курс «Прикладные задачи анализа данных»

Минимизация ошибок

Александр Дьяконов

30 октября 2020 года

План

Как настраиваться на конкретные функции

Задачи для решения

Минимизация конкретной функции ошибок на практике

1. Она минимизируется напрямую

RMSE – Ridge

2. Она может быть приближена (имитирована другой)

RLMSE

3. Реализация минимизации конкретной функции

- XGBoost – прописываем метрику и производные
- ниже расщепление в деревьях для AUC

4. Решаем одной, донастариваем на другую

При раннем останове смотрим на значение целевой функции

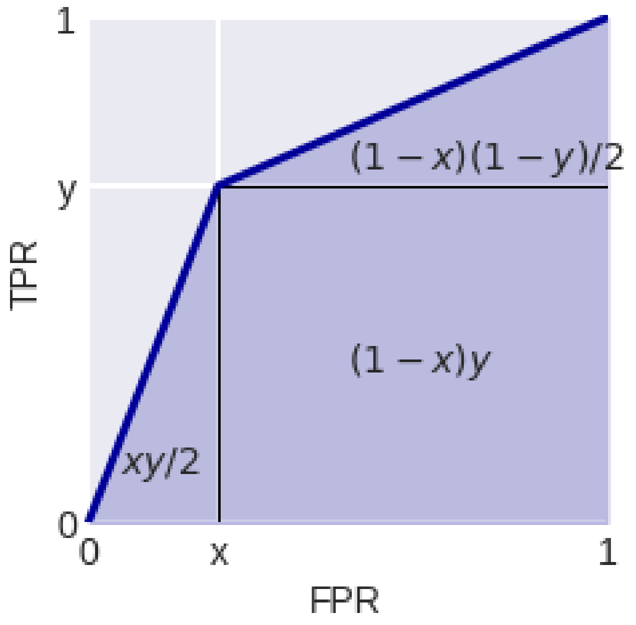
Идеология РП

Прямая настройка – идеология РП

$$F(\underbrace{B}_{\text{стандартная оптимизация}} \cdot \underbrace{C_c}_{\text{простое РП}}) \to \min_c$$

Как в задаче CrowdFlower (выбор порогов)

AUC ROC: Полностью бинарный случай $a\in\{0,1\}^m$, $y\in\{0,1\}^m$



$$S=\frac{xy}{2}+\frac{(1-x)(1-y)}{2}+(1-x)y$$

$$S=\frac{1-x+y}{2}$$

$$\mathrm{AUC}=\frac{1-\mathrm{FPR}+\mathrm{TPR}}{2}=\frac{1}{2}\left(1-\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}+\mathrm{TPR}\right)=$$

$$=\frac{1}{2}\left(1-\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}+\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\right)=\frac{1}{2}\left(\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}+\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\right)$$

AUC ROC: Полностью бинарный случай $a \in \{0,1\}^m$, $y \in \{0,1\}^m$

$$\mathrm{AUC} = \frac{\mathrm{R}_0 + \mathrm{R}_1}{2}$$

Среднее арифметическое полноты по классам 0 и 1…
это же сбалансированная точность!

А если выровнять мощности (как?),
то можно смотреть на точность…

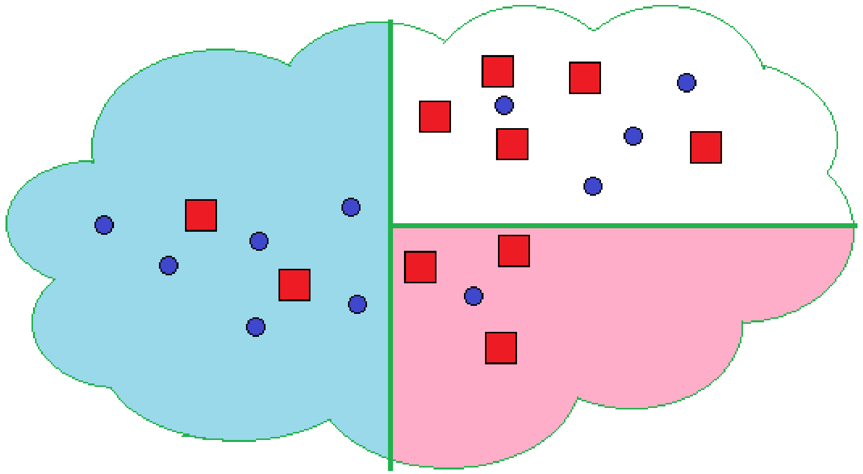
Максимизация AUC ~

$$\mathrm{TPR} - \mathrm{FPR} \to \max$$

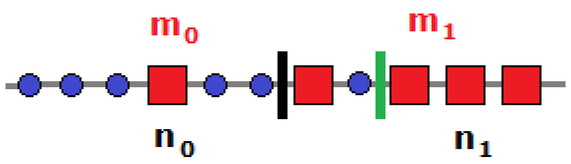
Реальный случай: Сбербанк

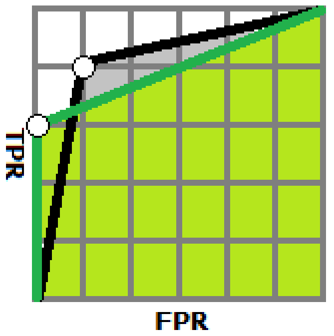
Использование «тайных знаний» на практике

Строим деревья в решающем лесе – хотим максимизировать AUC ROC



Хотим выбирать оптимальный порог



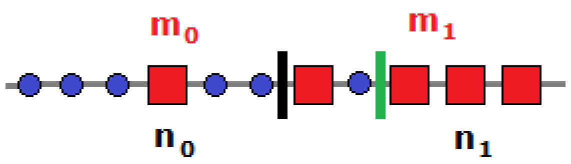


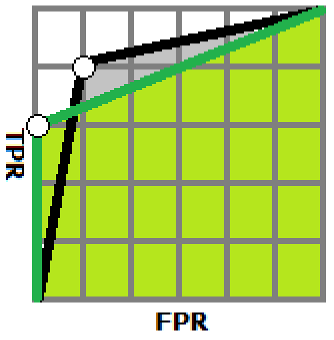
n_i – **числа точек в листах**

m_i – **числа объектов первого класса в листах**

$$m = m_1 + m_0$$
$$n = n_1 + n_0$$

Хотим выбирать оптимальный порог





$$AUC = \frac{1}{2}\left[\frac{m_1}{m} + \frac{n_0 - m_0}{n-m}\right] =$$

$$= \frac{1}{2}\left[\frac{m_1}{m} + \frac{(n-m)-(n_1-m_1)}{n-m}\right] =$$

$$= \frac{1}{2} + \frac{1}{2}\left[\frac{m_1}{m} - \frac{n_1-m_1}{n-m}\right]$$

Хотим выбирать оптимальный порог

Логично $|AUC - 0.5| \to \max$

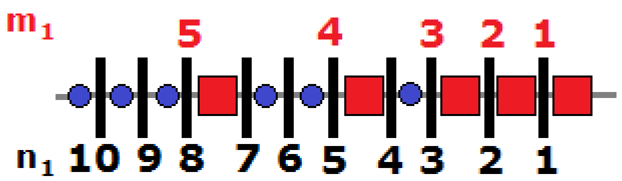
$$\left| \frac{m_1}{m} - \frac{n_1 - m_1}{n - m} \right| \to \max$$

Модуль разностей вероятностей классов «0», «1» в правом листе

$$\left| \frac{m_1 n - n_1 m}{m(n-m)} \right| \to \max$$

$$| m_1 n - n_1 m | \to \max$$

А ведь тогда просто реализовать перебор порогов в скриптовых языках



RF для AUC

Получили «новую» модель алгоритмов!

ДЗ Исследовать подобный критерий…
помогает ли в оптимизации AUC ROC?

Задача классификации {0,1} с ответами на [0,1]

Реальный случай

Пусть ошибка:

$$| y_i - a_i | \cdot \begin{cases} 0.8, & y_i = 1, \\ 0.2, & y_i = 0, \end{cases} \qquad (*)$$

где $y_i \in \{0,1\}$ **– верная классификация** i**-го объекта,**

$a_i \in [0,1]$ **– ответ нашего алгоритма.**

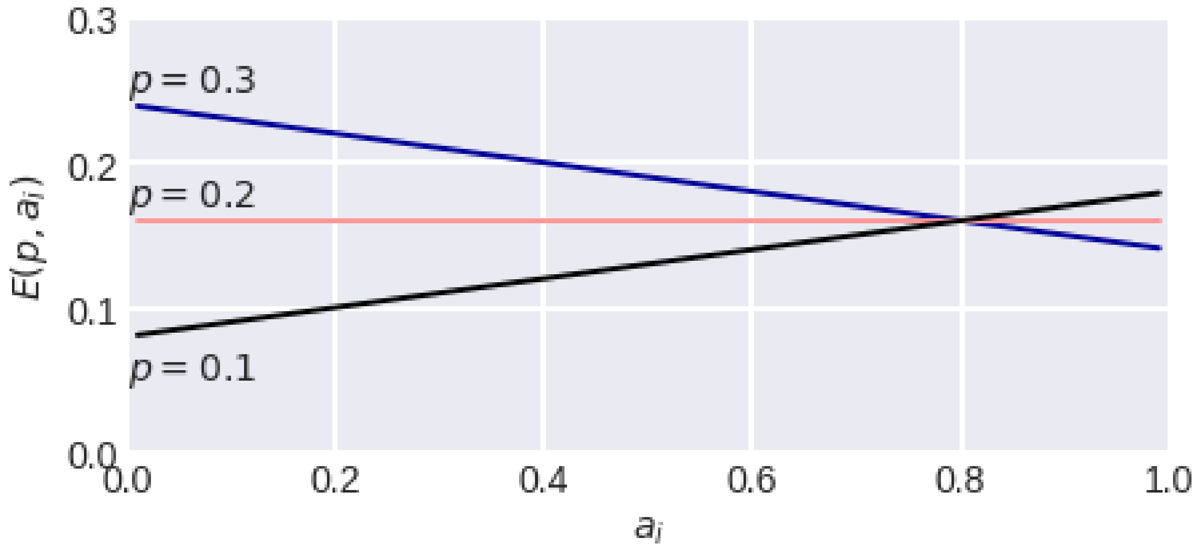
Заказчик: важно получать значения из отрезка [0,1] и интерпретировать как вероятности принадлежности к классу 1

Вычисление матожидания ошибки

Пусть i-й объект принадлежит к классу 1 с вероятностью p

Посчитаем матожидание нашей ошибки:

$$0.8|1-a_i|p+0.2|a_i|(1-p)=$$
$$=0.8p-0.8pa_i+0.2a_i-0.2pa_i=$$
$$=0.8p-(p-0.2)a_i$$



Вычисление матожидания ошибки

$$0.8p-(p-0.2)a_i \to \min$$

Оптимальное решение
(которое минимизирует матожидание ошибки)

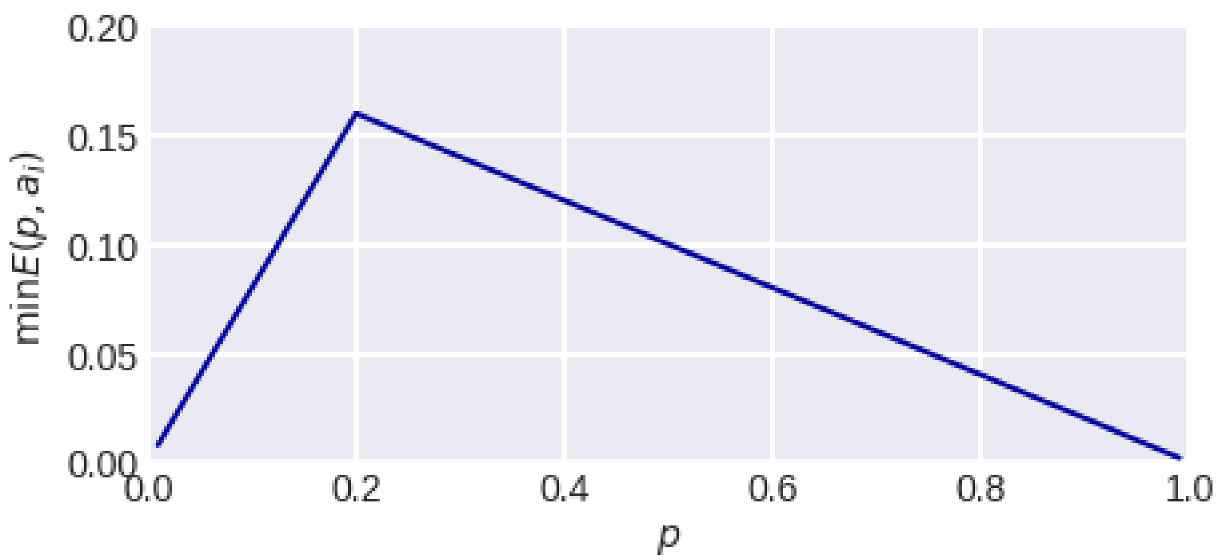
$$a_i=\begin{cases}0, & p<0.2,\\ 1, & p\ge 0.2.\end{cases}$$

Функционал (*) вынуждает нас
выдавать значения из множества {0,1}

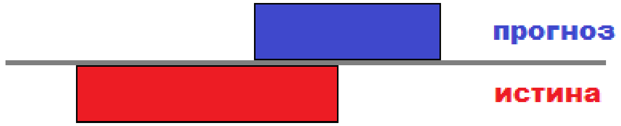
В чём ошибка заказчика, как исправить?

Неправильный выбор функционала

Интересно… матожидание ошибки (при оптимальном решении) в зависимости от p.



Задачи с интервальными признаками…
Как решать



Качество измеряем, например так:

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

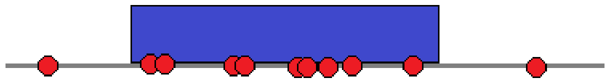
1 способ

Две задачи:
Целевой признак – начало интервала,
Целевой признак – конец интервала

- на практике работает не очень хорошо
- надо дорабатывать классические алгоритмы
(т.к. в случае начала интервала лучше занижать…)

2 способ

Целевой признак – середина интервала, плюс оцениваем отклонение от середины



- иногда противоречит природе данных (интервал заходит в отрицательную область)

Концепция решающего правила

Как всё-таки минимизировать нужный функционал…

1. **Есть предварительный ответ**

$$[a,b]$$

2. **Формируем окончательный параметрический…**

$$\left[\frac{a+b}{2}-\varepsilon\frac{b-a}{2},\frac{a+b}{2}-\varepsilon\frac{b+a}{2}\right]$$

3. **Настраиваем параметр**

Прямой перебор – явная минимизация

Можно и по-другому…

Но:

1. **Есть базовые алгоритмы (операторы)**
2. **Есть параметризованный способ перевода их ответов в нужные**
3. **Прямая минимизация функционала**

Из задачи Rossmann Store Sales

Root Mean Square Percentage Error (RMSPE)

$$\sqrt{\frac{1}{|\{i \mid y_i > 0\}|}\sum_{i:\, y_i>0}\left(\frac{a_i - y_i}{y_i}\right)^2}\ \text{и}$$

Оправдание деформации логарифмом…

Оправдание деформации логарифмом…

Ищем деформацию

$$\frac{a-y}{y} \approx F(a) - F(y)$$

чтобы функционал превратился в RMSE

$$\sqrt{\frac{1}{|\{i \mid y_i > 0\}|} \sum_{i: y_i > 0} (F(a_i) - F(y_i))^2}$$

Пусть $a = y + \delta$**, тогда**

$$\frac{\delta}{y} \approx F(y+\delta) - F(y) = F'\delta + o(\delta)$$

решим уравнение

$$\frac{\delta}{y} = F'\delta$$

Оправдание деформации логарифмом…

$$\frac{1}{y} = \frac{\partial F}{\partial y}$$

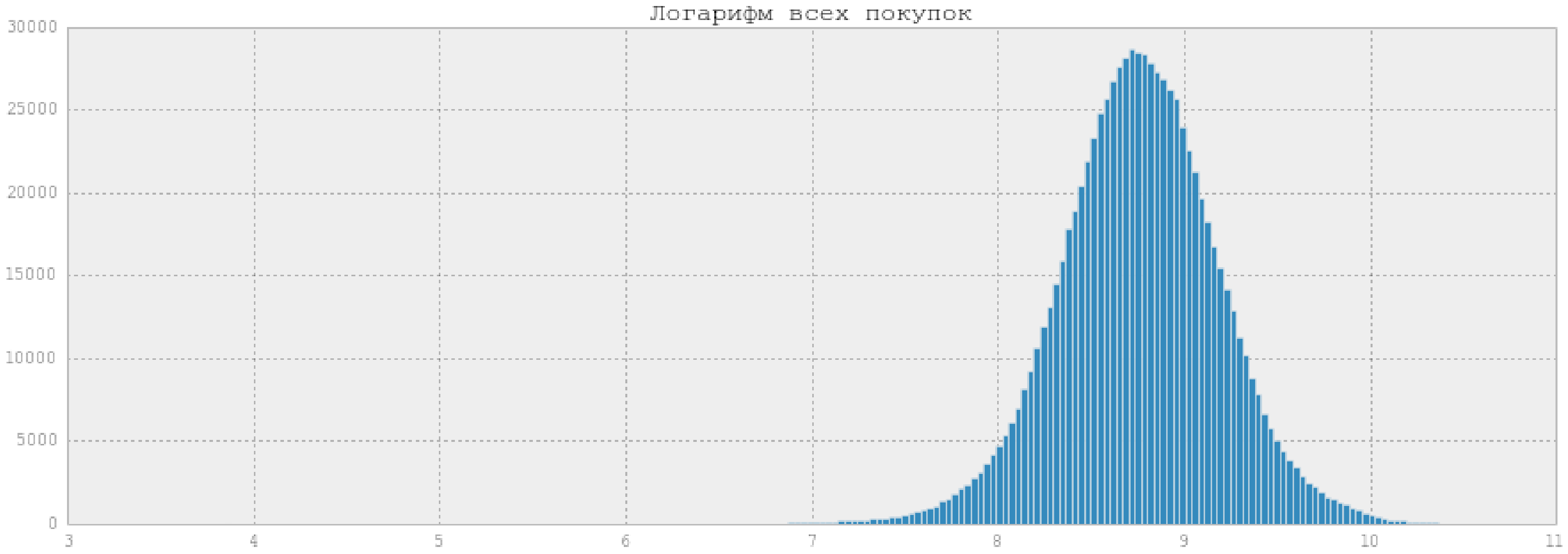
$$F(y) = \ln|y| + C$$

Выбираем деформацию $F(y) = \ln|y|$

Но, возможно, всё проще…

при логарифмировании отклонения похожи на нормальные

Распределения покупок



Метод градиентного спуска

Задача оптимизации функции ошибки от параметров алгоритма:

$$J(w) \to \min$$

$$w \leftarrow w - \alpha \left. \frac{\partial J}{\partial w} \right|_{w}$$

Возьмём конкретную задачу и метод

Качество: LOG LOSS

Метод: логистическая регрессия

(правильнее: сигмоида!)

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

$$a = \frac{1}{1+e^{-z}}$$

$z = z(w)$ – может как-то зависеть от параметров w.

На конкретном объекте:

$$J(w) = -\begin{cases} \log a, & y = 1, \\ \log(1-a), & y = 0. \end{cases}$$

Итак,

$$J(w) = -\begin{cases} \log\left(\dfrac{1}{1+e^{-z}}\right), & y=1, \\ \log\left(1-\dfrac{1}{1+e^{-z}}\right), & y=0. \end{cases}$$

$$J(w) = -\begin{cases} -\log\left(1+e^{-z}\right), & y=1, \\ -z-\log\left(1+e^{-z}\right), & y=0. \end{cases}$$

$$\frac{\partial \log\left(1+e^{-z}\right)}{\partial w} = -\frac{1}{1+e^{-z}} e^{-z} \frac{\partial z}{\partial w}$$

$$\frac{\partial J(w)}{\partial w} = -\frac{\partial z}{\partial w} \begin{cases} \dfrac{e^{-z}}{1+e^{-z}}, & y=1, \\ -1+\dfrac{e^{-z}}{1+e^{-z}}, & y=0. \end{cases}$$

Поэтому

$$\frac{\partial J(w)}{\partial w} = -\frac{\partial z}{\partial w}\begin{cases} 1-\dfrac{1}{1+e^{-z}}, & y=1, \\ 0-\dfrac{1}{1+e^{-z}}, & y=0. \end{cases} = -\frac{\partial z}{\partial w}(y-a)$$

Получаем формулу для коррекции весов:

$$w \leftarrow w + \alpha (y-a)\frac{\partial z}{\partial w}$$

Очень логичная: изменение зависит от величины ошибки

$$(y-a)$$

В классической логистической регрессии

$$a = \frac{1}{1 + e^{-\sum_{t=1}^{n} w_t [x]_t}}$$

(линейная комбинация признаков)

Поэтому

$$w \leftarrow w + \alpha(y - a)x$$

x **– признаковое описание объекта**

Вопрос с подвохом

Качество: logloss

Метод: линейная регрессия

$$J(w) = -\begin{cases} \log(z), & y = 1, \\ \log(1-z), & y = 0. \end{cases}$$

$$\frac{\partial J(w)}{\partial w} = -\frac{\partial z}{\partial w}\begin{cases} 1/z, & y = 1, \\ -1/(1-z), & y = 0, \end{cases} = \frac{1}{z+y-1}\frac{\partial z}{\partial w}$$

тогда

$$w \leftarrow w - \frac{\alpha}{z+y-1}\frac{\partial z}{\partial w}$$

Что смущает в этой формуле?

Почему так получилось?

Вопрос с подвохом

$$w \leftarrow w - \frac{1}{z+y-1}\frac{\partial z}{\partial w}$$

Коррекция происходит даже при абсолютно правильном ответе…

$$J(w) = -\begin{cases} \overbrace{\log(z)}^{\text{минимум не при } z=1}, & y=1, \\ \log(1-z), & y=0. \end{cases}$$

Нужны ещё ограничения

$$\min(\max(z, 0), 1)$$

В логистической регрессии

$$\frac{1}{1+e^{-z}} \in [0, 1]$$

Линейная регрессия с НСКО

$$J(w) = (w^{\mathrm{T}}x - y)^2 \to \min$$

x **– объект,**

y **– его регрессионная метка**

$$\frac{\partial J}{\partial w} = 2(w^{\mathrm{T}}x - y)x$$

$$w \leftarrow w - \alpha(w^{\mathrm{T}}x - y) \cdot x$$

Выберем α

Коррекция такая же как в логистической регрессии с logloss-ом!

Метод наискорейшего спуска

$$((w-\alpha(w^{\mathrm{T}}x-y)x)^{\mathrm{T}}x-y)^{2}\to\min$$

$$w^{\mathrm{T}}x-\alpha(w^{\mathrm{T}}x-y)x^{\mathrm{T}}x-y=0$$

$$w^{\mathrm{T}}x-y=\alpha(w^{\mathrm{T}}x-y)x^{\mathrm{T}}x$$

$$\alpha=\frac{1}{x^{\mathrm{T}}x}$$

Задача 1.

Качество: СКО

$$J = \frac{1}{m}\sum_{i=1}^{m}(y_i - a_i)^2$$

Метод: логистическая регрессия

$$a = \sigma(z) = \frac{1}{1+e^{-z}}$$

Вычислить формулу для коррекции весов методом стохастического градиентного спуска

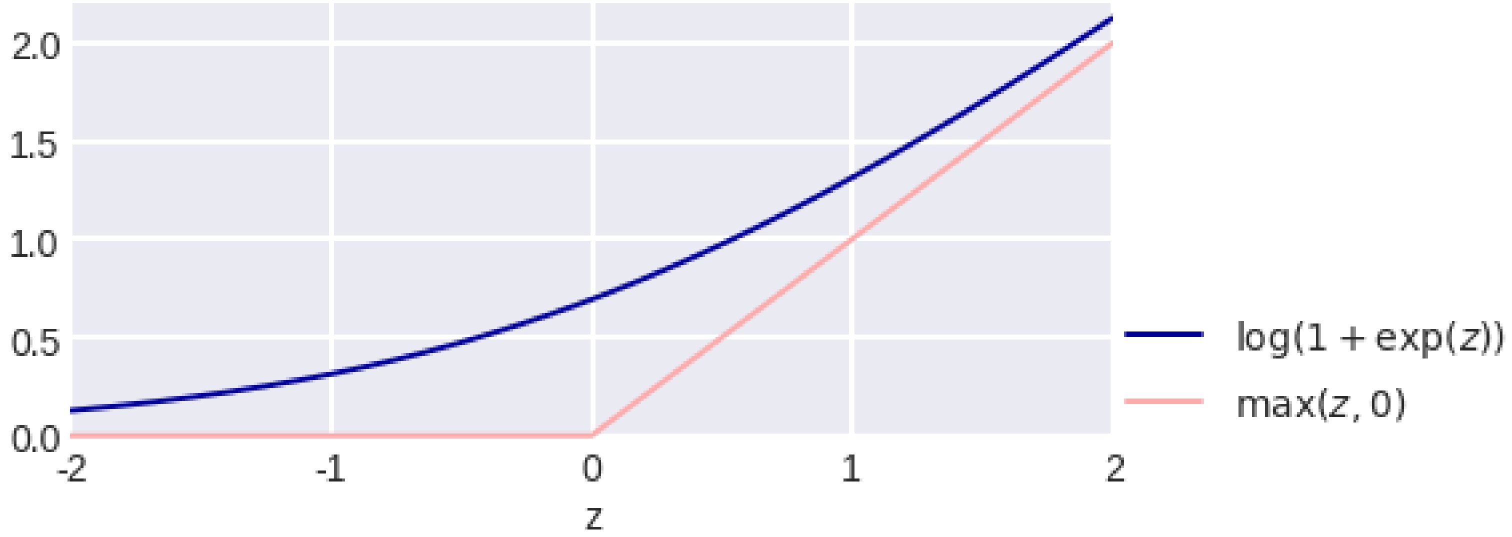
Задача 2.

Качество: СКО

$$J=\frac{1}{m}\sum_{i=1}^{m}(y_i-a_i)^2$$

Метод: $a=\ln(1+e^z)$

Вычислить формулу для коррекции весов методом стохастического градиентного спуска



Задача 2. Ответ

$$w \leftarrow w - \alpha \frac{(a-y)}{1+e^{-z}} \frac{\partial z}{\partial w}$$

Почти классический вариант (линейная регрессия + СКО), но с поправкой на отрицательной оси...

$$w \leftarrow w - \alpha \underbrace{(a-y)}_{\text{классика}} \underbrace{\sigma(z)}_{=\frac{1}{1+e^{-z}}} \frac{\partial z}{\partial w}$$

$$\frac{(a-y)}{1+e^{-z}} \approx \begin{cases} (a-y), & z >> 0 \\ 0, & z << 0 \end{cases}$$

Всё очень логично!

Задача 1. Ответ

$$w \leftarrow w - \alpha \cdot a(1-a)(y-a)\frac{\partial z}{\partial w}$$

рекомендация: $\frac{\partial a(w)}{\partial w} = a(1-a)$ (производная сигмоиды)

Вопрос: что плохого в это формуле?

Задача 1. Ответ

$$w \leftarrow w - \alpha \cdot \underbrace{a(1-a)}_{\text{что-то новое}} \underbrace{(y-a)}_{\text{классика}} \frac{\partial z}{\partial w}$$

Вопрос: **что плохого в это формуле?**

В случае полностью неправильного ответа, например

$$y = 0, a \approx 1$$

коррекции почти не будет:

$$a(1-a)(y-a) \approx 0$$

Вопрос: **что с этим делать?**

Задача 3. Вычислить Cohen's Kappa

	yes	no
yes	20	5
no	10	15

0.4

	yes	no
yes	45	15
no	25	15

~0.13

	yes	no
yes	30	10
no	10	5

~0.08

	yes	no
yes	30	20
no	10	15

~0.18

```
s = sum(H.iloc[i,i] for i in range(2)) # 35
n = H.values.sum() # 50
po = s / n # 0.7
pe = sum(H.iloc[i,:].sum() * H.iloc[:,i].sum() /
         n**2 for i in range(2)) # 0.5 = 36/n**2 + 64/n**2
(po - pe) / (1 - pe) # 0.4

from sklearn.metrics import cohen_kappa_score
cohen_kappa_score(a, y) # но это по ответам!
```

Задача 4.

Показать, что Каппа Коэна частный случай Weighted Kappa для 2х классов.

Напомним

$$\kappa = \frac{\mathrm{Accuracy} - \mathrm{Accuracy}_{\mathrm{chance}}}{1 - \mathrm{Accuracy}_{\mathrm{chance}}}$$

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

Будем считать, что матрица штрафов $\begin{bmatrix} 0 & w \\ w & 0 \end{bmatrix}$

Задача 4: решение

$$\kappa=\frac{\mathrm{Accuracy}-\mathrm{Accuracy}_{\mathrm{chance}}}{1-\mathrm{Accuracy}_{\mathrm{chance}}}=\frac{\dfrac{m_{00}+m_{11}}{m}-\dfrac{m_{0:}m_{:0}}{m^2}-\dfrac{m_{1:}m_{:1}}{m^2}}{1-\dfrac{m_{0:}m_{:0}}{m^2}-\dfrac{m_{1:}m_{:1}}{m^2}}=$$

$$=\frac{m(m_{00}+m_{11})-m_{0:}m_{:0}-m_{1:}m_{:1}}{m^2-m_{0:}m_{:0}-m_{1:}m_{:1}}=$$

$$=\frac{\begin{gathered}(m_{00}+m_{11})(m_{00}+m_{01}+m_{10}+m_{11})-\\-m_{00}m_{00}-m_{00}m_{10}-m_{01}m_{00}-m_{01}m_{10}-m_{11}m_{11}-m_{11}m_{01}-m_{10}m_{11}-m_{10}m_{01}\end{gathered}}{m_{0:}m_{:1}+m_{1:}m_{:0}}=$$

$$=\frac{2(m_{00}m_{11}-m_{10}m_{01})}{m_{0:}m_{:1}+m_{1:}m_{:0}}$$

Задача 4: решение

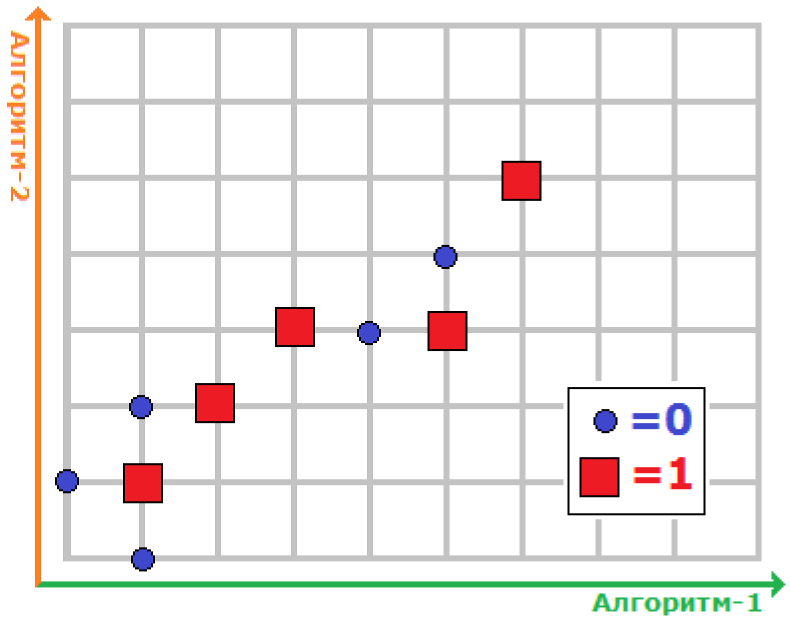
$$\kappa_w = 1 - \frac{wm_{01} + wm_{10}}{ws_{01} + ws_{10}} = 1 - \frac{wm_{01} + wm_{10}}{w\frac{m_{0:}m_{:1}}{m} + w\frac{m_{1:}m_{:0}}{m}} =$$

$$= \frac{m_{0:}m_{:1} + m_{1:}m_{:0} - m(m_{01} + m_{10})}{m_{0:}m_{:1} + m_{1:}m_{:0}} =$$

$$= \frac{2(m_{00}m_{11} - m_{10}m_{01})}{m_{0:}m_{:1} + m_{1:}m_{:0}}$$

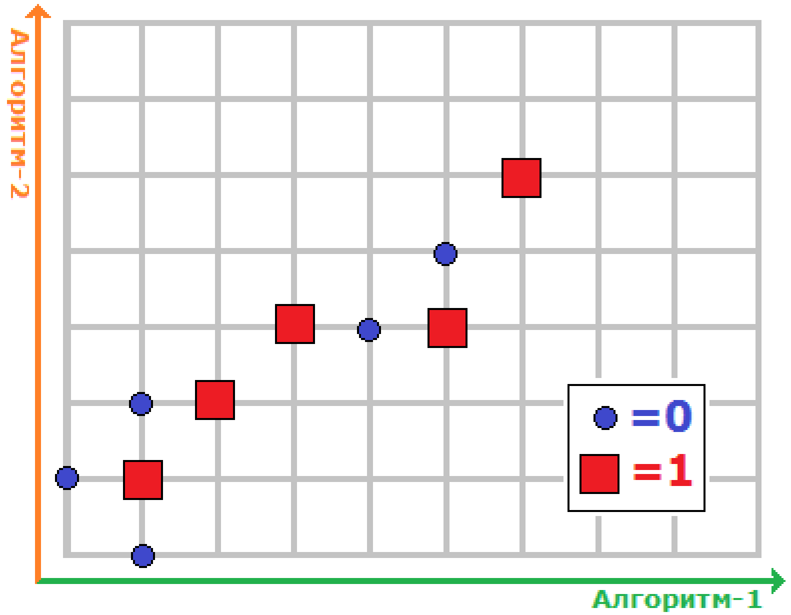
Задача 5.

Рассматривается задача классификации на два класса.
На рисунке показаны объекты в пространстве ответов двух алгоритмов.
Вычислить AUC ROC для алгоритмов.



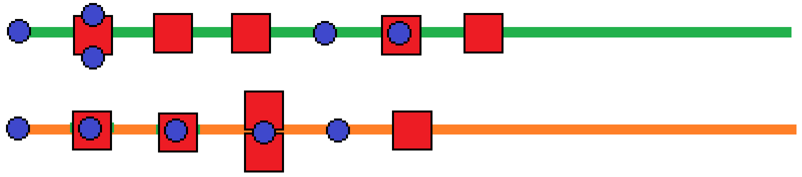
Задача 5 – Решение

1. Смотрим проекции на оси – ответы алгоритмов

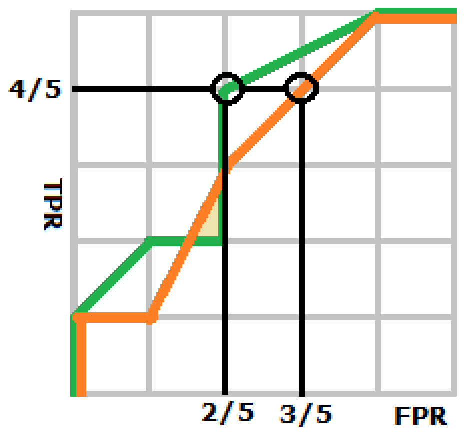


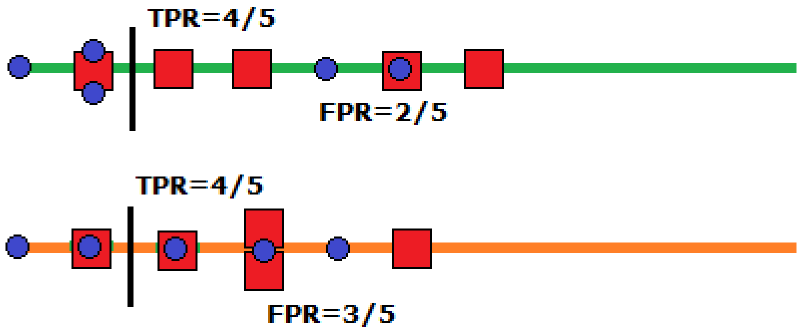
Первый алгоритм:

Второй алгоритм:

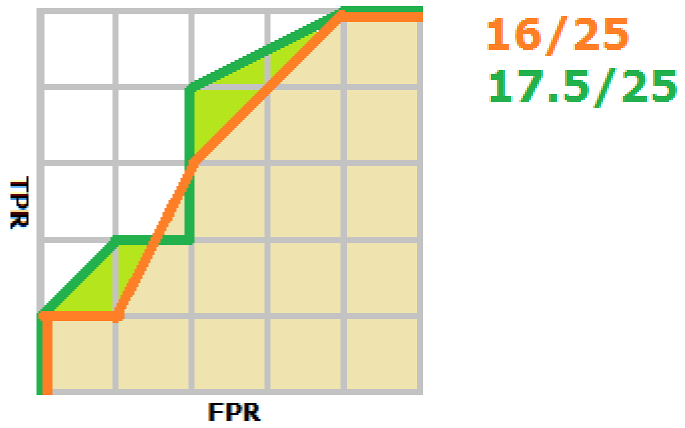


2. По проекциям строим ROC - кривые:





3. Вычисляем площади под ROC - кривыми:

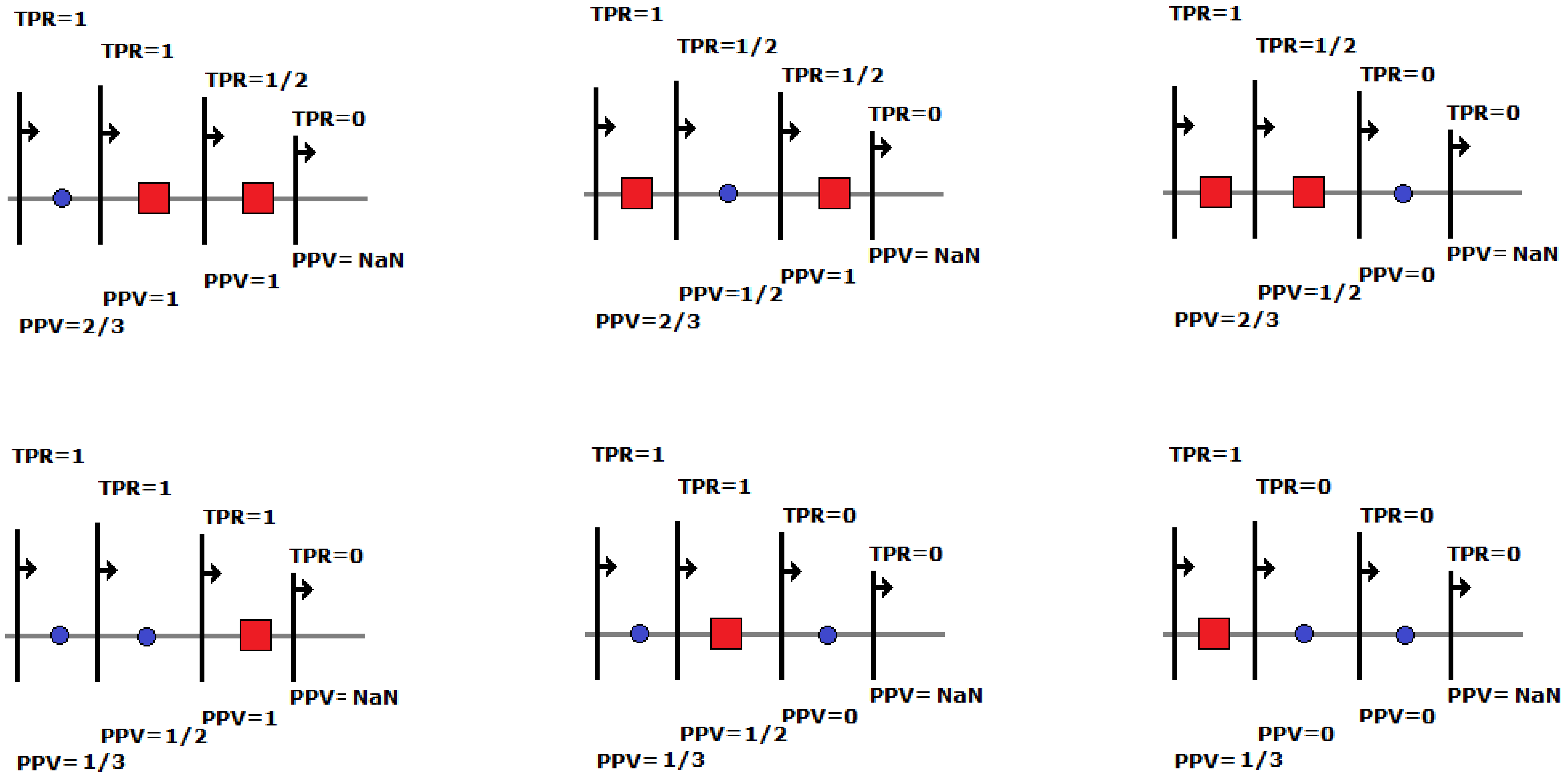


Задача 6

Какие значения F_1-меры могут быть у классификатора в задаче с двумя непересекающимися классами и тремя объектами?

Задача 6 – Решение.

Можно честно рассмотреть все возможные случаи:



Задача 6 – Решение.

Получаем, что F1-мера – среднее гармоническое чисел из пар (1, 1), (1/2, 1), (2/3, 1), (1/3, 1), (1/2, 1/2), (0, 0)

Все возможные значения F1-меры: 1, 0.8, 2/3, 0.5, 0

Но можно быстрее догадаться до ответа…

Задача 7

Вычислить ap@k:

ap@5(actual = [1, 2, 3], predict = [1, 4, 5, 2, 6, 3])

ap@3(actual = [1, 2, 3], predict = [1, 4, 5, 2, 6, 3])

ap@3(actual = [1], predict = [1, 2, 3, 4, 5, 6])

ap@3(actual = [1, 3], predict = [1, 2, 3, 4, 5, 6])

ap@2(actual = [1, 3], predict = [1, 2, 3, 4, 5, 6])

Задача 7

Решение:

ap@5(actual = [1, 2, 3], predict = [1, 4, 5, 2, 6, 3]) = 0.5

ap@3(actual = [1, 2, 3], predict = [1, 4, 5, 2, 6, 3]) = 1/3

ap@3(actual = [1], predict = [1, 2, 3, 4, 5, 6]) = 1

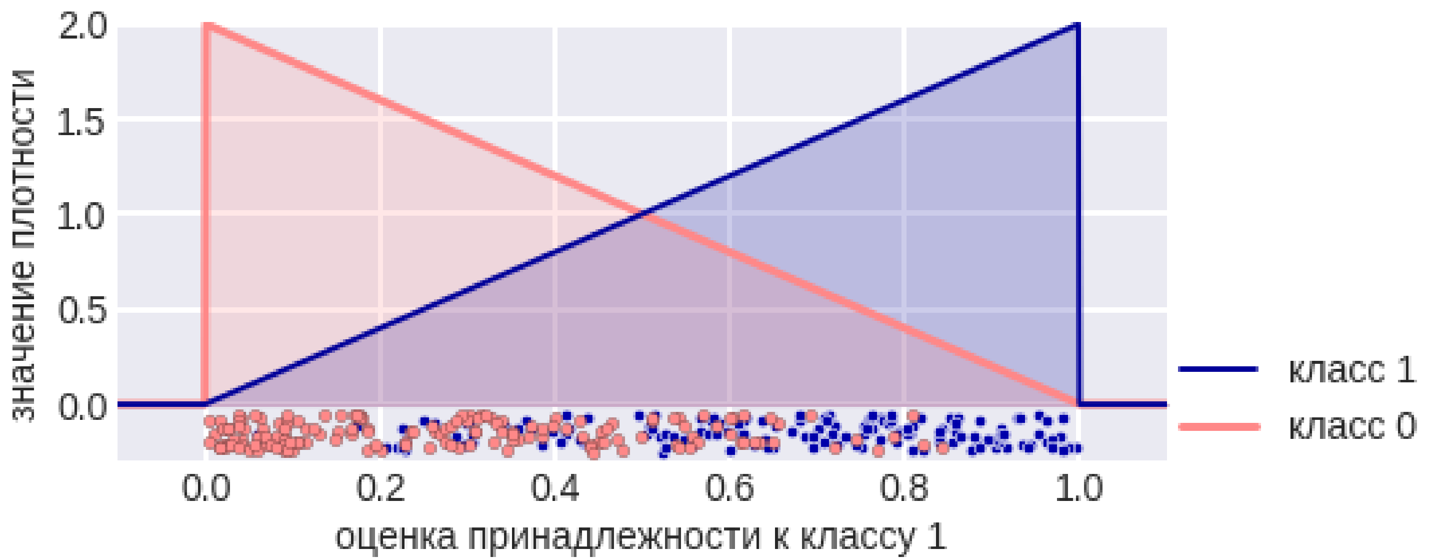
ap@3(actual = [1, 3], predict = [1, 2, 3, 4, 5, 6]) = 5/6

ap@2(actual = [1, 3], predict = [1, 2, 3, 4, 5, 6]) = 1/2

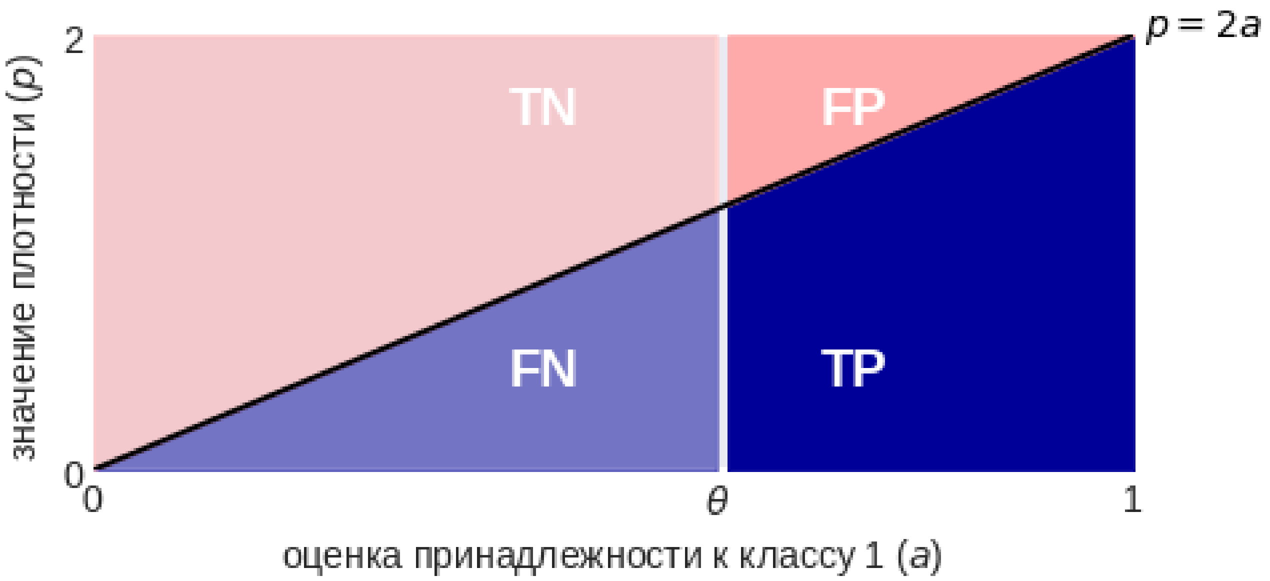
Задача 8

На ответах алгоритма $a(x) \in [0,1]$ объекты класса 0 распределены с плотностью $p_0(a) = 2 - 2a$, а объекты класса 1 – с плотностью $p_1(a) = 2a$.

Построить ROC-кривую и вычислить площадь под ней.



Задача 8 – решение



Решение

$$\mathrm{TPR}(\theta)=1-\frac{1}{2}\theta 2\theta=1-\theta^2$$

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

Задача 8 – решение

Площадь под параметрической кривой

$$\int_1^0 \mathrm{TPR}(\theta)\cdot \mathrm{FPR}'(\theta)\partial\theta = 2\int_0^1 (1-\theta^2)(1-\theta)\partial\theta$$

или

$$\mathrm{TPR} = 2\sqrt{\mathrm{FPR}} - \mathrm{FPR}.$$

$$\int_0^1 (2\sqrt{t} - t)\partial t = \frac{5}{6} \approx 0.83.$$

https://dyakonov.org/2017/07/28/auc-roc-%d0%bf%d0%bb%d0%be%d1%89%d0%b0%d0%b4%d1%8c-%d0%bf%d0%be%d0%b4-%d0%ba%d1%80%d0%b8%d0%b2%d0%be%d0%b9-%d0%be%d1%88%d0%b8%d0%b1%d0%be%d0%ba/

Задача 9

В описанной выше задаче вычислить:

$$R = 1 - \theta^2$$

$$P = (1+\theta)/2$$

$$F_1 = \frac{1-\theta^2}{1.5-\theta}$$

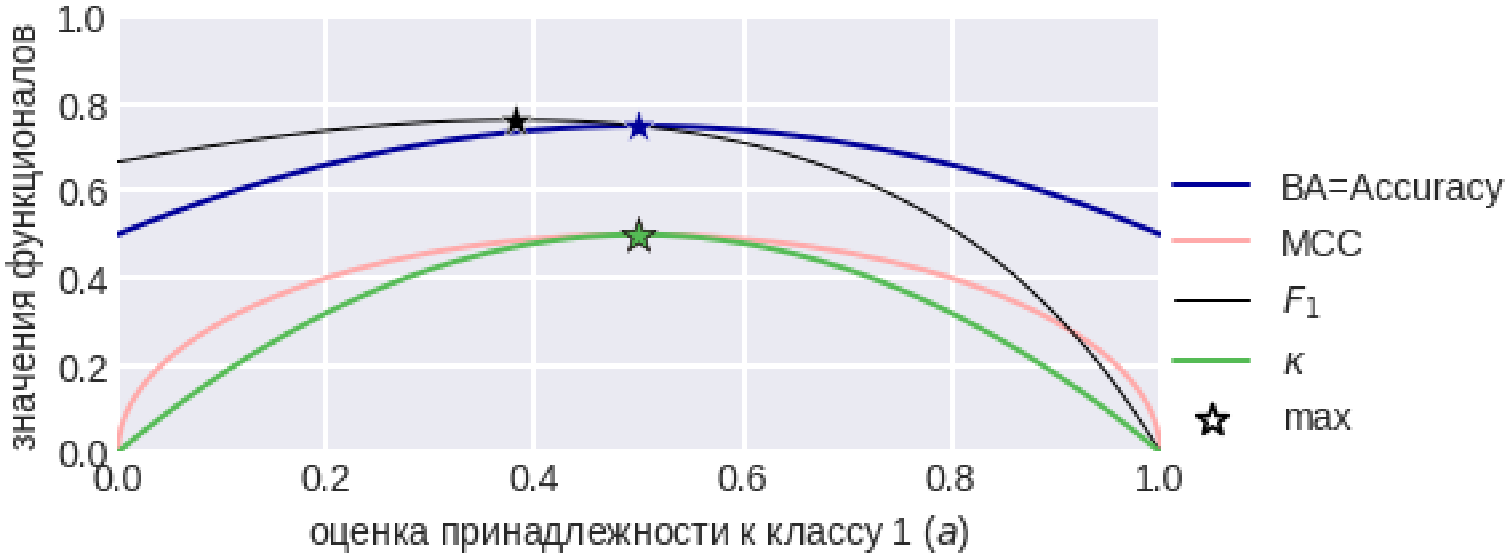
$$\mathrm{MCC} = \sqrt{\theta(1-\theta)}$$

$$\kappa = \frac{\dfrac{1+2\theta-2\theta^2}{2} - \dfrac{1}{2}}{1-\dfrac{1}{2}} = 2\theta(1-\theta)$$

$$\mathrm{BA{=}Accuracy} = \frac{(1-\theta^2)+(1-(1-\theta)^2)}{2} = \frac{1+2\theta-2\theta^2}{2}$$

Задача 9

построить графики



Задача 10 – задачи на вычисление

Вычислить коэффициент Мэттьюса

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

для следующих векторов меток и ответов

истина	ответ	MCC
[1,1,1,1,0,0,0,0]	[1,1,1,0,0,0,1,1]	0.258
[1,1,1,1,0,0,0,0]	[1,1,1,1,1,0,1,1]	0.378
[1,1,1,1,1,1,0,0]	[1,1,1,1,1,0,1,1]	-0.218

Задача 11 – Задачи на вычисление

Проверить иллюстрацию из лекции про многомерный AUC

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