курс «Машинное обучение»

Функции ошибки / функционалы качества
Часть 3: скоринговые функции и кривые в машинном обучении

Александр Дьяконов

03 марта 2021 года

План на эти несколько лекций

задача регрессии

задача бинарной классификации

- чёткая классификация
- скоринговые функции

кривые в ML

задача классификации с несколькими классами

задачи ранжирования

задачи кластеризации

Задача бинарной классификации

Теперь выдаём оценку принадлежности к классу 1

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

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

кроме меток {0, 1} возможны промежуточные значения

Приём: представление функции ошибки

Однозначное задание функции ошибки

«раздельная форма записи»

$$L(a,y)=\begin{cases}L(a,1), & y=1,\\ L(a,0), & y=0,\end{cases}$$

часто $L(a,1)=L(1-a,0)$

Используем представление

«совместная форма записи»

$$L(a,y)=yL(a,1)+(1-y)L(a,0)$$

Log Loss

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

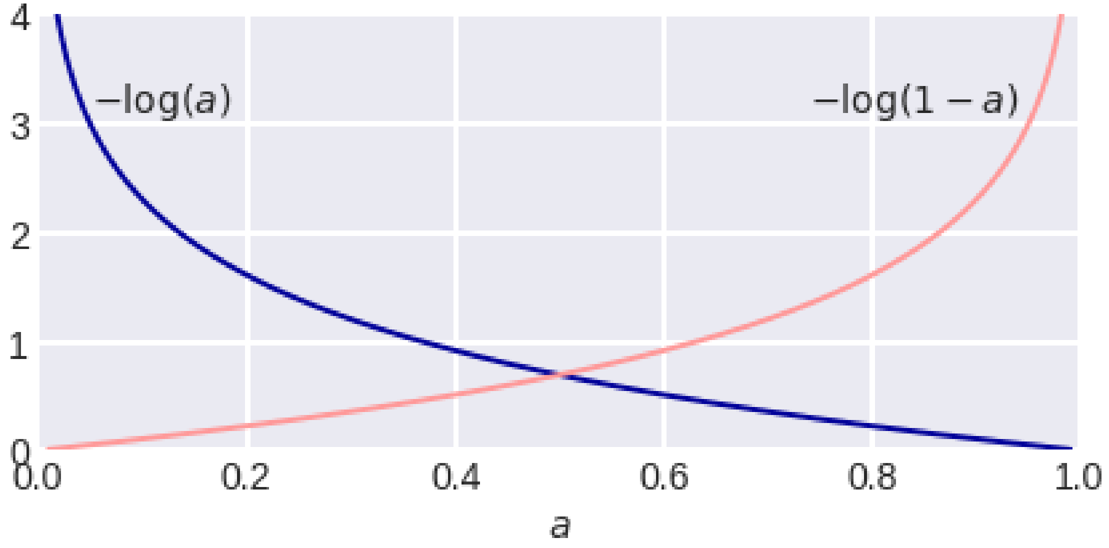
$$\text{logloss} = -\frac{1}{m}\sum_{i=1}^{m}(y_i \log a_i + (1-y_i)\log(1-a_i))$$

На что похоже?

Раздельная форма понятнее…

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

Нельзя ошибаться!



Откуда берётся Log Loss

Обучающая выборка ~ реализация обобщённой схемы Бернулли:

для x_i **генерируем**

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

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 для конечной выборки

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

$$-\frac{m_1}{m}\log a - \frac{m_0}{m}\log(1-a) \to \min_a$$

$$a = \frac{m_1}{m}$$

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! Так и должно быть, но не всегда бывает…

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

$$\left[0, -\frac{m_1}{m}\log\frac{m_1}{m} - \frac{m_0}{m}\log\frac{m_0}{m}\right]$$

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

Какие недостатки log_loss?

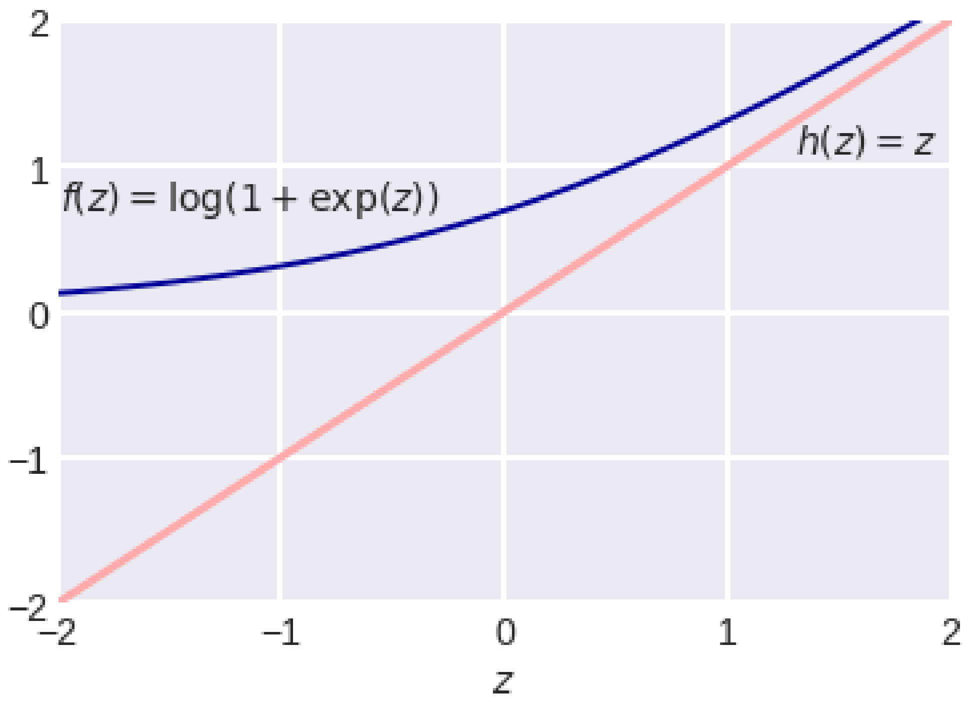
Его значение неинтерпретируемы…

Связь с логистической регрессией см. лекцию про минимизацию

Другая форма функционала

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

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



LogReg

$$\sum_i \log(1 + \exp(-y_i \cdot w^{\mathrm{T}} x_i)) \to \min$$

SVM – Hinge Loss

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

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

$$D_{\mathrm{KL}}(P \| Q) = \int p(z) \log \frac{p(z)}{q(z)} \partial z = \underbrace{H(p,q)}_{\text{кросс-энтропия}} - \underbrace{H(p)}_{\text{энтропия}}$$

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

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 – методы калибровки

Непараметрические

Histogram Binning

Isotonic Regression

Модификации бининга

Scaling-binning calibrator

Смешанные

Probability calibration trees

Параметрические

Калибровка Платта (Platt calibration)

Логистическая регрессия в пространстве логитов

Matrix and Vector Scaling

Beta calibration

Ансамблирование

Ensemble of near-isotonic regression (ENIR)

Bayesian Binning into Quantiles (BBQ)

DL

Temperature Scaling

Maximum Mean Calibration Error (MMCE)

Label smoothing, Entropy penalty, Focal loss

Drop Out

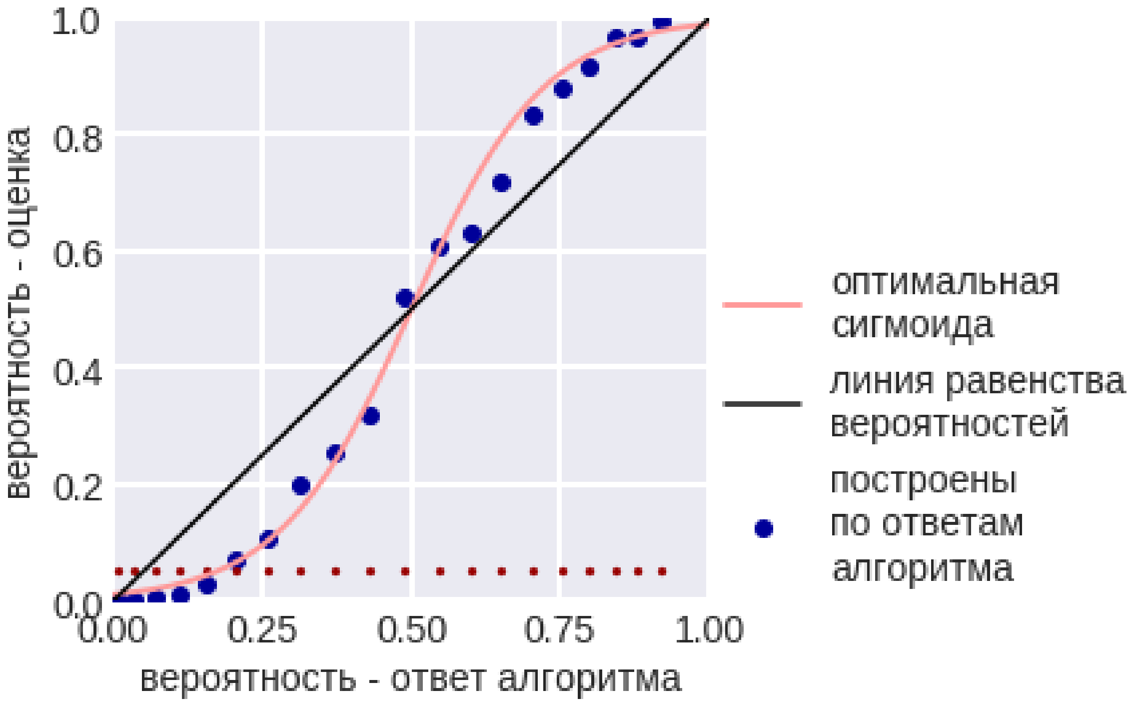
отдельная тема

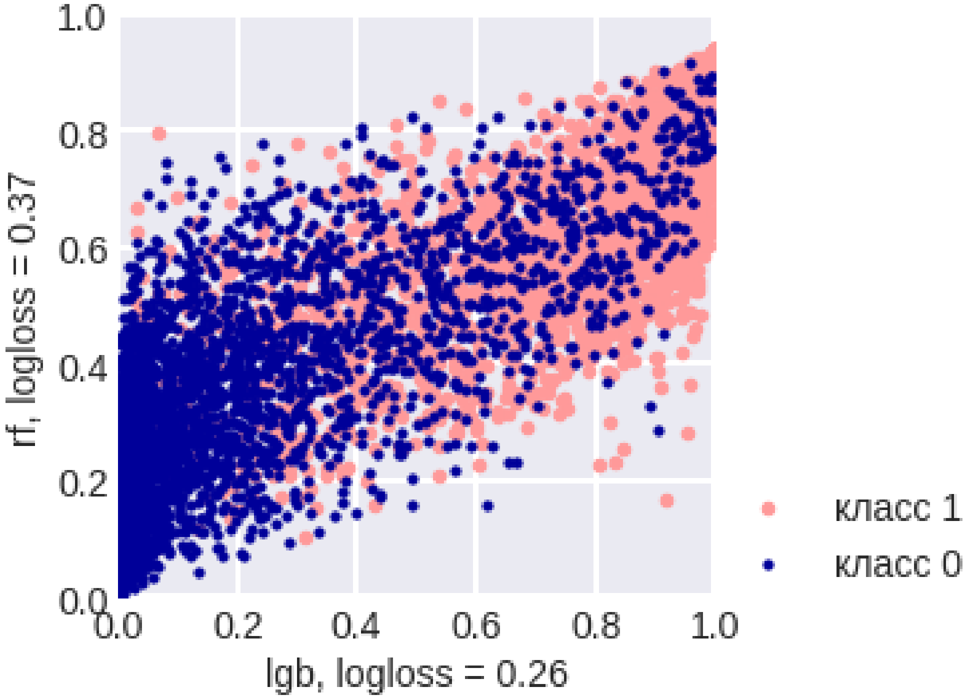
Настройка на Logloss – методы калибровки

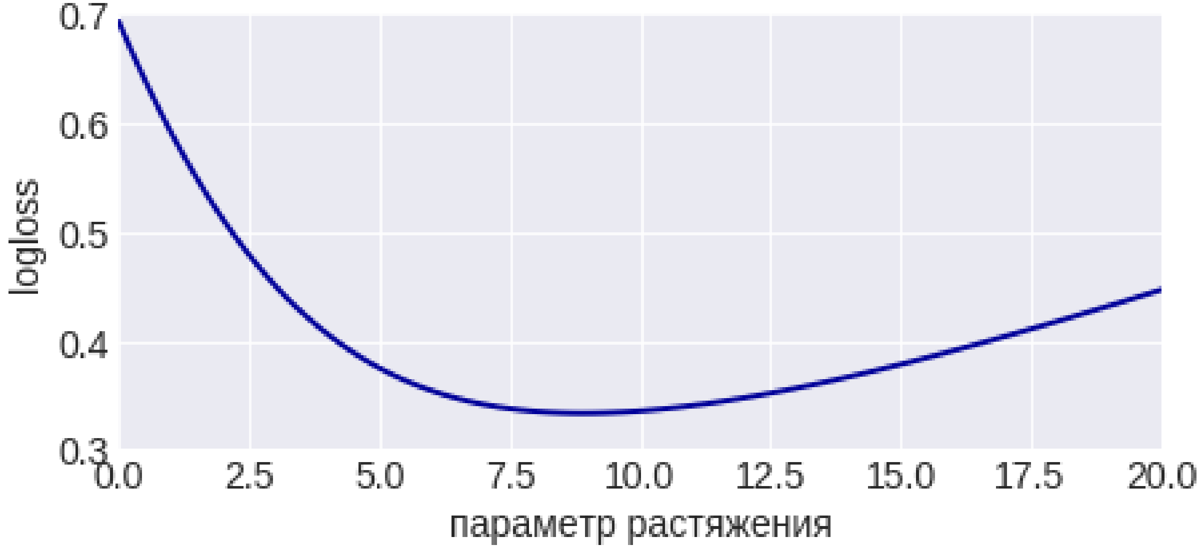
калибровка Платта (Platt calibration)

– для SVM

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







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

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

оптимальный ответ тоже $a=p$

Минимум матожидания (подставляем $a=p$, как делали в logloss):

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

критерий расщепления Джини минимизирует эту функцию ошибки!

это называется «Brier score»

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

Скоринговые ошибки

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

$L(y,a)$:

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

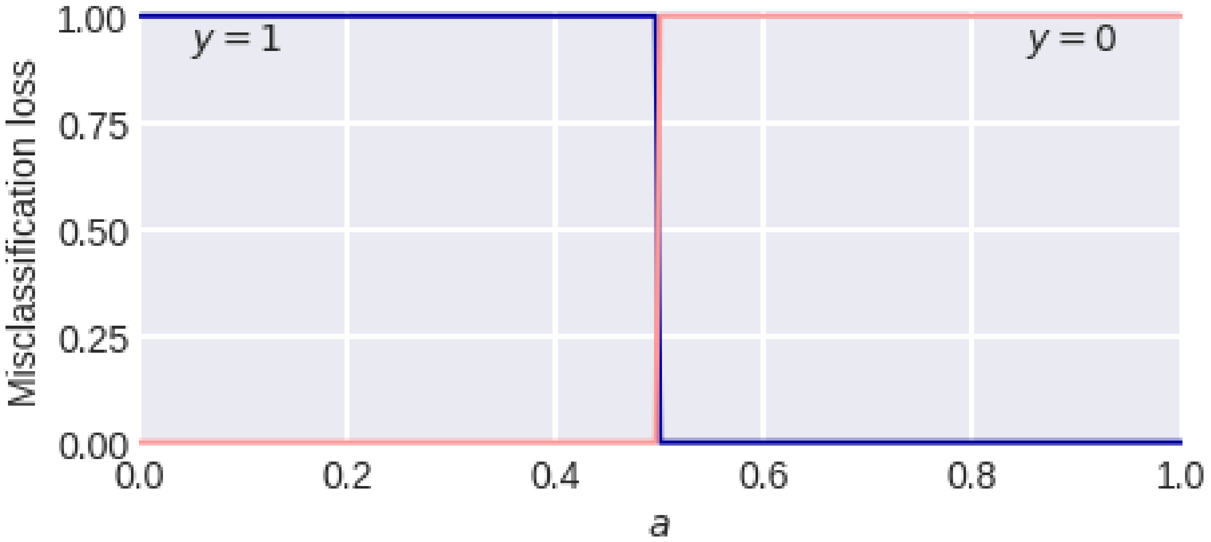
- **Log Loss**
- **MSE**
- **Exploss**
- **Misclassification Loss**

но не все…

~~**MAE**~~

Misclassification Loss

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



немного искусственная функция

почему?

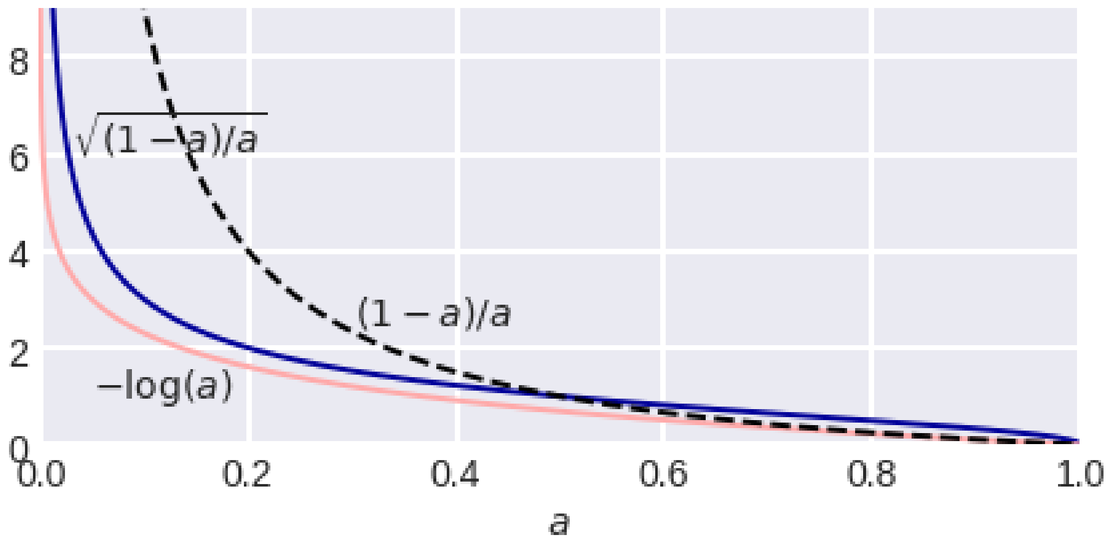
$$\mathbf{E}_y\,\mathrm{ME} = p\,I[a \le 0.5] + (1-p)\,I[a > 0.5],\ y \sim \mathrm{Bernoully}(p),$$

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

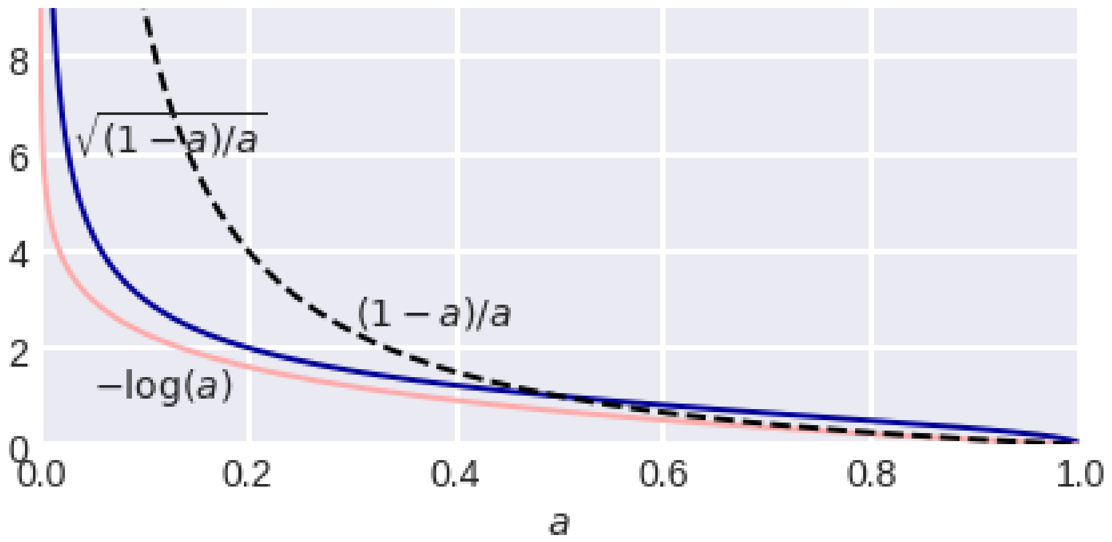


$$\text{exploss} = y\sqrt{\frac{1-a}{a}} + (1-y)\sqrt{\frac{a}{1-a}}$$

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

$$a = p$$

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?

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:

$$\exp(-ya)$$

(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 (–∞, +∞)

Exploss: почему логичная функция

Подставляем…

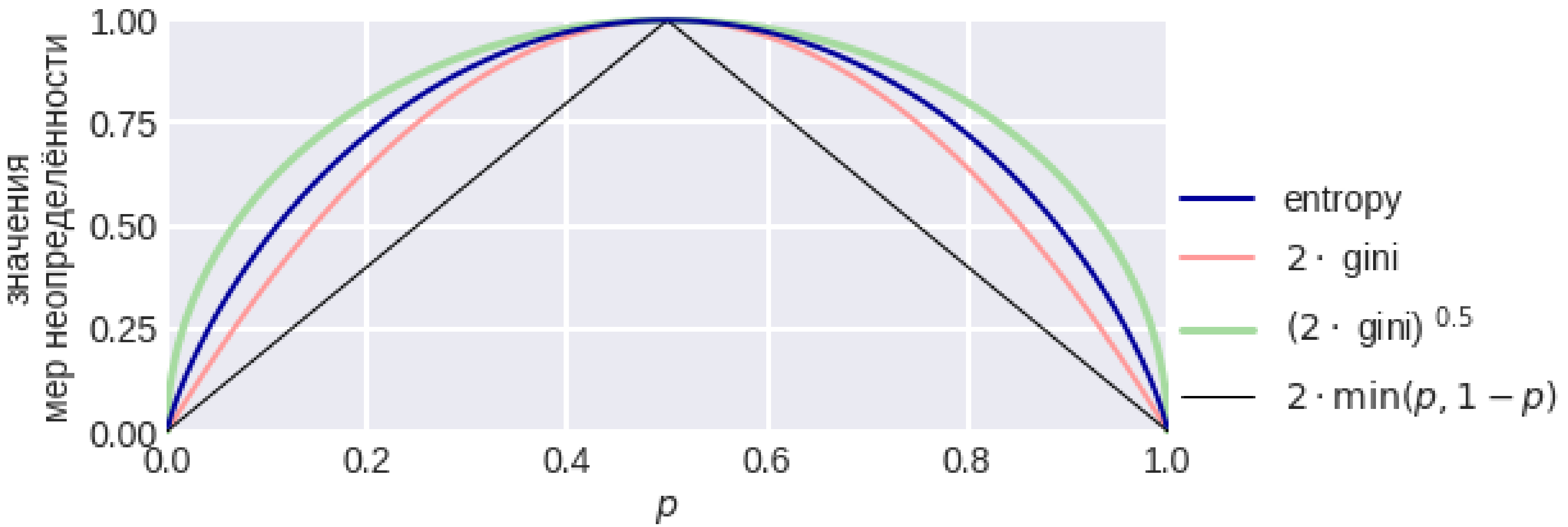
$$\exp(-ya) = \exp\left(-y\ln\sqrt{\frac{p}{1-p}}\right) =$$

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

– выражение exploss
(вместо ответов алгоритма там стоит вероятность)

Таким образом, это «естественная поправка» экспоненты,
если мы хотим ответы нашего алгоритма интерпретировать как вероятности
– перевод ответов в вероятностную шкалу (probability scale)

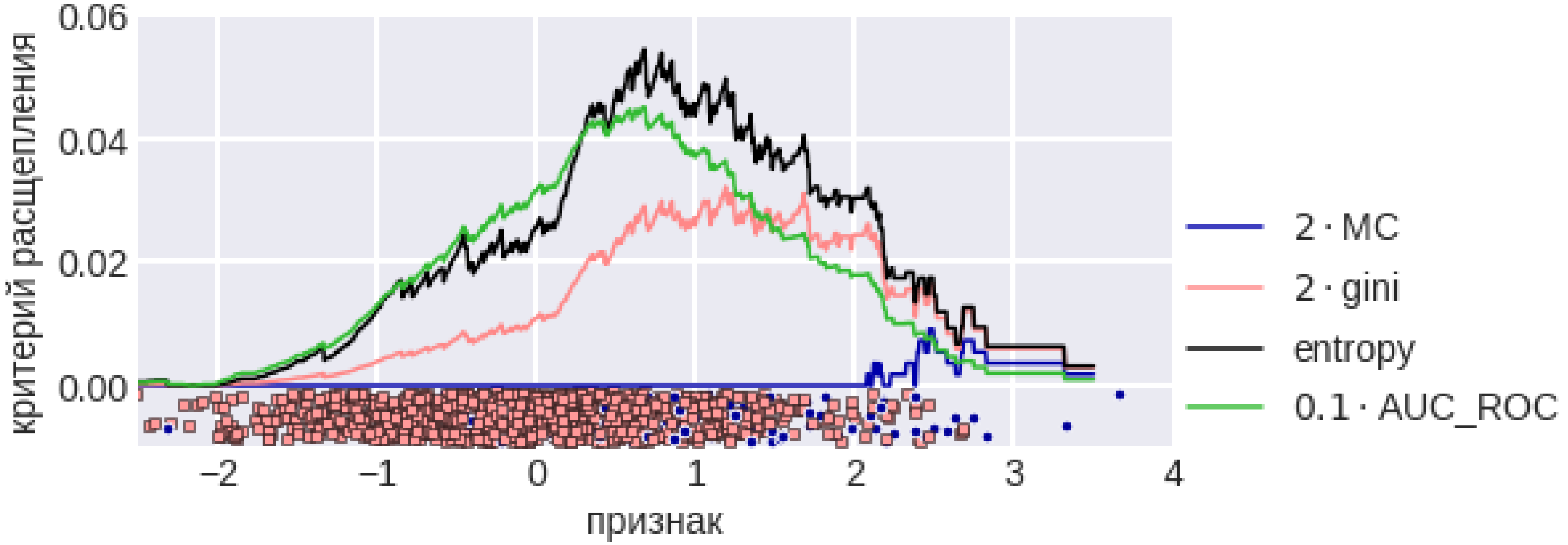
Критерии расщепления

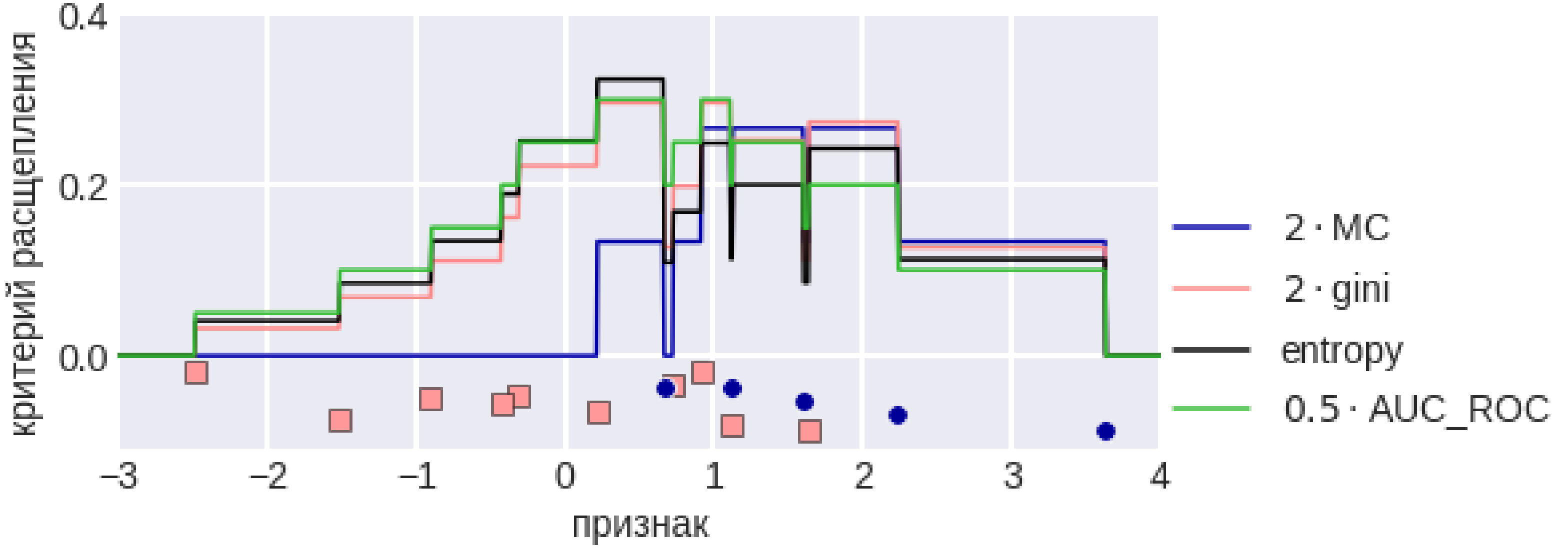


Любая скоринговая функция порождает информационную меру, которая может быть использована в критерии расщепления

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	$\min(p,1-p)$

Вспоминаем критерии расщепления





MC = Missclassification criteria

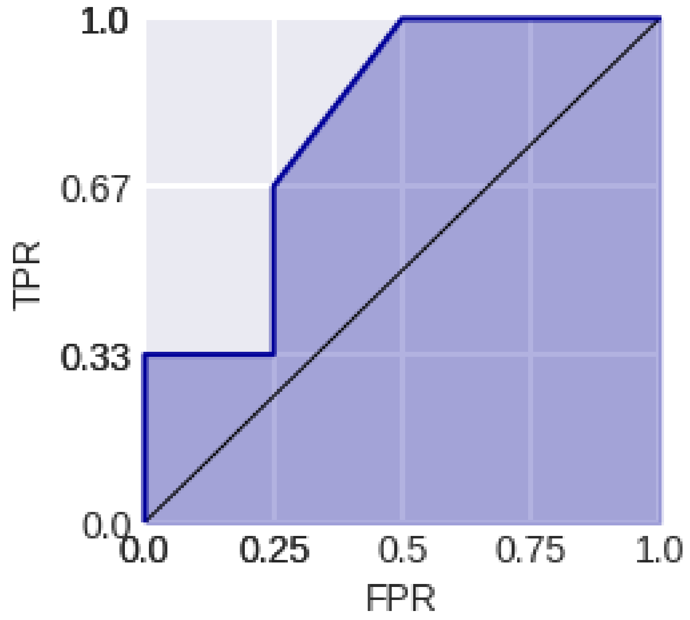
ROC и AUC ROC

ROC = receiver operating characteristic

Функционал зависит не от конкретных значений, а от их порядка

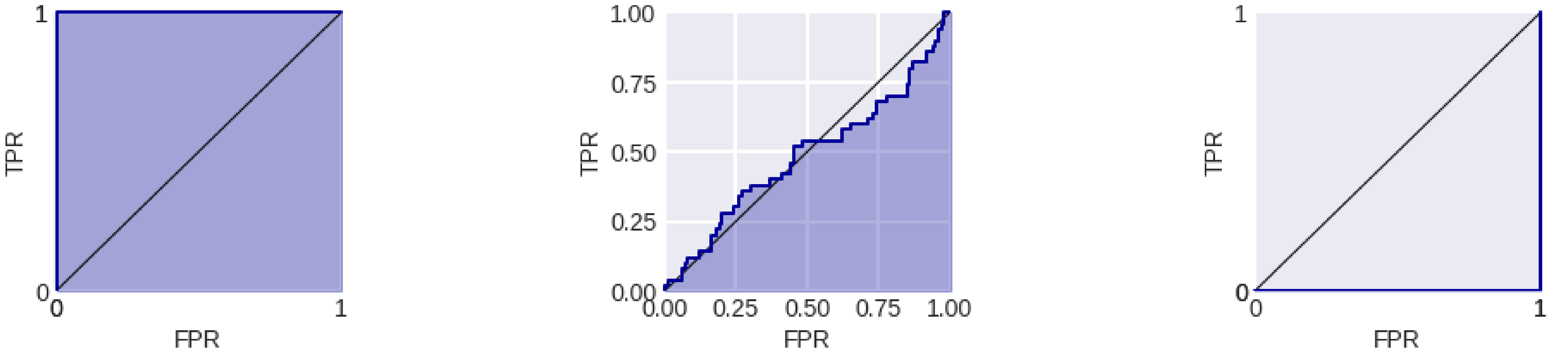
	оценка	класс
0	0.5	0
1	0.1	0
2	0.2	0
3	0.6	1
4	0.2	1
5	0.3	1
6	0.0	0

	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) алгоритм

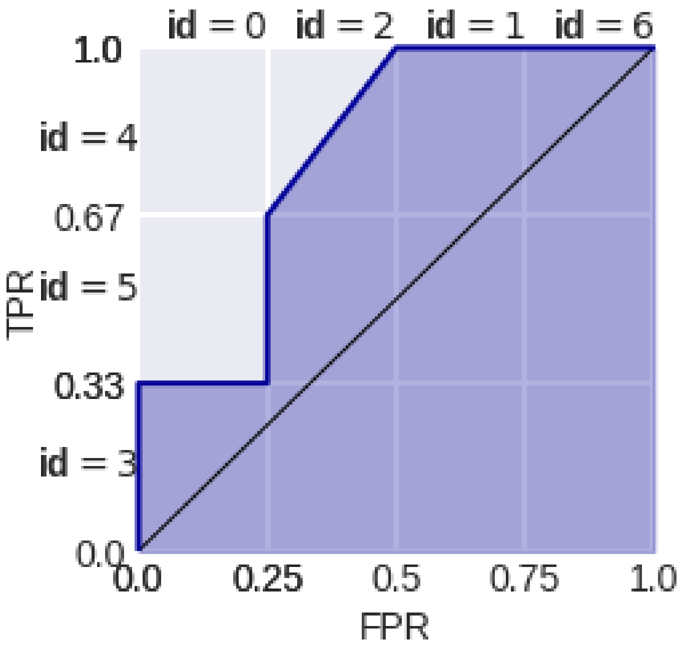
```
from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_test, a)
plt.plot(fpr, tpr, lw=3, c='#000099')
```

Смысл AUC

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

Это сложно объяснить заказчику!

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



	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

Чем хороша эта запись?
Что неправильно (требует пояснения) в формуле?

Смысл AUC

Чем хороша запись?

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

Можно обобщить, например, на регрессию.

Что неправильно (требует пояснения) в формуле?

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

Обобщения AUC

Иногда используют «естественные обобщения»:

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

Если есть веса объектов…

как обобщить AUC?

Обобщения AUC

Напишем, что есть FPR, TPR (эти формулы после бинаризации)

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

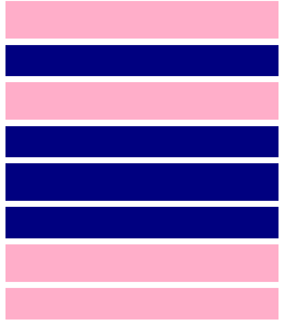
$$\mathrm{FPR} = \frac{\sum_{i=1}^{m} I[a_i = 1] I[y_i = 0]}{\sum_{i=1}^{m} I[y_i = 0]}$$

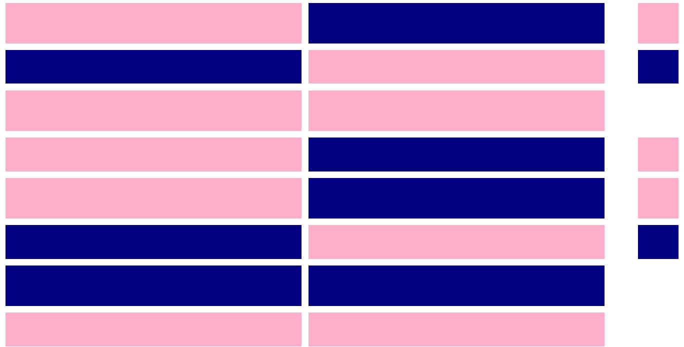
Теперь всё ясно…

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

Настройка RF/GBM на AUC ROC

Случай из жизни (Интернет-математика)



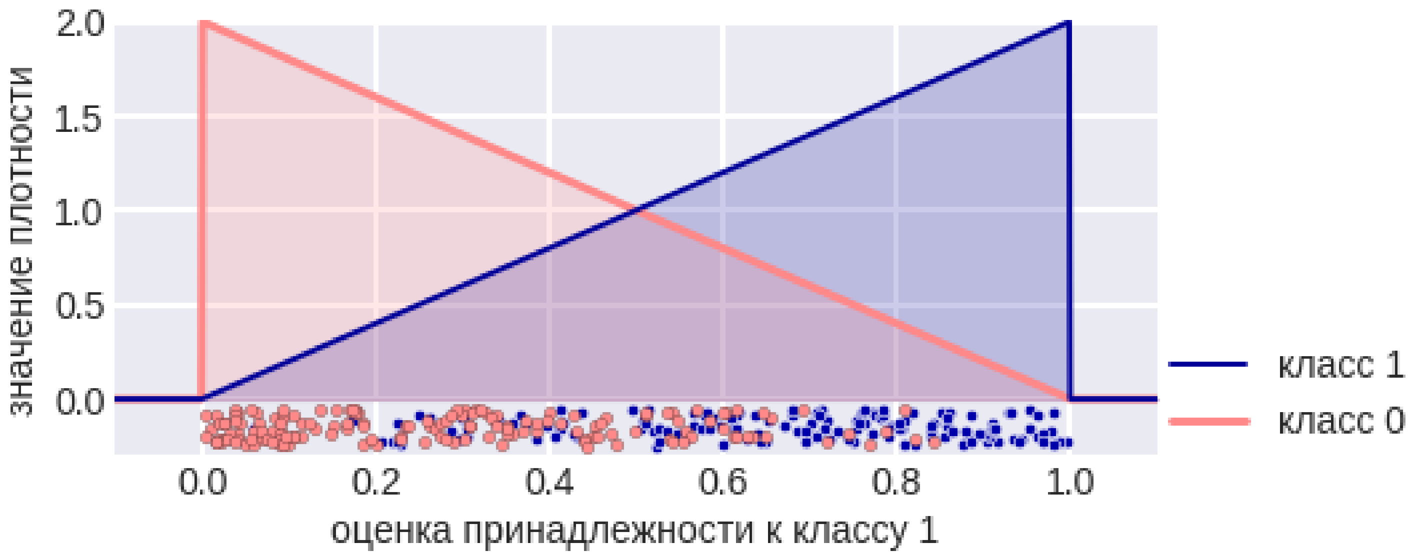


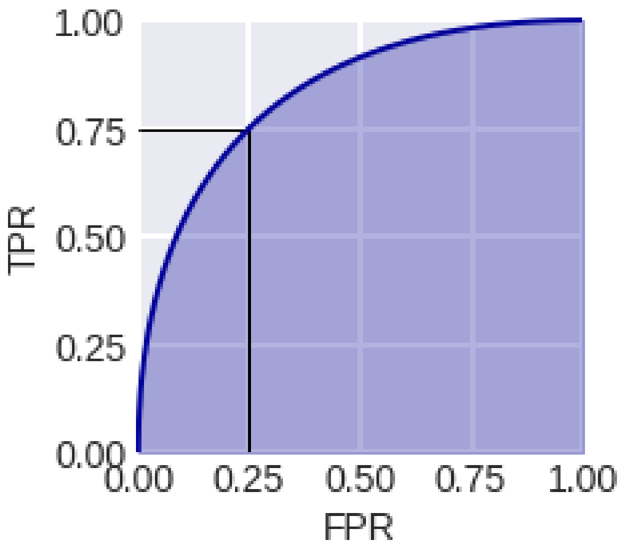
классификация $\rightarrow$ классификация пар

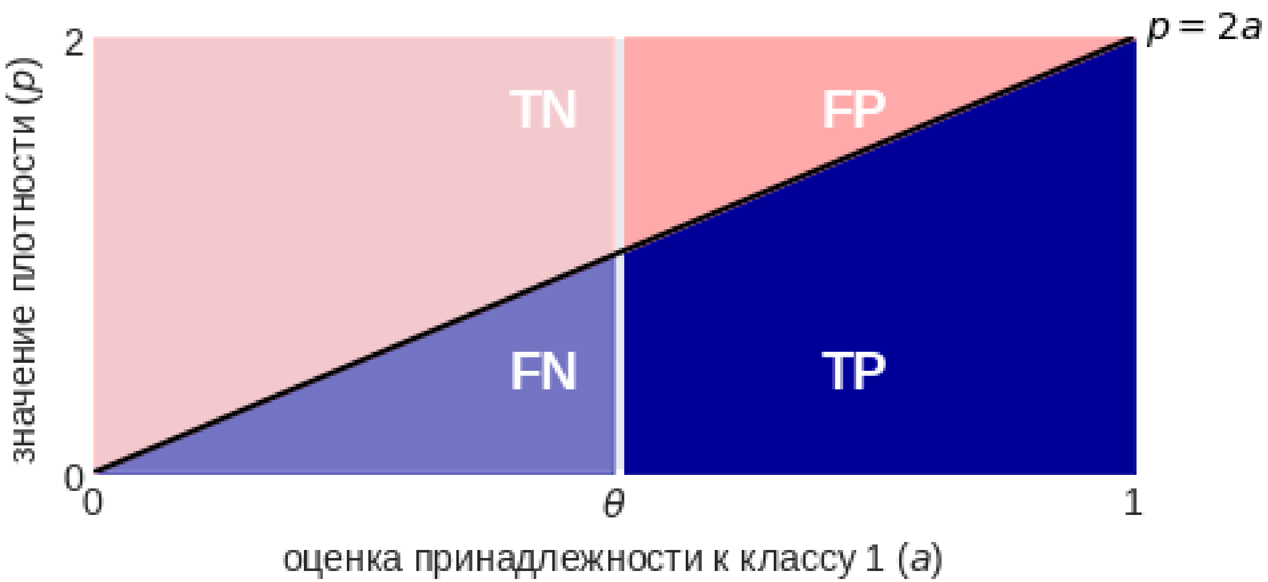
Можно дублировать,

Можно брать разности/отношения.

AUC ROC







$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{1-\theta^2}{1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = \frac{(1-\theta)^2}{1}$$

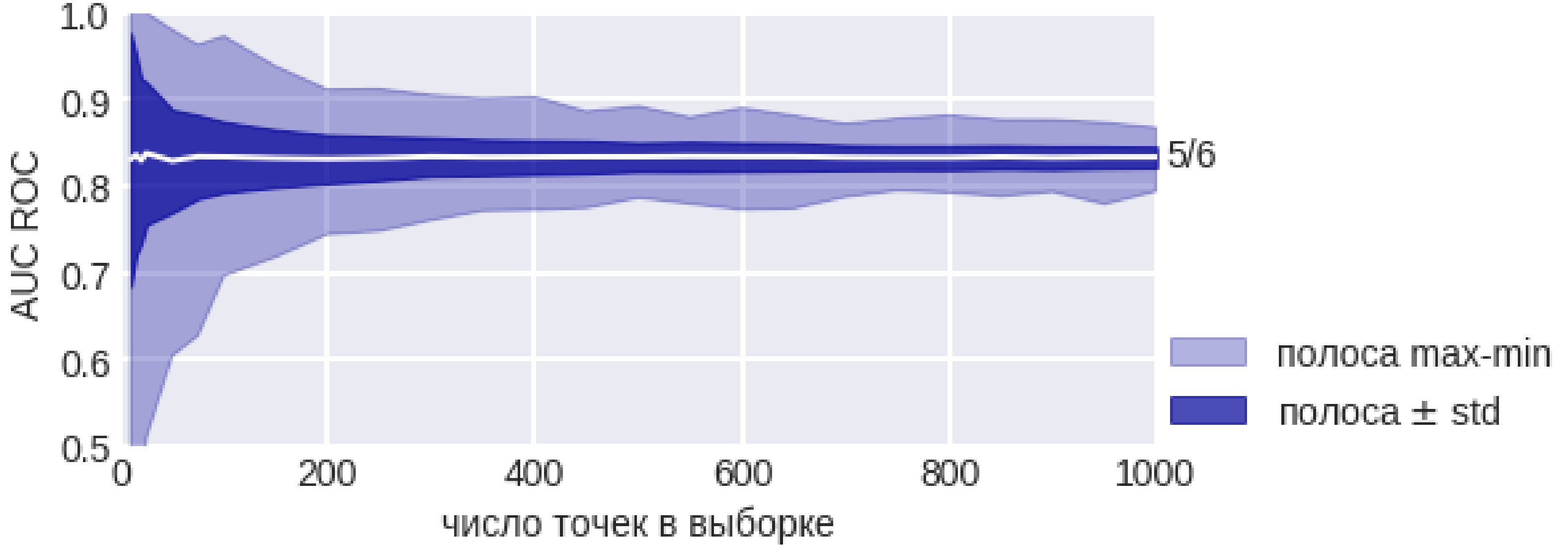
$$\mathrm{TPR} = 1-(1-\sqrt{\mathrm{FPR}})^2 = $$
$$= 2\sqrt{\mathrm{FPR}} - \mathrm{FPR}$$

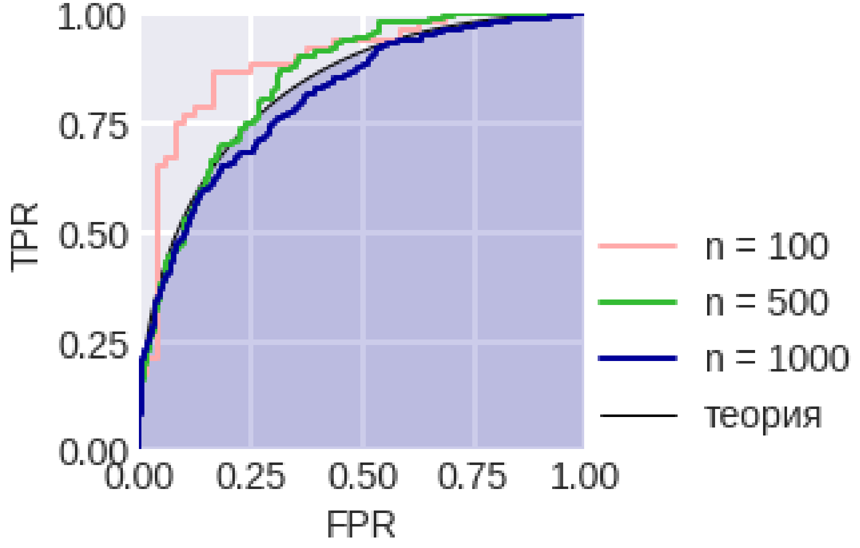
ROC – не всегда ступеньки!

ДЗ Как соотносятся AUC ROC и максимальная достижимая точность?

AUC ROC: эксперименты

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





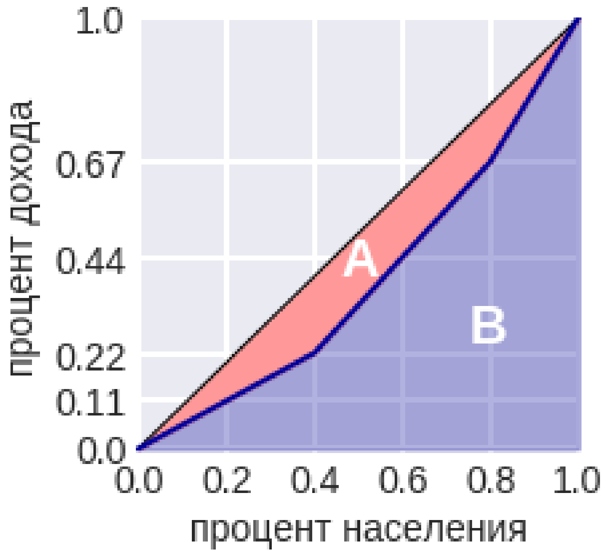
Для оценки AUC ROC маленькие выборки не подходят!

Плотности линейные, а ROC не линейная.

GINI

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

Кривая Лоренца

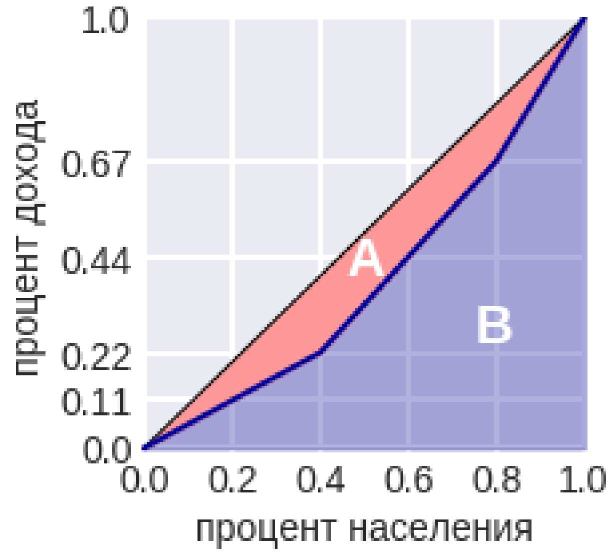


Пример для доходов: 1, 1, 2, 2, 3

40% населения имеют 2/9 дохода.

GINI

Вычисление



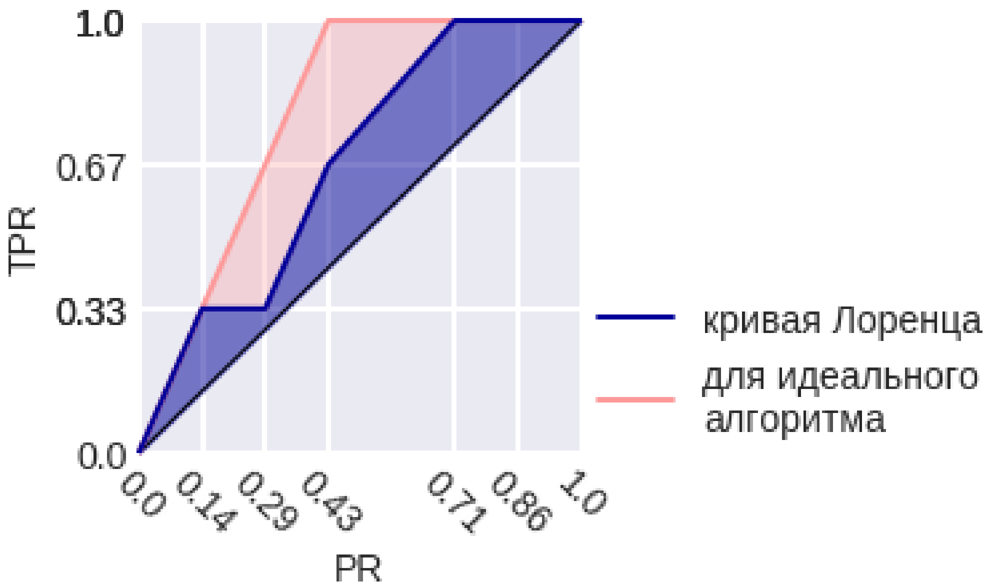
$$\text{gini} = \frac{A}{A+B} = 2A$$

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

не путать с Gini impurity

GINI в машинном обучении

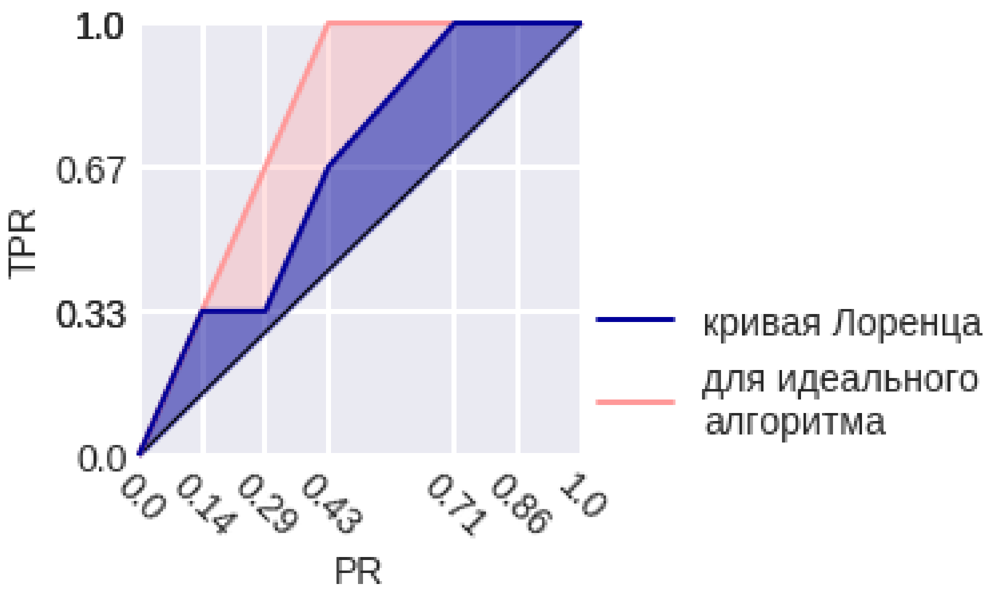
Кривая Лоренца (или CAP – Cumulative Accuracy Profile Curve)



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

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

GINI в машинном обучении



$$\mathrm{AUCROC} = \int_0^1 \mathrm{TPR}\,\partial\,\mathrm{FPR} = \int_0^1 \frac{\mathrm{TP}}{m_1}\,\partial\,\frac{\mathrm{FP}}{m_0} = \frac{1}{m_1 m_0}\int_0^1 \mathrm{TP}\,\partial\,\mathrm{FP}$$

$$\mathrm{gini} = \frac{\int_0^1 \mathrm{TPR}\,\partial\,\mathrm{PR} - 0.5}{0.5 m_0 / (m_0 + m_1)} = \frac{\int_0^1 \frac{\mathrm{TP}}{m_1}\,\partial\,\frac{\mathrm{FP}+\mathrm{TP}}{m_0 + m_1} - 0.5}{0.5 m_0 / (m_0 + m_1)}$$

GINI в машинном обучении

$$\mathrm{gini} = \frac{2}{m_1 m_0}\int\limits_0^1 \mathrm{TP}\,\partial(\mathrm{FP}+\mathrm{TP}) - \frac{m_0+m_1}{m_0} =$$

$$= 2\,\mathrm{AUCROC} + \frac{2}{m_1 m_0}\int\limits_0^1 \mathrm{TP}\,\partial\,\mathrm{TP} - \frac{m_1}{m_0} - 1$$

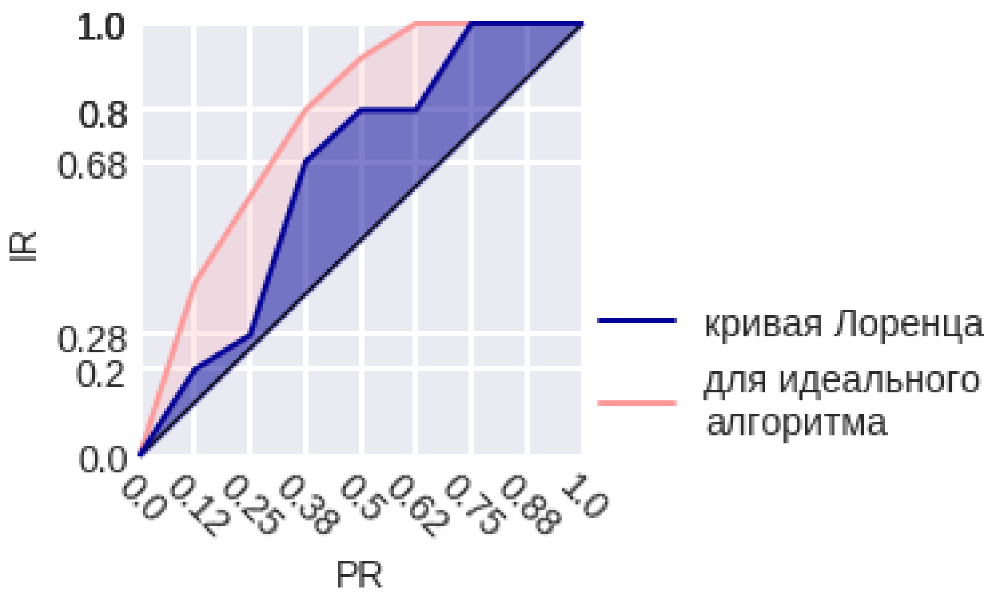
$$\mathrm{gini} = 2\,\mathrm{AUCROC} - 1$$

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

0.9 AUC = 0.8 gini

GINI в задаче регрессии

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



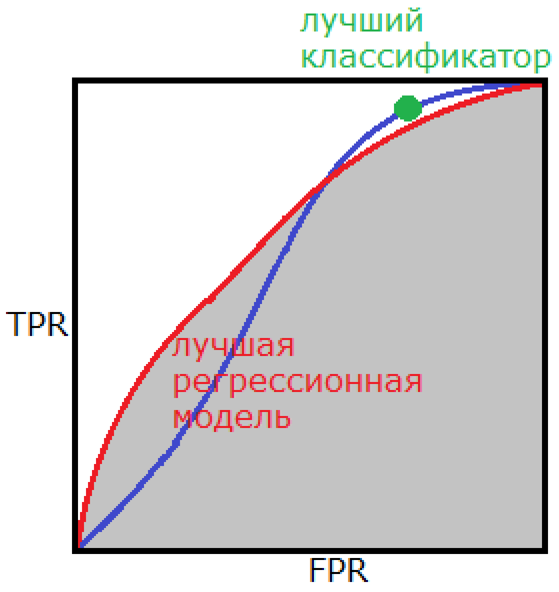
Идеальный алгоритм: 10, 5, 5, 3, 2, 0, 0, 0

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

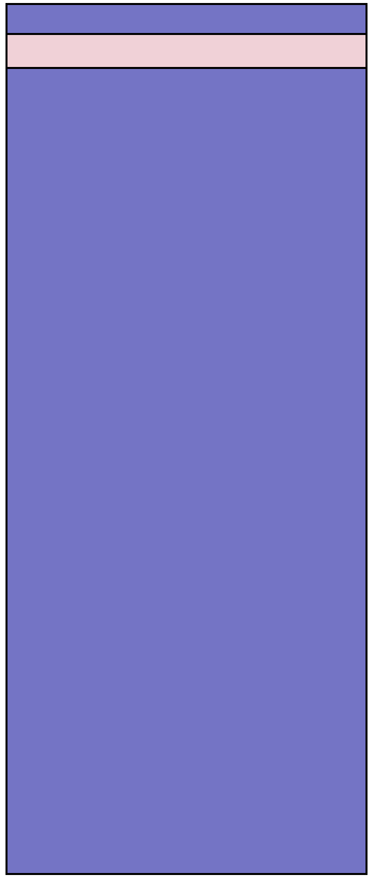
AUC ROC

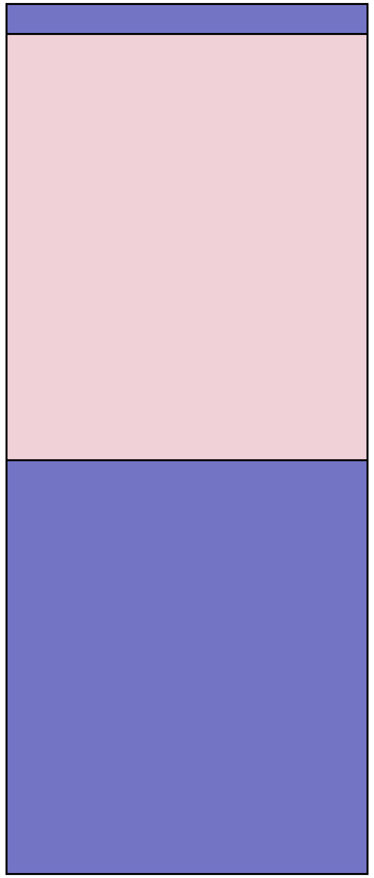
+ в задачах, где важен порядок
+ учитывает разную мощность классов (не зависит от пропорций)
+ не важны значения, важен порядок
+ можно использовать для оценки признаков

– «завышает» качество
– оценивает не конкретный классификатор, а регрессию
– сложно объяснить заказчику
– не путать классификацию и регрессию



AUC ROC и дисбаланс классов



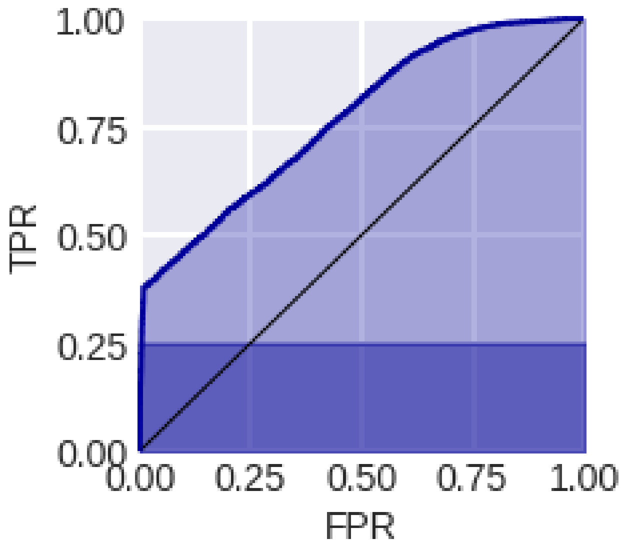


это пример «задачи поиска»,
а не задачи с дисбалансом

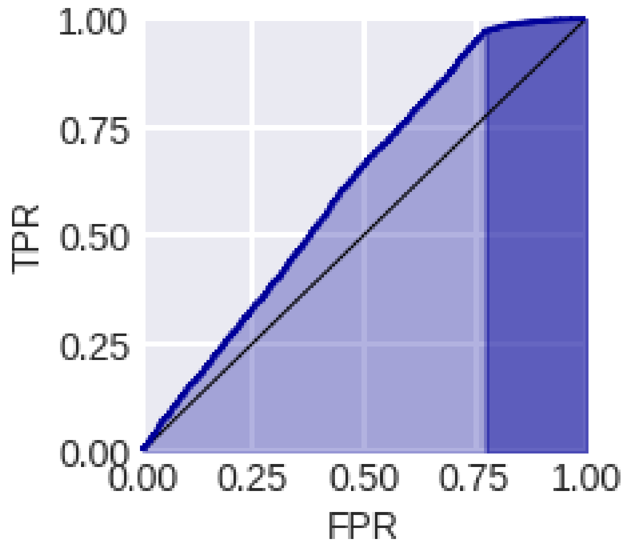
Маленький AUC не всегда плохо

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)

Максимизация AUC ROC

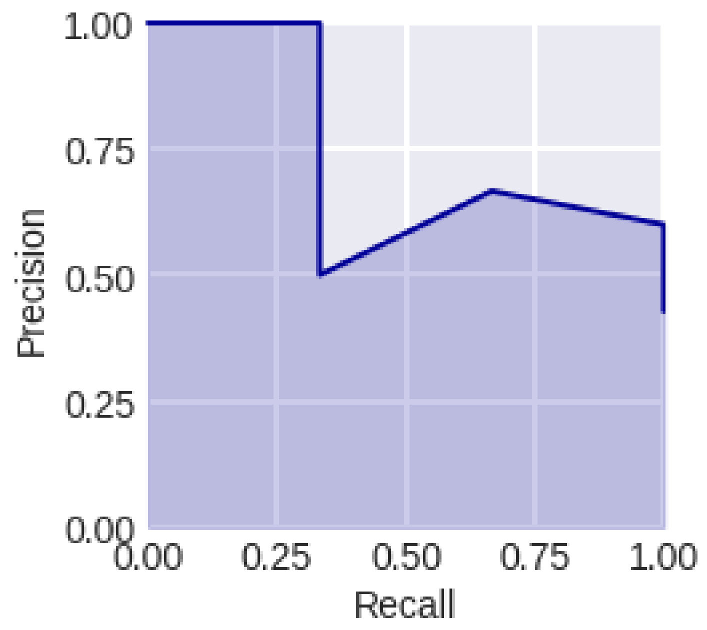
- замена индикаторных функций на дифференцируемые
- использование смысла функционала (переход к парам)
- ансамблирование с ранговой деформацией

Ещё примеры кривых… «полнота-точность»

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

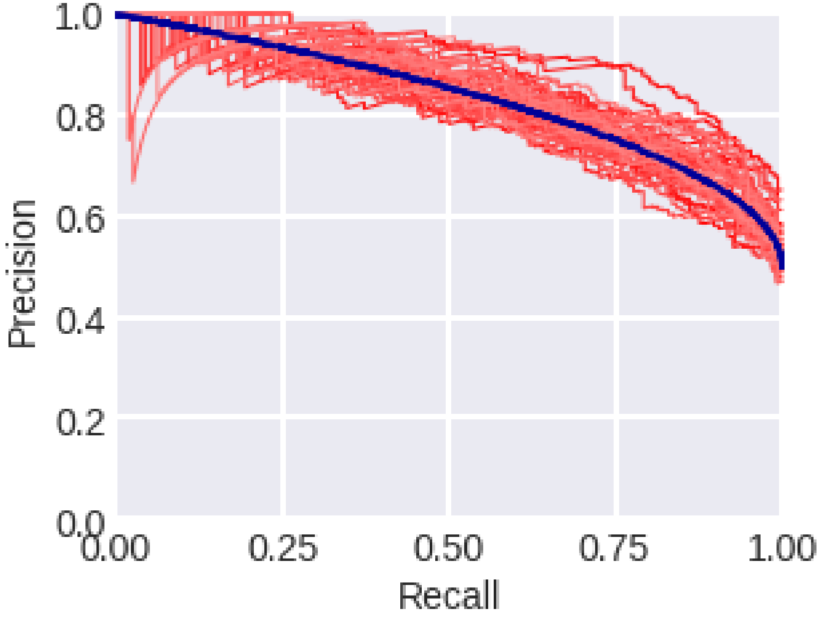
	оценка	класс
0	0.5	0
1	0.1	0
2	0.2	0
3	0.6	1
4	0.2	1
5	0.3	1
6	0.0	0

	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

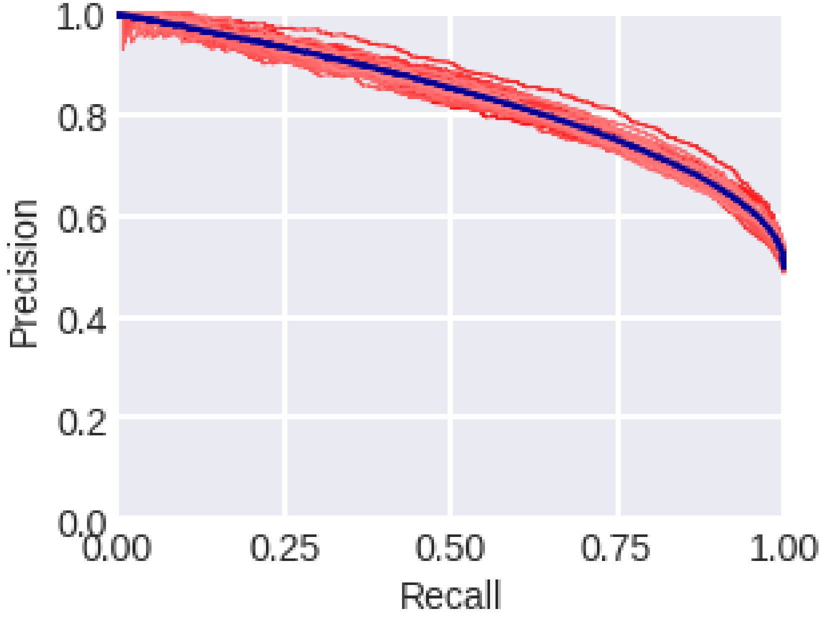



```
from sklearn.metrics import precision_recall_curve
precision, recall, thresholds = precision_recall_curve(y_test, a)
plt.plot(recall, precision)
# вычисление площади методом трапеций
from sklearn.metrics import auc
auc(recall, precision)
# или готовую функцию использовать
from sklearn.metrics import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

Кривая «полнота-точность»
в задаче с линейными плотностями



m = 300
AUC_PR = 0.839 ± 0.024

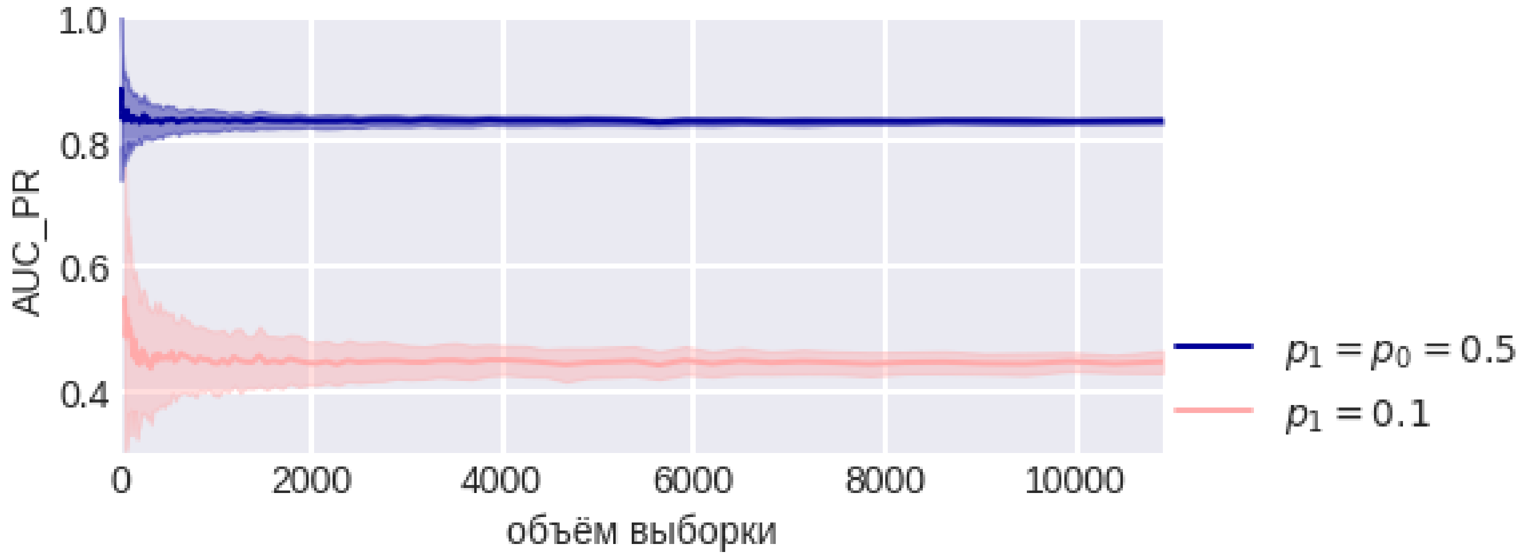


m = 3000
AUC_PR = 0.833 ± 0.012

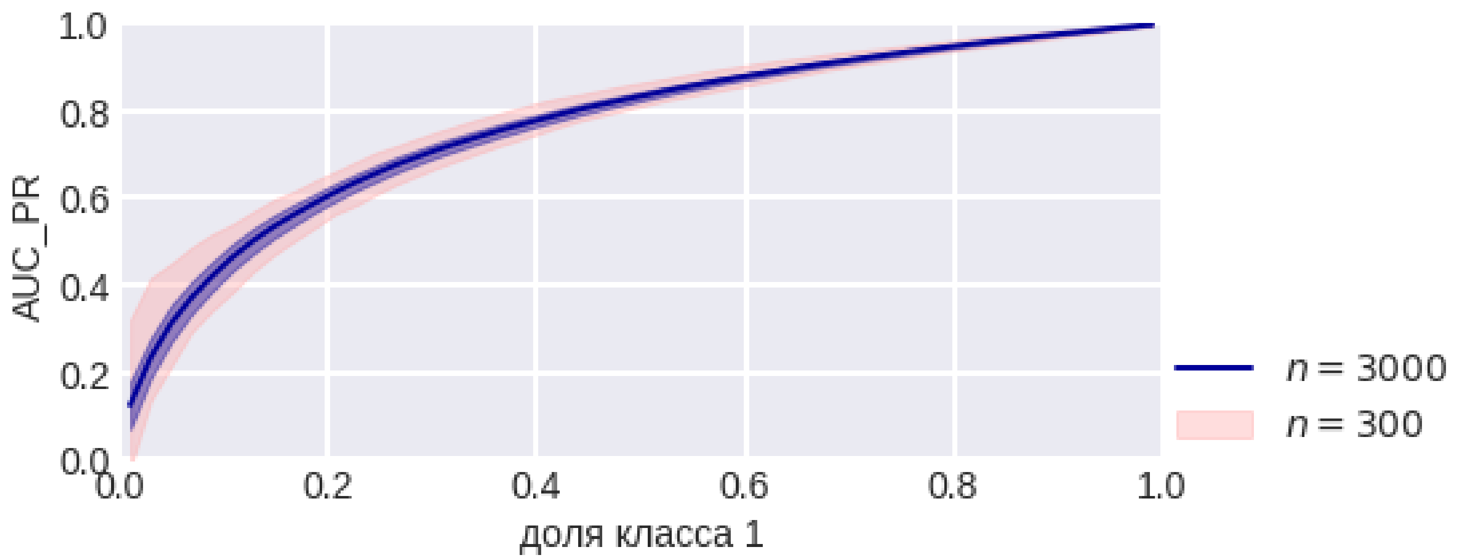
Генерируем выборки с соотв. распределениями и строим кривые

$$\mathrm{AUC}_{\mathrm{PR}} = \int_0^1 P\partial R = \int_0^1 \frac{1+\sqrt{1-R}}{2}\partial R = \frac{5}{6} = 0.83(3)$$

Кривая «полнота-точность»



Кривая «полнота-точность»

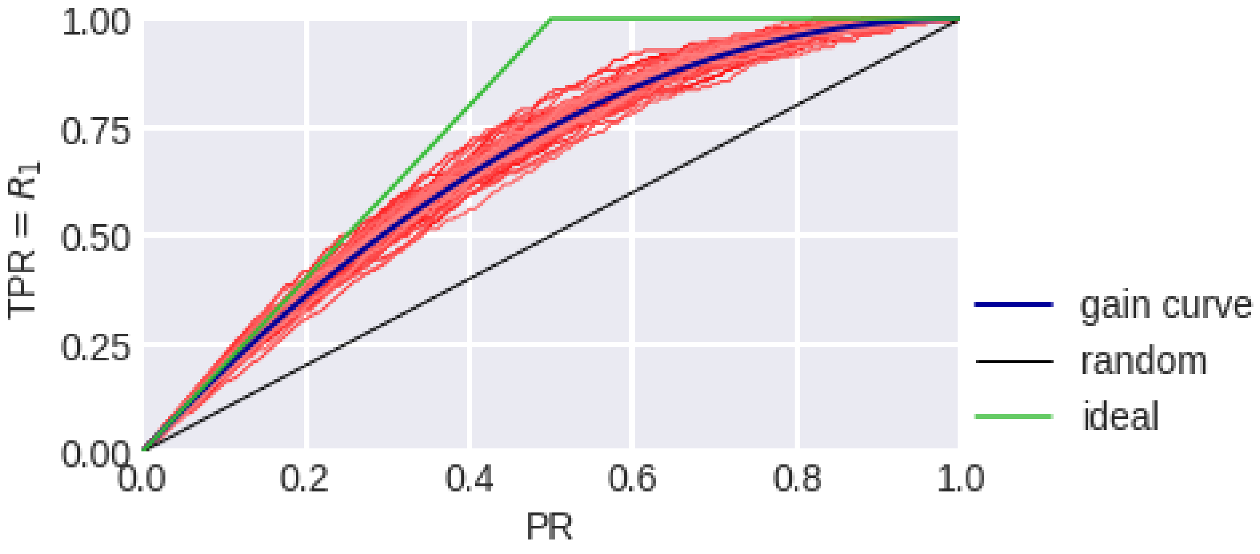


Есть зависимость от доли класса 1

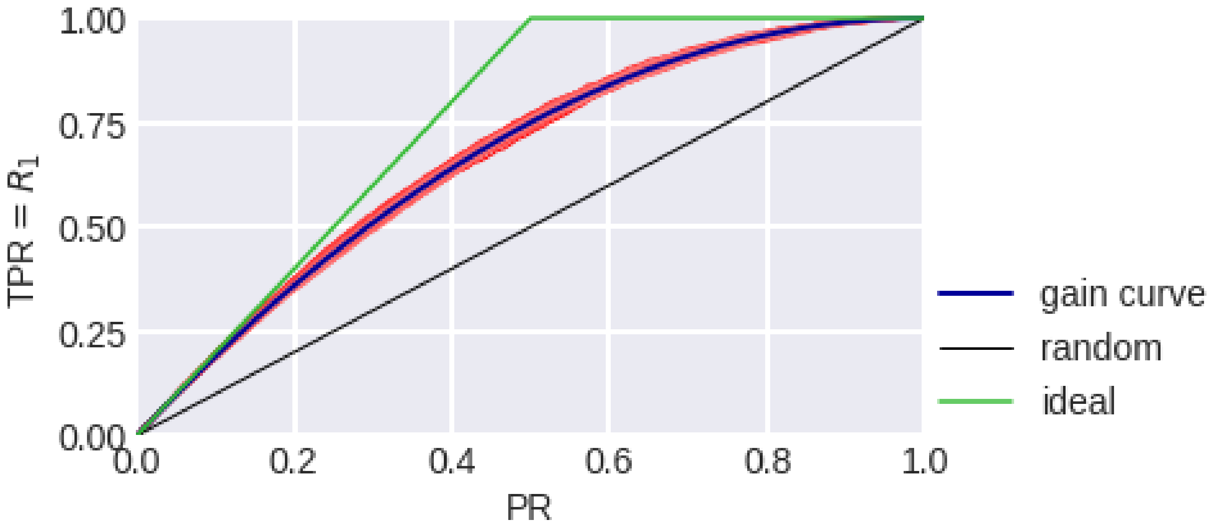
Дисперсия оценки тоже зависит от доли класса 1

Сильный дисбаланс класса 0 $\Rightarrow$ формально низкое качество и большая дисперсия

Ещё примеры кривых… Gain Curve (Chart)



m=300



m=3000

кривая в координатах

X: «доля, отнесённых к K_1»

$$\mathrm{PR} = \frac{|\{a(x) \geq \theta\}|}{|\{x\}|} = \frac{\mathrm{TP}+\mathrm{FP}}{m}$$

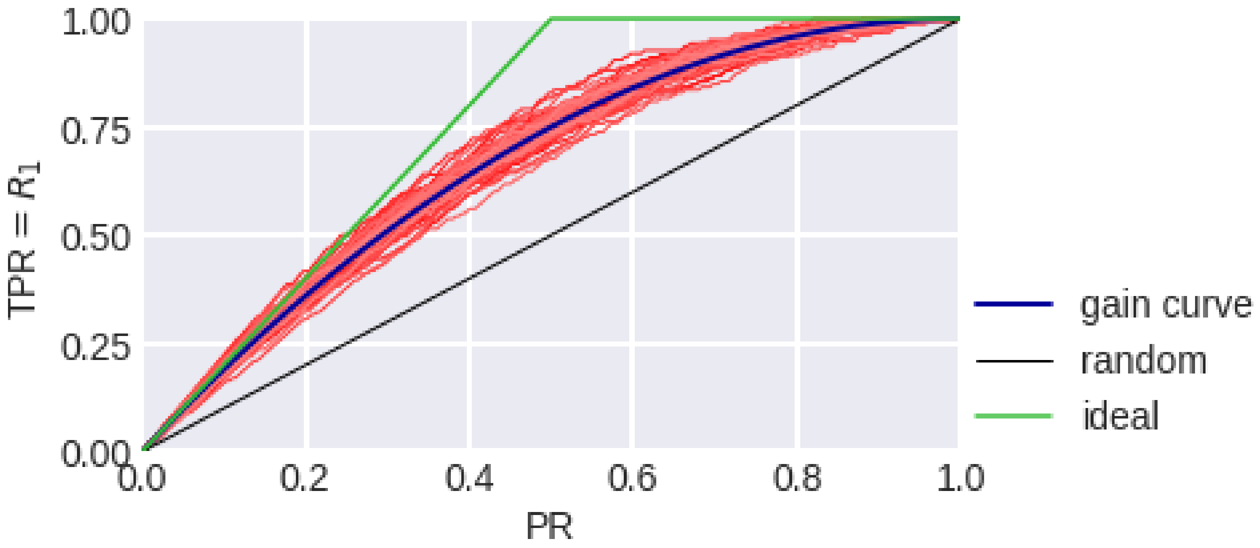
Y: «какой процент класса 1 отнесли к позитивному»

$$R_1 = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

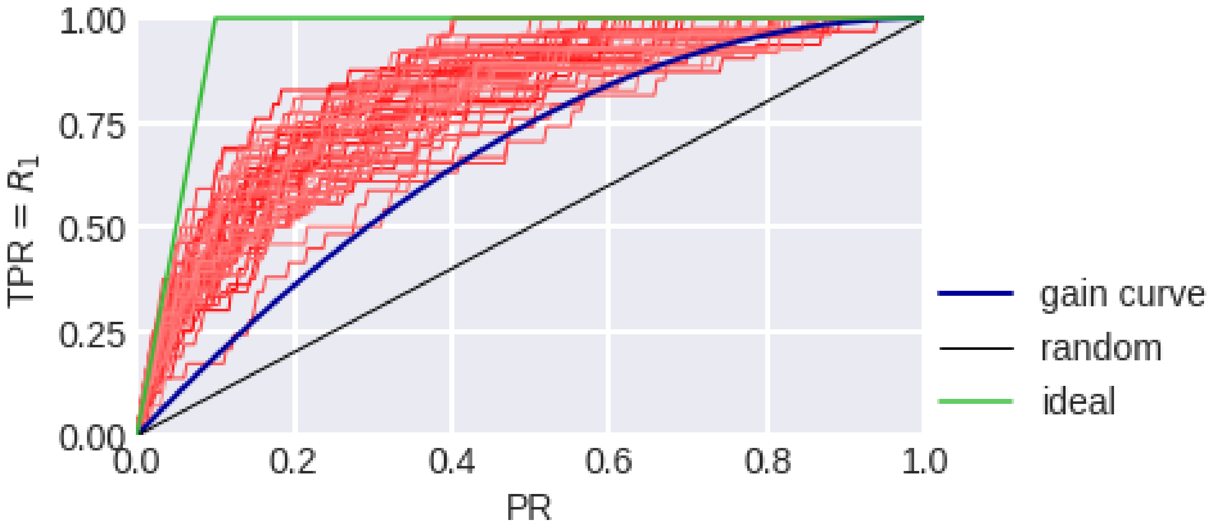
Диагональ – случайный алгоритм

это прямо кривая Лоренца

Ещё примеры кривых… Gain Curve (Chart)



m=300, классы равновероятны



m=300, P_1 = 0.1

кривая в координатах

X: «доля, отнесённых к K_1»

$$\mathrm{PR} = \frac{|\{a(x) \geq \theta\}|}{|\{x\}|} = \frac{\mathrm{TP} + \mathrm{FP}}{m}$$

Y: «какой процент класса 1 отнесли к позитивному»

$$R_1 = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Диагональ – случайный алгоритм

это прямо кривая Лоренца

Площадь под Gain-кривой

Мы доказали, что

$$\mathrm{gini} = \frac{\int\limits_0^1 \mathrm{TPR}\,\partial\,\mathrm{PR} - 0.5}{0.5 m_0 / (m_0 + m_1)} = 2\,\mathrm{AUC}_{ROC} - 1$$

отсюда получаем

$$\int\limits_0^1 \mathrm{TPR}\,\partial\,\mathrm{PR} = \frac{1}{2}\frac{m_0}{m_0 + m_1}(2\,\mathrm{AUC}_{ROC} - 1) + \frac{1}{2} =$$

$$= \frac{1}{m_0 + m_1}\left(m_0\,\mathrm{AUC}_{ROC} + \frac{m_1}{2}\right)$$

по смыслу это вероятность,
что у случайного позитивного объекта оценка выше,
чем у случайного

Gain Curve (Chart)

Это та же кривая Лоренца

А здесь

Miha Vuk, Tomaz Curk «ROC Curve, Lift Chart and Calibration Plot»

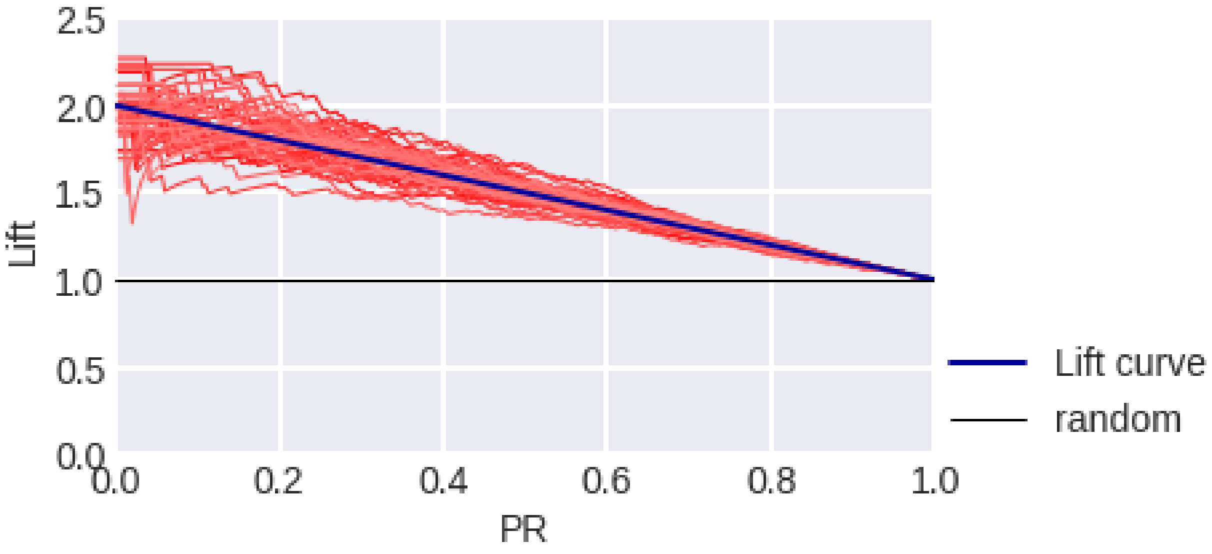
http://mrvar.fdv.uni-lj.si/pub/mz/mz3.1/vuk.pdf

это (по Y вместо TPR – TP) названо «Lift-кривой»

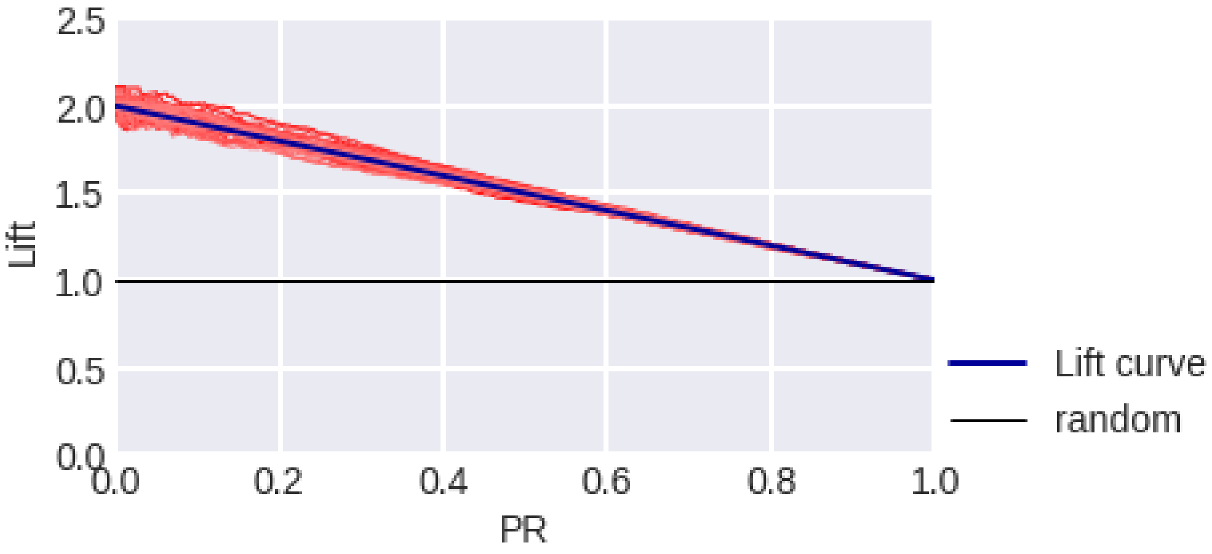
Смысл: если планируем обзвон аудитории
целевой признак – отклик на предложение

GC показывает как зависит покрытие целевой аудитории от масштаба обзвона

Ещё примеры кривых… Lift Curve (Chart)



m=300



m=3000

Насколько Gain Curve лежит выше диагонали

X: «доля, отнесённых к K_1»

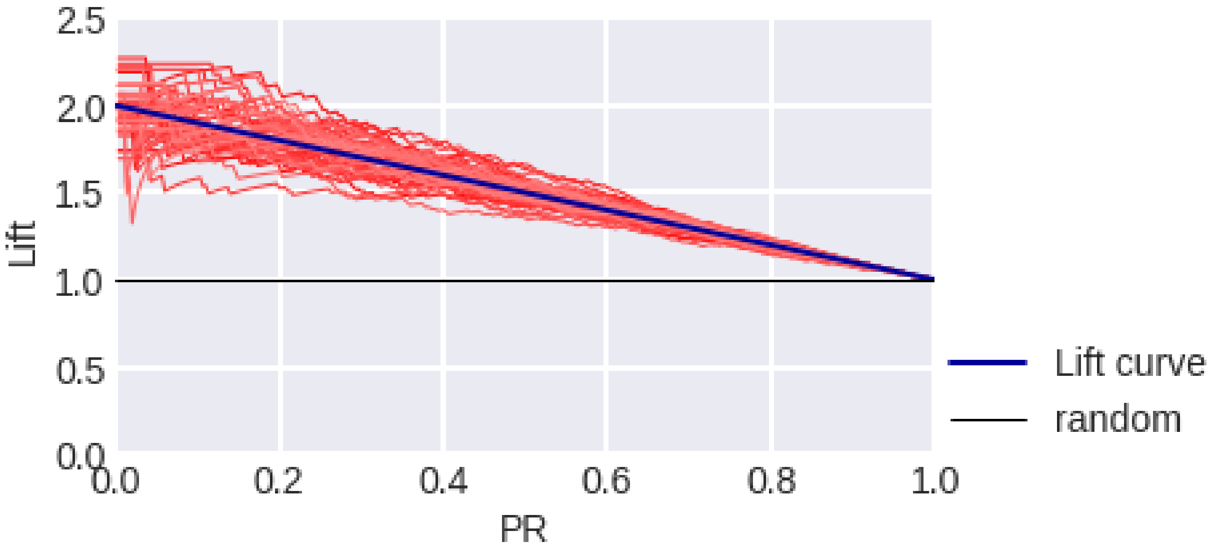
$$\mathrm{PR} = \frac{|\{a(x) \geq \theta\}|}{|\{x\}|} = \frac{\mathrm{TP} + \mathrm{FP}}{m}$$

Y: «отношение»

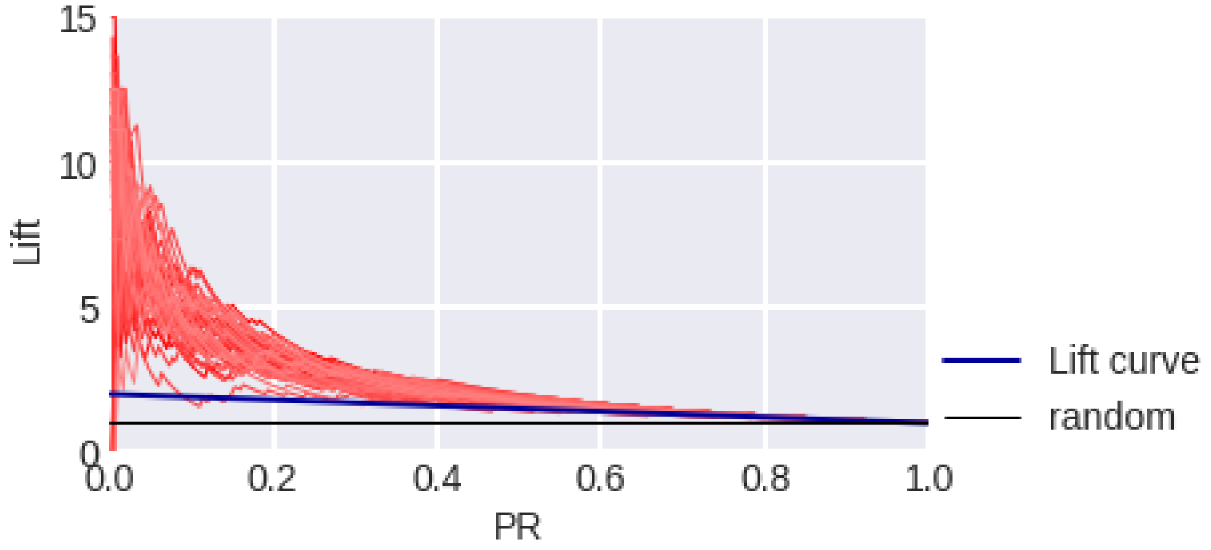
$$\mathrm{lift} = \frac{\mathrm{TPR}}{\mathrm{PR}}$$

Всегда стремится к 1

Ещё примеры кривых… Lift Curve (Chart)



m=300, классы равновероятны



m=300, P_1 = 0.1

Насколько Gain Curve лежит выше диагонали

X: «доля, отнесённых к K_1»

$$\mathrm{PR} = \frac{|\{a(x) \geq \theta\}|}{|\{x\}|} = \frac{\mathrm{TP} + \mathrm{FP}}{m}$$

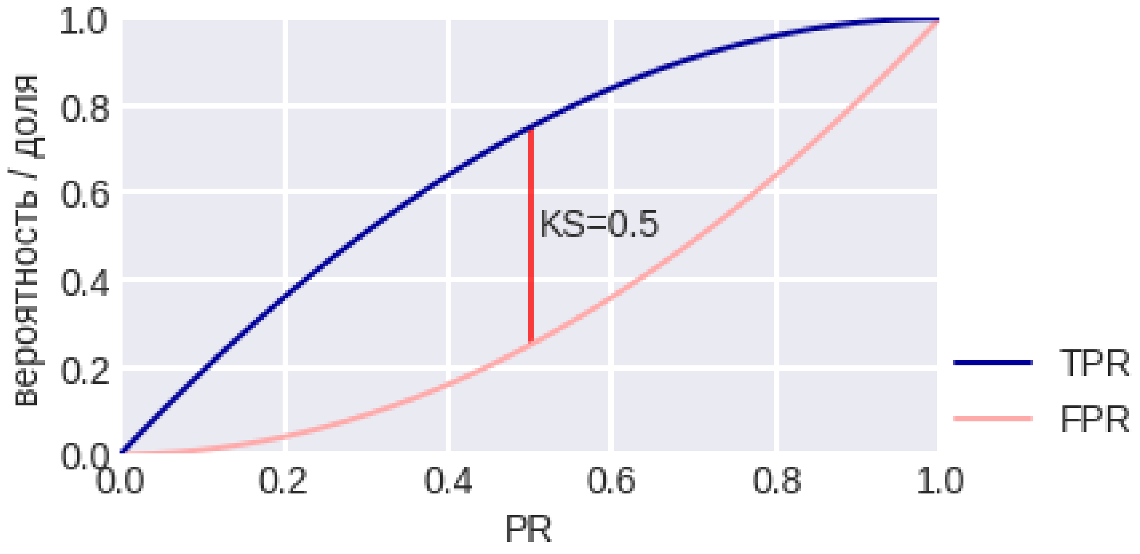
Y: «отношение»

$$\mathrm{lift} = \frac{\mathrm{TPR}}{\mathrm{PR}}$$

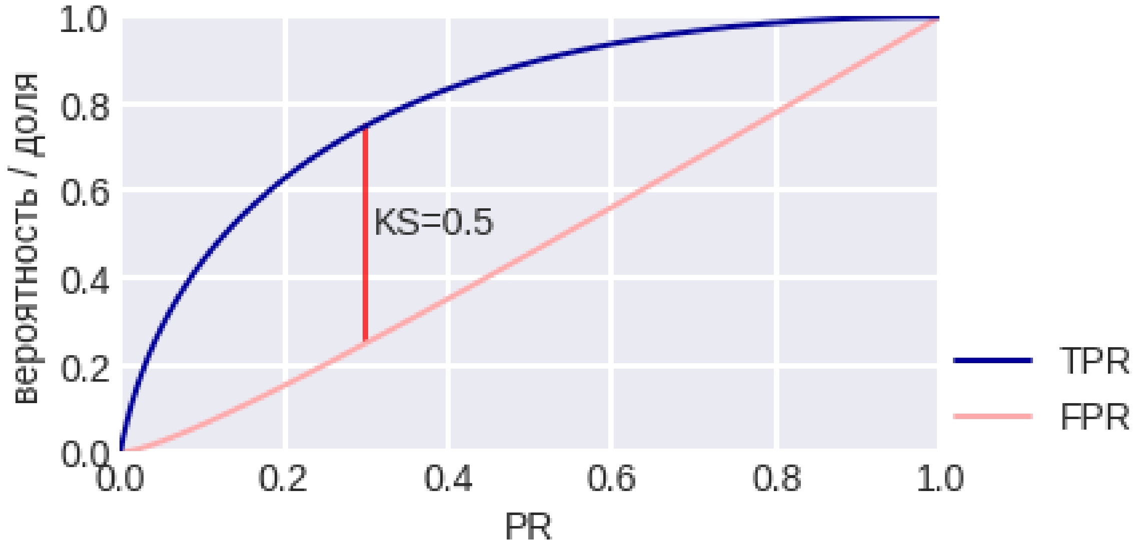
Всегда стремится к 1

Термин «Lift-Top-10%»

Ещё примеры кривых… Kolomogorov Smirnov chart



классы равновероятны



$P_1 = 0.1$

Строим кривые
«доля объектов из K_1 отнесённых к K_1»,
«доля объектов из K_0 отнесённых к K_1»

$$(\mathrm{PR}(\theta), \mathrm{TPR}(\theta))$$
$$(\mathrm{PR}(\theta), \mathrm{FPR}(\theta))$$

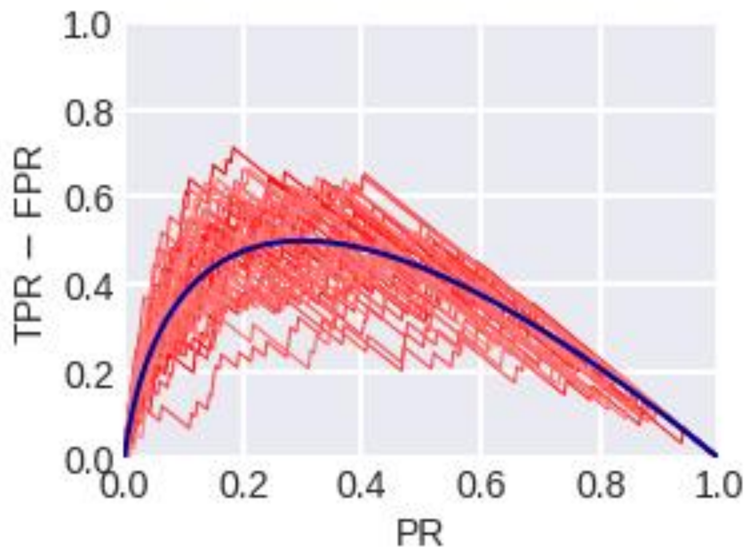
находим максимум разности между ними

Ещё примеры кривых... Kolomogorov Smirnov chart

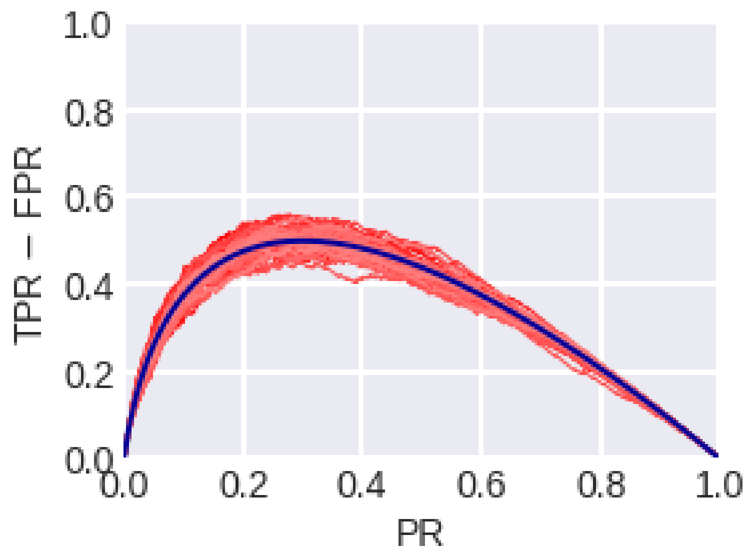
$$\mathrm{TPR}(\theta) - \mathrm{FPR}(\theta) \to \max$$

Интересно, что в модельной задаче $\mathrm{TPR}(\theta), \mathrm{FPR}(\theta)$ не зависят от баланса классов, а вот K-S chart зависит... почему?

Доказать, что при $p_1 = 0.1$ на KSC максимальная разница в точке 0.3



300 объектов



3 000 объектов

Ещё примеры кривых…

Есть реализации

https://github.com/reiinakano/scikit-plot

Profit Analysis: The Gains Table

N	%	cum_%	Prob	N_1	%_1	cum_N1	cum_%1	N_0	%_0	cum_N0	cum_%0	K-S	Lift
11238	10.0%	10.0%	0.229	2572	49.0%	2572	49.0%	8666	8.1%	8666	8.1%	40.9%	4.902
11237	10.0%	20.0%	0.081	912	17.4%	3484	66.4%	10325	9.6%	18991	17.7%	48.7%	3.320
11238	10.0%	30.0%	0.050	565	10.8%	4049	77.2%	10673	10.0%	29664	27.7%	49.5%	2.572
11237	10.0%	40.0%	0.037	413	7.9%	4462	85.0%	10824	10.1%	40488	37.8%	47.2%	2.126
11238	10.0%	50.0%	0.025	282	5.4%	4744	90.4%	10956	10.2%	51444	48.0%	42.4%	1.808
11237	10.0%	60.0%	0.018	197	3.8%	4941	94.2%	11040	10.3%	62484	58.3%	35.8%	1.569
11237	10.0%	70.0%	0.013	146	2.8%	5087	97.0%	11091	10.4%	73575	68.7%	28.3%	1.385
11238	10.0%	80.0%	0.008	94	1.8%	5181	98.7%	11144	10.4%	84719	79.1%	19.7%	1.234
11237	10.0%	90.0%	0.005	51	1.0%	5232	99.7%	11186	10.4%	95905	89.5%	10.2%	1.108
11238	10.0%	100.0%	0.001	15	0.3%	5247	100.0%	11223	10.5%	107128	100.0%	0.0%	1.000

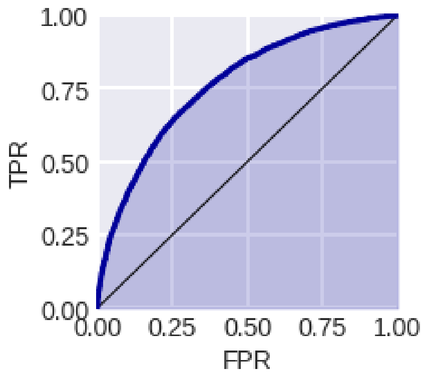
теперь можно задать стоимость обзвона (пусть =1$),
доход с отклика (пусть =5$)

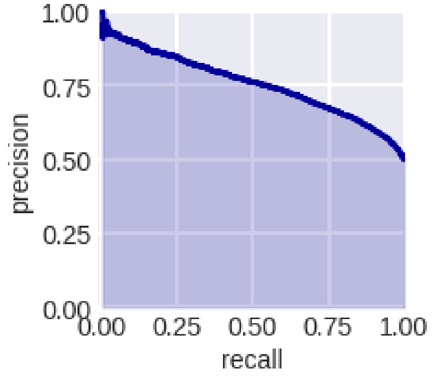
Если обзвонить 10%, траты = 11 238$, доход = 2572*5 = 12 860$
прибыль = 1 622$

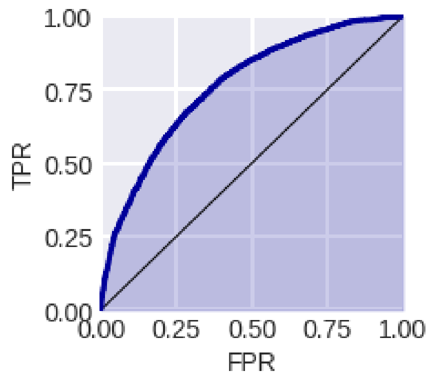
https://towardsdatascience.com/how-to-determine-the-best-model-6b9c584d0db4

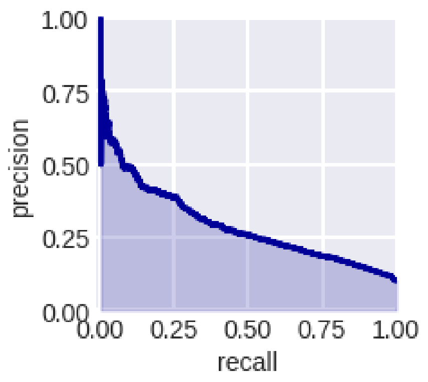
Сравнение метрик в задачах классификации

Модельные задачи

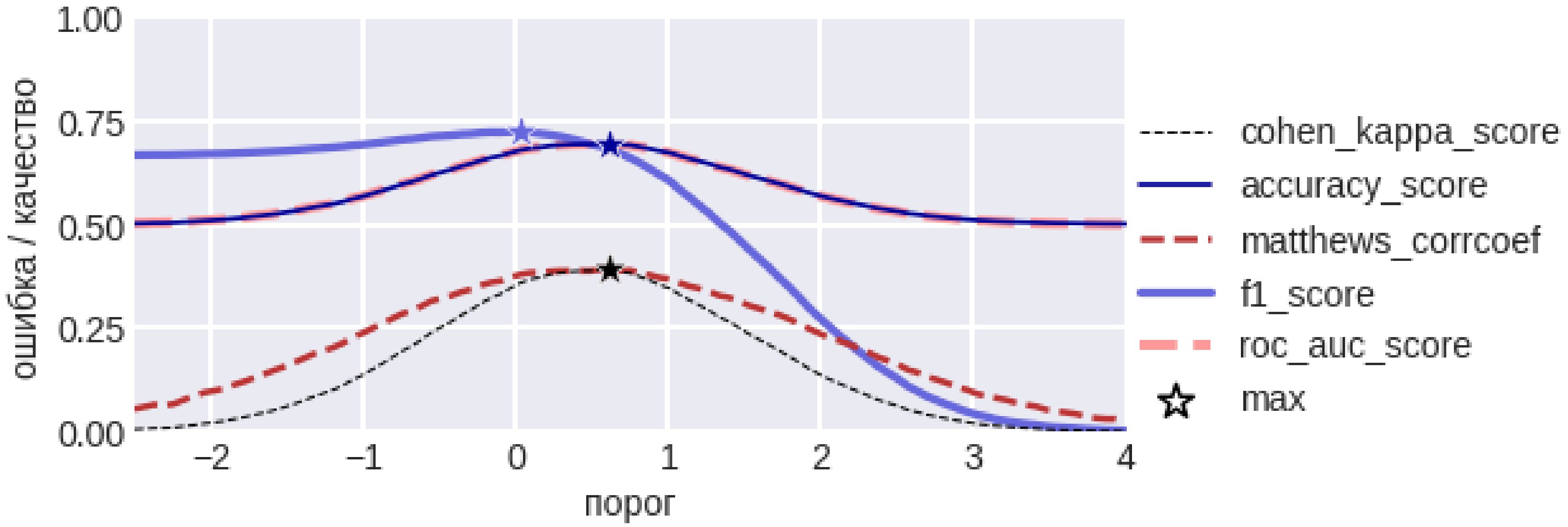


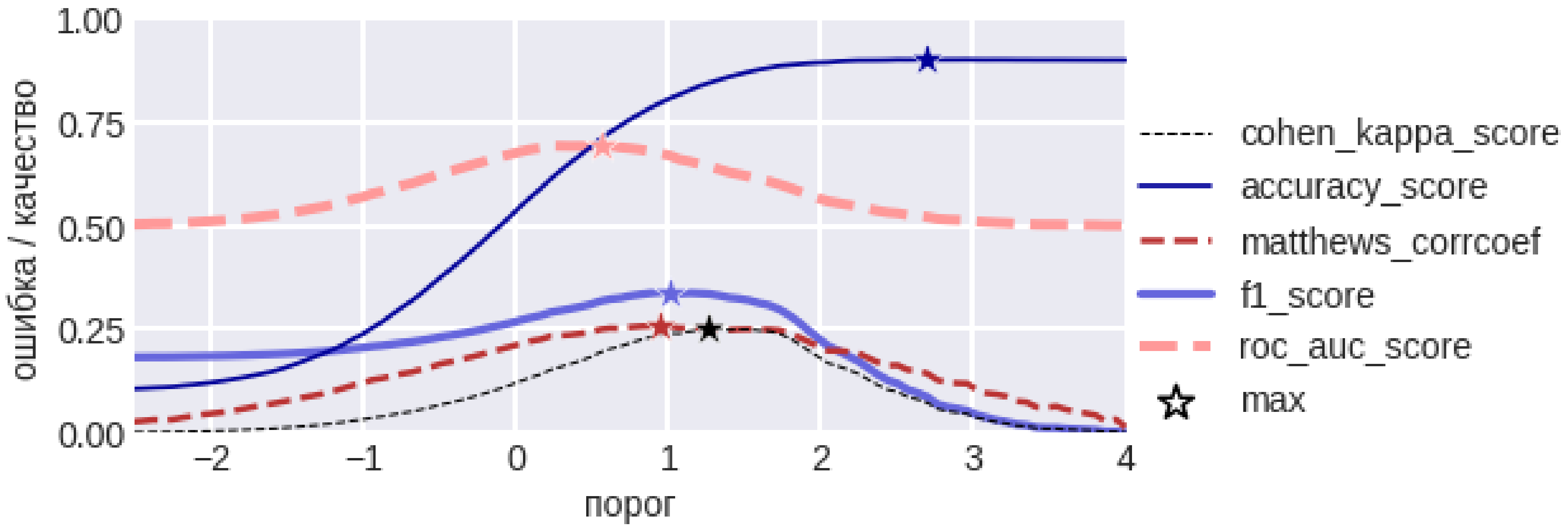






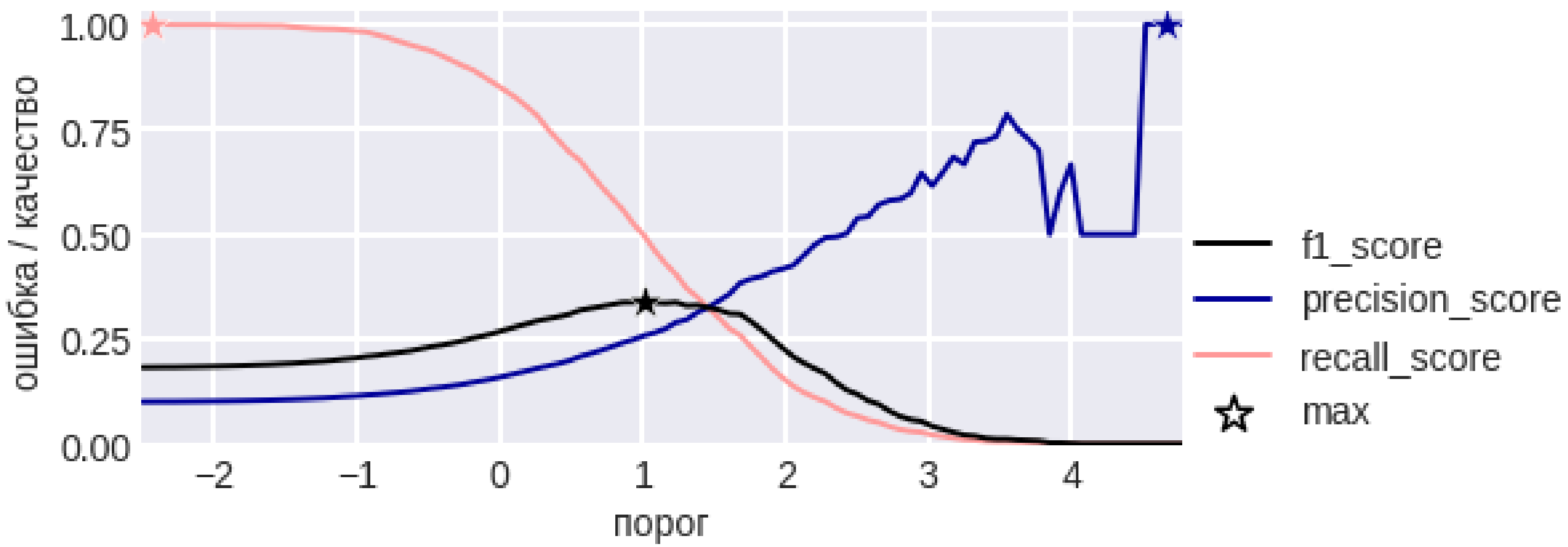
Сравнение метрик в задачах классификации





Сравнение метрик в задачах классификации

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



Почему прыгает точность?

Итог

Ищите матожидание ошибки!

Пробуйте константные решения.

Функции ошибки / качества можно обобщать!

Скоринговые задачи

Log Loss	AUC ROC
MSE	Gini
Exp Loss	AUC PR
ME	

Ещё кривые

Gain Curve	Lift Curve	K-S Chart

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

Интерактивная ROC-кривая
http://www.navan.name/roc/

Логистическая функция ошибки
https://dyakonov.org/2018/03/12/логистическая-функция-ошибки/

Кривые в машинном обучении
https://dyakonov.org/2019/08/29/кривые-в-машинном-обучении/

Калибровки
https://dyakonov.org/2020/03/27/проблема-калибровки-уверенности/