Лекция 4
Линейные модели классификации. Часть 1.

Кантонистова Е.О.

ВШЭ, 2018

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ (НАПОМИНАНИЕ)

Обучающая выборка $\{(x_i, y_i)\}, x_i \in \mathbb{R}^n, y_i \in \mathbb{R}$

- Модель линейной регрессии:

$$a(x,w) = (x,w) \ = \sum_{i=1}^{n} w_j x_j$$

- Функция потерь – квадратичная:

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

- Метод обучения – метод наименьших квадратов:

$$Q(w) = \sum_{i=1}^{n} (a(x_i, w) - y_i)^2 \to min$$

БИНАРНАЯ КЛАССИФИКАЦИЯ

Обучающая выборка $\{(x_i, y_i)\}, x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$

- Модель линейного классификатора:

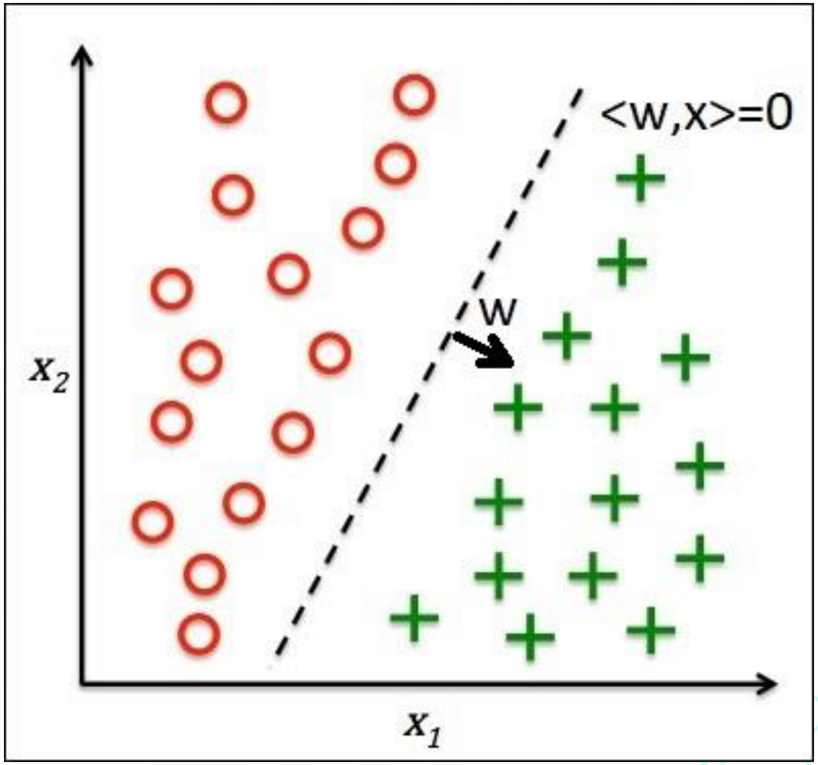
$$\boldsymbol{a(x, w) = sign(x, w) = sign(\sum_{j=1}^{n} w_j x_j)}$$

Уравнение

$(x, w) = \sum_{j=1}^{n} w_j x_j = 0$

– уравнение гиперплоскости

с нормалью w.



ОБУЧЕНИЕ КЛАССИФИКАТОРА

- Обучение - минимизация доли ошибок классификатора:

$$Q(a,X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) \neq y_i] = \frac{1}{l}\sum_{i=1}^{l}[sign(w,x_i) \neq y_i] \to min$$

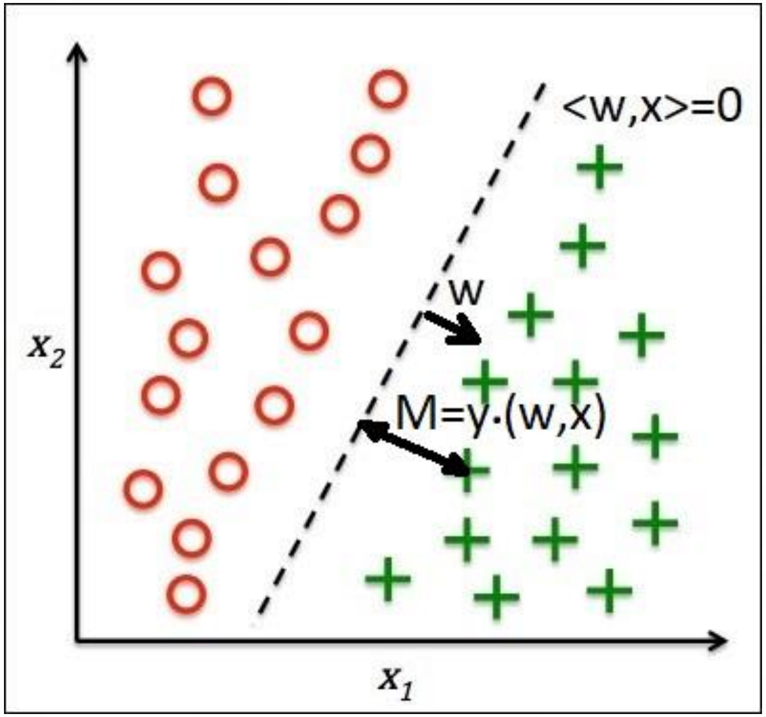
Функционал Q можно переписать в виде:

$$Q(a,X) = \frac{1}{l}\sum_{i=1}^{l}[y_i \cdot (w,x_i) < 0] = \frac{1}{l}\sum_{i=1}^{l}[M_i < 0] \to min$$

- $M_i = y_i \cdot (w,x_i)$ - ***отступ***

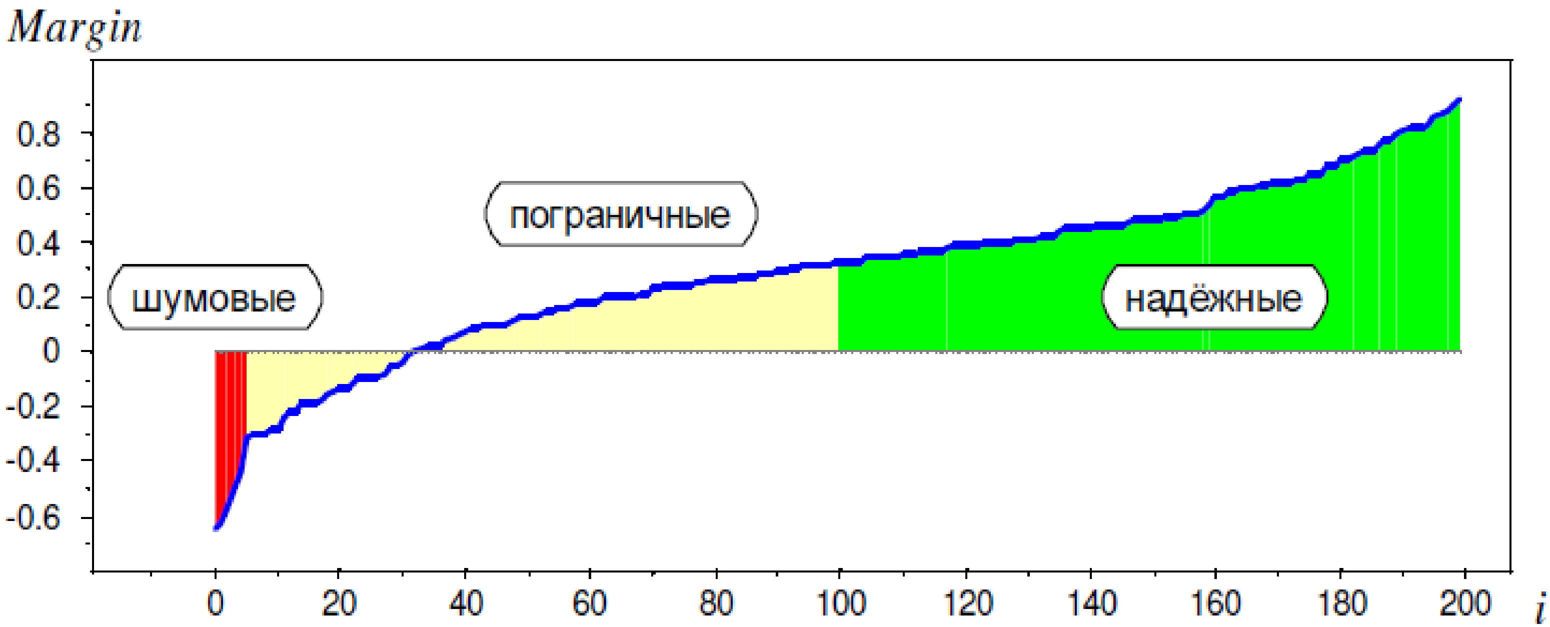
ОТСТУП (MARGIN)

- Знак отступа M говорит о корректности классификации (M>0 – объект классифицирован верно, M<0 – неверно)
- Абсолютная величина отступа M обозначает степень уверенности классификатора в ответе (чем ближе M к нулю, тем меньше уверенность в ответе)



ОТСТУП (MARGIN)

Ранжирование объектов по возрастанию отступа:



ОБУЧЕНИЕ КЛАССИФИКАТОРА

Обучающая выборка $\{(x_i, y_i)\}, x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$

- Модель линейного классификатора:

$$a(x, w) = sign(x, w) \ = sign(\sum_{i=1}^{n} w_j x_j)$$

- Функция потерь – бинарная:

$$L(a, y) = [a \neq y] = [a \cdot y < 0]$$

- Метод обучения – *минимизация эмпирического риска*:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} [y_i \cdot (w, x_i) < 0] \to min$$

ВЕРХНИЕ ОЦЕНКИ ЭМПИРИЧЕСКОГО РИСКА

- $L(a, y) = L(M) = [M < 0]$ – разрывная функция потерь

Оценим

$\boldsymbol{L(M)} \leq \ \boldsymbol{\tilde{L}(M)}$, где $\tilde{L}(M)$ - непрерывная или гладкая функция потерь.

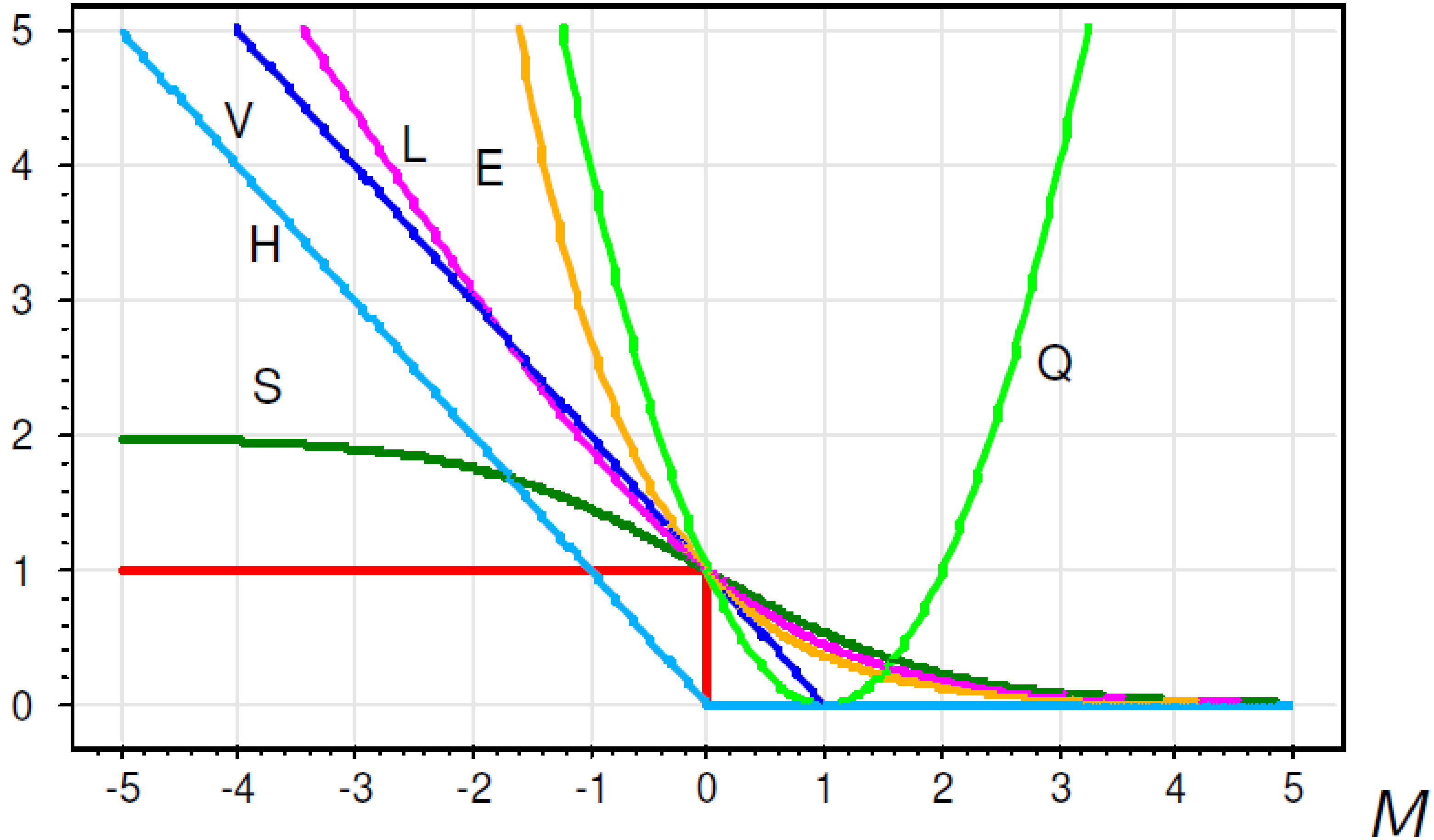
- Тогда

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l} L(y_i \cdot (w, x_i)) \leq \frac{1}{l}\sum_{i=1}^{l} \tilde{L}\big(y_i \cdot (w, x_i)\big) \to min$$

ФУНКЦИИ ПОТЕРЬ

- $L(M) = \log(1 + e^{-M})$ – логистическая функция потерь
- $\mathrm{V}(M) = (1 - \mathrm{M})_+ = \max(0{,}1 - M)$ – кусочно-линейная функция потерь (SVM)
- $H(M) = (-\mathrm{M})_+ = \max(0, -M)$ – кусочно-линейная функция потерь (персептрон)
- $E(M) = e^{-M}$ - экспоненциальная функция потерь
- $S(M) = \frac{2}{1+e^{-M}}$ - сигмоидная функция потерь
- $[M < 0]$ – пороговая функция потерь

ФУНКЦИИ ПОТЕРЬ



ОПТИМИЗАЦИЯ ФУНКЦИОНАЛА ПОТЕРЬ

- Нахождение минимума функции потерь происходит с помощью метода градиентного спуска:

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla Q(w^{(k-1)})$$

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

- Accuracy – доля правильных ответов:

$$accuracy(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i]$$

Недостаток: при сильно несбалансированной выборке не отражает качество работы алгоритма

МАТРИЦА ОШИБОК

Матрица ошибок (confusion matrix):

		Actual Value	
		positives	negatives
Predicted Value	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

$$\boldsymbol{accuracy = (TP + TN)/(TP + FP + FN + TN)}$$

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ: PRECISION, RECALL

- **Precision (точность):**

$$Precision(a, X) = \frac{TP}{TP + FP}$$

Показывает, насколько можно доверять классификатору при $a(x) = 1$

PRECISION: ПРИМЕР

Модель $a_1(x)$:

$\text{precision}(a_1, X) = 0.8$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$a(x) = 1$ Получили кредит	80	20
$a(x) = -1$ Не получили кредит	20	80

Модель $a_2(x)$:

$\text{precision}(a_2, X) = 0.96$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$a(x) = 1$ Получили кредит	48	2
$a(x) = -1$ Не получили кредит	52	98

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ: PRECISION, RECALL

- Precision (точность):

$$\boldsymbol{Precision(a, X) = \frac{TP}{TP + FP}}$$

Показывает, насколько можно доверять классификатору при $a(x) = 1$

- **Recall (полнота):**

$$\boldsymbol{Recall(a, X) = \frac{TP}{TP + FN}}$$

Показывает, как много объектов положительного класса находит классификатор

RECALL: ПРИМЕР

Модель $a_1(x)$:

$\mathrm{rec}all(a_1, X) = 0.8$

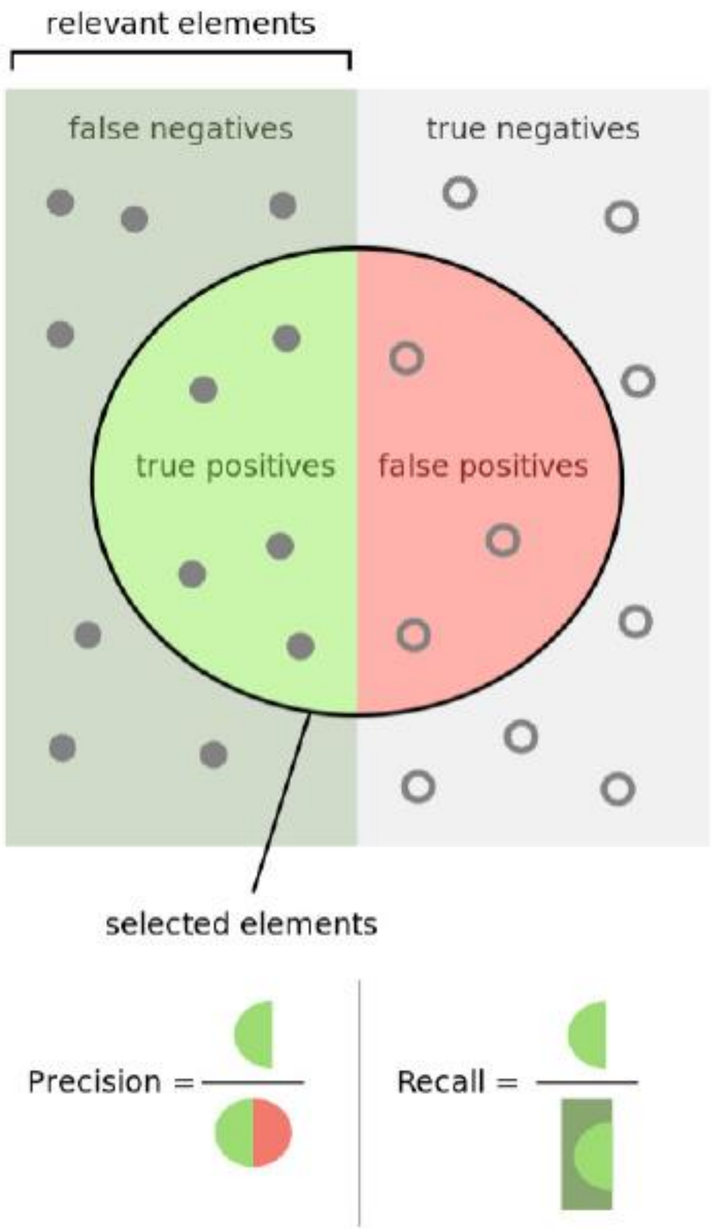
	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	80	20
$\mathbf{a(x) = -1}$ Не получили кредит	20	80

Модель $a_2(x)$:

$\mathrm{recall}(a_2, X) = 0.48$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	48	2
$\mathbf{a(x) = -1}$ Не получили кредит	52	98

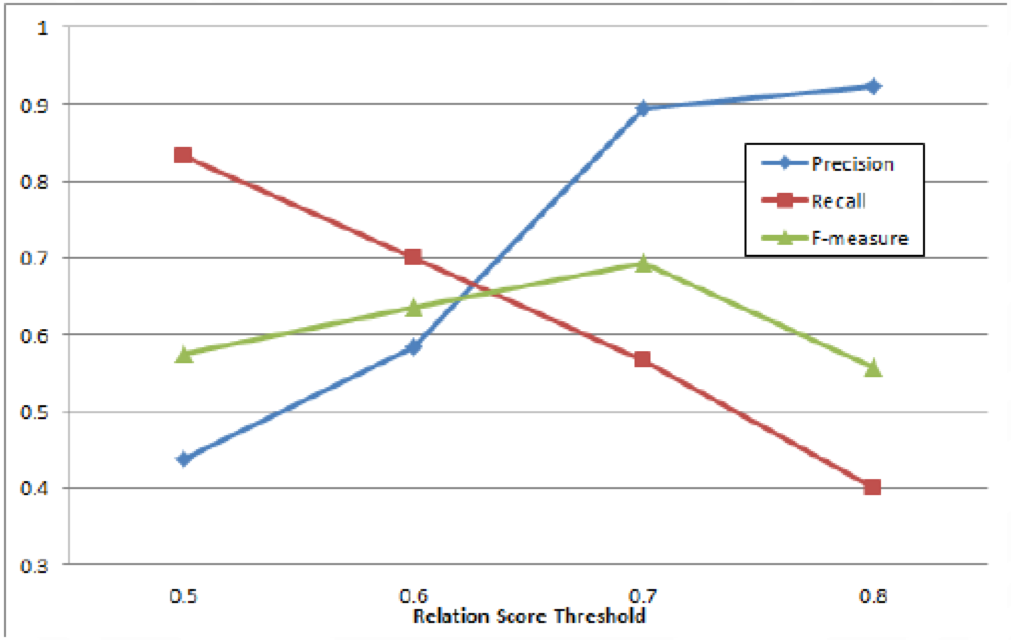
ТОЧНОСТЬ И ПОЛНОТА



F-МЕРА

F-мера – это метрика качества, учитывающая и точность, и полноту

$$F(a, X) = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$



КАК РЕГУЛИРОВАТЬ ТОЧНОСТЬ И ПОЛНОТУ

Обобщенная форма записи классификатора:

$$a(x) = [b(x) > t], t \in \mathbb{R}$$

В случае линейного классификатора

$$a(x) = [(w, x) > 0],$$

$$b(x) = (w, x), t = 0$$

- Регулировать точность и полноту можно путем изменения порога t.

Интегральная метрика: ROC-AUC

Хотим измерить качество всего семейства классификаторов

$$a(x) = [b(x) > t], t \in \mathbb{R}$$

(без фиксации порога t).

Для этого будем использовать метрику AUC

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

ROC-КРИВАЯ

Для каждого значения порога t вычислим:

- False Positive Rate (доля неверно принятых объектов):

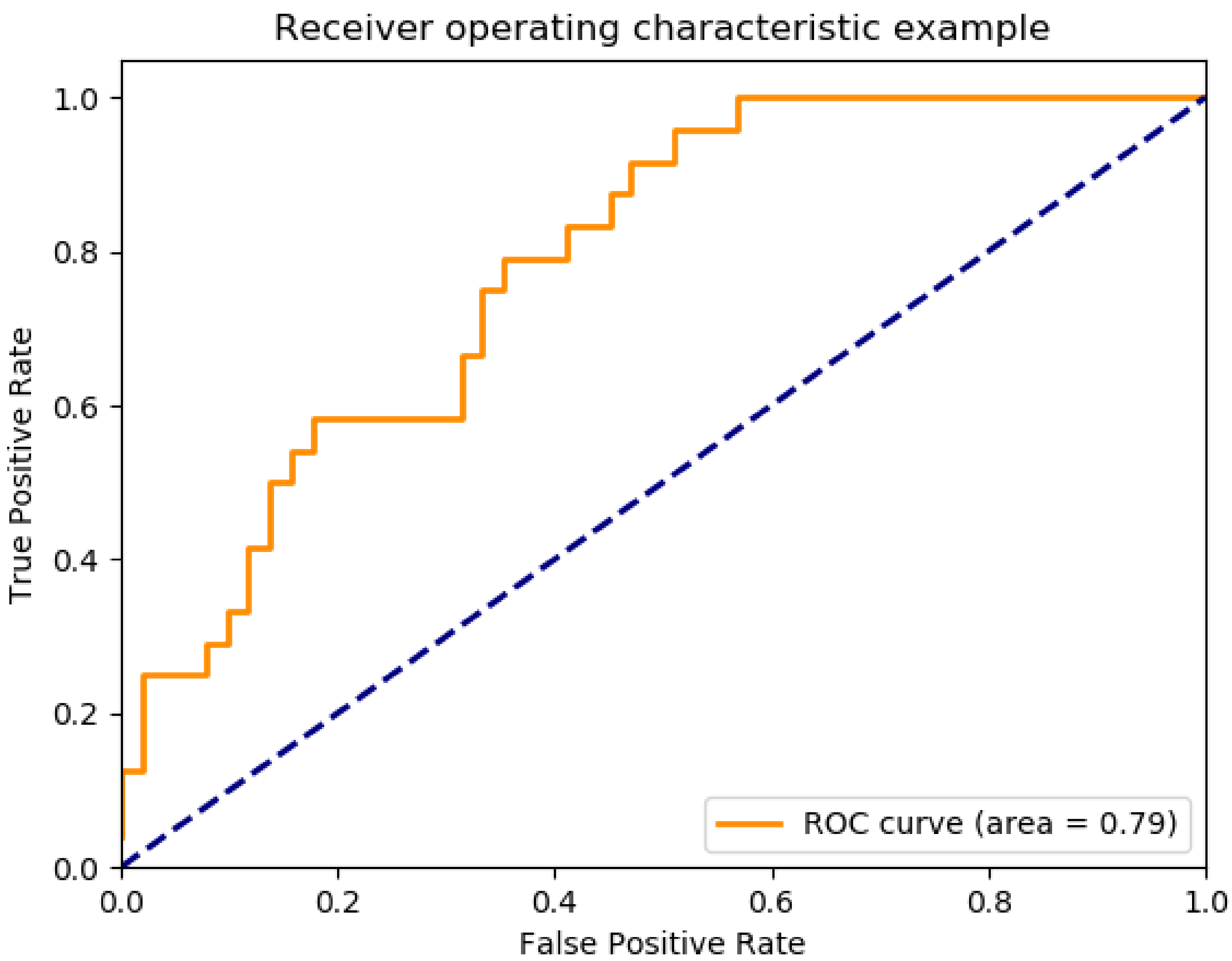
$$FPR = \frac{FP}{FP+TN} = \frac{\sum_i [y_i = -1][a(x_i) = +1]}{\sum_i [y_i = -1]}$$

- True Positive Rate (доля верно принятых объектов):

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

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

ROC-КРИВАЯ



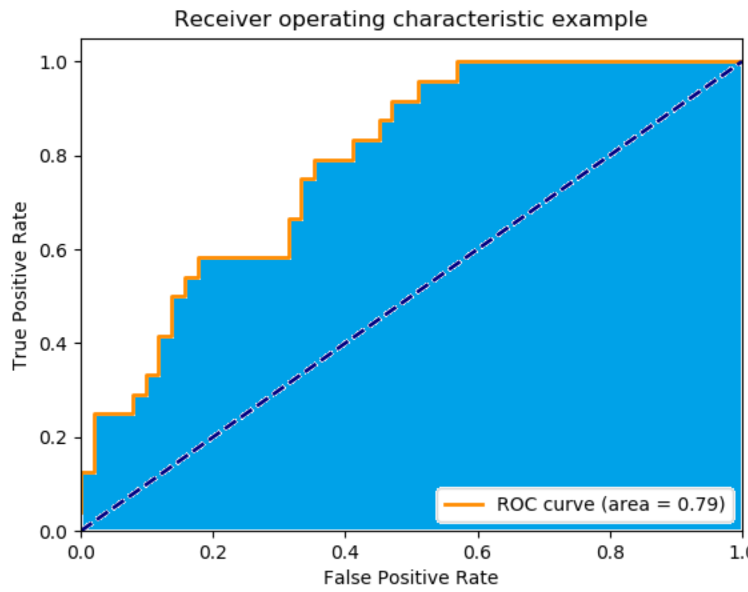
Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.79)

ROC-КРИВАЯ. AUC.

- Каждая точка на ROC-кривой соответствует классификатору с фиксированным значением порога t.
- Всего различных порогов $l + 1$, где l – количество объектов.

AUC – 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: $\boldsymbol{t = 0.7}$, то есть

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

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

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

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: $\boldsymbol{t = 0.7}$, то есть

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

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

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

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

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:

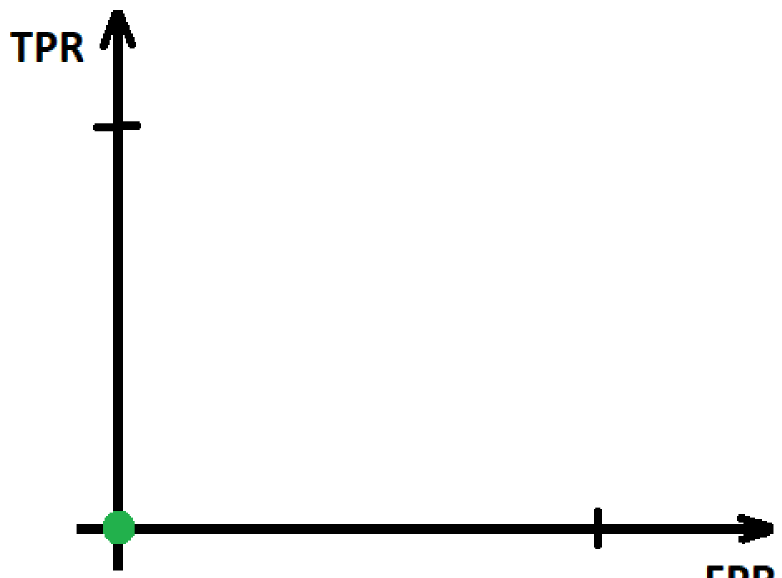
$(0.7, 0.4, 0.2, 0.1, 0.05)$

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	$\mathbf{0.7}$	0.05
y	-1	$+1$	-1	$\mathbf{+1}$	$+1$

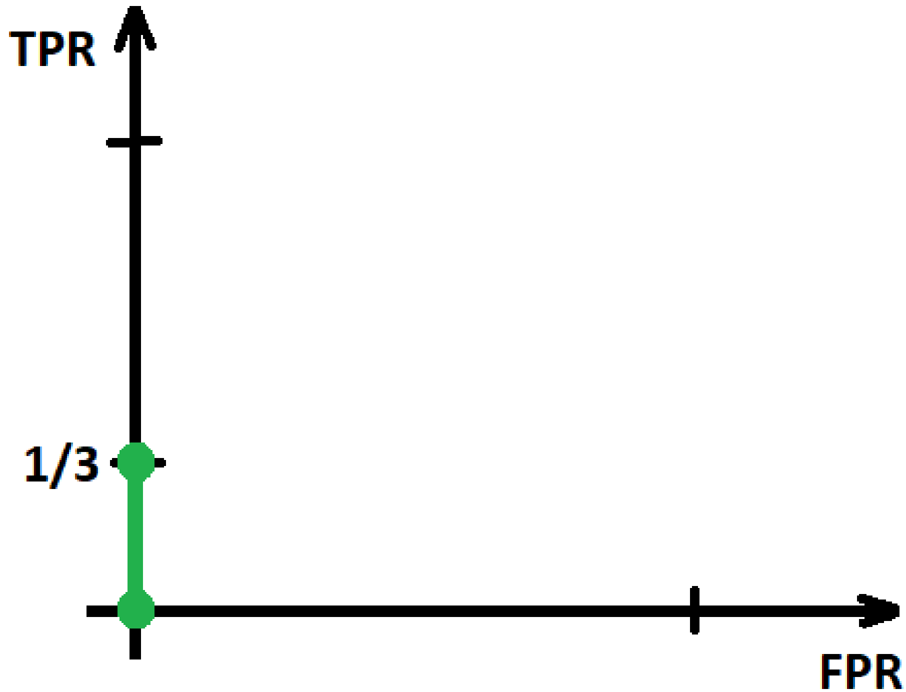
- Упорядочим объекты по

убыванию предсказаний: $(0.7, 0.4, 0.2, 0.1, 0.05)$

2 шаг: $\boldsymbol{t = 0.4}$, 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	$+1$	$+1$

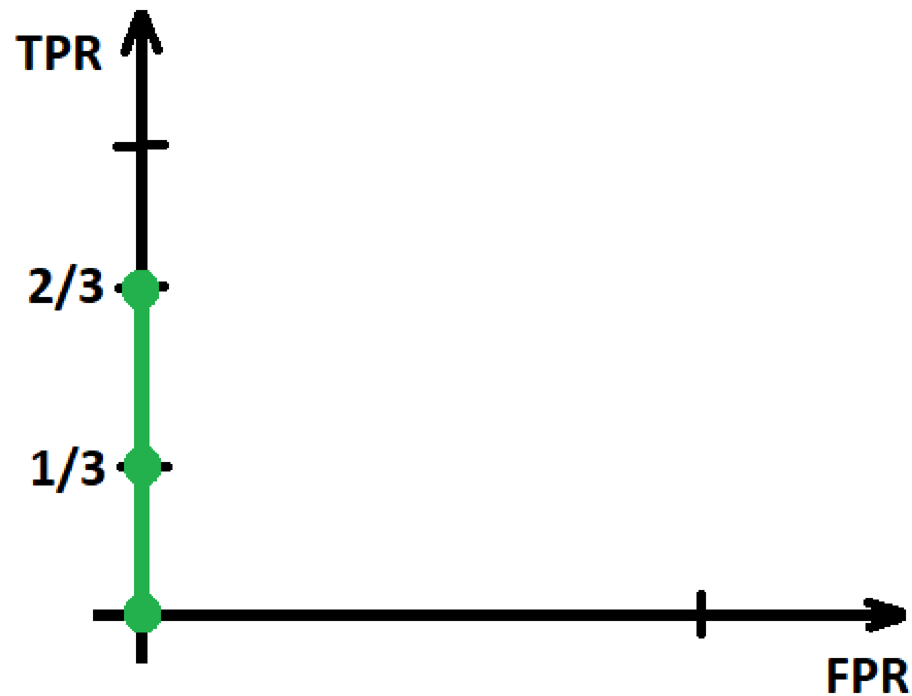
- Упорядочим объекты по

убыванию предсказаний:
$(0.7, 0.4, 0.2, 0.1, 0.05)$

3 шаг: $\boldsymbol{t = 0.2}$, 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	$\mathbf{+1}$	$+1$

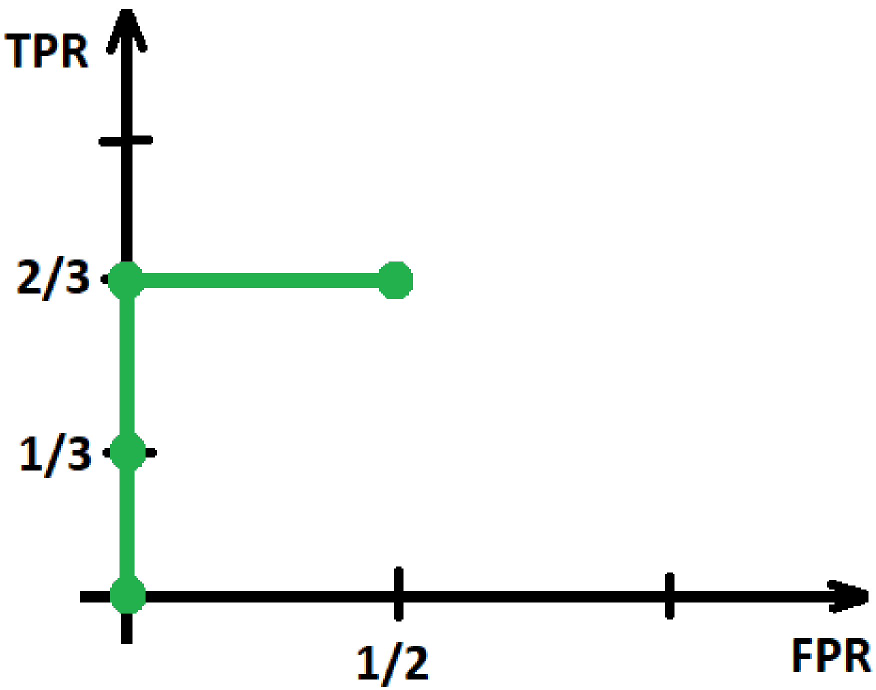
- Упорядочим объекты по

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	$\mathbf{-1}$	$\mathbf{+1}$	$\mathbf{-1}$	$\mathbf{+1}$	$+1$

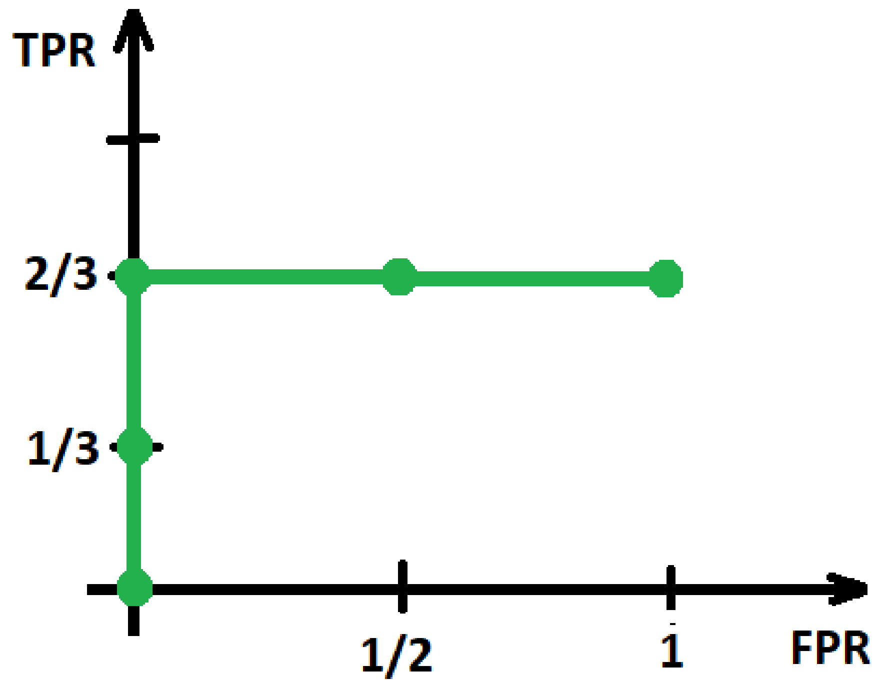
- Упорядочим объекты по

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	$+1$	$+1$

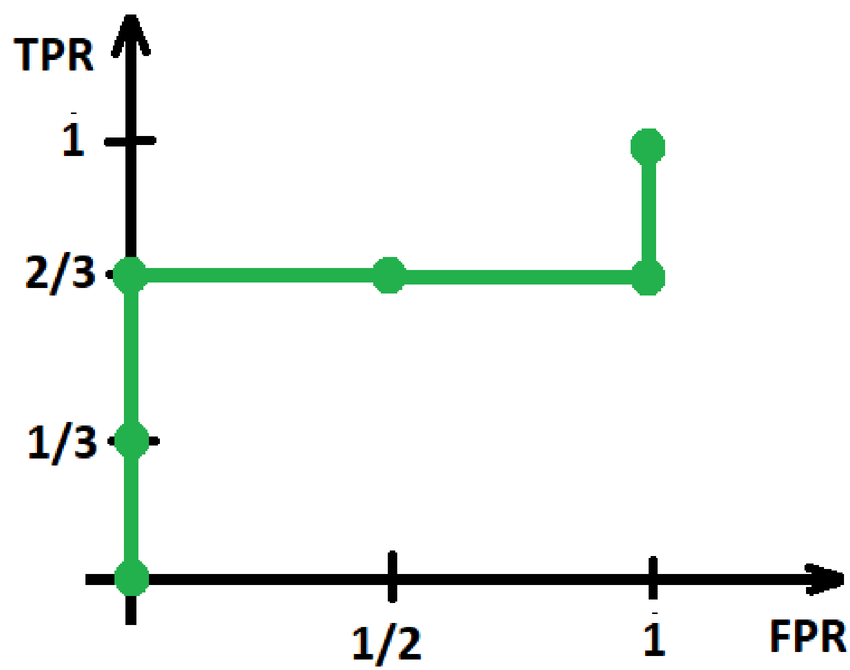
- Упорядочим объекты по

убыванию предсказаний:
$(0.7, 0.4, 0.2, 0.1, 0.05)$

5 шаг: $\boldsymbol{t = 0}$, то есть
$a(x) = [b(x) > 0]$

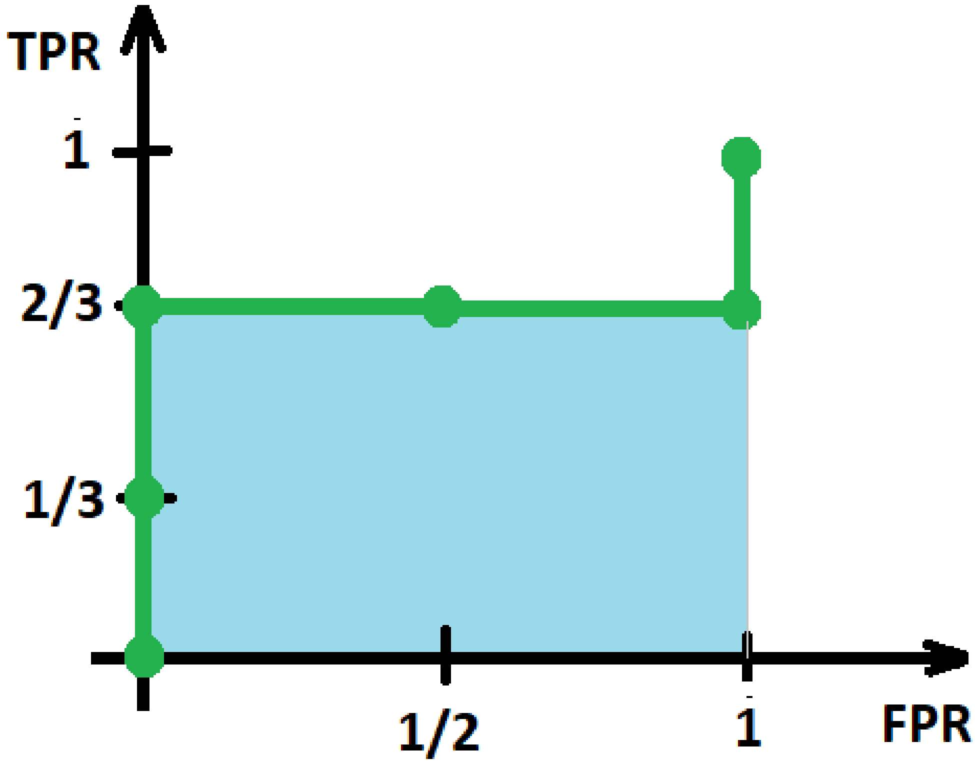
$TPR = \frac{3}{3+0} = 1,$

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



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

$$AUC = 2/3$$

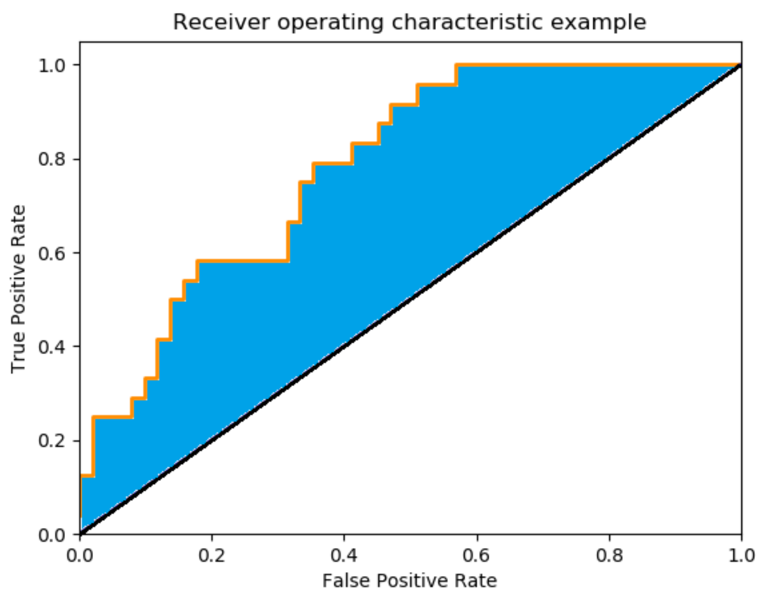


ИНДЕКС ДЖИНИ

Индекс Джини:

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

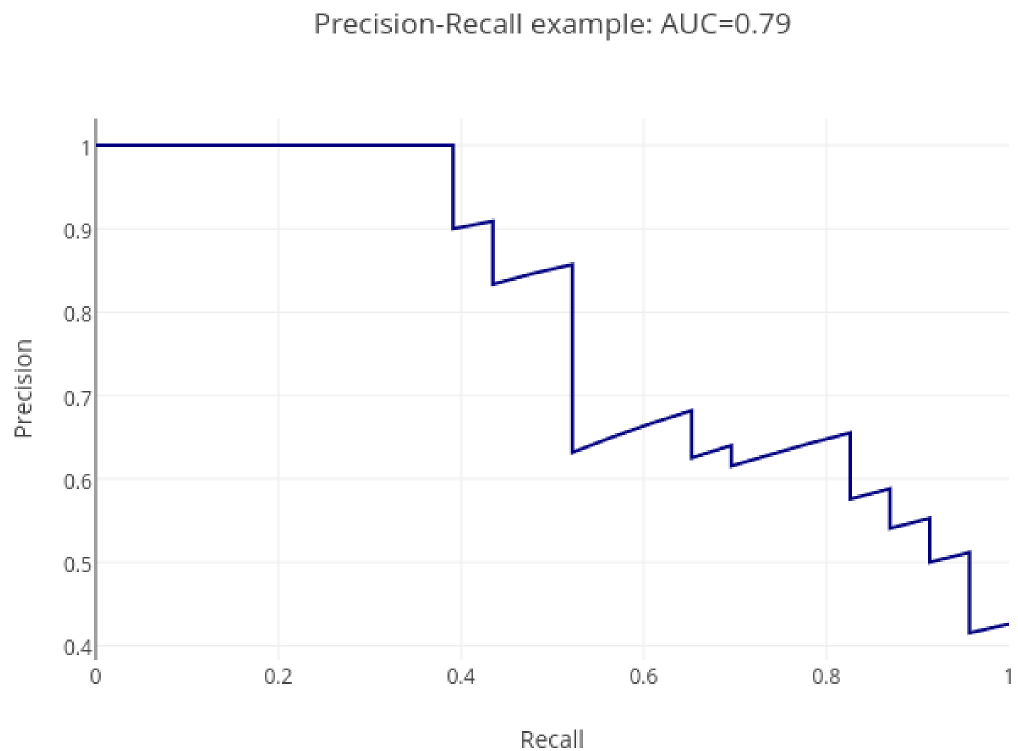
- Индекс Джини – это удвоенная площадь между главной диагональю и ROC-кривой.



PRECISION-RECALL КРИВАЯ

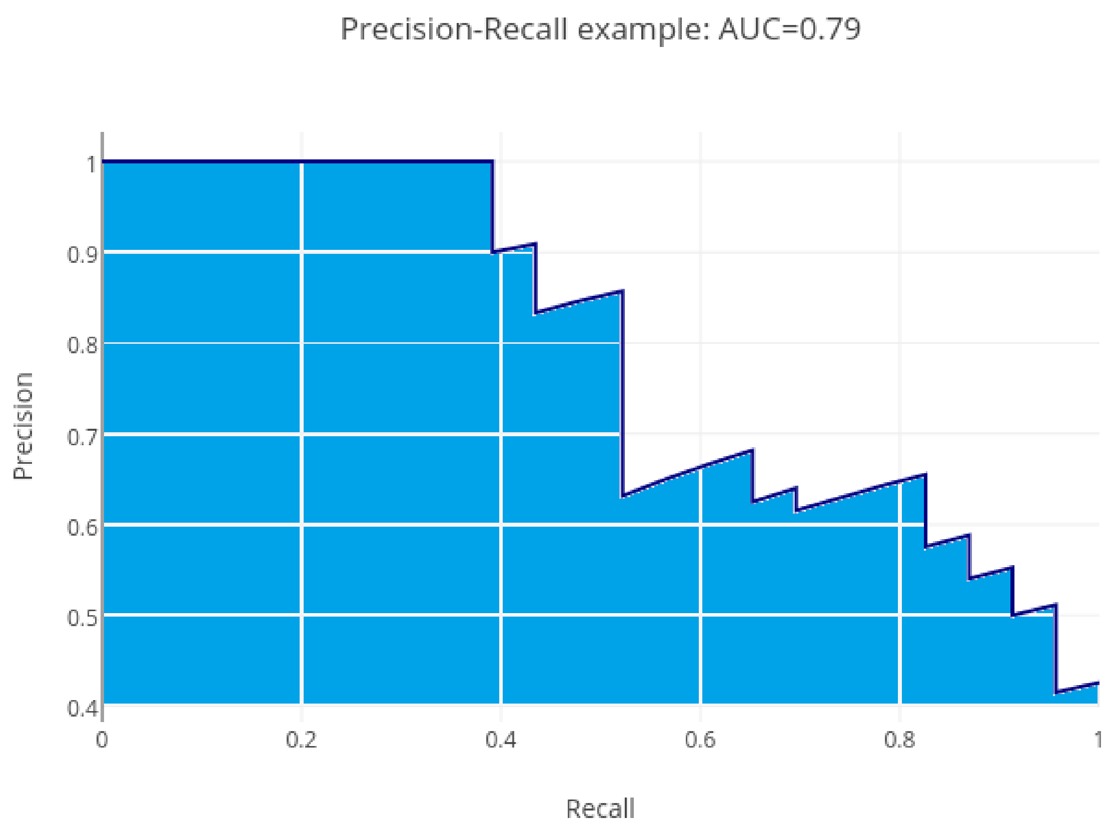
- В случае малой доли объектов положительного класса AUC-ROC может давать неадекватно хороший результат

Precision-Recall кривая:



AUC-PR

AUC-PR – площадь под PR-кривой



ВИДЫ ЛИНЕЙНЫХ КЛАССИФИКАТОРОВ

В этой лекции:

- Персептрон
- Логистическая регрессия (начало)

ПЕРСЕПТРОН РОЗЕНБЛАТТА

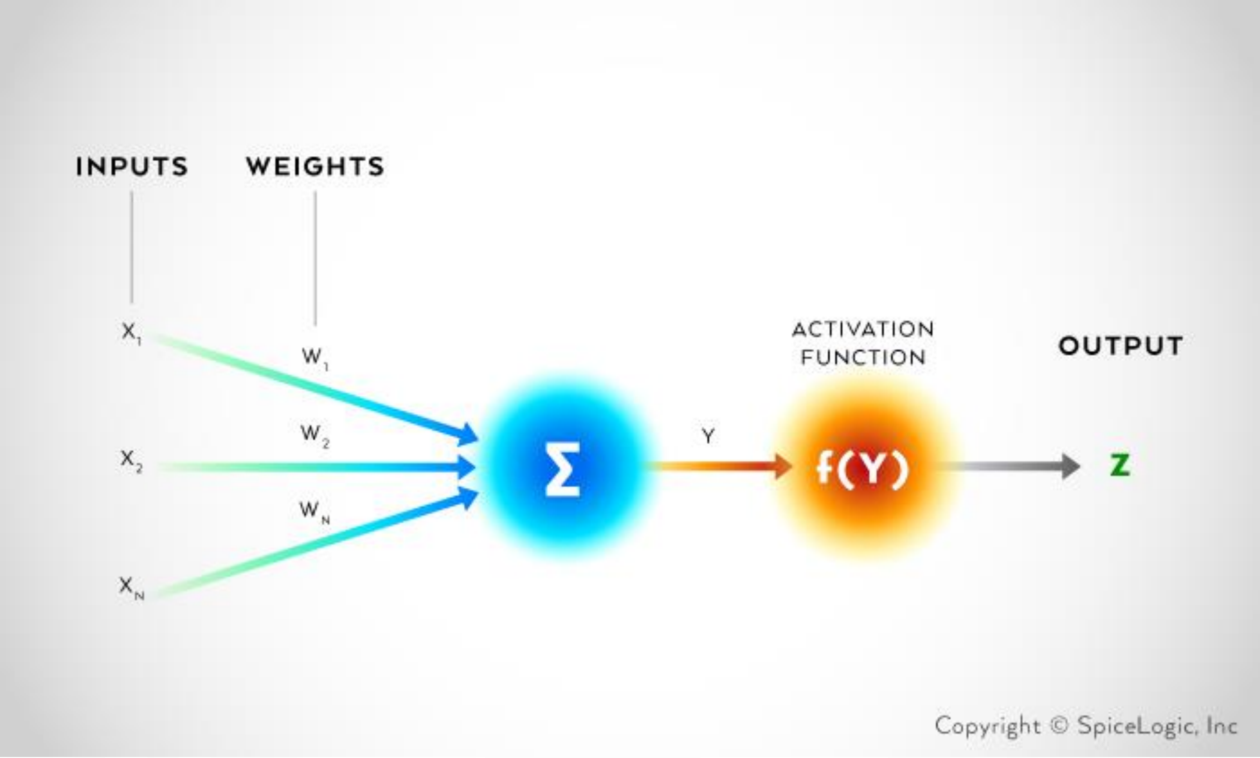
- Задача классификации с двумя классами $y_i \in \{0,1\}$
- Признаки объектов – бинарные: $x_i^j \in \{0,1\}$

Алгоритм: $\boldsymbol{a(x,w) = [w_1 x_1 + \cdots + w_n x_n > 0] = [(w,x) > 0]}$

ПЕРСЕПТРОН РОЗЕНБЛАТТА

- Задача классификации с двумя классами $y_i \in \{0,1\}$
- Признаки объектов – бинарные: $x_i^j \in \{0,1\}$

Алгоритм: $\boldsymbol{a(x,w) = [w_1 x_1 + \cdots + w_n\, x_n > 0] = [(w,x) > 0]}$



ПЕРСЕПТРОН РОЗЕНБЛАТТА

- Задача классификации с двумя классами $y_i \in \{0,1\}$
- Признаки объектов – бинарные: $x_i^j \in \{0,1\}$

Алгоритм: $a(x, w) = [(w, x) > 0]$

Обучение алгоритма, правило Хебба (исправление весов в случае ошибки):

- $a(x_i, w) = y_i$ => w менять не нужно
- $a(x_i, w) = 0, \mathrm{y_i} = 1 => w_j := w_j + h \cdot x_i^j$
- $a(x_i, w) = 1, \mathrm{y_i} = 0 => w_j := w_j - h \cdot x_i^j$

ПЕРСЕПТРОН РОЗЕНБЛАТТА

- Задача классификации с двумя классами $y_i \in \{0,1\}$
- Признаки объектов – бинарные: $x_i^j \in \{0,1\}$

Алгоритм: $a(x,w) = [(w,x) > 0]$

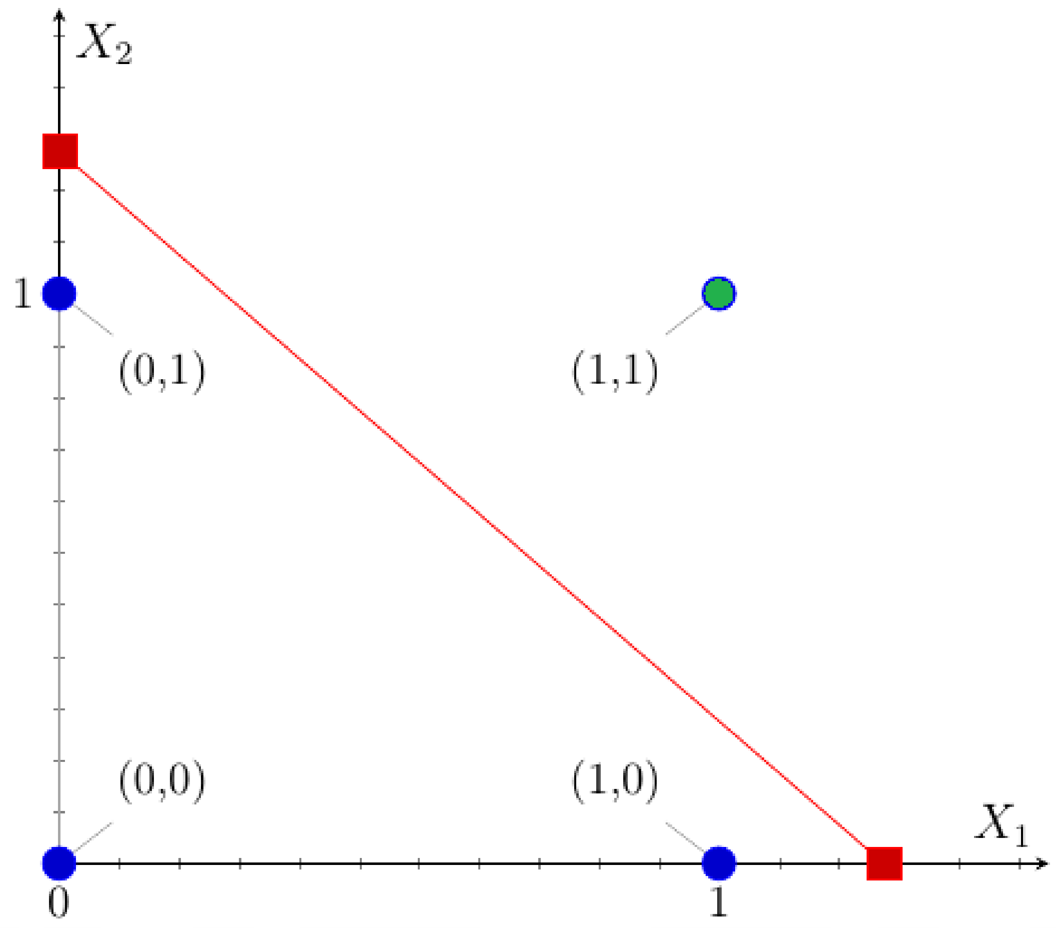
Обучение алгоритма (исправление весов в случае ошибки):

- $a(x_i,w) = y_i$ => w менять не нужно
- $a(x_i,w) = 0, \mathrm{y_i} = 1 => w_j := w_j + h \cdot x_i^j$
- $a(x_i,w) = 1, \mathrm{y_i} = 0 => w_j := w_j - h \cdot x_i^j$

Общая формула коррекции весов:

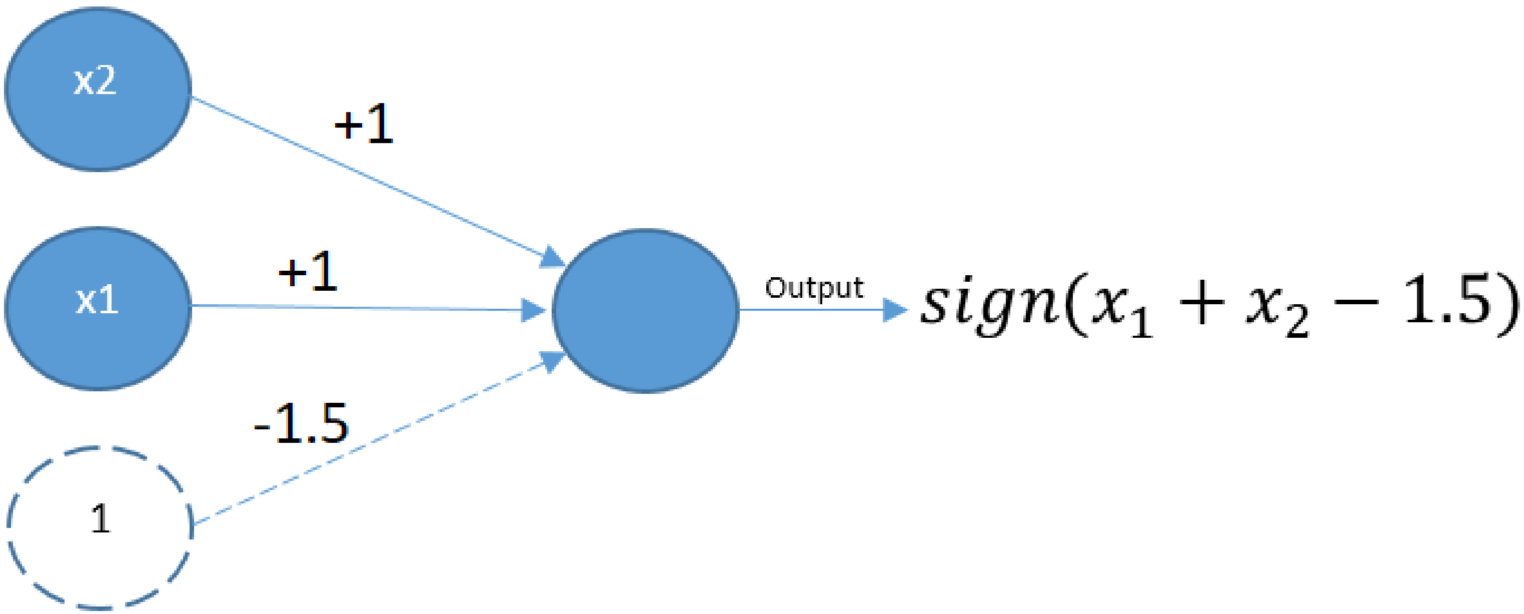
$$\boldsymbol{w := w - h(a(x_i,w) - y_i)x_i}$$

Пример: реализация логического AND с помощью персептрона

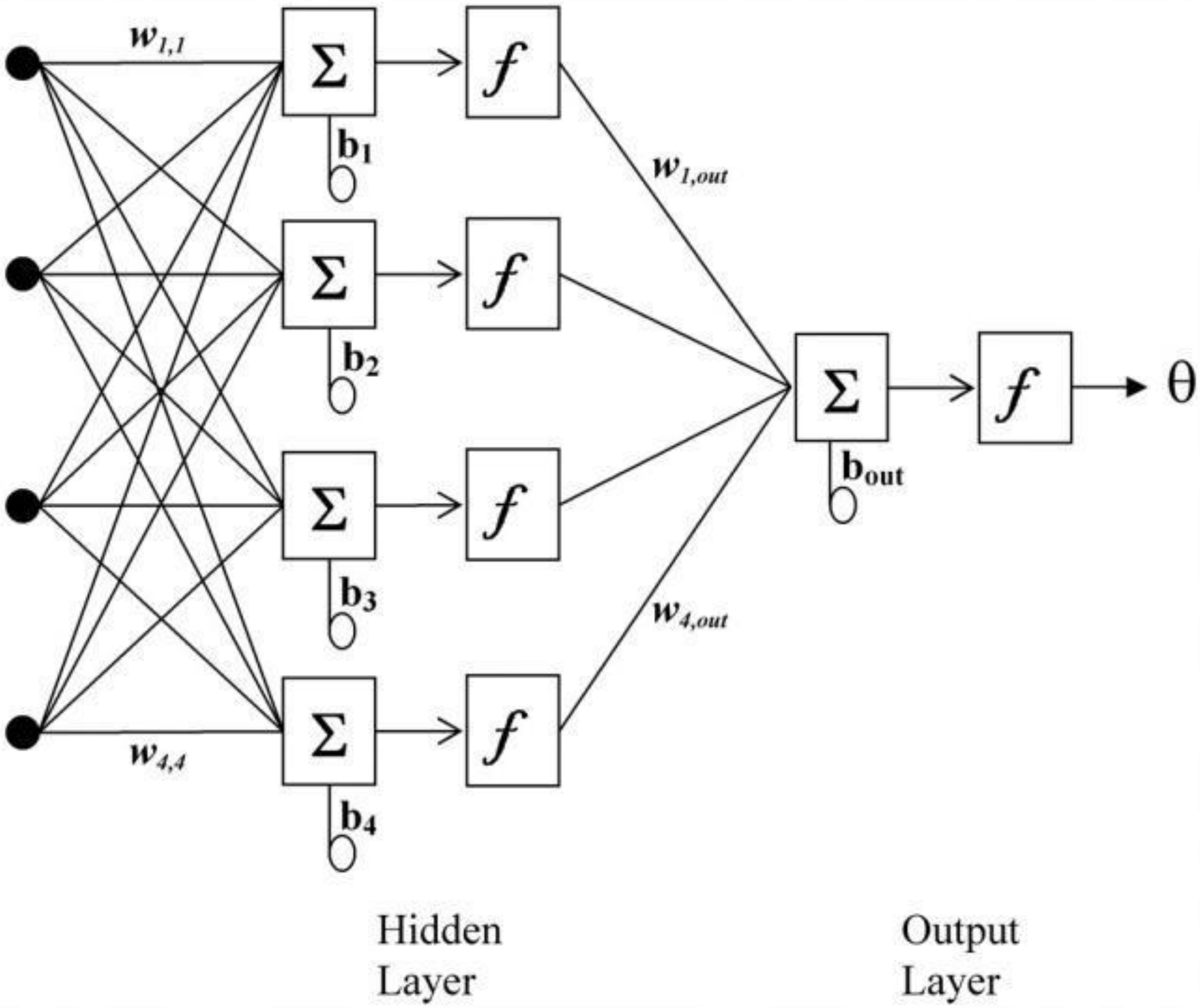


ПРИМЕР: РЕАЛИЗАЦИЯ ЛОГИЧЕСКОГО AND С ПОМОЩЬЮ ПЕРСЕПТРОНА

$$a(x, w) = sign(x_1 + x_2 - 1.5)$$



ПРИМЕР ДВУХСЛОЙНОГО ПЕРСЕПТРОНА



ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не классы, а вероятности классов.

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

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

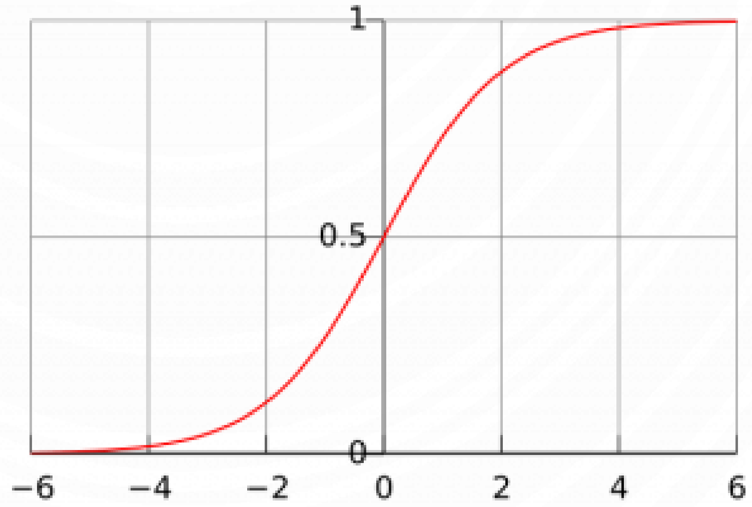
ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не классы, а вероятности классов.

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

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

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



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

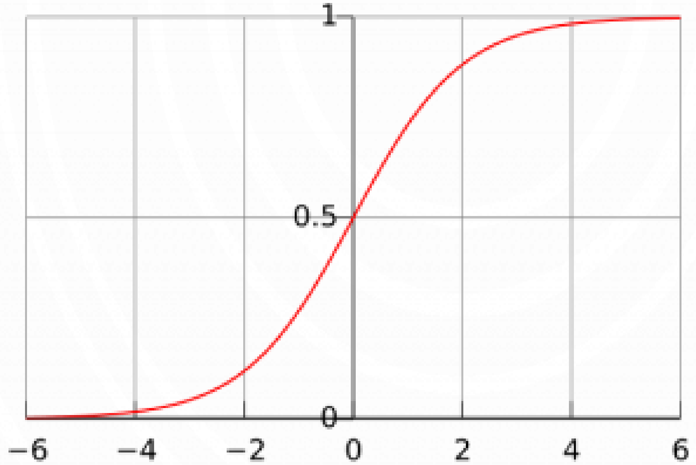
ВЕРОЯТНОСТНЫЙ СМЫСЛ

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

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

РАЗДЕЛЯЮЩАЯ ГРАНИЦА

Предсказываем $y = +1$, если $a(x, w) \geq 0.5$.



$a(x, w) = g(w^T x) \geq 0.5$, если $w^T x \geq 0$.

Получаем, что

- $y = +1 \text{ при } w^T x \geq 0$
- $y = -1 \text{ при } w^T x < 0$,

т.е. $w^T x = 0$ – разделяющая гиперплоскость.

Логистическая регрессия

Логистическая регрессия - это линейный классификатор!

ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Если взять квадратичную функцию потерь $L(a, y) = (a - y)^2$,

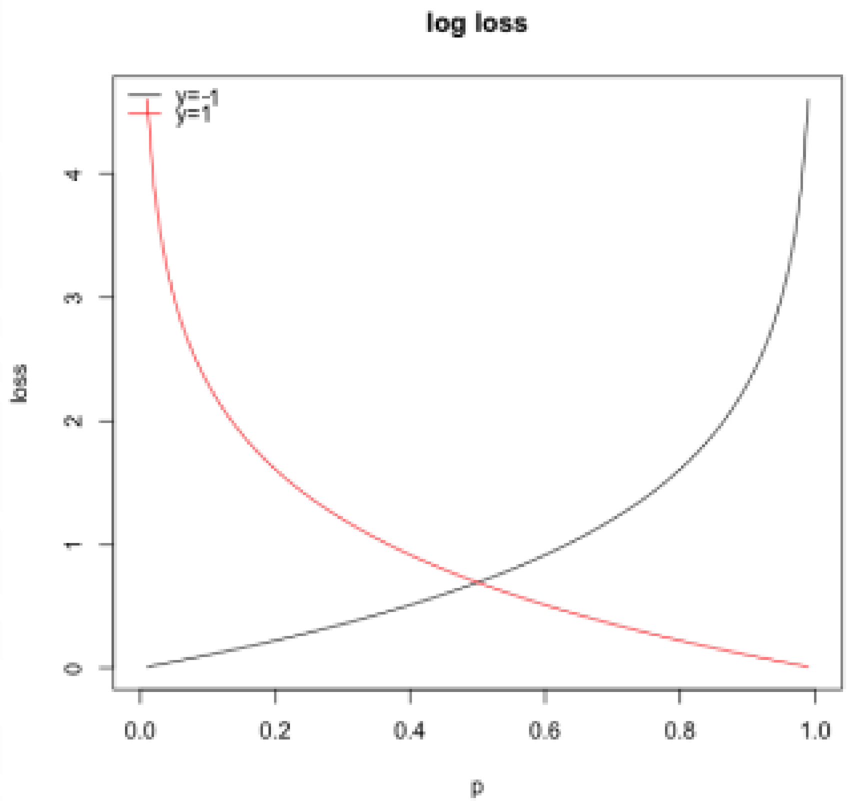
то возникнут проблемы:

- $Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{1}{1+e^{-w^T x}} - y\right)^2$ - не выпуклая функция (можем не попасть в глобальный минимум при оптимизации)
- На совсем неправильном предсказании маленький штраф (пусть предсказали вероятность 0% на объекте класса $y = +1$, тогда штраф всего $(1 - 0)^2 = 1$)

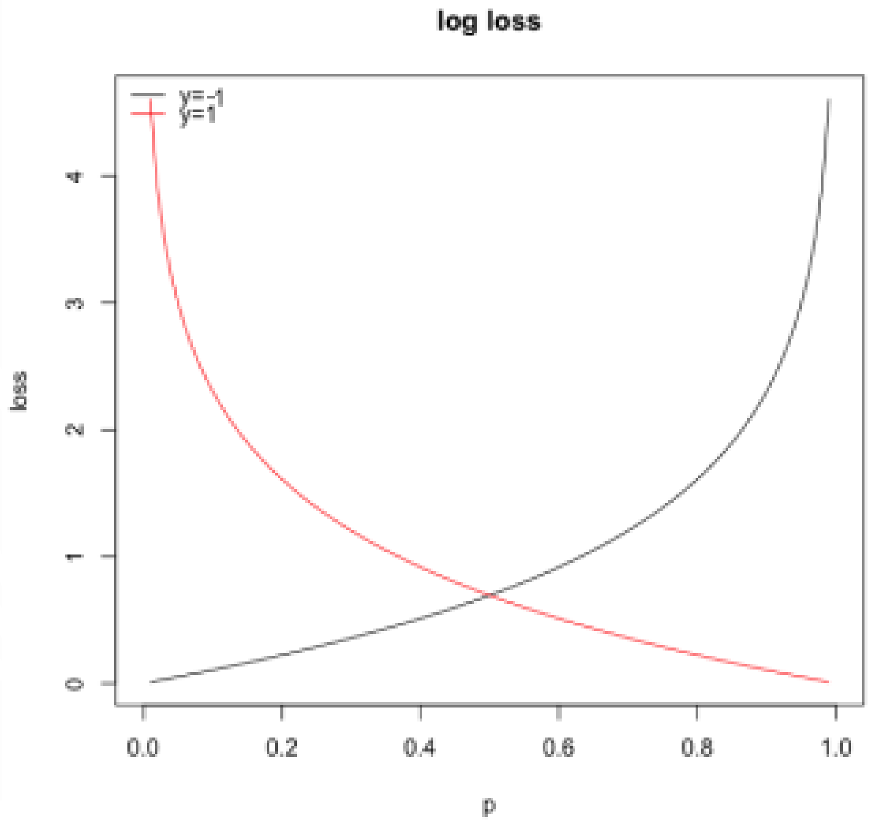
ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Возьмем логистическую функцию потерь (log-loss):

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



ЛОГИСТИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ



- если $a(x, w) = 1$ и $y = +1$, то штраф $L(a, y) = 0$
- если $a(x, w) \to 0$, а $y = +1$, то штраф $L(a, y) \to +\infty$