

Линейные методы классификации

МАКСИМОВСКАЯ
АНАСТАСИЯ

Задача классификации

Задача классификации — задача, в которой имеется множество объектов, разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся.

Линейный классификатор — алгоритм классификации, основанный на построении линейной разделяющей поверхности.

Бинарная классификация — это задача классификации элементов заданного множества в две группы.

Целевая переменная в задаче классификации – класс, к которому принадлежит наблюдение.

Основные термины

- $\mathbb{X} = \mathbb{R}^d$ – пространство объектов
- $\mathbb{Y} = \{+1, -1\}$ – множество допустимых ответов
- « + 1» – положительный класс
- « - 1» – отрицательный класс

Линейная модель

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j$$

Параметры линейной регрессии – веса (коэффициенты w_j). Вес w_0 называется свободным коэффициентом или сдвигом (bias). Заметим, что после знака суммы написано скалярное произведение. Также добавим в выборку w_{d+1} признак, равный единице, тогда необходимость в свободном коэффициенте отпадет. Перепишем формулу в более компактном виде:

$$a(x) = \langle w, x \rangle$$

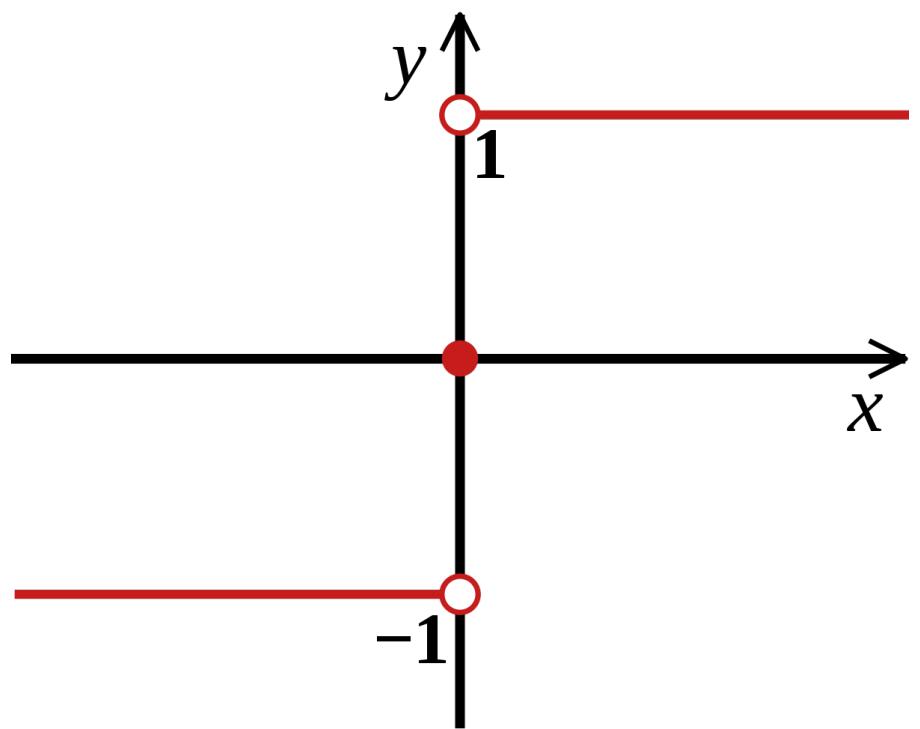
Линейный классификатор

$$a(x) = \langle w, x \rangle$$

Что можем сделать, чтобы стало -1 или +1? Взять знак!

$$a(x) = \text{sign} \langle w, x \rangle$$

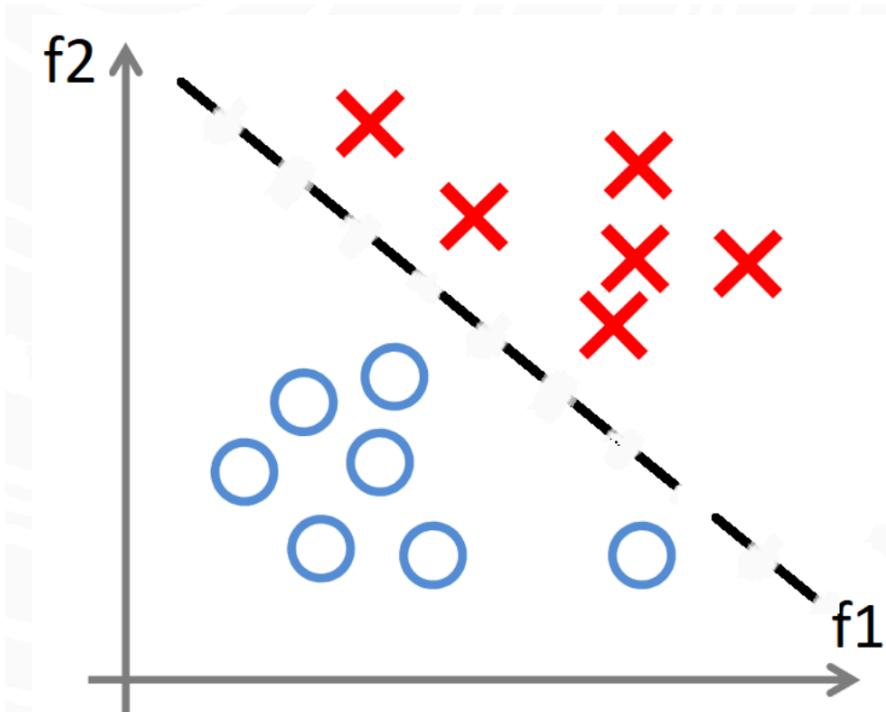
Напоминание



$$sgn(x) = \begin{cases} 1 & if \quad x > 0, \\ 0 & if \quad x = 0, \\ -1 & if \quad x < 0 \end{cases}$$

Геометрическая интерпретация

Уравнение вида $\langle w, x \rangle = 0$ определяет гиперплоскость (прямая на картинке).



- Если $\sum_{j=1}^n w_j x_j > 0$, то $\text{sign}(\sum_{j=1}^n w_j x_j) = +1$ – объект будет отнесен к положительному классу
- Если $\sum_{j=1}^n w_j x_j < 0$, то $\text{sign}(\sum_{j=1}^n w_j x_j) = -1$ – объект будет отнесен к отрицательному классу

Обучение линейного классификатора

Возьмем функционал качества error rate:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

Его будем минимизировать:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i] = \frac{1}{\ell} \sum_{i=1}^{\ell} [\text{sign}\langle w, x_i \rangle \neq y_i] \rightarrow \min_w$$

Обучение линейного классификатора

- Функционал дискретный
- Не можем минимизировать градиентными методами или вывести аналитическое решение
- Попробуем перейти к минимизации гладкого функционала

Обучение линейного классификатора

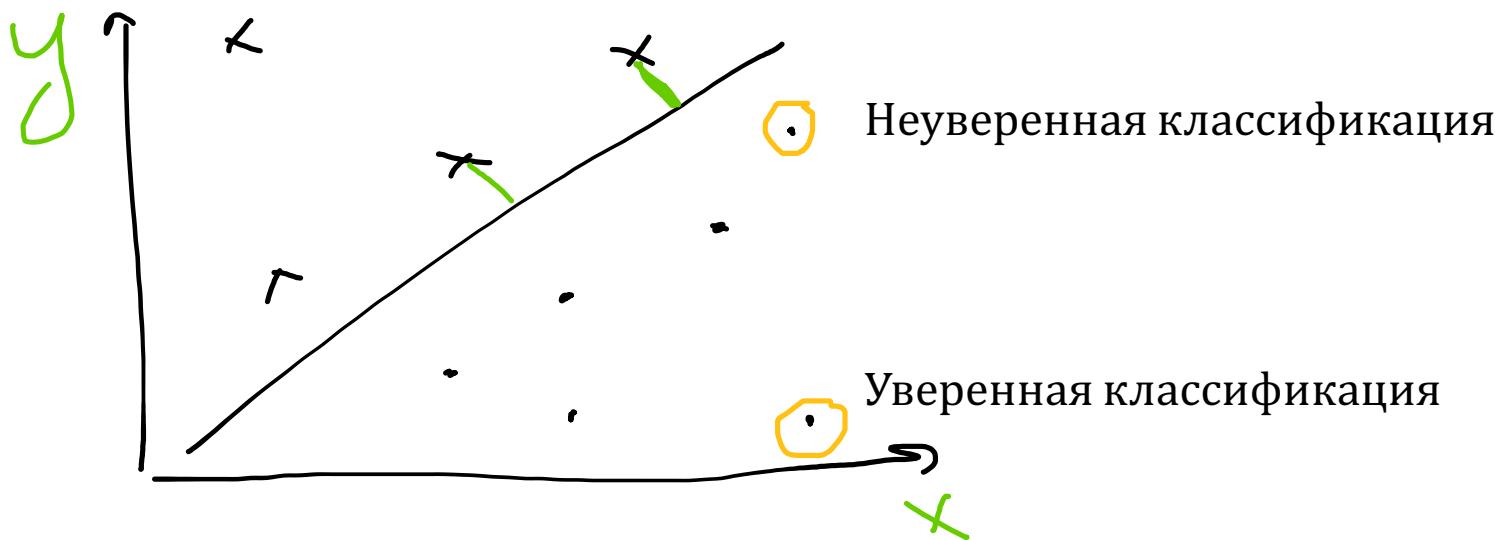
Перепишем наш функционал:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [y_i \underbrace{\langle w, x_i \rangle}_{M_i} < 0] \rightarrow \min_w$$

- $M = y \langle w, x \rangle$ – Отступ
- $M > 0$ – целевая переменная и скалярное произведение одного знака, значит, класс правильно предсказан
- $M < 0$ – целевая переменная и скалярное произведение не одного знака, значит, класс неверно предсказан

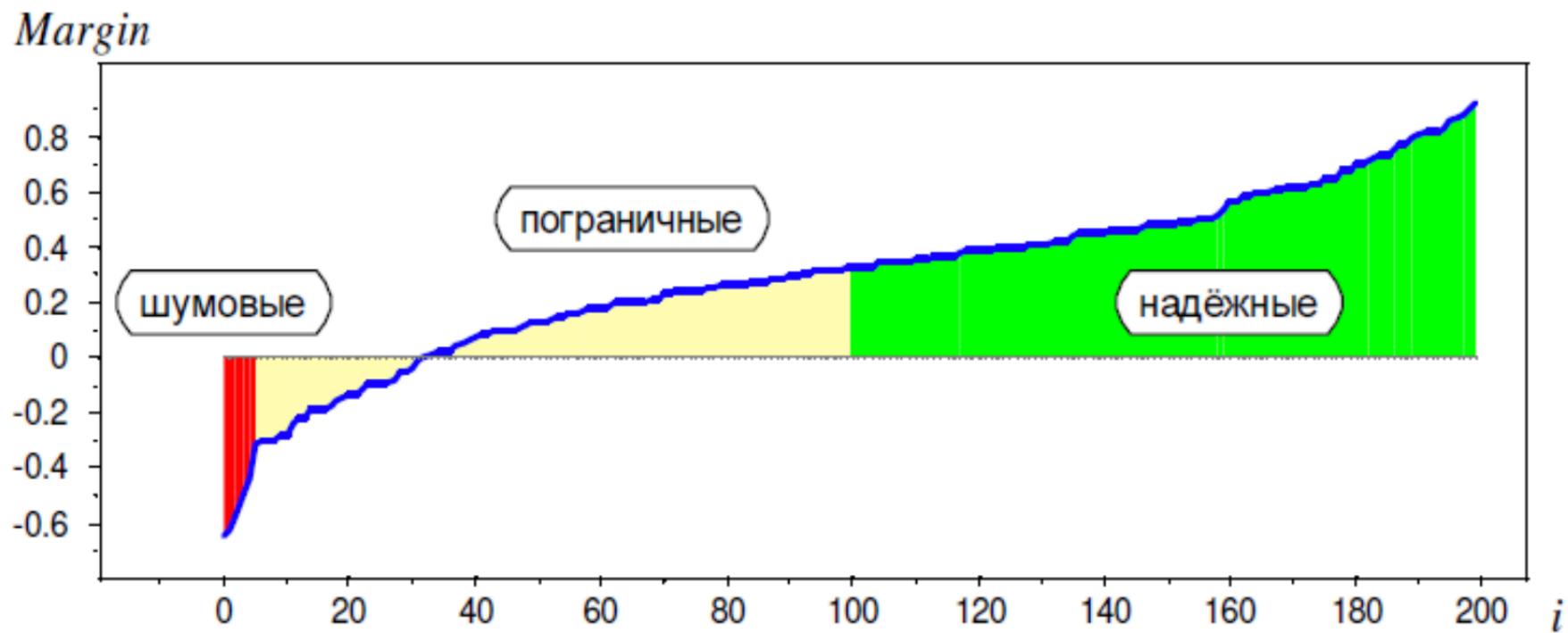
Обучение линейного классификатора

- Абсолютная величина отступа $|M|$ – по сути, расстояние до разделяющей гиперплоскости
- Абсолютная величина отступа отражает уверенность прогноза



Обучение линейного классификатора

Ранжирование объектов по возрастанию отступа:



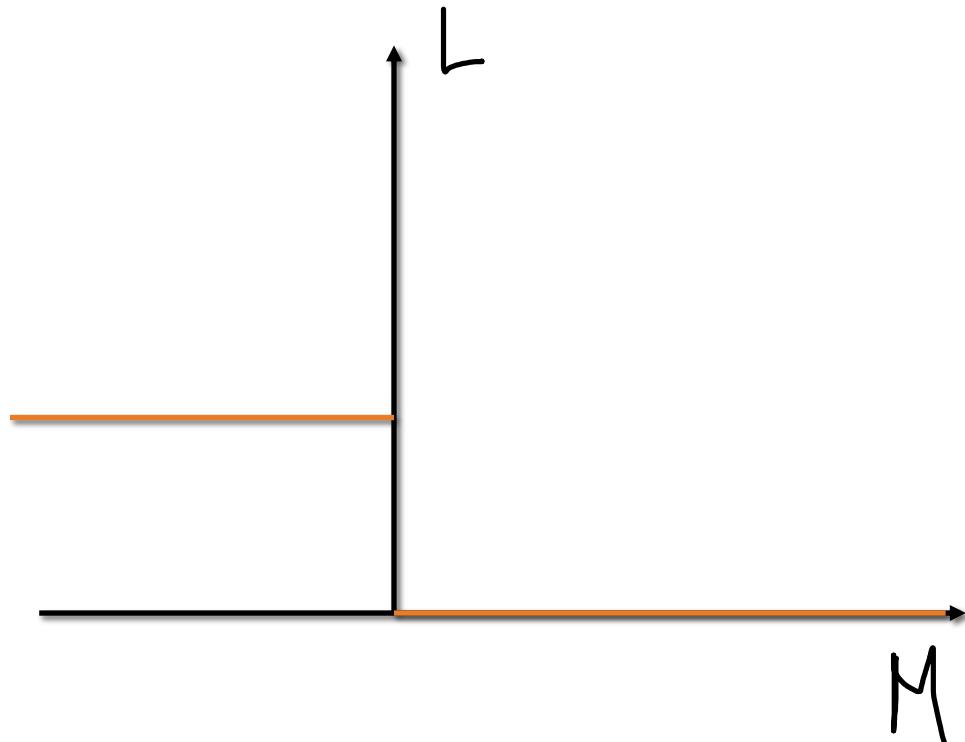
Обучение линейного классификатора

- Если отступ очень большой и отрицательный – модель ошиблась, но очень уверена в своем прогнозе
- Зачастую это выброс

Обучение линейного классификатора

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0]$$

Данный функционал оценивает ошибку алгоритма
с помощью функции потерь $L(M) = [M < 0]$



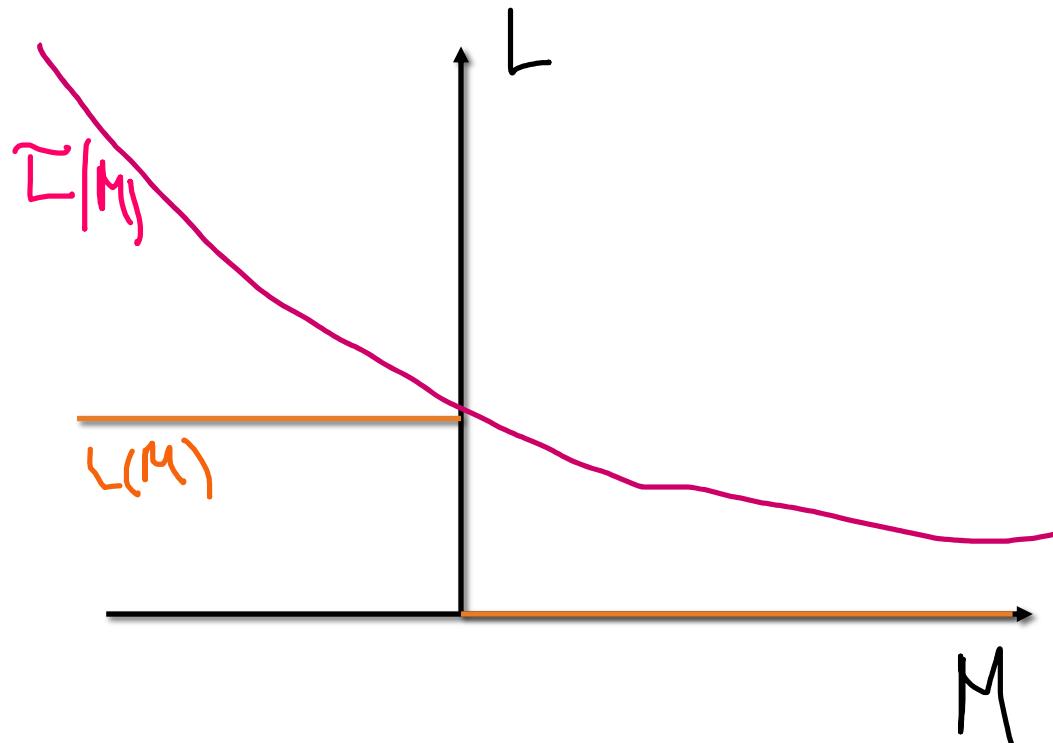
Обучение линейного классификатора

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0]$$

Данный функционал оценивает ошибку алгоритма с помощью функции потерь $L(M) = [M < 0]$

Хотим гладкий функционал – оценим эту функцию сверху:

$$L(M) \leq \tilde{L}(M)$$



Обучение линейного классификатора

- После этого можно получить верхнюю оценку на функционал:

$$Q(a, X) \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(y_i \langle w, x_i \rangle) \rightarrow \min_w$$

Обучение линейного классификатора

- $\tilde{L}(M)$ выберем гладкой и будем оптимизировать

$$0 \leq \frac{1}{l} \sum_{i=1}^l [y_i < w_i x_i > < 0] \leq \frac{1}{l} \sum_{i=1}^l \tilde{L}(y_i < w_i x_i >) \rightarrow \min_w$$

- Теперь верхнюю оценку можно будет минимизировать с помощью, например, градиентного спуска
- Если верхнюю оценку удастся приблизить к нулю, то и доля неправильных ответов тоже будет близка к нулю

Обучение линейного классификатора

Несколько примеров:

1. $\tilde{L}(M) = \log(1 + e^{-M})$ — логистическая функция потерь L
2. $\tilde{L}(M) = (1 - M)_+ = \max(0, 1 - M)$ — кусочно-линейная функция потерь (используется в методе опорных векторов) ✓
3. $\tilde{L}(M) = (-M)_+ = \max(0, -M)$ — кусочно-линейная функция потерь (соответствует персептрону Розенблатта) Н
4. $\tilde{L}(M) = e^{-M}$ — экспоненциальная функция потерь E
5. $\tilde{L}(M) = 2/(1 + e^M)$ — сигмоидная функция потерь S

Обучение линейного классификатора

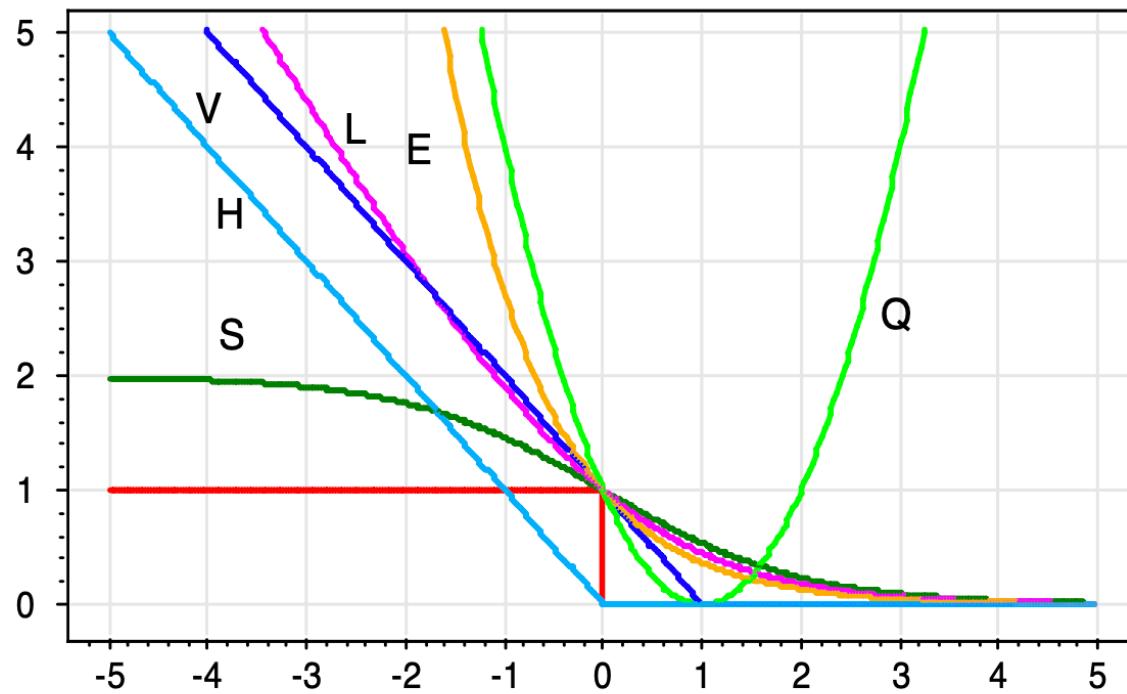


Рис. 1. Верхние оценки на пороговую функцию потерь.

Метрики качества

Метрики качества

- Accuracy – доля правильных ответов

$$\text{accuracy}(a, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i].$$

99.5% - 0
0.5% - 1

$$\hat{y} = 0$$

$$\text{acc} = 99.5\% = 0.995$$

Метрики качества

α_1 и α_2

ϵ_1 и ϵ_2 , $\epsilon_2 > \epsilon_1$

η_1

20%

50%

0.1%

η_2

10%

25%

0.01%

$r_1 - r_2$

10%

25%

0.09%

$\frac{r_1 - r_2}{r}$

5%

5%

9%

Метрики качества

- Матрица ошибок (confusion matrix):

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False negative (FN)	True Negative (TN)

Метрики качества

- Перепишем в контексте матрицы ошибок (confusion matrix):

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Метрики качества

- Точность (precision) и полнота (recall):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}};$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Метрики качества

H_1 : есть беременность; H_0 : нет беременности

Истинный
позитив, верна
 H_1



Ложный
позитив,
ошибка I
рода,
ложная
тревога



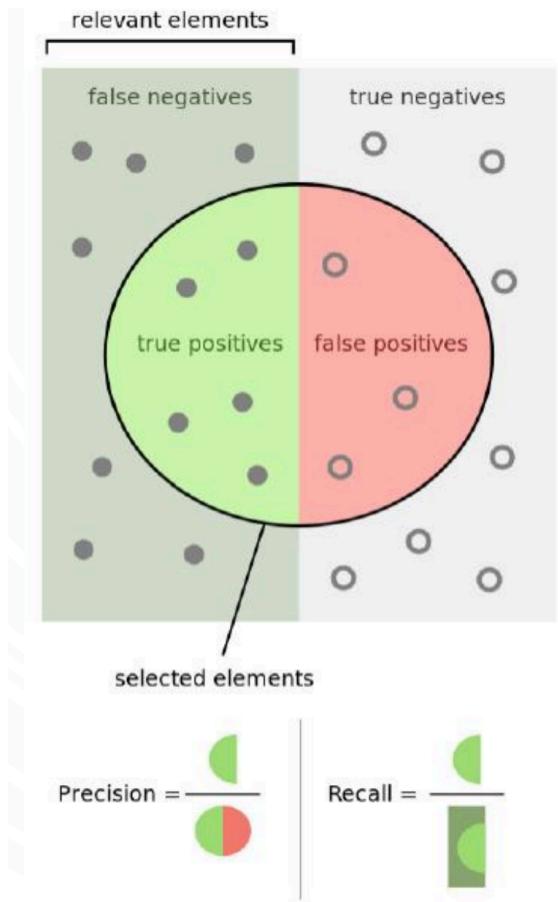
Ложный
негатив,
ошибка II
рода,
халатная
беспечность



Истинный
негатив,
верна H_0



Точность и полнота



Метрики качества

- F-мера:

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Метрики качества

- Для задач, связанных с выбором подмножества (выделение лояльных клиентов банка, например) можно использовать прирост концентрации (lift). Если при рассылке предложений о кредите клиентам из подмножества и всем клиентам будет получаться одна и та же доля откликнувшихся, то подмножество не будет представлять особой ценности.

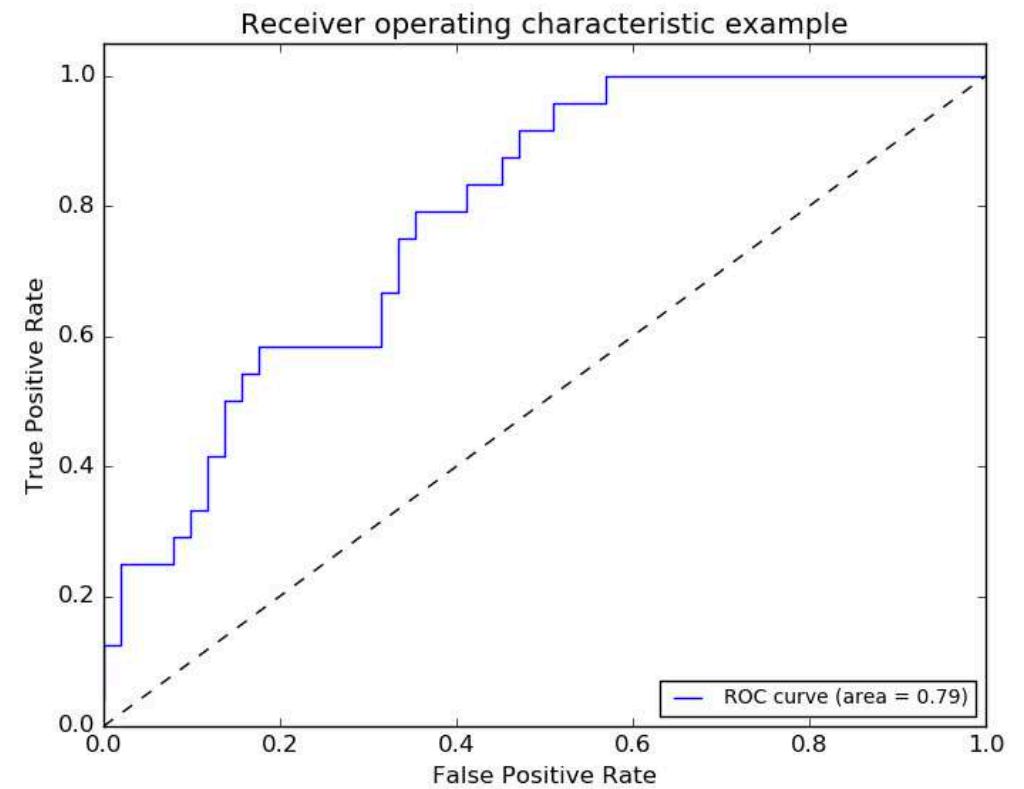
$$\text{lift} = \frac{\text{precision}}{(\text{TP} + \text{FN})/\ell}$$

- Улучшение доли положительных объектов в данном подмножестве относительно доли в случайно выбранном подмножестве такого же размера.

Метрики качества

- ROC-AUC (Area under receiver operating characteristic):

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}};$$
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$



Метрики качества

- Индекс Джини:

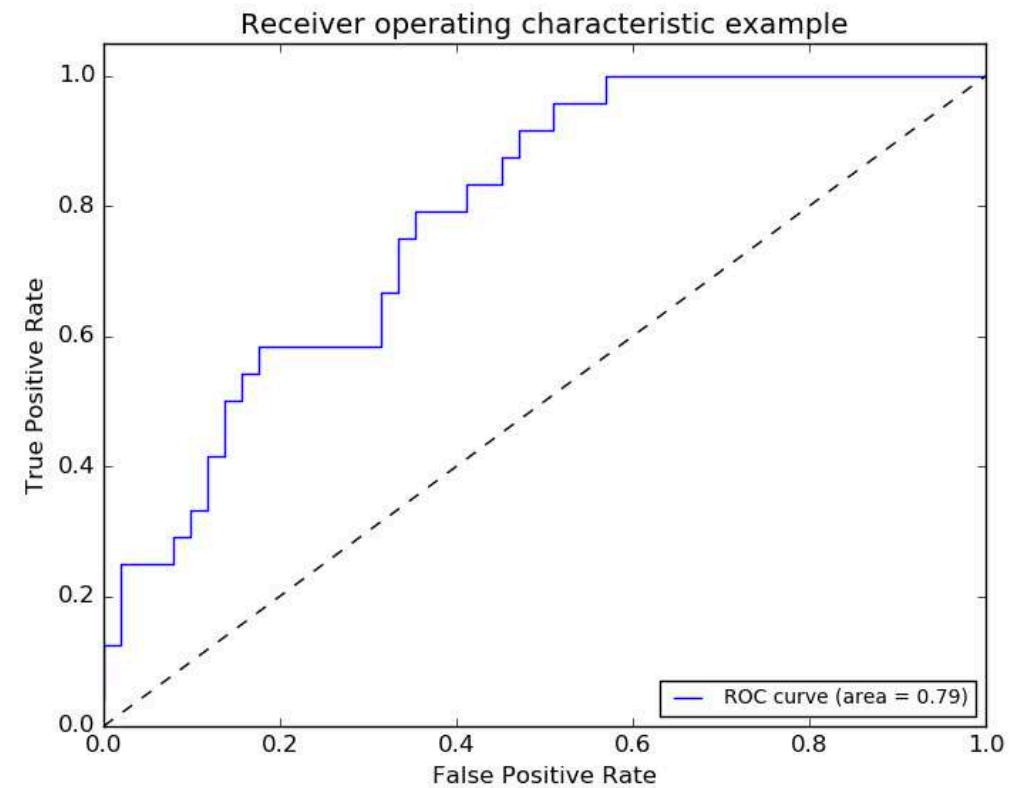
$$FPR = \frac{FP}{FP + TN};$$

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

$$Gini = 2AUC - 1$$

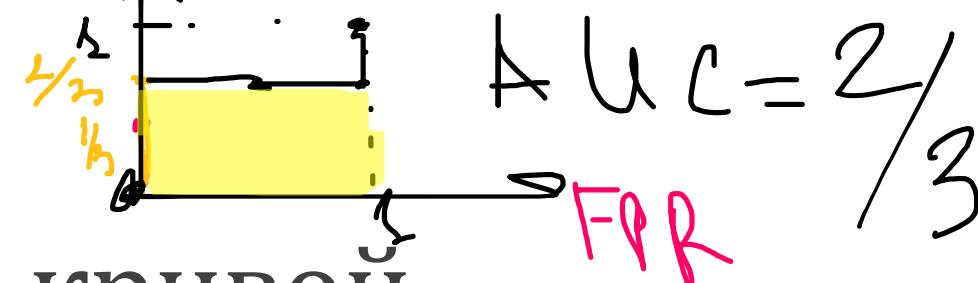
- Это площадь между ROC-кривой и диагональю, соединяющей точки $(0,0)$ и $(1,1)$.

Крупная статья про Gini: [URL](#)



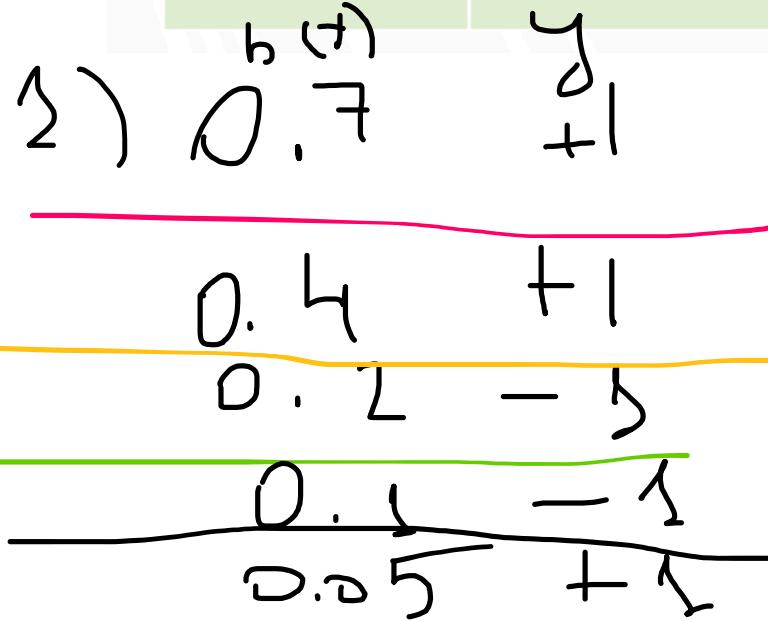
Задачи

2.5 0.05 $\frac{2}{3}$ 1
2.6 0 1 1



Пример построения ROC-кривой

$b(x)$	0.2	0.4	0.1	0.7	0.05
y	-1	+1	-1	+1	+1



2.1 $t = 0.7 \quad \square (x)_z \sqcup b(x) > t$

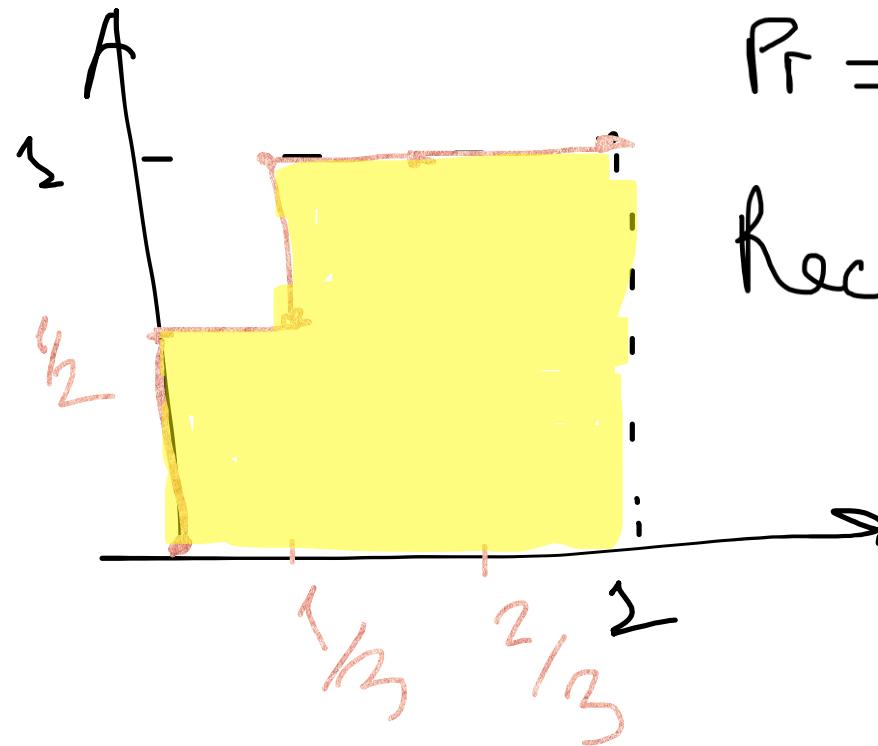
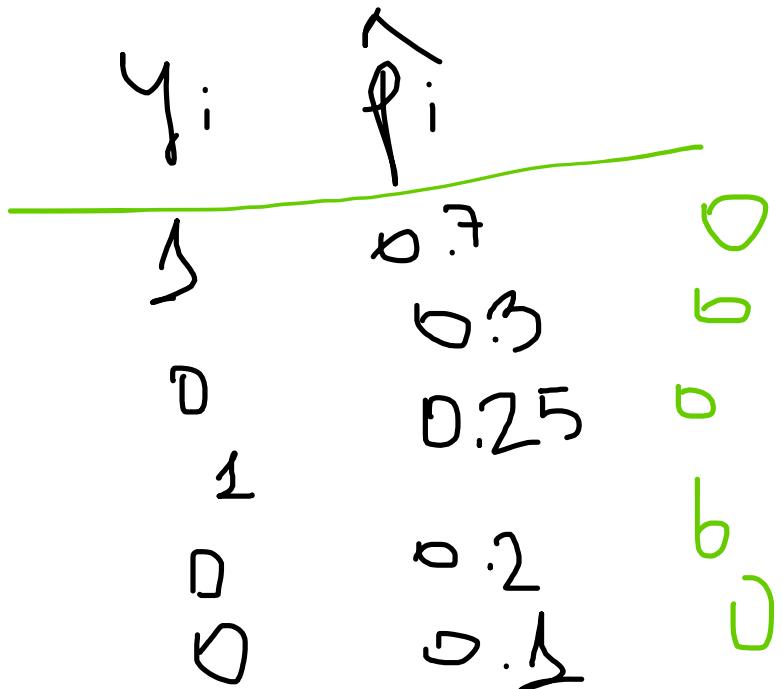
2.2 $t = 0.4 \quad TPR = \frac{0}{0+3} = 0 \quad FPR = 0$

2.3 $t = 0.2 \quad \frac{1}{1+2} = \frac{1}{3} \cdot 0$

2.4 $t = 0.1 \quad \frac{2}{2+1} = \frac{2}{3} \cdot 0$

$\frac{2}{3}; \frac{1}{2}$

Пример построения ROC-кривой



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

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

Пример построения PR-кривой

$\geq t$	f_t	Rec
0.1	$\frac{1}{5}$	1
0.2	$\frac{1}{2}$	1
0.25	$\frac{1}{3}$	1
0.3	0.5	0.5
0.7	1	0.5
0.8	0	0

$$Pr = \frac{1}{TP + FN}$$
$$Rec = \frac{TP}{TP + FN}$$

