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:

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

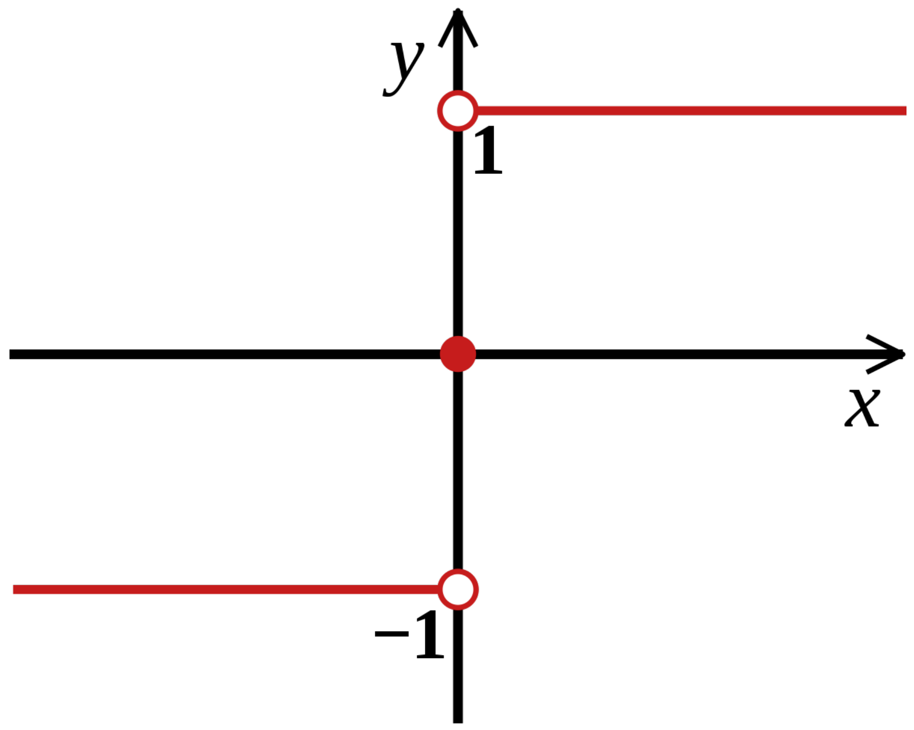
Линейный классификатор

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

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

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

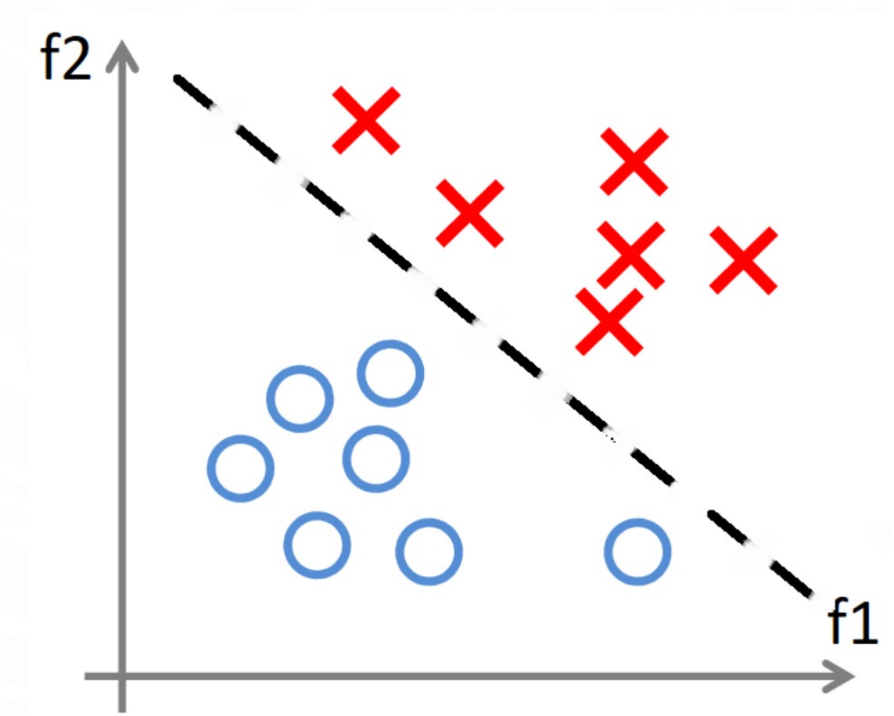
Напоминание



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

Геометрическая интерпретация

Уравнение вида $< w, x > = 0$ 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:

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

Обучение линейного классификатора

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

Обучение линейного классификатора

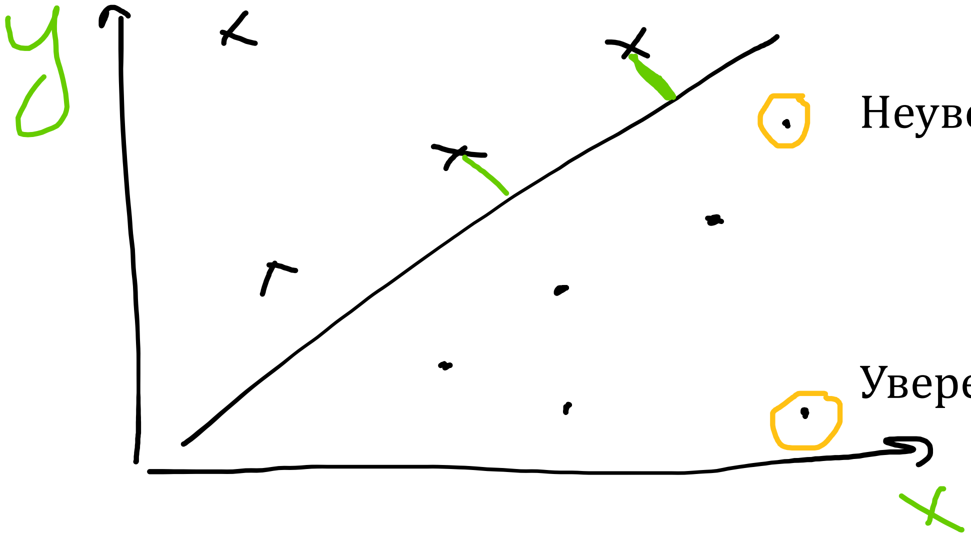
Перепишем наш функционал:

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

- $M = y < w, x >$ – 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

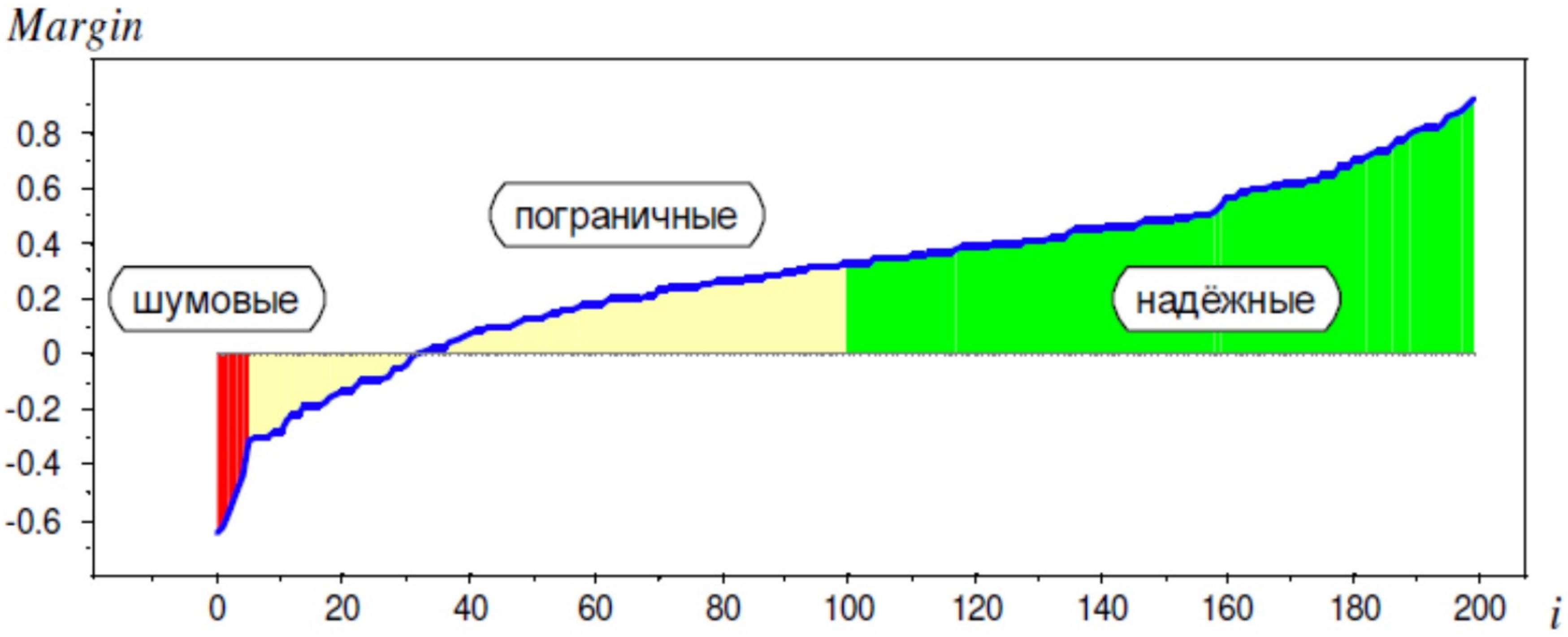


Неуверенная классификация

Уверенная классификация

Обучение линейного классификатора

Ранжирование объектов по возрастанию отступа:



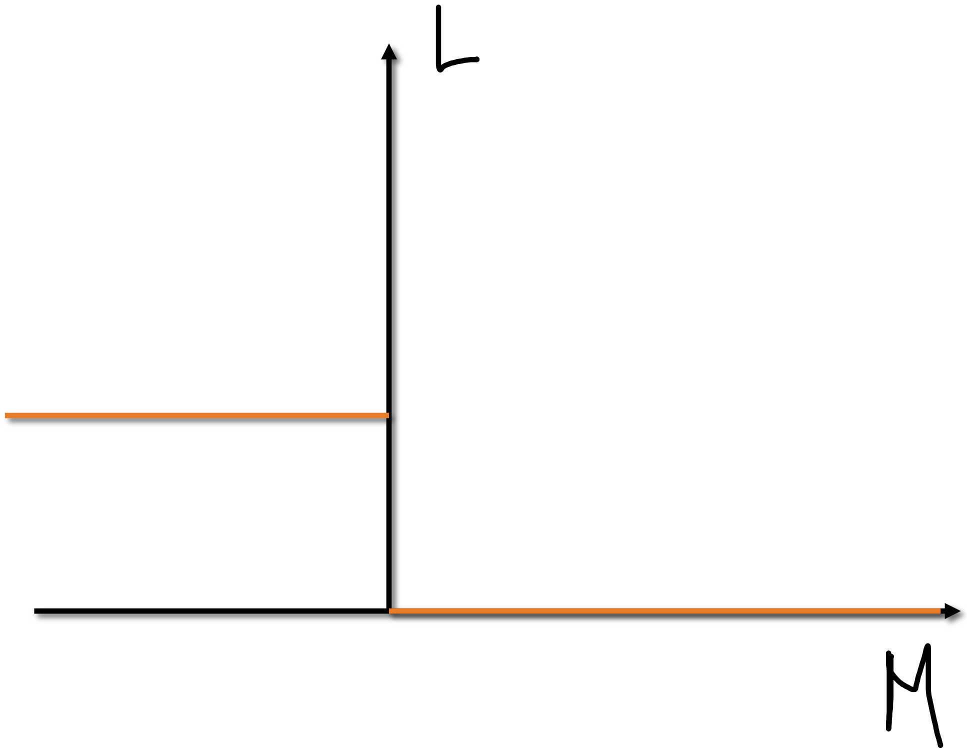
Обучение линейного классификатора

- 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 $L(M) = [M < 0]$



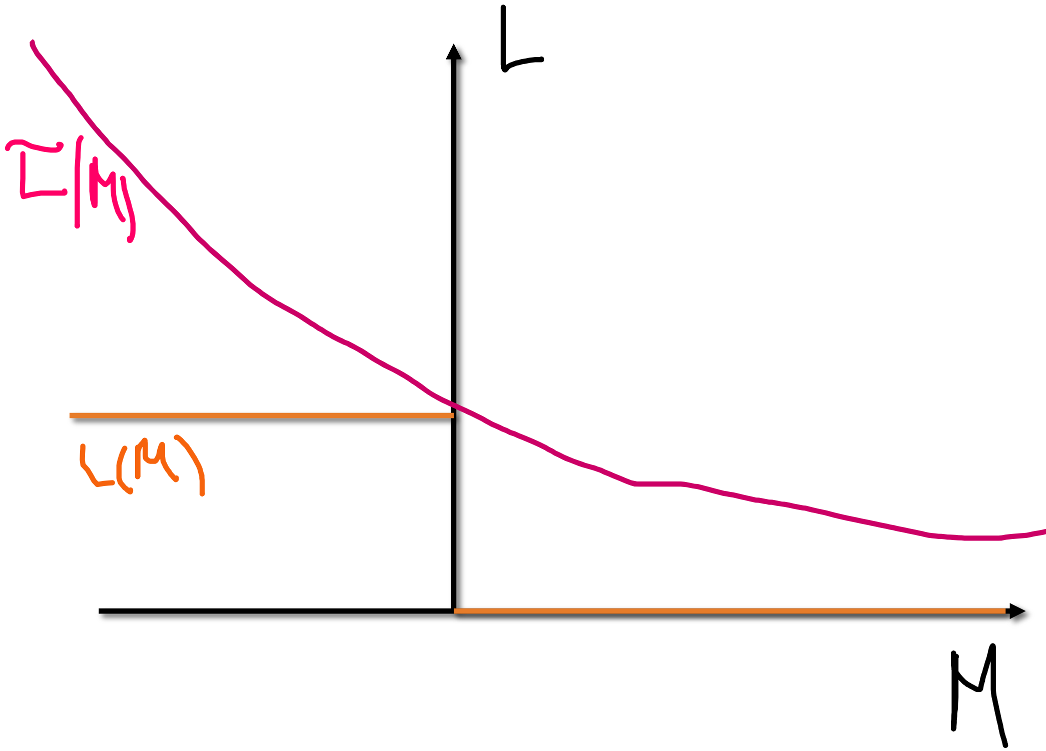
Обучение линейного классификатора

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

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:

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

Обучение линейного классификатора

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

$$0 \le \frac{1}{l}\sum_{i=1}^{l}[y_i < w_i x_i > < 0] \le \frac{1}{l}\sum_{i=1}^{l}\tilde{L}(y_i < w_i x_i >) \to \min_{w}$$

- 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. $\tilde{L}(M) = \log\left(1 + e^{-M}\right)$ — логистическая функция потерь L
2. $\tilde{L}(M) = (1 - M)_+ = \max(0, 1 - M)$ — кусочно-линейная функция потерь (используется в методе опорных векторов) V
3. $\tilde{L}(M) = (-M)_+ = \max(0, -M)$ — кусочно-линейная функция потерь (соответствует персептрону Розенблатта) H
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

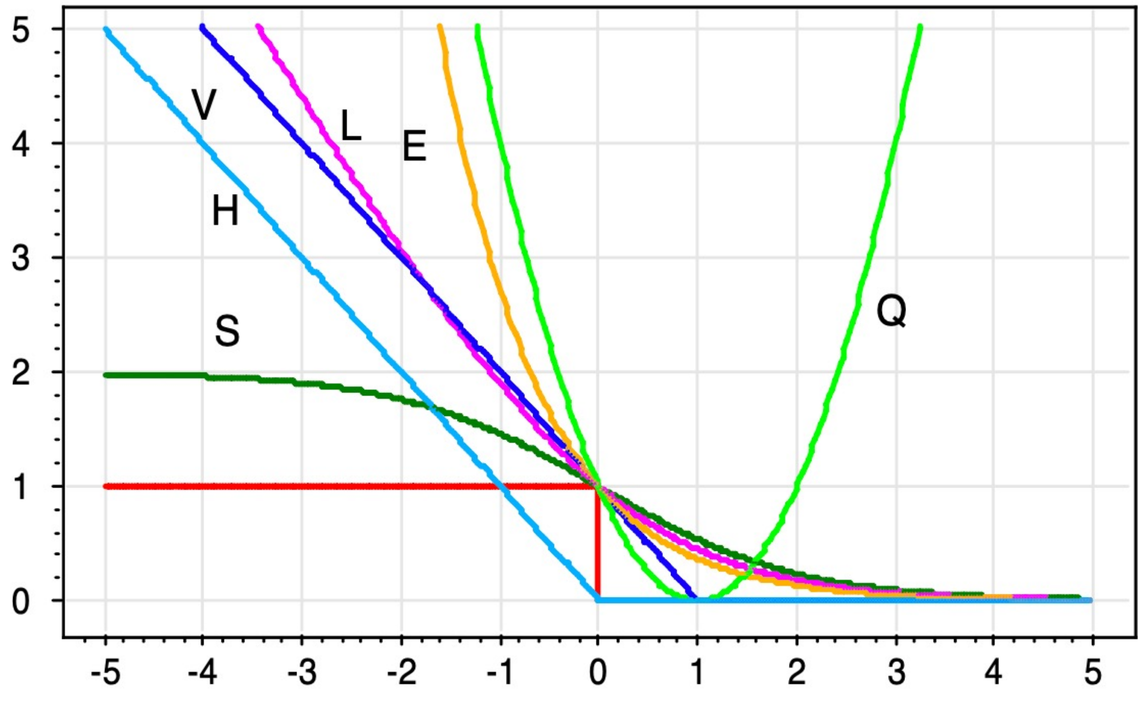


Рис. 1. Верхние оценки на пороговую функцию потерь.

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

$y = 0$

Acc = 99.5% = 0,995

Метрики качества

a_1 и a_2

r_1 и r_2, $r_2 > r_1$

r_1	r_2	$r_1 - r_2$	$\frac{r_1 - r_2}{r_1}$
20%	10%	10%	50
50%	25%	25%	5
0.1%	0.01%	0.09%	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):

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

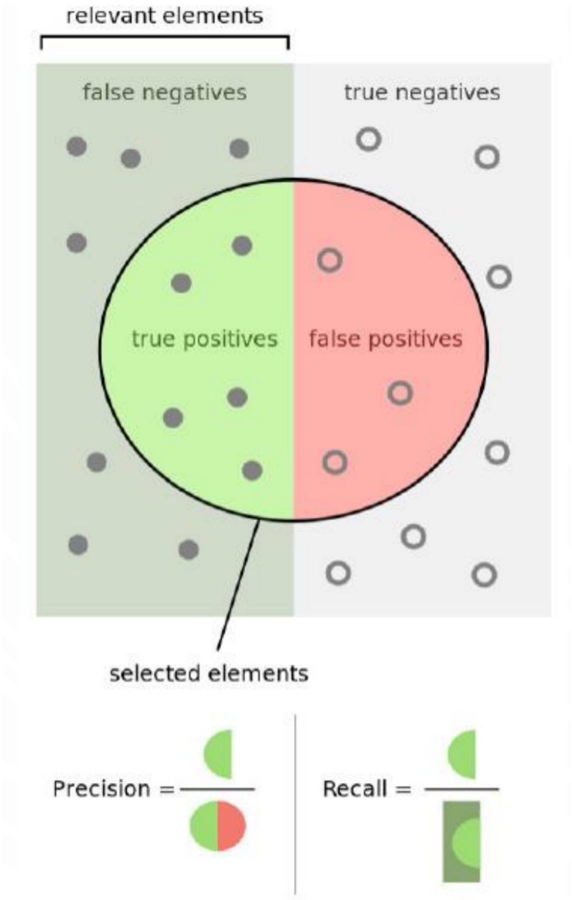
Метрики качества

- 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

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.

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

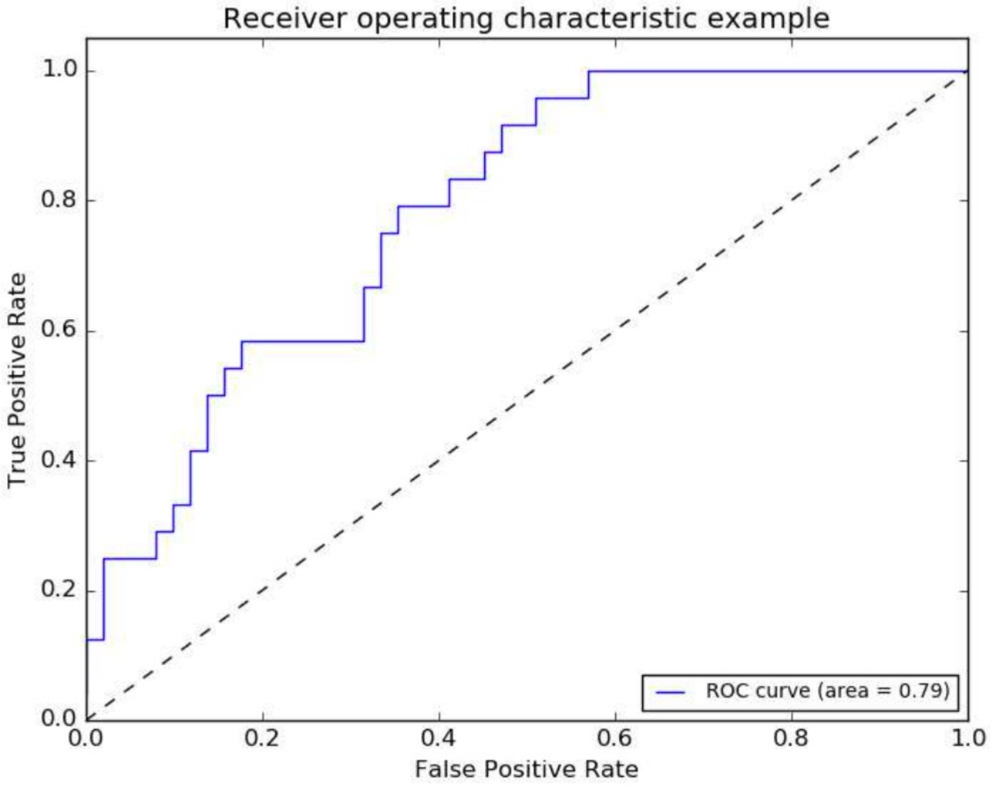
- Улучшение доли положительных объектов в данном подмножестве относительно доли в случайно выбранном подмножестве такого же размера.

Метрики качества

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

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

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



Метрики качества

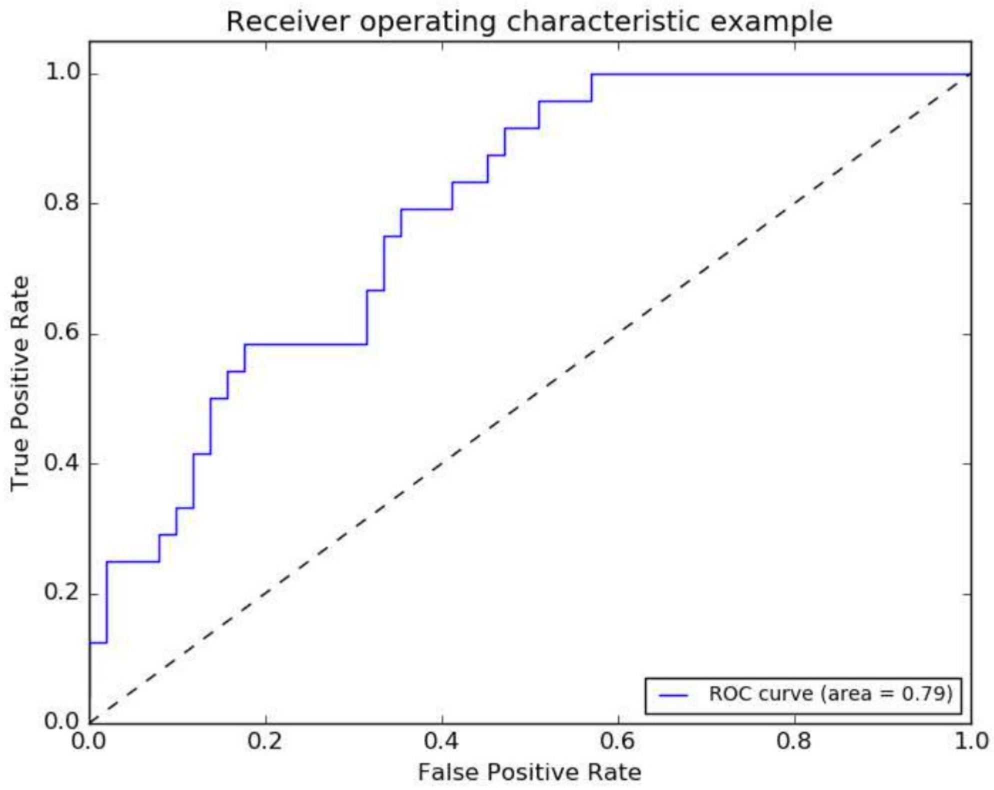
- Индекс Джини:

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

$$Gini = 2AUC - 1$$

- Это площадь между ROC-кривой и

диагональю, соединяющей точки (0,0) и (1, 1).

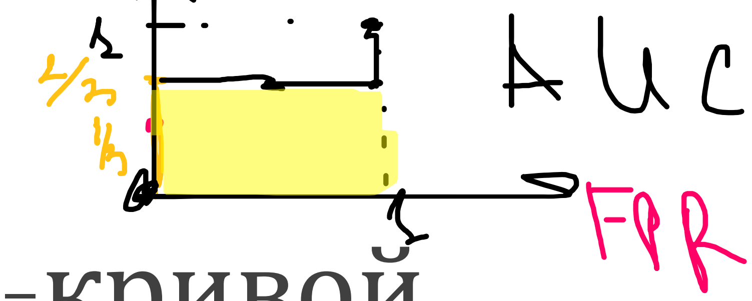


Крутая статься про Gini: [URL]

Задачи

2.5   0.05   2/3   1

2.6   0   1   1



$AUC = 2/3$

Пример построения ROC-кривой

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

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

2.1 $t = 0.7$ $a(x) = [b(x) > t]$

$TPR = \frac{0}{0+3} = 0$ $FPR = 0$

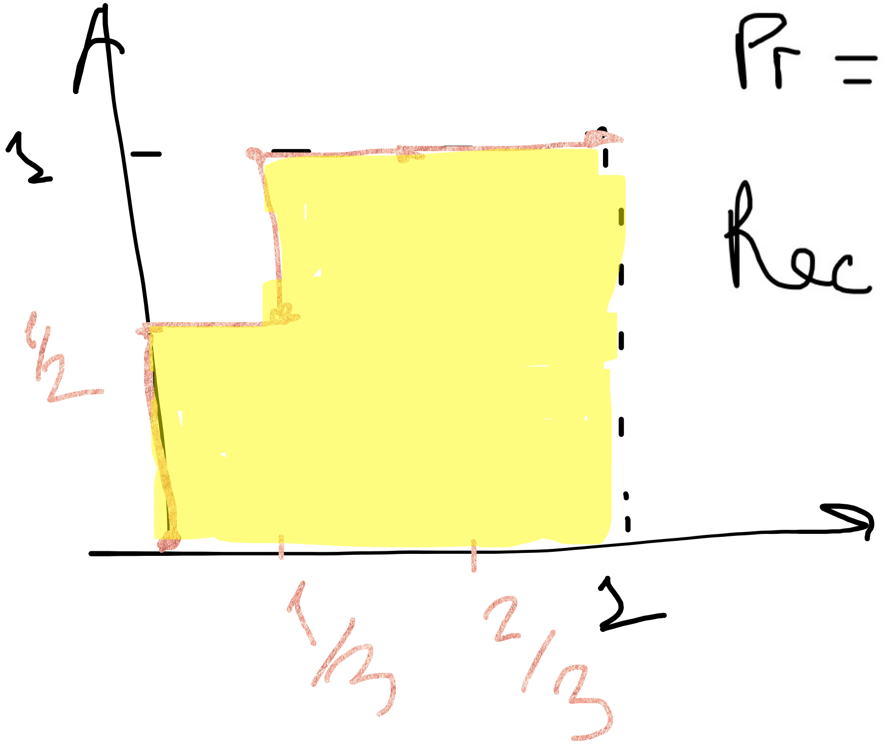
2.2 $t = 0.4$; $\frac{1}{1+2} = \frac{1}{3}$; 0

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

2.4 $t = 0.1$ 2/3 ; 1/2

Пример построения ROC-кривой

y_i	$\hat{p}_i$	
1	0.7	0
	0.3	0
0	0.25	0
1		b
0	0.2	
0	0.1	0



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

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

Пример построения PR-кривой

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

$$Pr = \frac{1}{TP + FP}$$

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