Линейная классификация - 2

Елена Кантонистова

ПЛАН ЛЕКЦИИ

1) Метрики качества классификации
2) Логистическая регрессия (вероятностное обоснование)

ВСПОМИНАЕМ БАЗОВЫЕ МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

- Какие метрики помните?
- Какие у них есть особенности?

ИНТЕГРАЛЬНЫЕ МЕТРИКИ КЛАССИФИКАЦИИ

ИНТЕГРАЛЬНАЯ МЕТРИКА: ROC-AUC

Хотим измерить качество всего семейства классификаторов независимо от выбранного порога.

Для этого будем использовать метрику AUC

AUC *– Area Under ROC Curve (площадь под ROC-кривой)*

ROC-AUC: ИНТУИЦИЯ

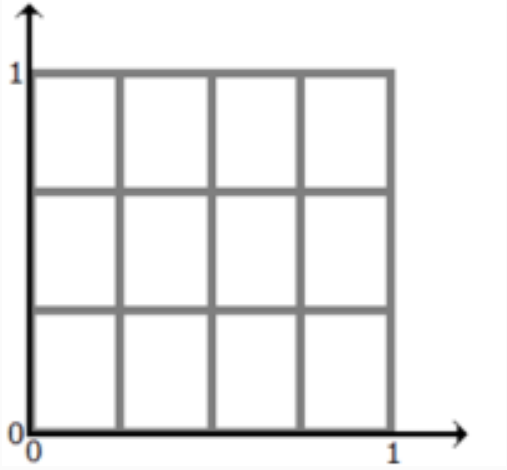
- Пример:

p	класс
0.5	0
0.1	0
0.25	0
0.6	1
0.2	1
0.3	1
0.0	0

p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0

ROC-AUC: АЛГОРИТМ

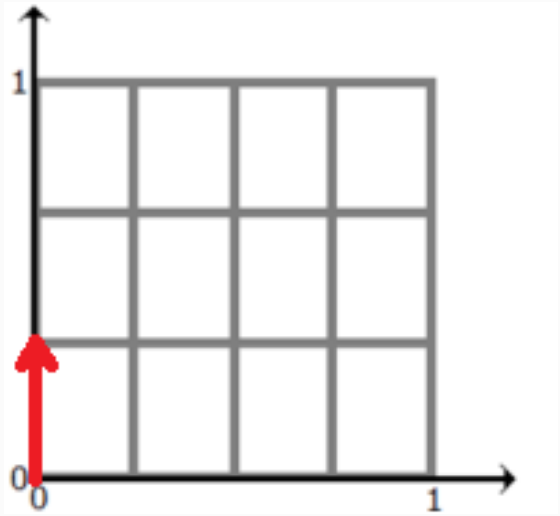
- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1



ROC-AUC: АЛГОРИТМ

- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1

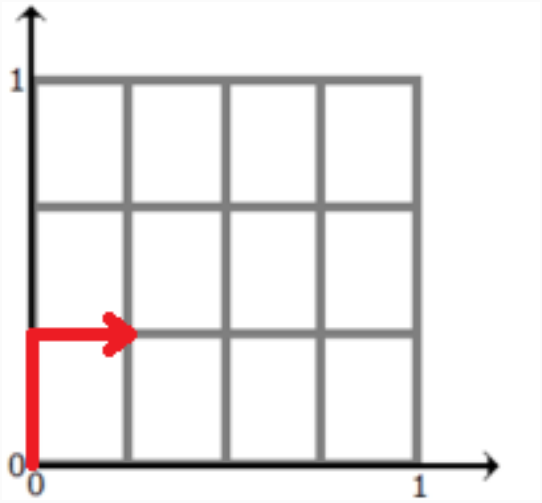
p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



ROC-AUC: АЛГОРИТМ

- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1

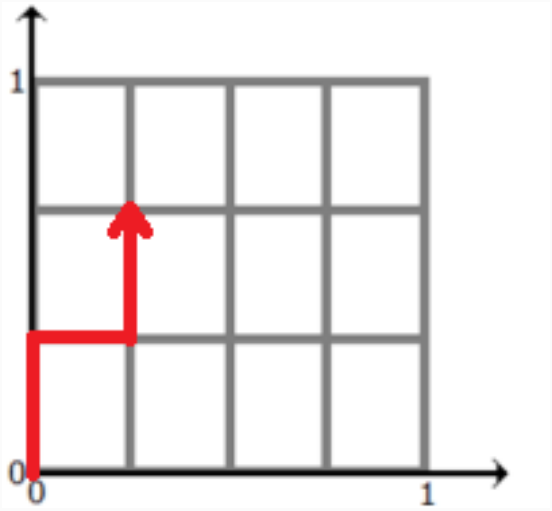
p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



ROC-AUC: АЛГОРИТМ

- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1

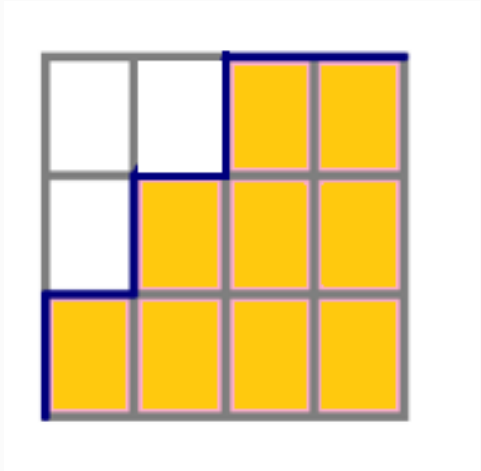
p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



ROC-AUC: АЛГОРИТМ

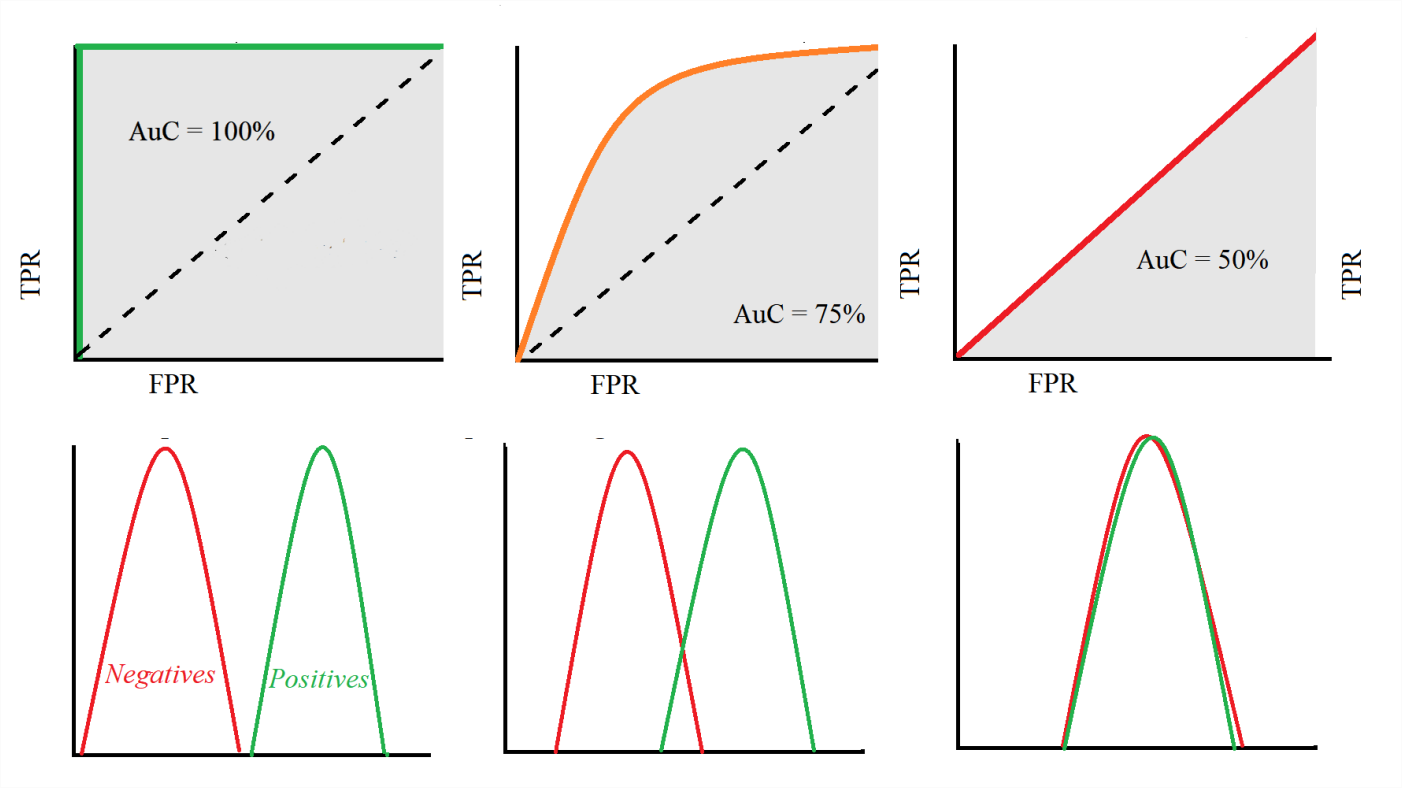
- Пойдем по отсортированной таблице по столбцу класс сверху вниз
- Будем стартовать из точки (0,0) на квадрате. И если мы встречаем 1, сдвигаемся на одну клеточку вверх, а если 0 - то вправо
- В итоге мы придём в точку (1,1).

p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



Полученная кривая называется ROC-кривой, а метрика, равная площади под ней - AUC-ROC.

ROC-AUC: ПРИМЕРЫ



ROC-КРИВАЯ (ФОРМАЛЬНО)

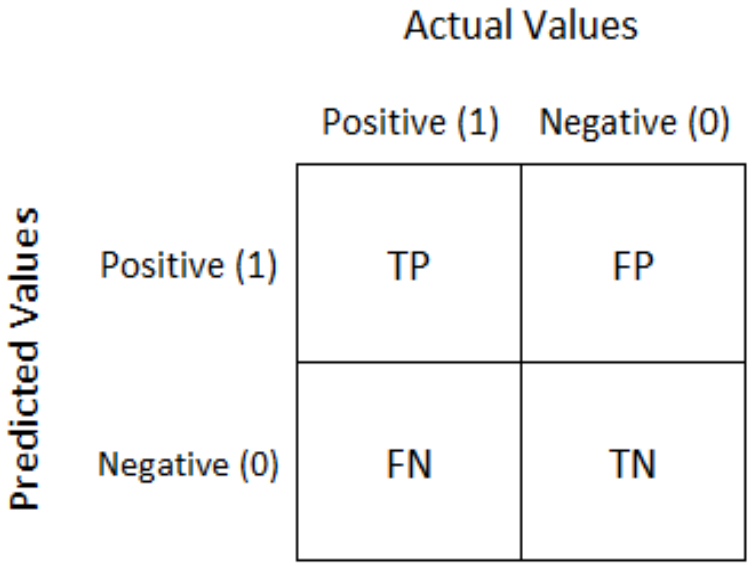
Для каждого значения порога t вычислим:

- False Positive Rate (доля неверно принятых объектов отрицательного класса):

$$\boldsymbol{FPR = \frac{FP}{FP + TN} = \frac{\sum_i [y_i = -1][a(x_i) = +1]}{\sum_i [y_i = -1]}}$$

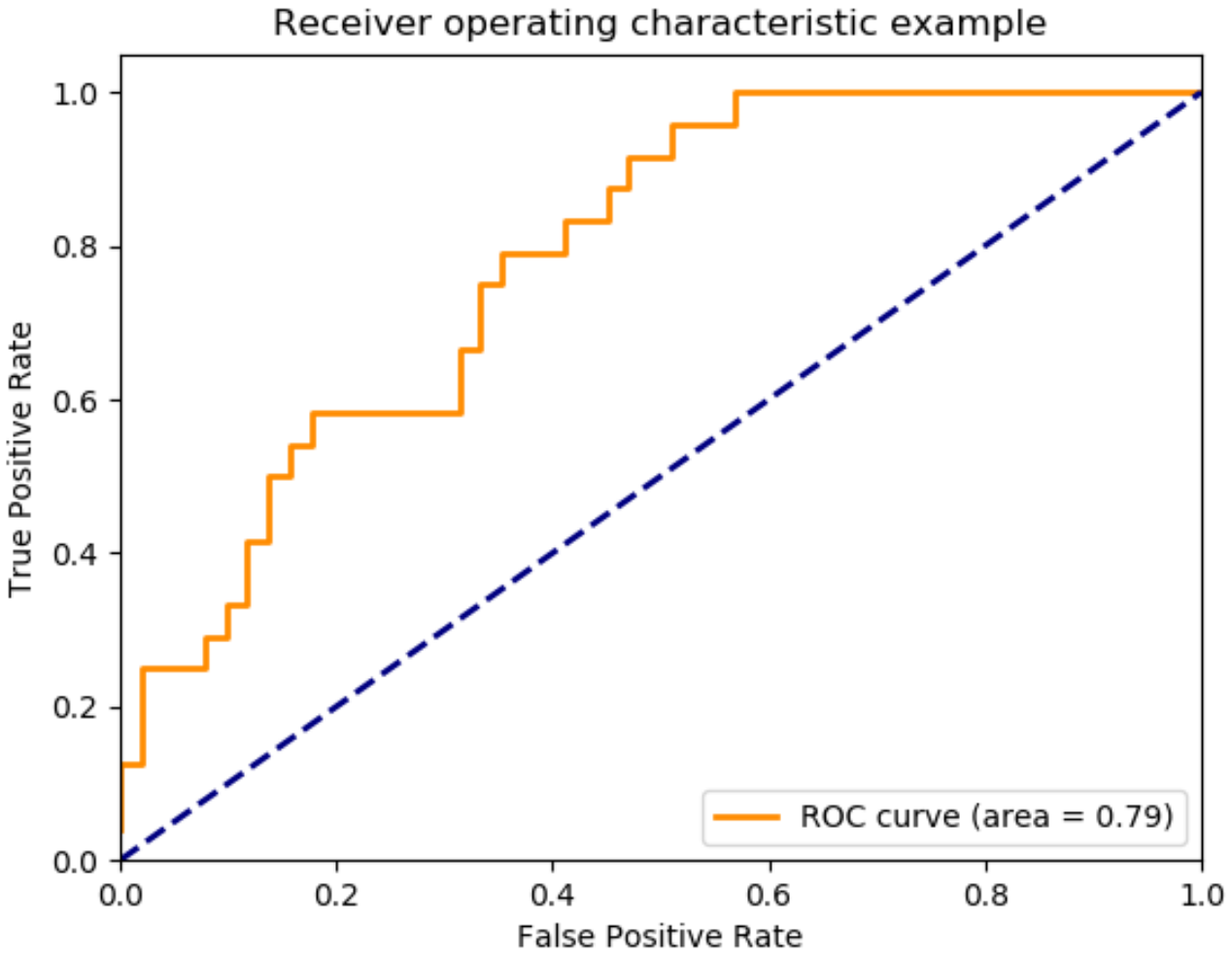
- True Positive Rate (доля верно принятых объектов положительного класса):

$$\boldsymbol{TPR = \frac{TP}{TP+FN} = \frac{\sum_i [y_i=+1][a(x_i)=+1]}{\sum_i [y_i=+1]}}.$$



ROC-КРИВАЯ

Кривая, состоящая из точек с координатами (FPR,TPR) для всех возможных порогов – это и есть ROC-кривая.

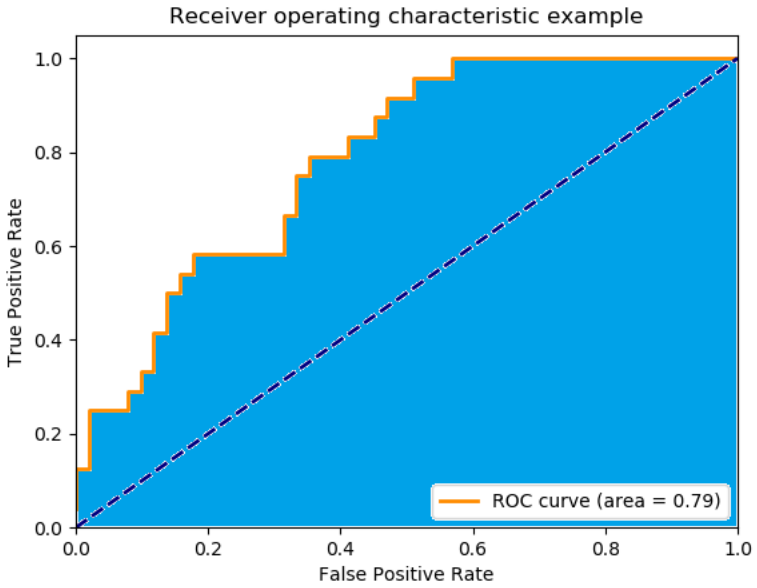


ROC-КРИВАЯ. AUC.

AUC (Area Under Curve) – площадь под ROC-кривой.

$$AUC \in [0; 1].$$

- *Чему равен AUC при идеальной классификации?*
- *Чему равен AUC при случайной классификации?*



ROC-КРИВАЯ. AUC.

AUC (Area Under Curve) – площадь под ROC-кривой.

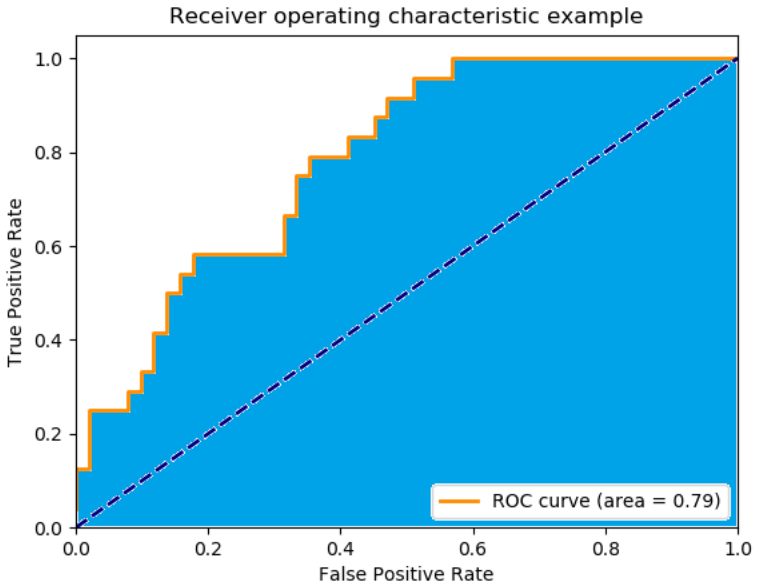
$$AUC \in [0; 1].$$

- $AUC = 1$ –

идеальная классификация

- $AUC = 0.5$ –

случайная классификация



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Пусть есть выборка из 5 объектов и следующие предсказания классификатора оценки принадлежности к классу +1:

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

ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Пусть есть выборка из 5 объектов и следующие предсказания классификатора оценки принадлежности к классу +1:

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

- Упорядочим объекты по убыванию предсказаний:

$(0.7, 0.4, 0.2, 0.1, 0.05)$

1 шаг: $t = 0.7$, то есть

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

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

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

ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Пусть есть выборка из 5 объектов и следующие предсказания классификатора оценки принадлежности к классу +1:

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

- Упорядочим объекты по убыванию предсказаний:

$(0.7, 0.4, 0.2, 0.1, 0.05)$

1 шаг: $t = 0.7$, то есть

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

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

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

$$TPR = \frac{0}{0+3} = 0, \quad FPR = \frac{0}{0+2} = 0.$$

ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Пусть есть выборка из 5 объектов и следующие предсказания классификатора оценки принадлежности к классу +1:

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

- Упорядочим объекты по убыванию предсказаний:

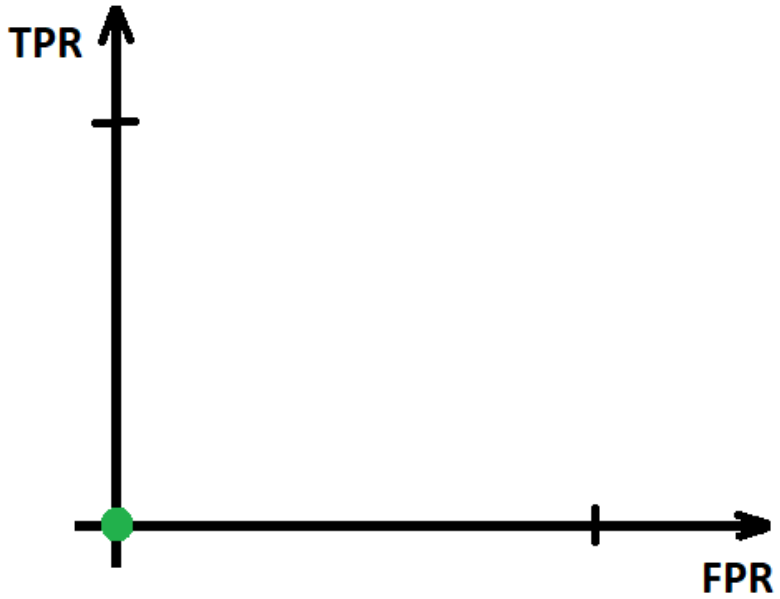
$(0.7, 0.4, 0.2, 0.1, 0.05)$

1 шаг: $\boldsymbol{t = 0.7}$, то есть

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

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

$$FPR = \frac{0}{0+2} = 0.$$



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Оценки принадлежности к классу +1:

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

- Упорядочим объекты по

убыванию предсказаний:

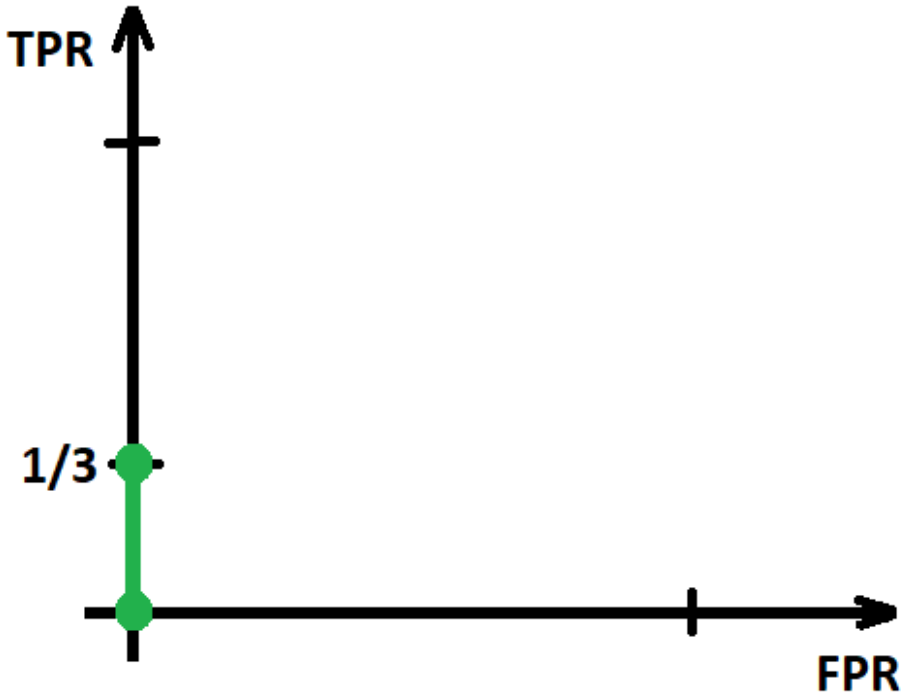
$(0.7, 0.4, 0.2, 0.1, 0.05)$

2 шаг: $t = 0.4$, то есть

$a(x) = [b(x) > 0.4]$

$$TPR = \frac{1}{1+2} = \frac{1}{3},$$

$$FPR = \frac{0}{0+2} = 0.$$



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Оценки принадлежности к классу +1:

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

- Упорядочим объекты по

убыванию предсказаний:

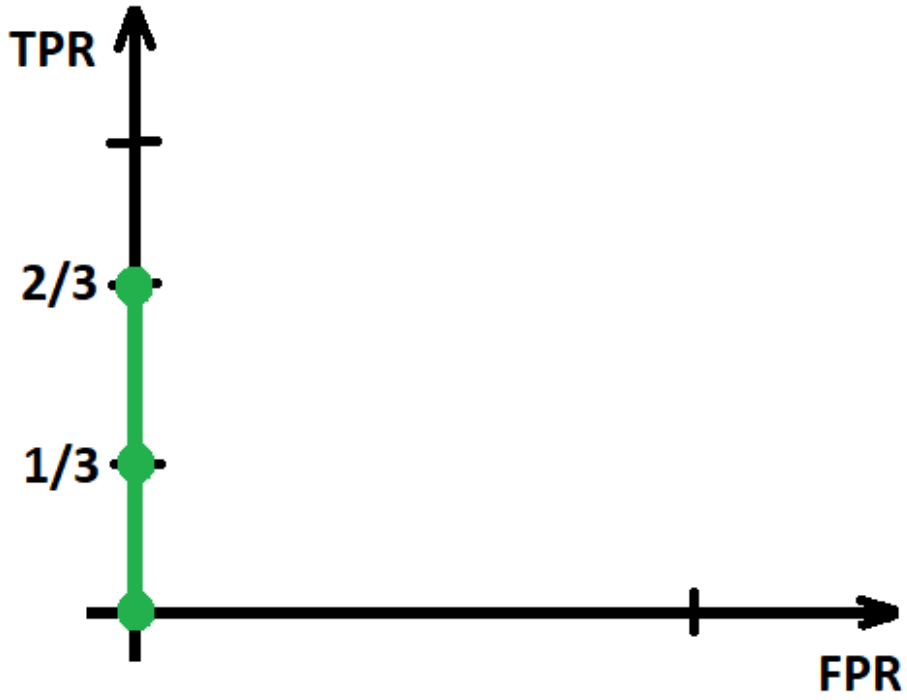
$(0.7, 0.4, 0.2, 0.1, 0.05)$

3 шаг: $t = 0.2$, то есть

$a(x) = [b(x) > 0.2]$

$TPR = \frac{2}{2+1} = \frac{2}{3}$,

$FPR = \frac{0}{0+2} = 0.$



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Оценки принадлежности к классу +1:

$b(x)$	$\mathbf{0.2}$	$\mathbf{0.4}$	0.1	$\mathbf{0.7}$	0.05
y	$\mathbf{-1}$	$\mathbf{+1}$	-1	$\mathbf{+1}$	$+1$

- Упорядочим объекты по

убыванию предсказаний:

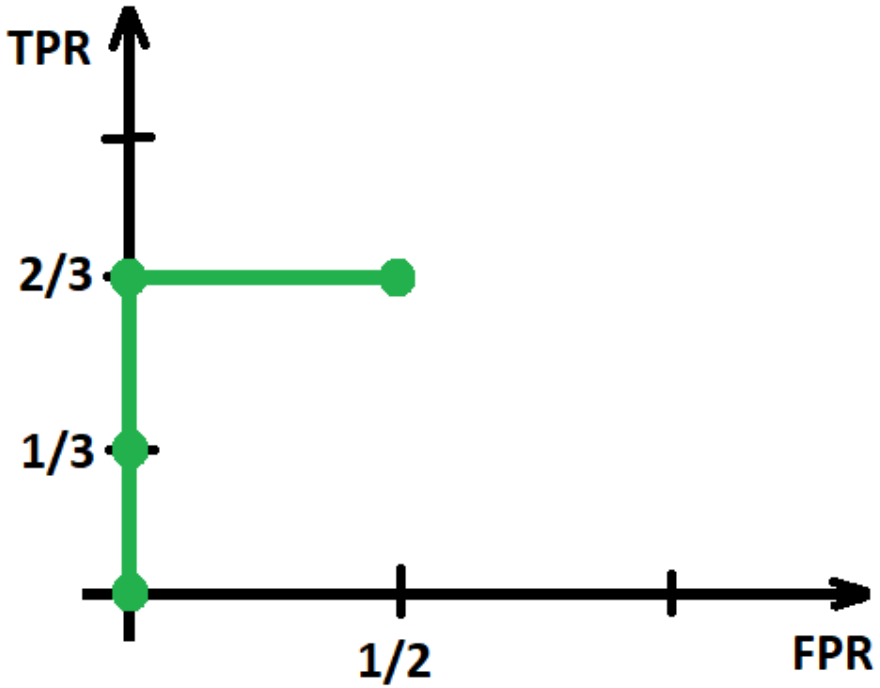
$(0.7, 0.4, 0.2, 0.1, 0.05)$

4 шаг: $\boldsymbol{t = 0.1}$, то есть

$a(x) = [b(x) > 0.1]$

$$TPR = \frac{2}{2+1} = \frac{2}{3},$$

$$FPR = \frac{1}{1+1} = \frac{1}{2}.$$



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Оценки принадлежности к классу +1:

$b(x)$	$\mathbf{0.2}$	$\mathbf{0.4}$	$\mathbf{0.1}$	$\mathbf{0.7}$	0.05
y	$\mathbf{-1}$	$\mathbf{+1}$	$\mathbf{-1}$	$\mathbf{+1}$	+1

- Упорядочим объекты по

убыванию предсказаний:

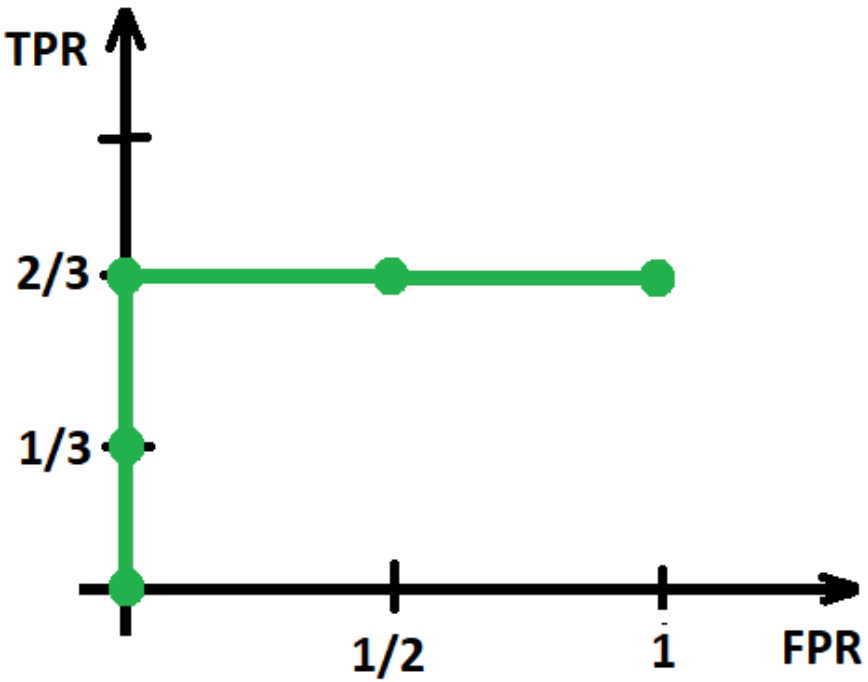
$(0.7,0.4,0.2,0.1,0.05)$

5 шаг: $\boldsymbol{t = 0.05}$, то есть

$a(x) = [b(x) > 0.05]$

$TPR = \frac{2}{2+1} = \frac{2}{3}$,

$FPR = \frac{2}{2+0} = 1.$



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

- Оценки принадлежности к классу +1:

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

- Упорядочим объекты по

убыванию предсказаний:

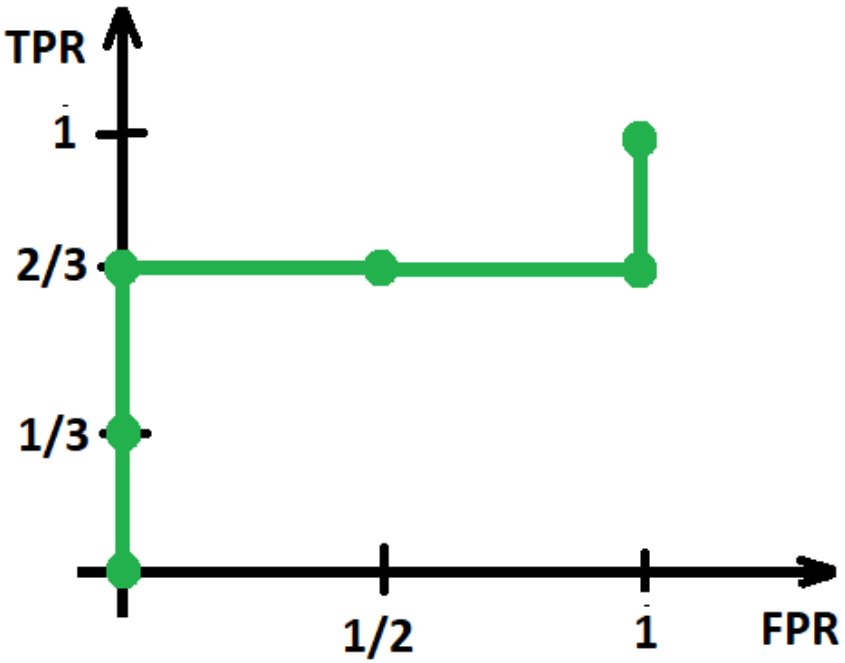
$(0.7,0.4,0.2,0.1,0.05)$

5 шаг: $t = 0$, то есть

$a(x) = [b(x) > 0]$

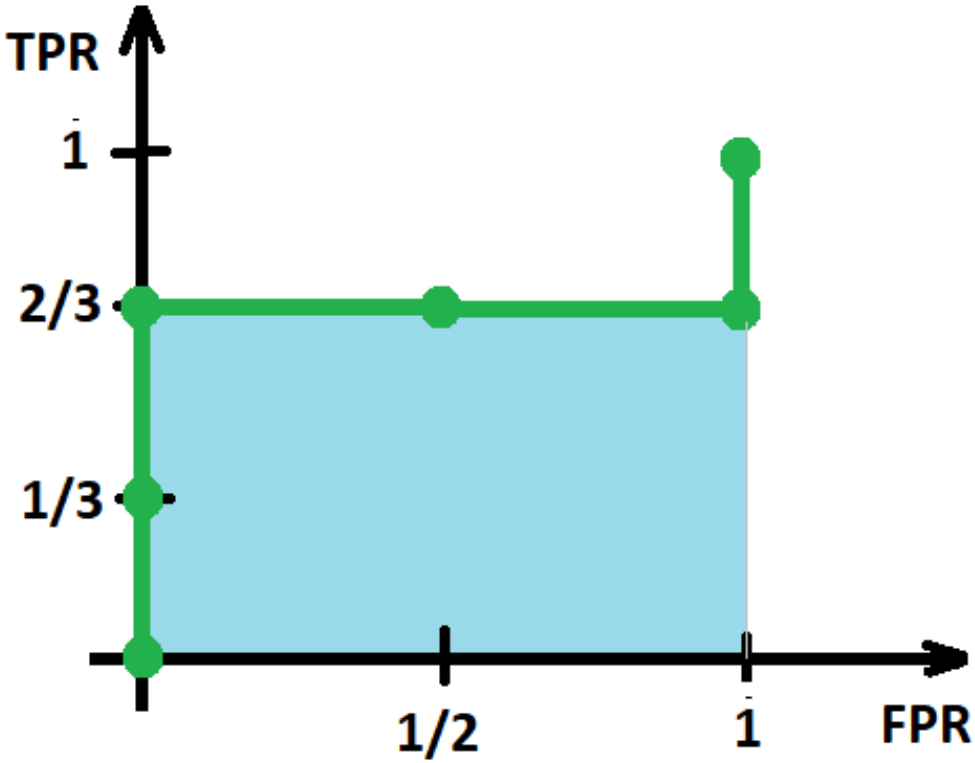
$$TPR = \frac{3}{3+0} = 1,$$

$$FPR = \frac{2}{2+0} = 1.$$



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

$$AUC = 2/3$$

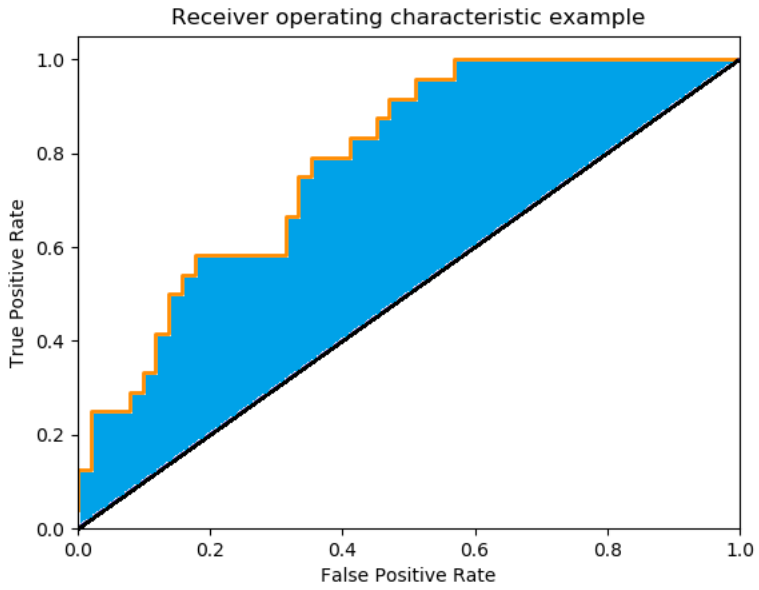


ИНДЕКС ДЖИНИ

Индекс Джини:

$$Gini = 2 \cdot AUC - 1$$

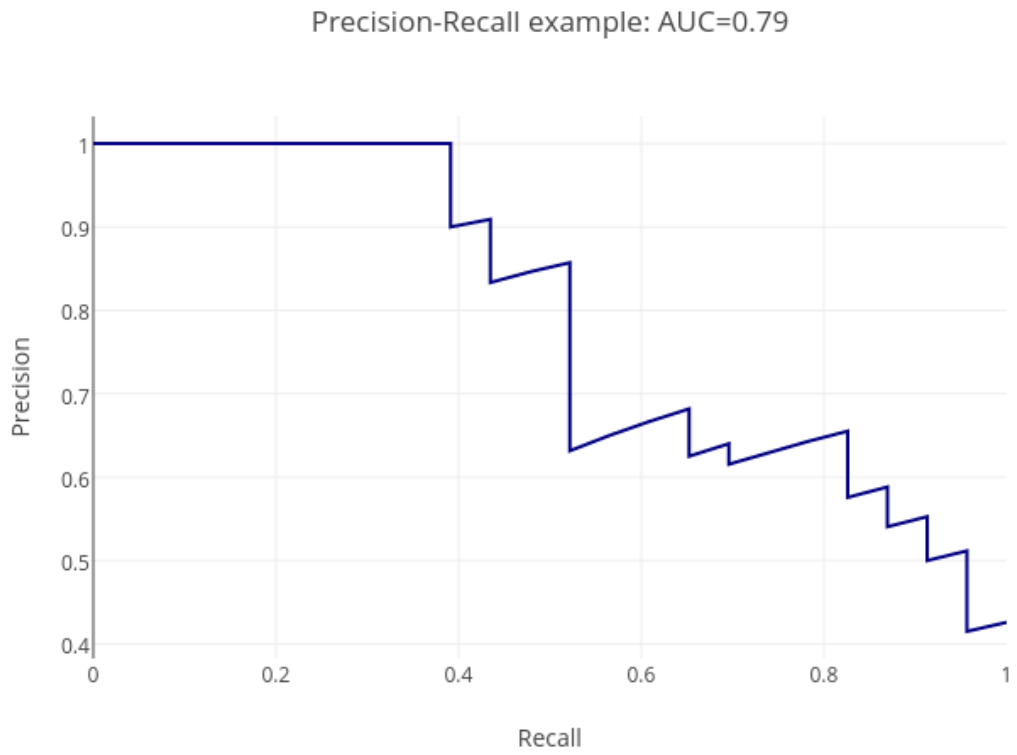
- Индекс Джини – это удвоенная площадь между главной диагональю и ROC-кривой.



PRECISION-RECALL КРИВАЯ

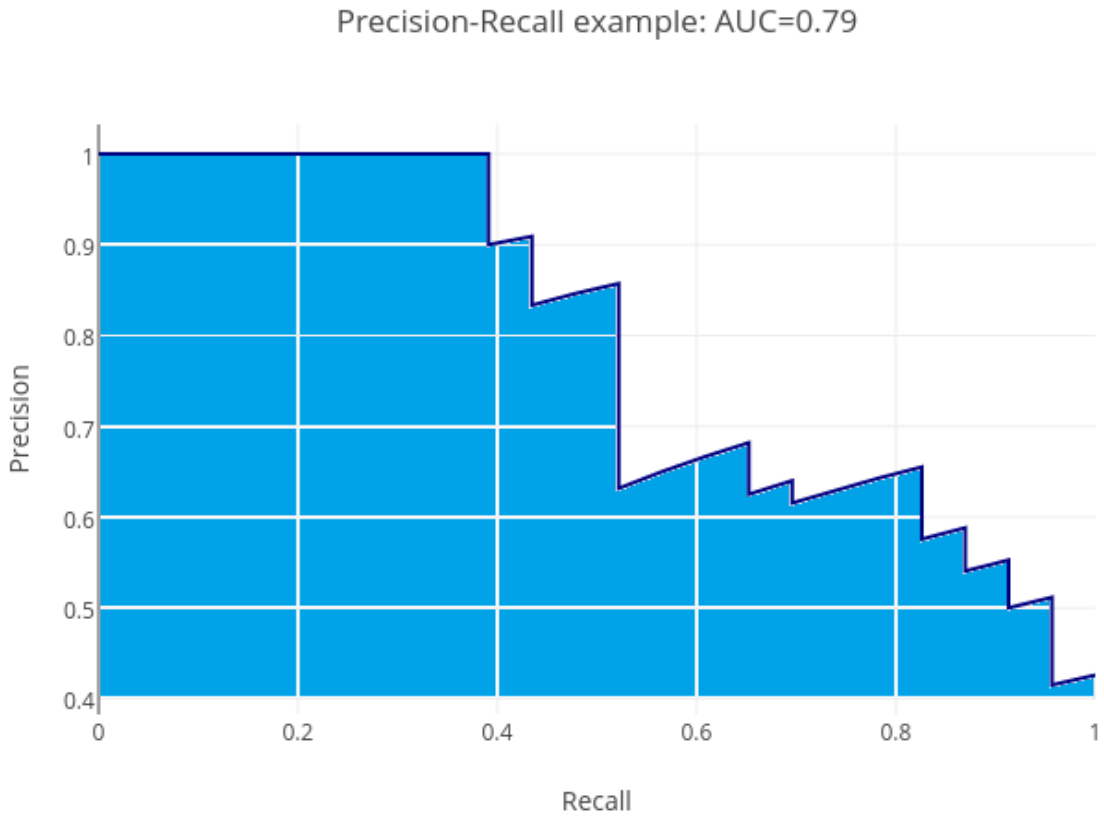
- В случае малой доли объектов положительного класса AUC-ROC может давать неадекватно хороший результат

Precision-Recall кривая:



AUC-PR

AUC-PR – площадь под PR-кривой



ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

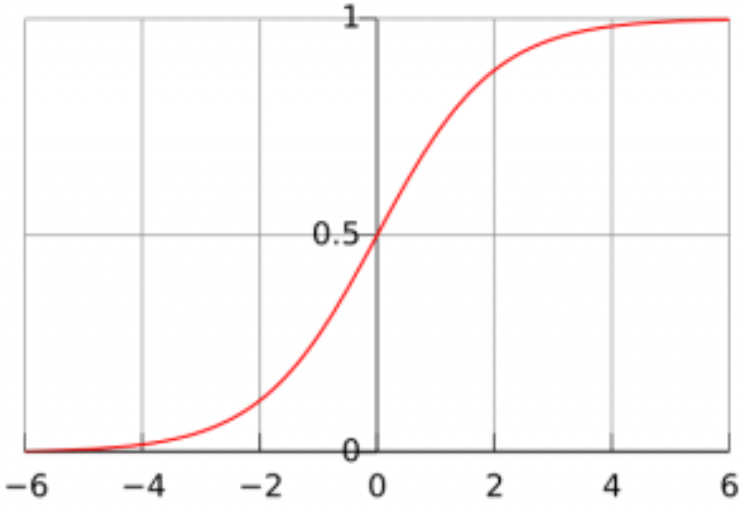
ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не классы, а вероятности классов.

- Линейная регрессия: $a(x, w) = (x, w) = w^T x \in \mathbb{R}$
- Логистическая регрессия: $\boldsymbol{a(x, w) = \sigma(w^T x)}$,

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоида (логистическая функция),

$\sigma(z) \in (0; 1)$.



Логистическая регрессия: $\boldsymbol{a(x, w) = \frac{1}{1+e^{-w^T x}}}$

ВЕРОЯТНОСТНЫЙ СМЫСЛ

Утверждение. $a(x, w)$ – вероятность того, что $y = +1$ на объекте x, т.е.

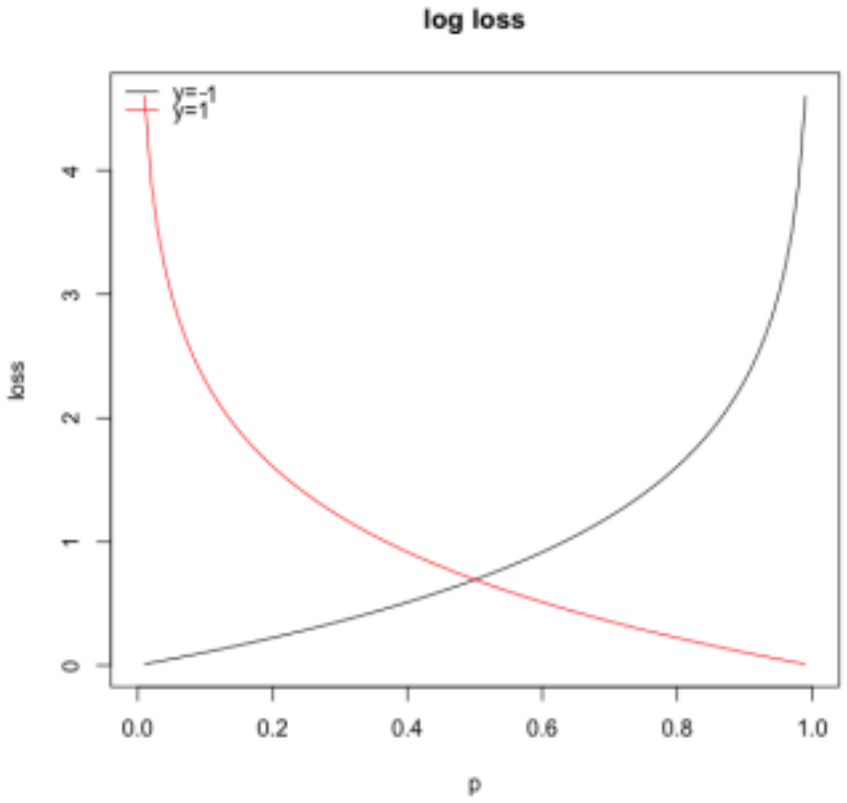
$$\boldsymbol{a(x, w) = P(y = +1 | x; w)}$$

Доказательство. Дальше в лекции.

ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Возьмем логистическую функцию потерь (***log-loss***):

$$Q(w) = -\sum_{i=1}^{l}([y_i = +1] \cdot \log(a(x_i, w)) + [y_i = -1] \cdot \log(1 - a(x_i, w)))$$



ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ: ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

Предположение: В каждой точке x пространства объектов задана вероятность $p(y = +1|x)$

Объекты с одинаковым признаковым описанием могут иметь разные значения целевой переменной.

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

Предположение: В каждой точке x пространства объектов задана вероятность $p(y = +1|x)$

Объекты с одинаковым признаковым описанием могут иметь разные значения целевой переменной.

Цель: построить алгоритм $b(x)$, в каждой точке x предсказывающий $p(y = +1|x)$.

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

Предположение: В каждой точке x пространства объектов задана вероятность $p(y = +1|x)$

Объекты с одинаковым признаковым описанием могут иметь разные значения целевой переменной.

Цель: построить алгоритм $b(x)$, в каждой точке x предсказывающий $p(y = +1|x)$.

Комментарий: *пока что мы будем решать задачу в общем виде, то есть у нас нет ограничений на вид алгоритма $b(x)$ и на вид функции потерь $L(y, b)$.*

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

- Пусть объект x встречается в выборке n раз с ответами $\{y_1, \dots, y_n\}$. Хотим, чтобы алгоритм выдавал вероятность положительного класса:

$$b_*(x) = \operatorname*{argmin}_{b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, b) \approx p(y = +1|x)$$

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

- Пусть объект x встречается в выборке n раз с ответами $\{y_1, \dots, y_n\}$. Хотим, чтобы алгоритм выдавал вероятность положительного класса:

$$b_*(x) = \operatorname*{argmin}_{b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, b) \approx p(y = +1|x)$$

По закону больших чисел при $n \to \infty$ получаем

$$b_*(x) = \operatorname*{argmin}_{b \in \mathbb{R}} E[L(y, b)|x]$$

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

- Пусть объект x встречается в выборке n раз с ответами $\{y_1, \dots, y_n\}$. Хотим, чтобы алгоритм выдавал вероятность положительного класса:

$$b_*(x) = \underset{b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, b) \approx p(y = +1|x)$$

По закону больших чисел при $n \to \infty$ получаем

$$b_*(x) = \underset{b \in \mathbb{R}}{\operatorname{argmin}} E[L(y, b)|x]$$

Отсюда получаем ***условие на функцию потерь***:

$$\boldsymbol{\operatorname{argmin} E[L(y, b)|x] = p(y = +1|x)}$$

ФУНКЦИИ ПОТЕРЬ

Подходят:

- **Квадратичная**

$$L(y,z) = (y-z)^2$$

- **Логистическая (log-loss)**

$$L(y,z) = [y=+1]\cdot\log\big(b(x,w)\big) + [y=-1]\cdot\log(1-b(x,w))$$

Не подходят:

- **Модуль**

$$L(y,z) = |y-z|$$

ПРАВДОПОДОБИЕ И LOG-LOSS

- Вероятности, которые выдает алгоритм $b(x)$, должны согласовываться с выборкой
- Вероятность того, что в выборке встретится объект x с классом y:

$$b(x)^{[y=+1]} \cdot \big(1 - b(x)\big)^{[y=-1]}$$

ПРАВДОПОДОБИЕ И LOG-LOSS

- Вероятности, которые выдает алгоритм $b(x)$, должны согласовываться с выборкой
- Вероятность того, что в выборке встретится объект x с классом y:

$$b(x)^{[y=+1]} \cdot \big(1 - b(x)\big)^{[y=-1]}$$

Правдоподобие выборки:

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]}$$

ФУНКЦИЯ ПОТЕРЬ ДЛЯ ОБУЧЕНИЯ

- Для нахождения оптимальных параметров алгоритма можно воспользоваться методом максимума правдоподобия (ММП):

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]} \to \max_b$$

ФУНКЦИЯ ПОТЕРЬ ДЛЯ ОБУЧЕНИЯ

- Для нахождения оптимальных параметров алгоритма можно воспользоваться методом максимума правдоподобия (ММП):

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]} \to \max_b$$

- Прологарифмируем правдоподобие и поставим перед ним минус, получим следующую эквивалентную задачу:

$$-\sum_{i=1}^{l} ([y_i = +1] \log b(x_i) + [y_i = -1] \log(1 - b(x_i))) \to \min_b$$

ФУНКЦИЯ ПОТЕРЬ ДЛЯ ОБУЧЕНИЯ

- Для нахождения оптимальных параметров алгоритма можно воспользоваться методом максимума правдоподобия (ММП):

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]} \to \max_b$$

- Прологарифмируем правдоподобие и поставим перед ним минус, получим следующую эквивалентную задачу:

Это log-loss!

$$-\sum_{i=1}^{l} ([y_i = +1] \log b(x_i) + [y_i = -1] \log(1 - b(x_i))) \to \min_b$$

ФУНКЦИЯ ПОТЕРЬ ДЛЯ ОБУЧЕНИЯ

- Для нахождения оптимальных параметров алгоритма можно воспользоваться методом максимума правдоподобия (ММП):

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]} \to \max_b$$

- Прологарифмируем правдоподобие и поставим перед ним минус, получим следующую эквивалентную задачу:

Это log-loss!

$$-\sum_{i=1}^{l} ([y_i = +1] \log b(x_i) + [y_i = -1] \log(1 - b(x_i))) \to \min_b$$

Вывод: *логистическая функция потерь корректно предсказывает вероятности.*

ВЫБОР АЛГОРИТМА $b(x)$

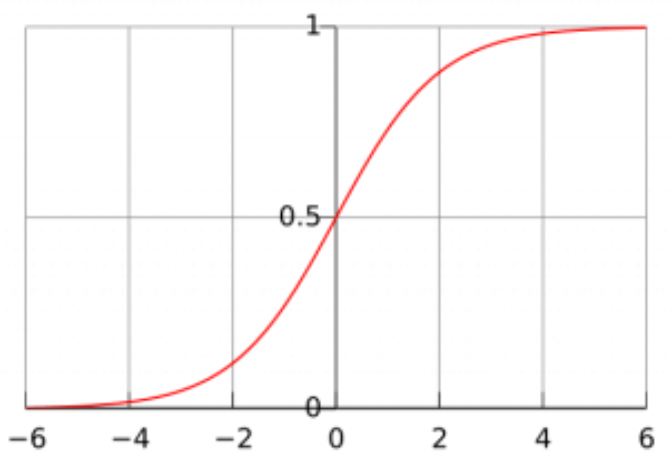
- Хотим, чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$.

ВЫБОР АЛГОРИТМА $b(x)$

- Хотим, чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$.
- Можно взять $b(x) = \sigma(w^T x)$, где σ – любая монотонно неубывающая функция с областью значений $[0, 1]$.

ВЫБОР АЛГОРИТМА $b(x)$

- Хотим, чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$.
- Можно взять $b(x) = \sigma(w^T x)$, где σ – любая монотонно неубывающая функция с областью значений $[0, 1]$.
- Возьмем ***сигмоиду***: $\sigma(z) = \frac{1}{1+e^{-z}}$



СМЫСЛ (w, x) В ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

- Логистическая регрессия в каждой точке x предсказывает вероятность того, что x принадлежит положительному классу $p(y = +1|x)$.
- То есть $p(y = +1|x) = \frac{1}{1+e^{-w^T x}}$. Отсюда можно выразить $(w, x) = w^T x$:

$$(w, x) = w^T x = \log \frac{p(y = +1|x)}{p(y = -1|x)}$$

СМЫСЛ (w, x) В ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

- Логистическая регрессия в каждой точке x предсказывает вероятность того, что x принадлежит положительному классу $p(y = +1|x)$.
- То есть $p(y = +1|x) = \frac{1}{1+e^{-w^T x}}$. Отсюда можно выразить $(w, x) = w^T x$:

$$(w, x) = w^T x = \log \frac{p(y = +1|x)}{p(y = -1|x)}$$

- Величина $\log \frac{p(y=+1|x)}{p(y=-1|x)}$ называется ***логарифм отношения шансов (log odds)***. Из формулы видно, что величина может принимать любое значение.

ЛОГАРИФМИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ

Утверждение. Логарифмическая функция потерь может быть записана в виде

$$\boldsymbol{L(b, X) = \sum_{i=1}^{l} \log(1 + e^{-y_i(w,x)})}$$

Идея доказательства:

Подставляем явный вид сигмоиды в логарифмическую функцию потерь:

$$-\sum_{i=1}^{l}([y_i = +1]\log\sigma(w^T x_i) + [y_i = -1]\log(1 - \sigma(w^T x_i))) \to \min_w$$