn.metrics import balanced_accuracy_score
```

	score
cohen_kappa_score	0.24
accuracy_score	0.81
matthews_corrcoef	0.26
f1_score	0.34
roc_auc_score	0.67
balanced_accuracy_score	0.67

Совет

Ищите матожидание!

Пробуйте константные решения.

Многоклассовая задача «Multi-label»

матрица классификаций

$$\| y_{ij} \|_{m \times l}$$

	class 1	class 2	class 3
0	1	0	0
1	0	1	0
2	0	0	1
3	1	1	0

матрица ответов

$$\| a_{ij} \|_{m \times l}$$

	class 1	class 2	class 3
0	0.75	0.00	0.25
1	0.00	0.50	0.25
2	0.25	1.00	0.25
3	0.00	0.25	0.75

По сути, надо сравнить матрицы на похожесть

- **можно сравнивать матрицы (~МИКРО-подход)**
- **можно сравнивать строки матриц (~ПО ОБЪЕКТАМ) (и усреднять)**
- **можно сравнивать столбцы матриц (~МАКРО-подход)**

Многоклассовая задача: Hamming Loss

(функционал качества)

Число (процент) ошибок

$$a_{ij} \in \{0,1\}$$

$$\mathrm{HL} = \frac{1}{ml} \sum_{i=1}^{m} \sum_{j=1}^{l} I[y_{ij} = a_{ij}]$$

(1 – точность)

Многоклассовая задача: Log Loss (cross-entropy)

Естественное обобщение логистической ошибки

$$a_{ij} \in [0,1]$$

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

(тонкость: лучше для непересекающихся классов)

Многоклассовая задача: Mean Probability Rate

(это функционал качества, $a_{ij} \in [0,1]$ ~ распределение)

Есть макро-версия, см. дальше

$$\mathrm{MPR} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{l} y_{ij}a_{ij}$$

$$\mathrm{MAPR} = \frac{1}{l}\sum_{j=1}^{l}\frac{\sum_{i=1}^{m} y_{ij}a_{ij}}{\sum_{i=1}^{m} y_{ij}}$$

Многоклассовая задача: MSE, MAE

$$\mathrm{MSE} = \frac{1}{ml}\sum_{i=1}^{m}\sum_{j=1}^{l}(y_{ij} - a_{ij})^2$$

$$\mathrm{MAE} = \frac{1}{ml}\sum_{i=1}^{m}\sum_{j=1}^{l}|\, y_{ij} - a_{ij}\,|$$

это всё вариации на тему схожести / различия бинарного и вещественного вектора

Многоклассовый AUCROC: Макро-усреднение

$$\mathrm{AUC} = \frac{1}{l}\sum_{j=1}^{l}\mathrm{AUC}_j$$

AUC_j **– значение функционала в задаче бинарной классификации «j-й класс, не j-й класс».**

$$\{(x_i, I[y_j = 1])\}_{i=1}^{m}$$

Многоклассовый AUCROC: Весовое макро-усреднение

$$\mathrm{AUC} = \frac{\sum_{j=1}^{l} P_j\,\mathrm{AUC}_j}{\sum_{j=1}^{l} P_j}$$

P_j **– вероятность j-го класса**

(процент «1» в столбце матрицы классификации)

Многоклассовый AUCROC: Микро-усреднение

значение функционала в задаче

$$\{((x_i, j), I[y_j = 1])\}_{i=1, j=1}^{m, l}$$

«вытягиваем матрицу ответов в вектор»

Многоклассовый AUCROC: Усреднение по объектам

$$\mathrm{AUC} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{AUC}'_i$$

AUC'_i **– значение функционала в задаче**

$$\{((x_i, j), I[y_j = 1])\}_{j=1}^{l}$$

«решение задачи по строкам»

Многомерный AUC: минутка кода

```
from sklearn.metrics import roc_auc_score

roc_auc_score(y, a, average='macro')
# эквивалентно:
auc_pclass = [roc_auc_score(y[:,i], a[:,i]) for i in range(l)]
auc_pclass, mean(auc_pclass)

roc_auc_score(y, a, average='micro')
# эквивалентно:
roc_auc_score(y.ravel(), a.ravel())

roc_auc_score(y, a, average='weighted')
# эквивалентно:
w = y.sum(axis=0)
sum(np.array(auc_pclass) * w) / sum(w)

roc_auc_score(y, a, average='samples')
# эквивалентно:
auc_pinstance = [roc_auc_score(y[i,:], a[i,:]) for i in range(m)]
auc_pinstance, mean(auc_pinstance)
```

Многомерный AUC ROC

матрица классификаций

	class 1	class 2	class 3
0	1	0	0
1	0	1	0
2	0	0	1
3	1	1	0

macro	micro	weighted	samples
0.49	0.53	0.52	0.56

матрица ответов

	class 1	class 2	class 3
0	0.75	0.00	0.25
1	0.00	0.50	0.25
2	0.25	1.00	0.25
3	0.00	0.25	0.75

	class 0	class 1	class 2
AUC_per_class	0.62	0.5	0.33
P_per_class	0.50	0.5	0.25

	class 0	class 1	class 2	class 3
AUC_per_instance	1.0	1.0	0.25	0.0

Как на практике использовать особенности функционалов

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

Точность: сравнение макро- и микро- усреднения

$$\mathrm{P}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j}$$

$$\mathrm{P}_{\text{macro-mean}} = \frac{1}{l}\sum_{j=1}^{l} \mathrm{P}_j$$

$$\mathrm{P}_{\text{micro-mean}} = \frac{\sum_{j=1}^{l} \mathrm{TP}_j}{\sum_{j=1}^{l} \mathrm{TP}_j + \sum_{j=1}^{l} \mathrm{FP}_j}$$

Кстати, макро-усреднение делают по-разному.

Часто: среднее геометрическое.

Точность: сравнение макро- и микро- усреднения

При вычислении каких-то функционалов, например точности

МАКРО	• вычислить точность для каждого класса • усреднить полученные точности `(H.TP / (H.TP + H.FP)).mean()`	0.344

МИКРО	• вычислить точность сразу для всех классов
	
	`TP = H.TP.sum() + H.TP.sum()`
	`FP = H.FP.sum() + H.FP.sum()`
	`P = TP / (TP + FP)`
		0.246

	TP	FP
class 1	2	2
class 2	5	10
class 3	10	40

	P
class 1	0.50
class 2	0.33
class 3	0.20

Точность: сравнение макро- и микро- усреднения

	TP	FP
class 1	2	2
class 2	5	10
class 3	100	400

не изменились точности по классам ⇒ не изменилась макро-точность 0.344

изменились TP, FP по классам ⇒ микро-точность смещается в сторону «большего» класса: 0.206 **(вместо** 0.246**)**

где какое усреднение лучше использовать?

Совет: смотреть дисперсию показателей по классам

F-мера – ещё больше вариантов усреднения

Макро F-мера

$$F_{\text{macro-mean}} = \frac{1}{l}\sum_{j=1}^{l} F_j$$

на основе других макро-параметров

$$F_{\text{macro-mean}} = \frac{2}{\dfrac{1}{P_{\text{macro-mean}}} + \dfrac{1}{R_{\text{macro-mean}}}}$$

ДЗ провести сравнение!

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
...
```

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

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

Сбалансированная точность «Balanced accuracy»

– макро-усреднение полноты

Сбалансированная точность (accuracy) не есть усреднение точностей (precision)

$$\mathrm{BA} = \frac{1}{l}\sum_{j=1}^{l} R_j = \frac{1}{l}\sum_{j=1}^{l} \frac{\sum_{t=1}^{m} I[y_t = j] I[a_t = j]}{\sum_{t=1}^{m} I[y_t = j]}$$

```
from sklearn.metrics import balanced_accuracy_score
```

Другие (неэквивалентные) определения:

$$\mathrm{BA} = \frac{1}{l}\sum_{j=1}^{l} \min[P_j, R_j]$$

$$\mathrm{BA} = \frac{1}{l}\sum_{j=1}^{l} \min[\mathrm{sens}_j, \mathrm{spec}_j]$$

ДЗ Сравните разные подходы

Оценка результатов поиска/рекомендаций

Задача с бинарной релевантностью

$$x_1 \prec x_2 \prec \ldots \prec x_q$$

$y_i = 1$ – **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

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

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):

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

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

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

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

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

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

Mean Average Precision

– усреднение ap@n по всем пользователям

Concordant – Discordant ratio

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

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

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

$x_1,\ldots,x_n$ – **упорядоченный** **список ответов**

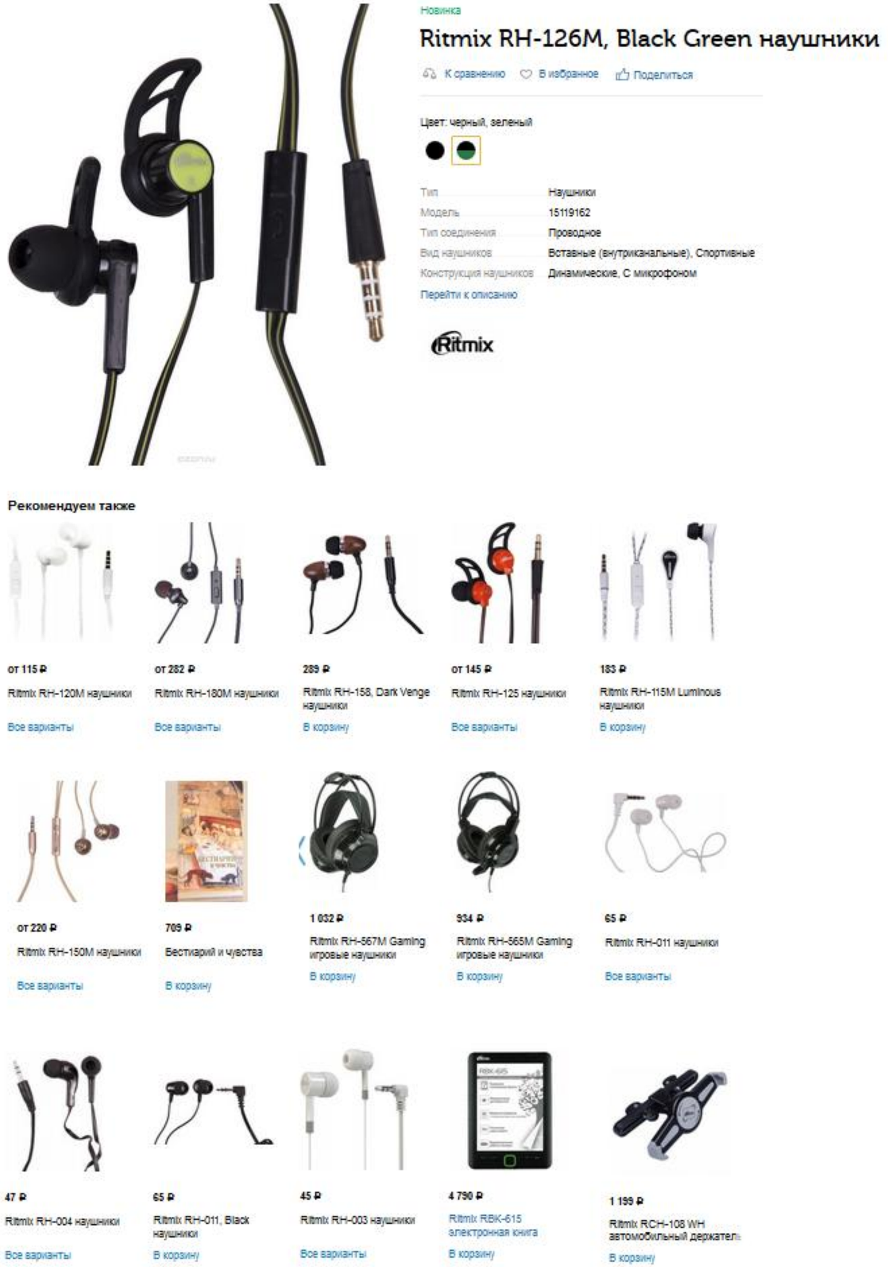
$x'_1,\ldots,x'_m$ – **все релевантные**

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

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

когда логично применить?

Рекомендации



Mean Reciprocal Rank (MRR)

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

$$RR = \frac{1}{\min\{i : y_i = 1\}}$$

Часто оптимизируют именно его!

Классические функционалы в поиске

Случай небинарной релевантности

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

$$y_1, \ldots, y_q$$

Cumulative Gain at n

$$\mathrm{CG@}n = y_1 + \ldots + y_n$$

Discounted Cumulative Gain at n

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

Ещё вариант:

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

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

$$x_1 \prec x_2 \prec \ldots \prec x_q$$

Правильный порядок

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

Надо сравнить:

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

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

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)

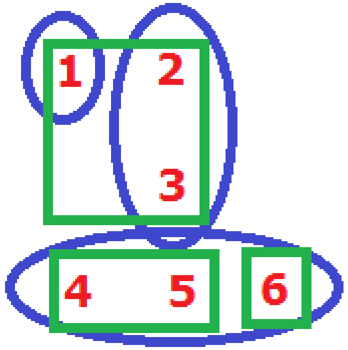
$$ERR@n = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{k}p_k\prod_{i<k}(1-p_i)$$

Как интерпретировать?

Редакторское расстояние

Операции

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



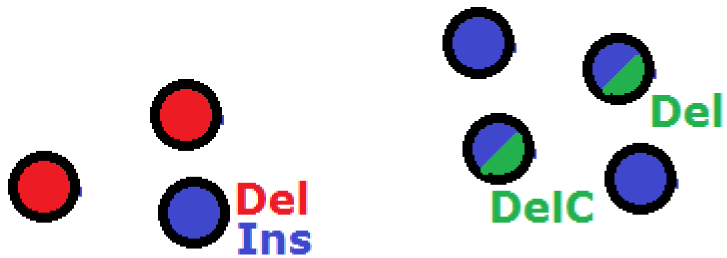
1 2 3;4 5;6
1 2 3; 4 5 [delC]
2 3; 4 5 [del]
2 3; 4 5; 1 [insC]
2 3; 4 5 6; 1 [ins]

	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

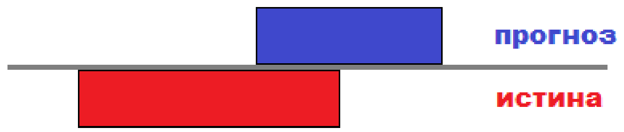
$\Rightarrow$ осторожный алгоритм



- Многое зависит от операций…

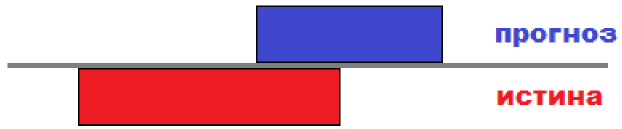
Задача с «неклассическим целевым вектором»

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 (Sörensen)

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

коэффициент Кульчинского (Kulczinsky)

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

коэффициент Отиаи (Ochiai)

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

Меры включения

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

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

Как решать задачи с интервалами? Потом вернёмся…

Литература

Tom Fawcett An introduction to ROC analysis // Pattern Recognition Letters V.27 № 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