



Занятие 4

Линейные модели

классификации. Часть 1.

Блуменау М.И.

На основе материалов Кантонистовой Е.О.

ВШЭ, 2025

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ (НАПОМИНАНИЕ)

Обучающая выборка:

пусть x – объект (x_1, x_2, \dots, x_l - его признаки), а y – ответ на объекте (произвольное число), n – количество объектов.

Модель линейной регрессии:

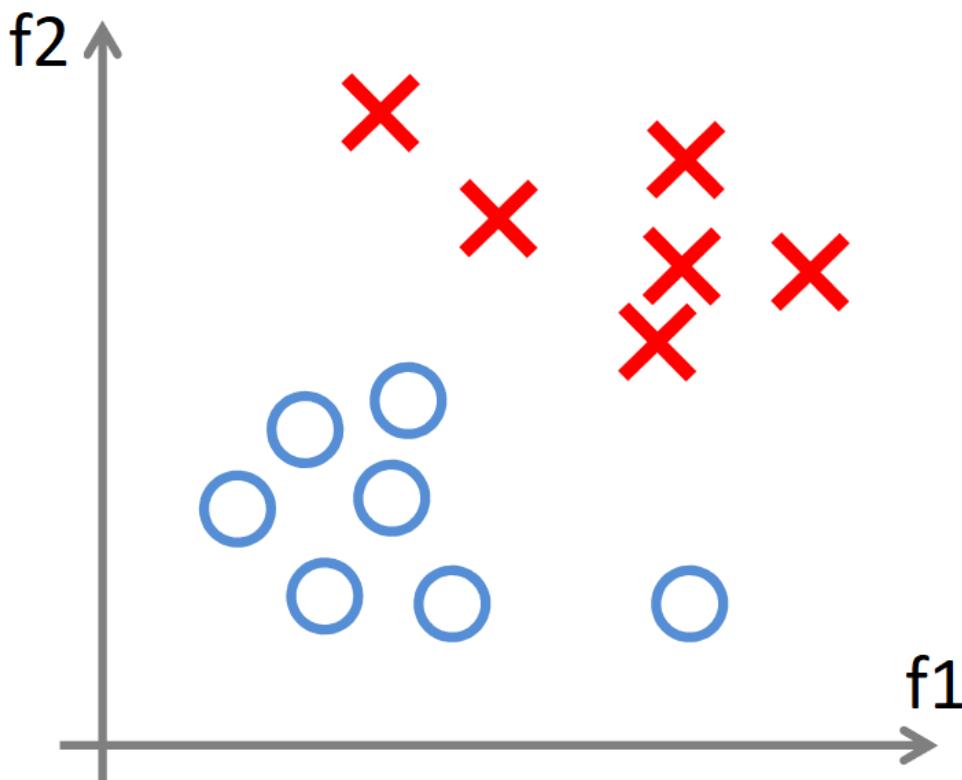
$$a(x, w) = \sum_{i=1}^l w_j x_j$$

- Метод обучения – метод наименьших квадратов
(минимизируем разность между предсказанием и правильным ответом):

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

БИНАРНАЯ КЛАССИФИКАЦИЯ

y_1, y_2, \dots, y_n - ответы (+1 или -1).



Как выглядит модель линейного классификатора: $a(x, w) = ?$

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

$$a(x, w) = \textcolor{red}{sign}\left(\sum_{j=1}^l w_j x_j\right)$$

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

$$a(x, w) = \text{sign} \left(\sum_{j=1}^l w_j x_j \right)$$

- если $\sum_{j=1}^l w_j x_j > 0$, то $\text{sign}(\sum_{j=1}^l w_j x_j) = +1$, то есть объект отнесён к положительному классу
- если $\sum_{j=1}^l w_j x_j < 0$, то $\text{sign}(\sum_{j=1}^l w_j x_j) = -1$, то есть объект отнесён к отрицательному классу
- значит, $\sum_{j=1}^l w_j x_j = 0$ – уравнение разделяющей границы между классами. Это уравнение плоскости (или прямой в двумерном случае), поэтому классификатор является линейным.

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

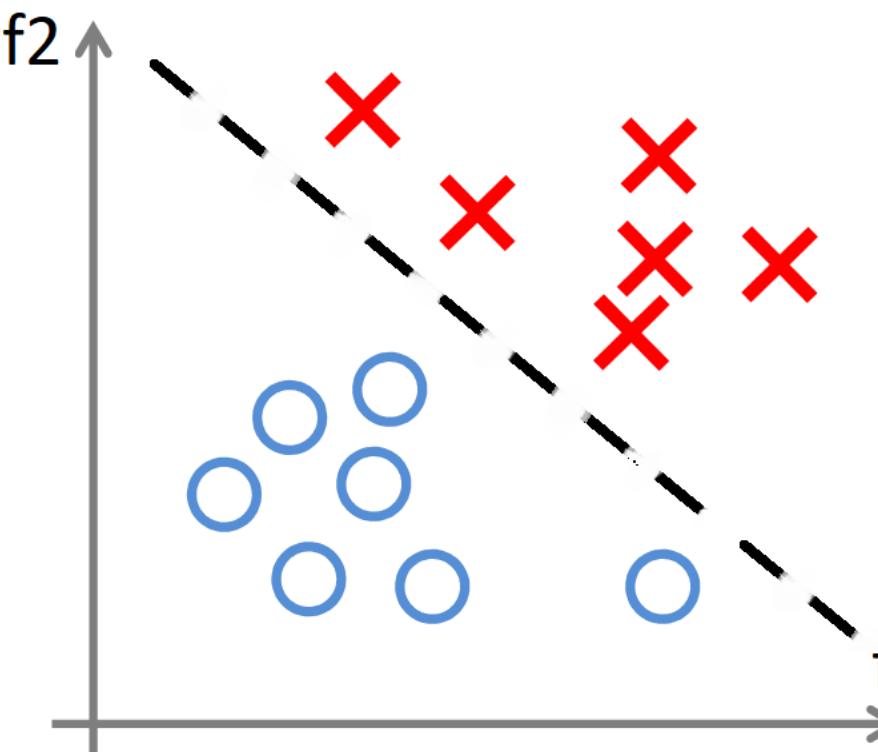
$$a(x, w) = \text{sign}\left(\sum_{j=1}^l w_j x_j\right)$$

Уравнение

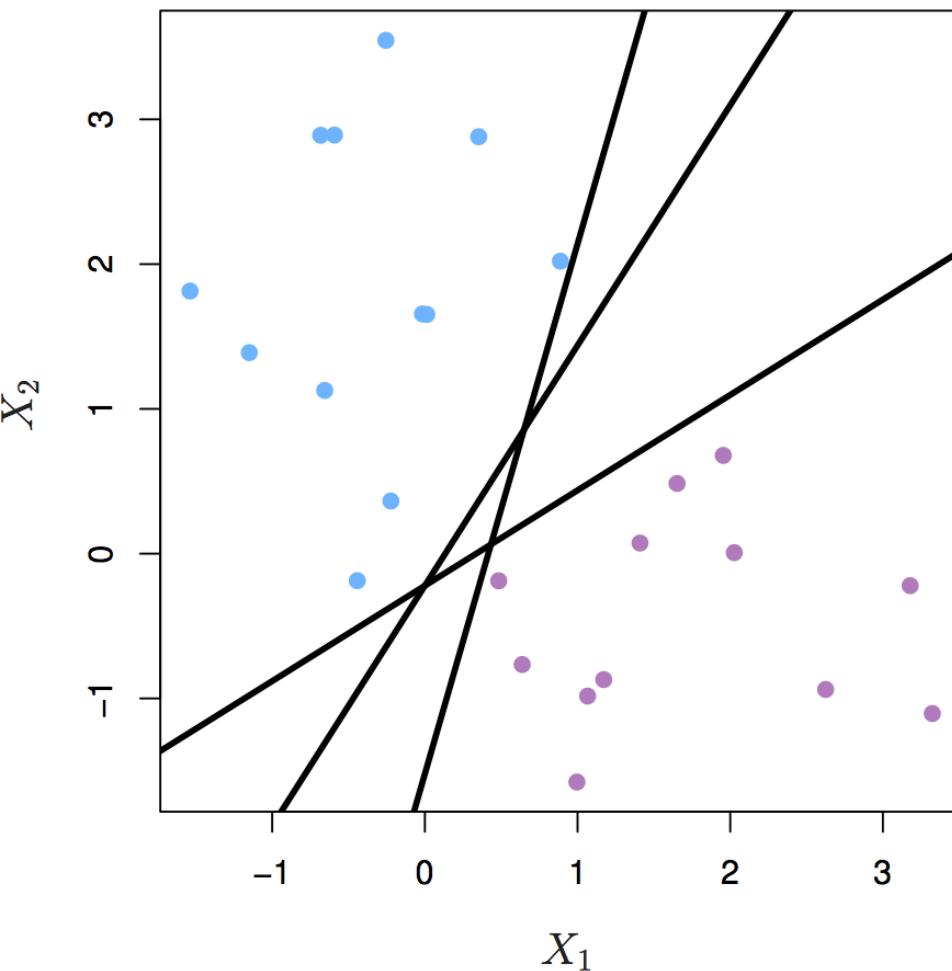
$$\sum_{j=1}^l w_j x_j = 0$$

– уравнение плоскости

(или прямой).



КАК ВЫБРАТЬ ПРЯМУЮ?



ОБУЧЕНИЕ КЛАССИФИКАТОРА

- Обучение - минимизация доли ошибок классификатора:

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

где $[a(x_i) \neq y_i] = 1$, если предсказание на объекте неверное, то есть $a(x_i) \neq y_i$, и 0 иначе.

- Обозначим $M_i = y_i \cdot (w, x_i)$ - *отступ* на i -м объекте.

Утверждение. Решение задачи (*) эквивалентно решению задачи

$$Q(a, X) = \frac{1}{n} \sum_{i=1}^n [M_i < 0] \rightarrow \min$$

ДОКАЗАТЕЛЬСТВО УТВЕРЖДЕНИЯ

- Обучение - минимизация доли ошибок классификатора:

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

Функционал Q можно переписать в виде:

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

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

ОТСТУП (MARGIN)

Знак отступа $M = y \cdot (w, x)$ говорит о корректности классификации на объекте:

Случай неверной классификации (предсказание не совпадает с правильным ответом):

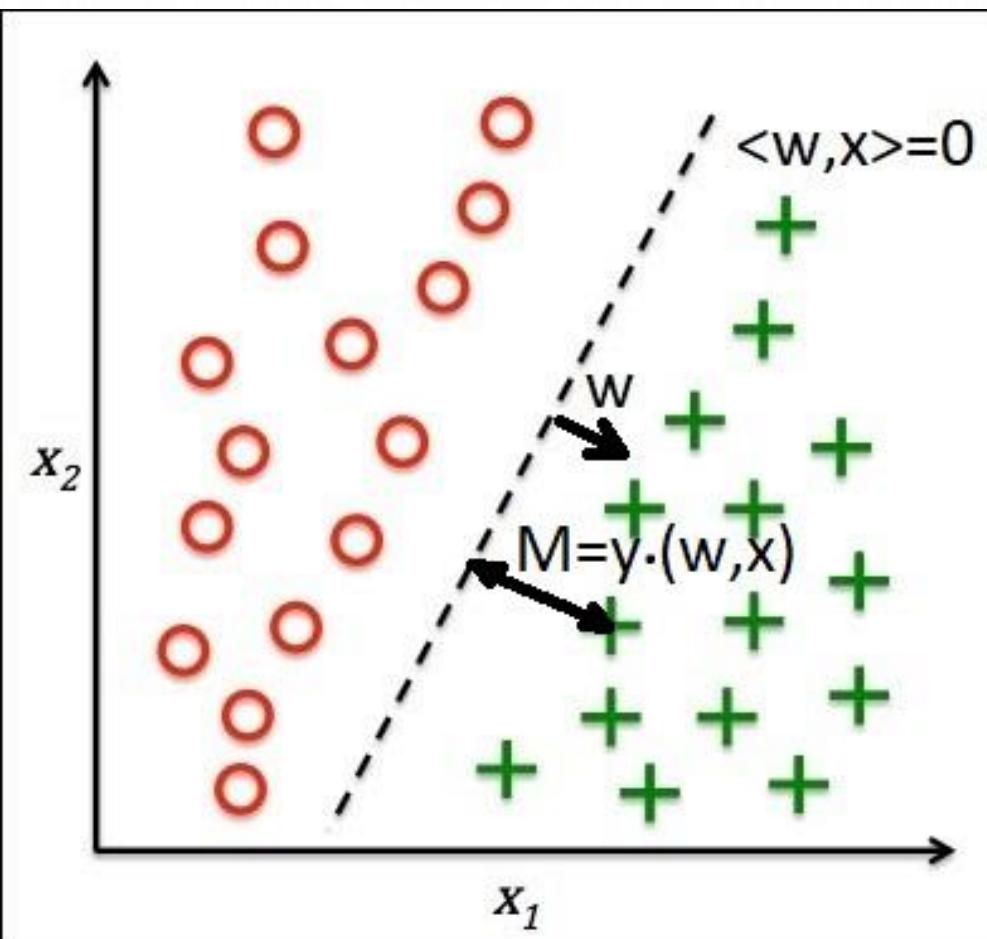
- Если $(w, x) > 0$ (то есть объект отнесён к классу +1), а $y = -1$, то $M = y \cdot (w, x) < 0$.
- Аналогично, если $(w, x) < 0$, а $y = +1$, то $M = y \cdot (w, x) < 0$.

Случай верной классификации:

- Если $(w, x) > 0$ и $y = +1$ или $(w, x) < 0$ и $y = -1$ получаем $M = y \cdot (w, x) > 0$.

ОТСТУП (MARGIN)

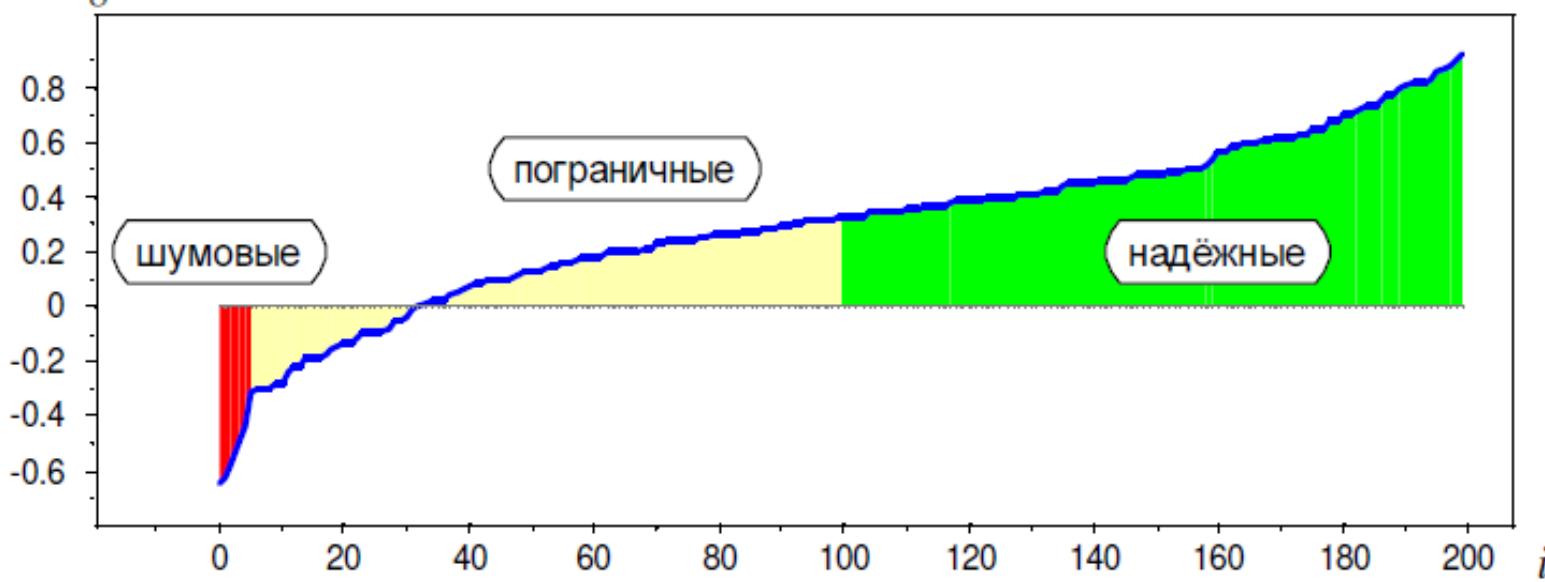
Абсолютная величина отступа M обозначает степень уверенности классификатора в ответе (чем ближе M к нулю, тем меньше уверенность в ответе)



ОТСТУП (MARGIN)

Ранжирование объектов по возрастанию отступа:

Margin

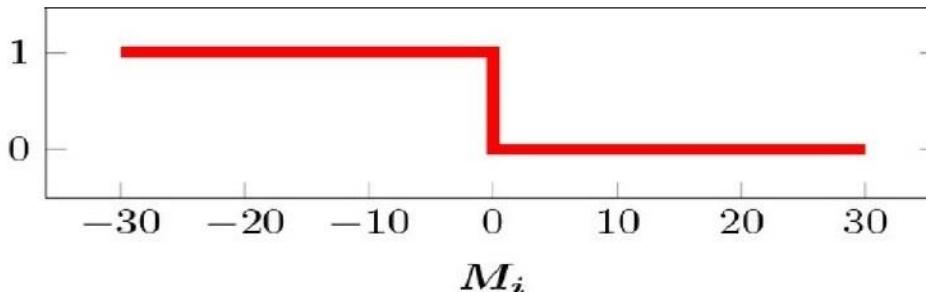


ПОРОГОВАЯ ФУНКЦИЯ ПОТЕРЬ

Ранее мы показали, что обучение классификатора – это минимизация *пороговой функции потерь*:

$$Q(a, X) = \frac{1}{n} \sum_{i=1}^n [M_i < 0] \rightarrow \min$$

- Пороговая функция потерь разрывна, и этот факт сильно затрудняет процесс минимизации.



- Для решения этой проблемы используют *другие функции потерь – непрерывные или гладкие, как правило, являющиеся верхними оценками пороговой функции*.

ПОРОГОВАЯ ФУНКЦИЯ ПОТЕРЬ

Ранее мы показали, что обучение классификатора – это минимизация *пороговой функции потерь*:

$$Q(a, X) = \frac{1}{n} \sum_{i=1}^n [M_i < 0] \rightarrow \min$$

- Пороговая функция потерь разрывна, и этот факт сильно затрудняет процесс минимизации.
- Для решения этой проблемы используют другие функции потерь – непрерывные или гладкие, как правило, являющиеся верхними оценками пороговой функции.
- Задача минимизации некоторой функции потерь называется *минимизацией эмпирического риска* (сама функция потерь – эмпирический риск).

ВЕРХНИЕ ОЦЕНКИ ЭМПИРИЧЕСКОГО РИСКА

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

Оценим

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

- Тогда

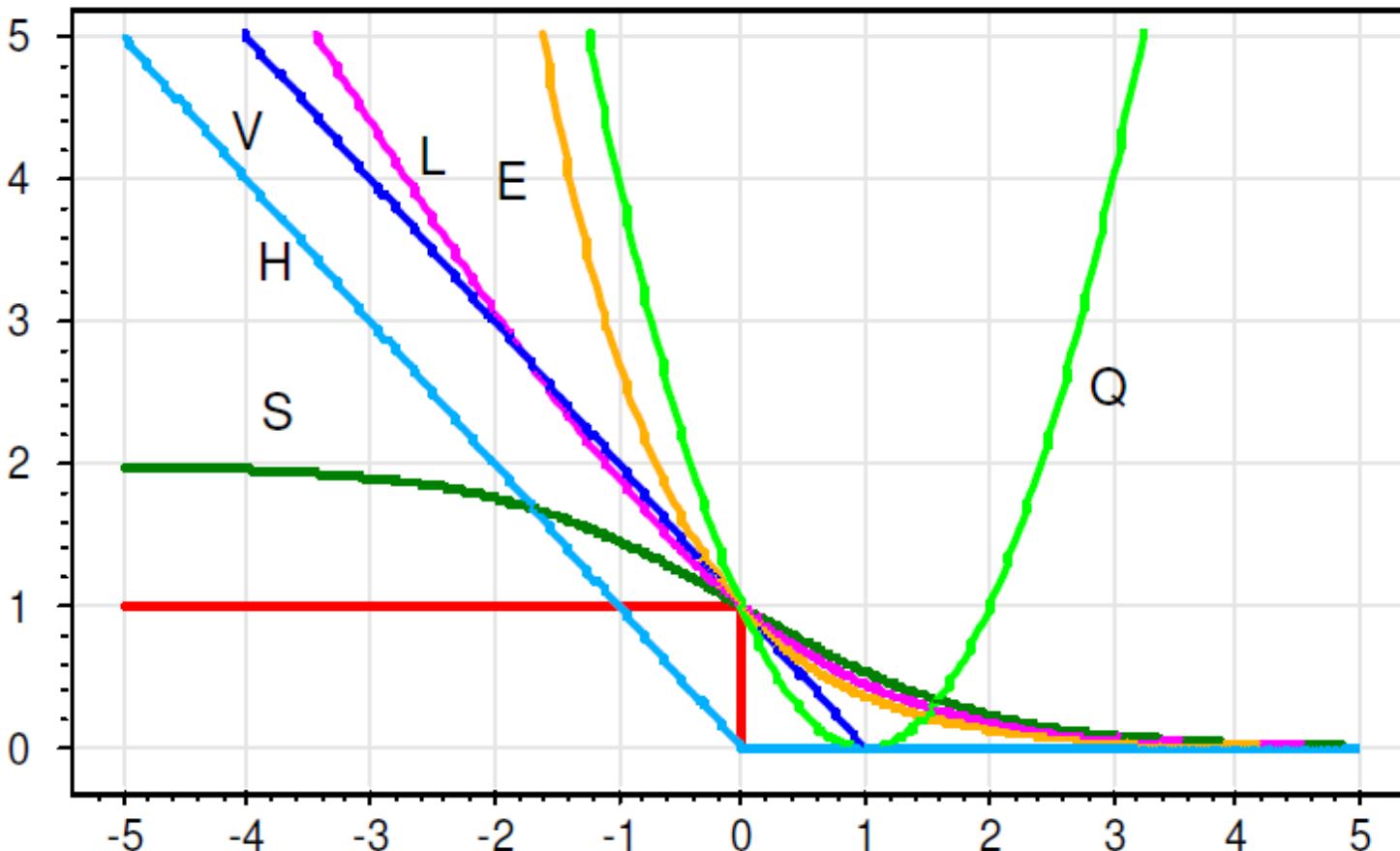
$$Q(a, X) = \frac{1}{n} \sum_{i=1}^n L(y_i \cdot (w, x_i)) \leq \frac{1}{n} \sum_{i=1}^n \tilde{L}(y_i \cdot (w, x_i)) \rightarrow \min$$

ФУНКЦИИ ПОТЕРЬ

Минимизируя различные функции потерь, получаем разные результаты. Поэтому разные функции потерь определяют различные классификаторы.

- $L(M) = \log(1 + e^{-M})$ – логистическая функция потерь
- $V(M) = (1 - M)_+ = \max(0, 1 - M)$ – кусочно-линейная функция потерь (метод опорных векторов)
- $H(M) = (-M)_+ = \max(0, -M)$ – 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



M

НАПРИМАНИЕ: ОПТИМИЗАЦИЯ ФУНКЦИОНАЛА ПОТЕРЬ

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

$$\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \eta \cdot \nabla Q(\mathbf{w}^{(k-1)})$$

А теперь давайте посмотрим на то, что мы будем кодить!

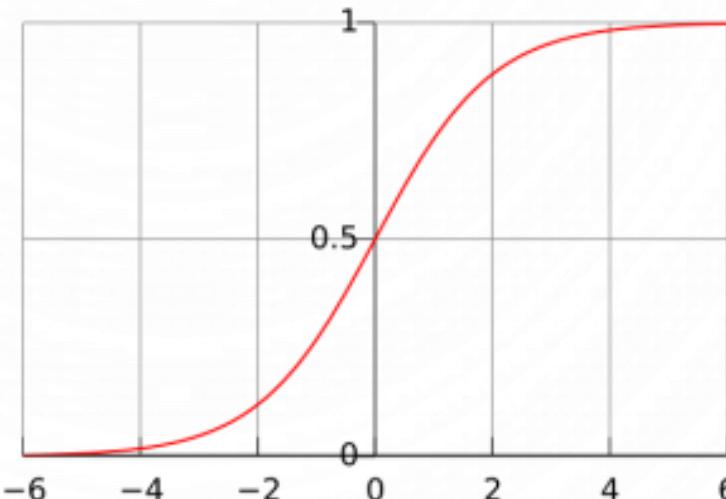
ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не классы, а вероятности классов.

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

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

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



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

ВЕРОЯТНОСТНЫЙ СМЫСЛ

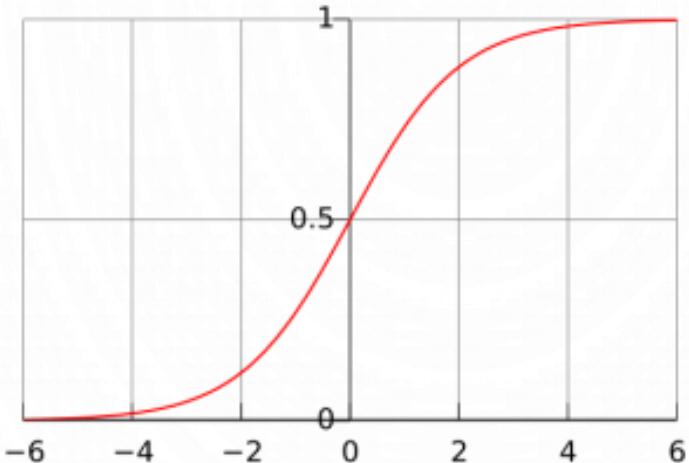
Утверждение. $a(x, w)$ – вероятность того, что $y = +1$ на объекте x , т.е.

$$a(x, w) = P(y = +1|x; w)$$

Доказательство. Через неделю 😊

РАЗДЕЛЯЮЩАЯ ГРАНИЦА

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



$$a(x, w) = \sigma(w^T x) \geq 0.5, \text{ если } w^T x \geq 0.$$

Получаем, что

- $y = +1$ при $w^T x \geq 0$
- $y = -1$ 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

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

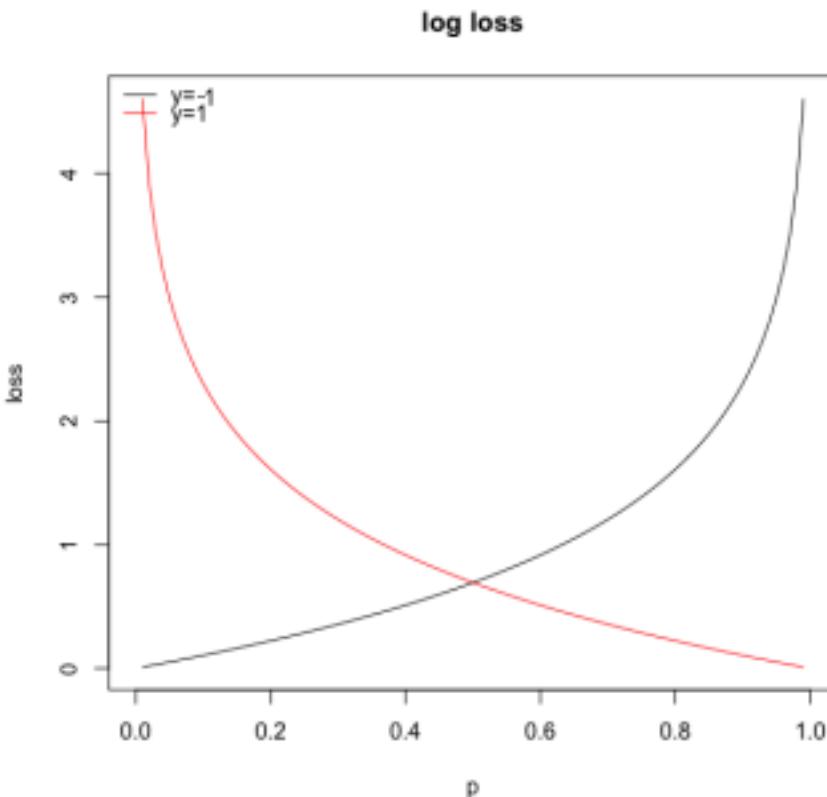
то возникнут проблемы:

- $Q(a, X) = \frac{1}{l} \sum_{i=1}^l \left(\frac{1}{1+e^{-w^T x_i}} - y_i \right)^2$ - 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 $(1 - 0)^2 = 1$)

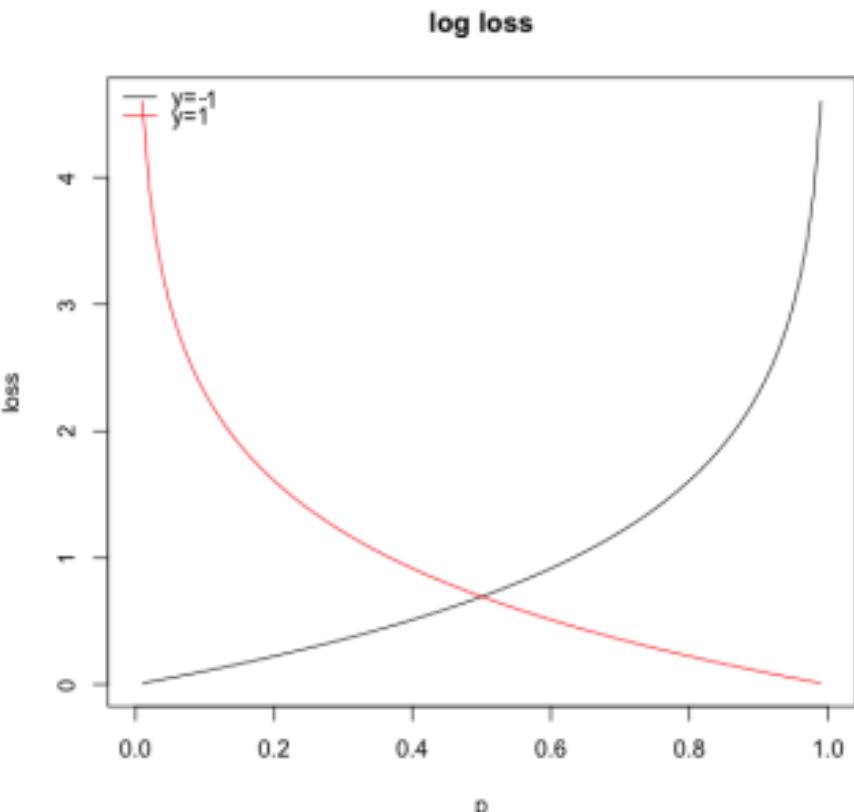
ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

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

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



ЛОГИСТИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ



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

МЕТРИКИ КАЧЕСТВА

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

- Accuracy – доля правильных ответов:

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

*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

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	
	20	80	
$a(x) = -1$ Не получили кредит			

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	
	52	98	
$a(x) = -1$ Не получили кредит			

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ: PRECISION, RECALL

- Precision (точность):

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

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

- Recall (полнота):

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

Показывает, как много объектов положительного класса находит классификатор.

RECALL: ПРИМЕР

Модель $a_1(x)$:

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

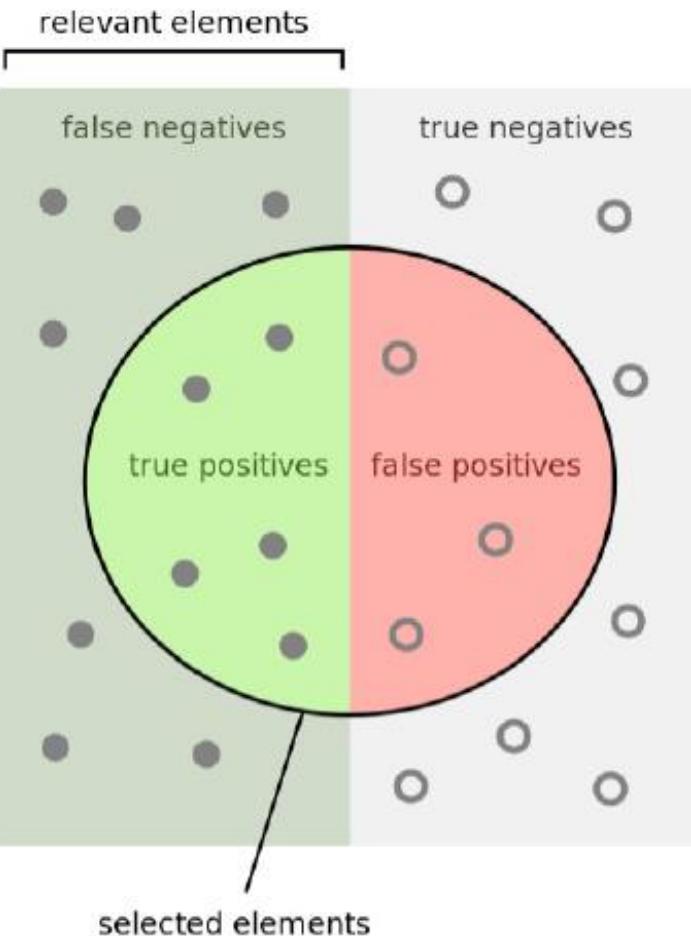
	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$a(x) = 1$ Получили кредит	80	20
$a(x) = -1$ Не получили кредит	20	80

Модель $a_2(x)$:

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

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

ТОЧНОСТЬ И ПОЛНОТА



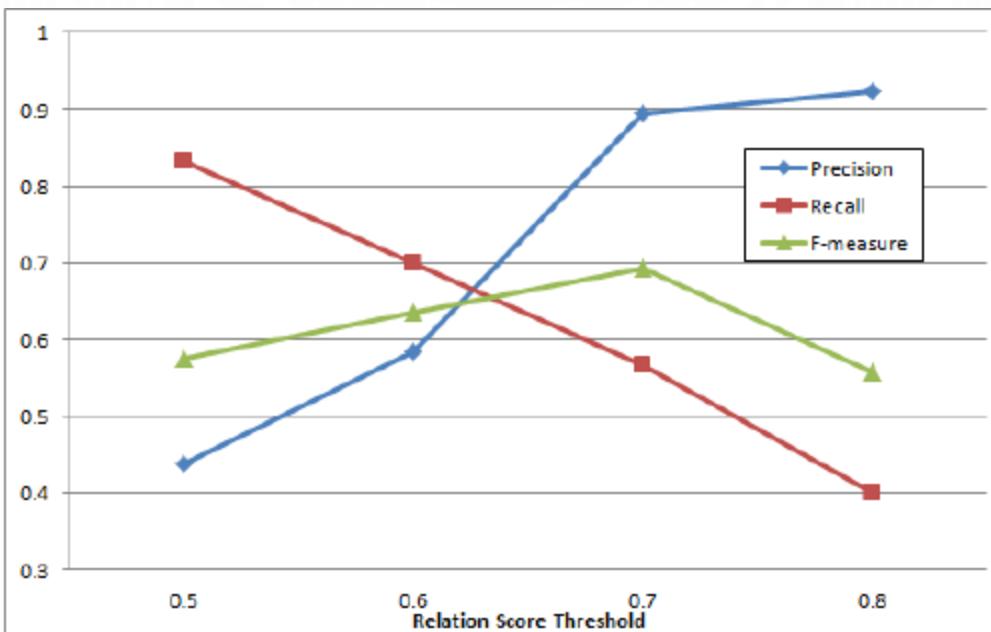
$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

F-МЕРА

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

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



РЕГУЛИРУЕМ ТОЧНОСТЬ И ПОЛНОТУ

Пусть $p(x)$ - уверенность классификатора в том, что объект x относится к классу +1, $p(x) \in [0; 1]$.

Обычно если $p(x) > 0.5$, то мы относим объект к положительному классу, а иначе – к отрицательному.

Можно изменять этот порог, то есть вместо 0.5 брать другое число из отрезка $[0; 1]$.

РЕГУЛИРУЕМ ТОЧНОСТЬ И ПОЛНОТУ

Пусть $p(x)$ - уверенность классификатора в том, что объект x относится к классу +1, $p(x) \in [0; 1]$.

Обычно если $p(x) > 0.5$, то мы относим объект к положительному классу, а иначе – к отрицательному.

Можно изменять этот порог, то есть вместо 0.5 брать другое число из отрезка $[0; 1]$.

Путем изменения порога t можно регулировать точность и полноту:

➤ Чему будут равны точность и полнота при $t = 0$?

РЕГУЛИРУЕМ ТОЧНОСТЬ И ПОЛНОТУ

Пусть $p(x)$ - уверенность классификатора в том, что объект x относится к классу +1, $p(x) \in [0; 1]$.

Обычно если $p(x) > 0.5$, то мы относим объект к положительному классу, а иначе – к отрицательному.

Можно изменять этот порог, то есть вместо 0.5 брать другое число из отрезка $[0; 1]$.

Путем изменения порога t можно регулировать точность и полноту:

- при $t = 0$ мы все объекты относим к положительному классу, то есть полнота = 1, а точность маленькая.
- **При увеличении t полнота уменьшается** (могут появиться объекты положительного класса, которые мы не нашли), **а точность возрастает** (появляются объекты положительного класса).

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-КРИВАЯ

Для каждого значения порога t вычислим:

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

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

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

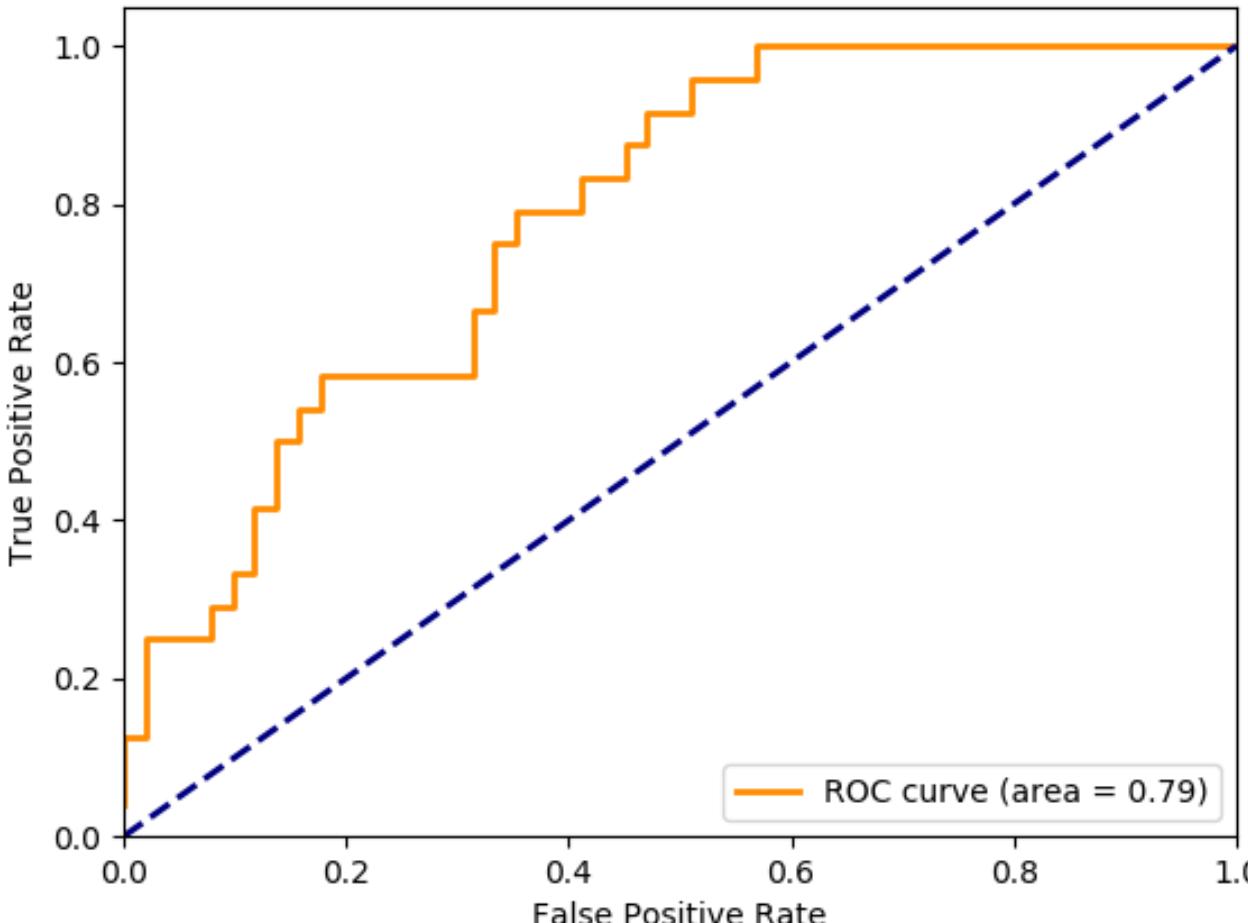
$$TPR = \frac{TP}{TP+FN} = \frac{\sum_i [y_i = +1] [a(x_i) = +1]}{\sum_i [y_i = +1]}.$$

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

ROC-КРИВАЯ

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

Receiver operating characteristic example

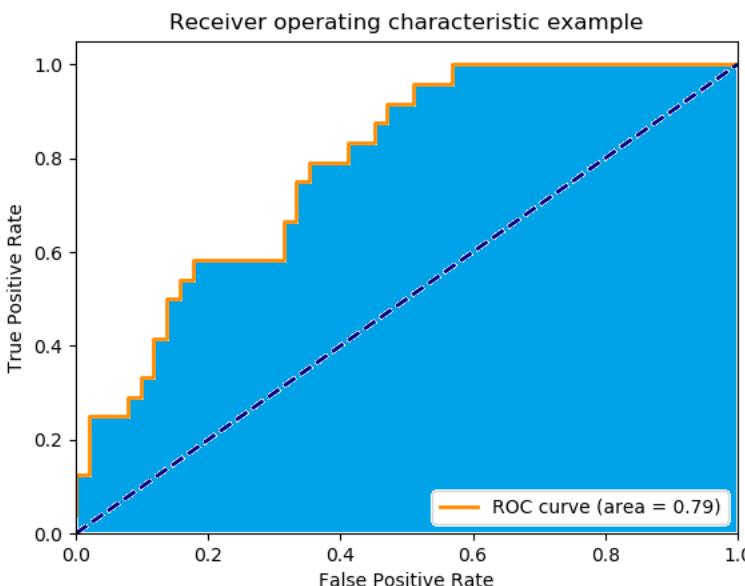


ROC-КРИВАЯ. AUC.

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.

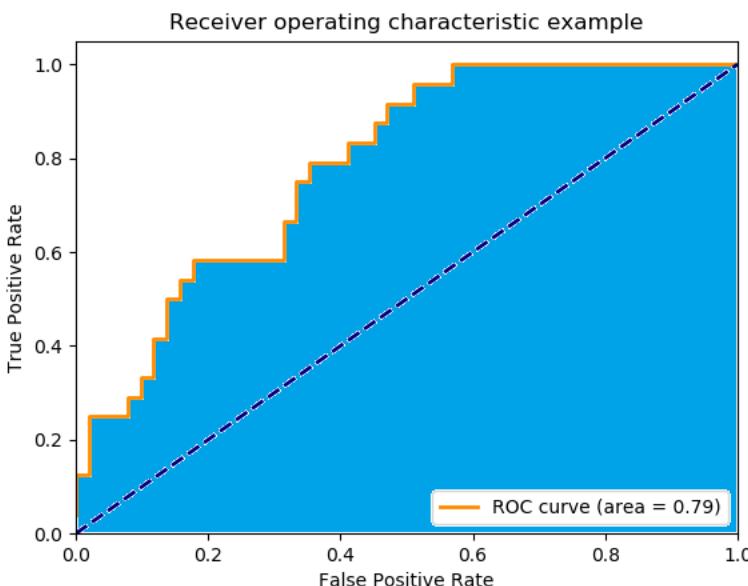
$$AUC \in [0; 1].$$

- $AUC = 1$ –

иdealная 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

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

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

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

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

ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

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

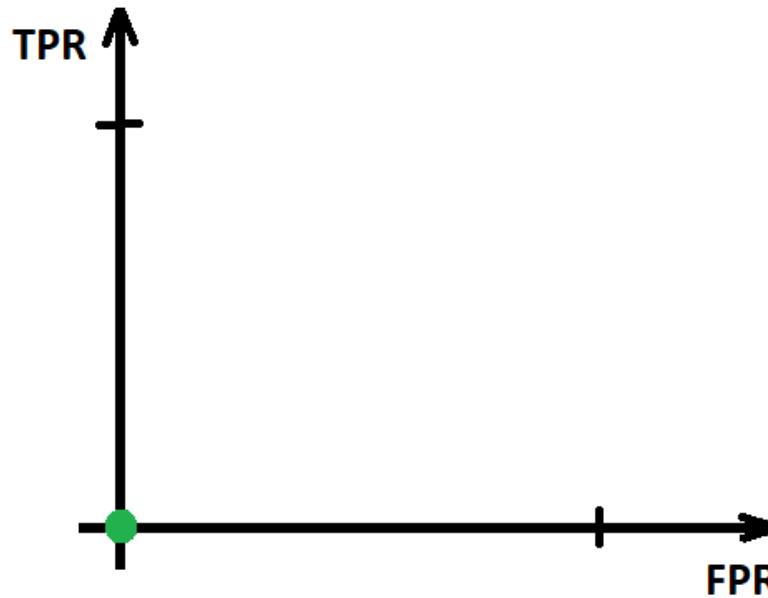
$b(x)$	0.2	0.4	0.1	0.7	0.05
y	-1	+1	-1	+1	+1

- Упорядочим объекты по убыванию предсказаний:
 $(0.7, 0.4, 0.2, 0.1, 0.05)$

1 шаг: $t = 0.7$, то есть
 $a(x) = [b(x) > 0.7]$

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

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



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:

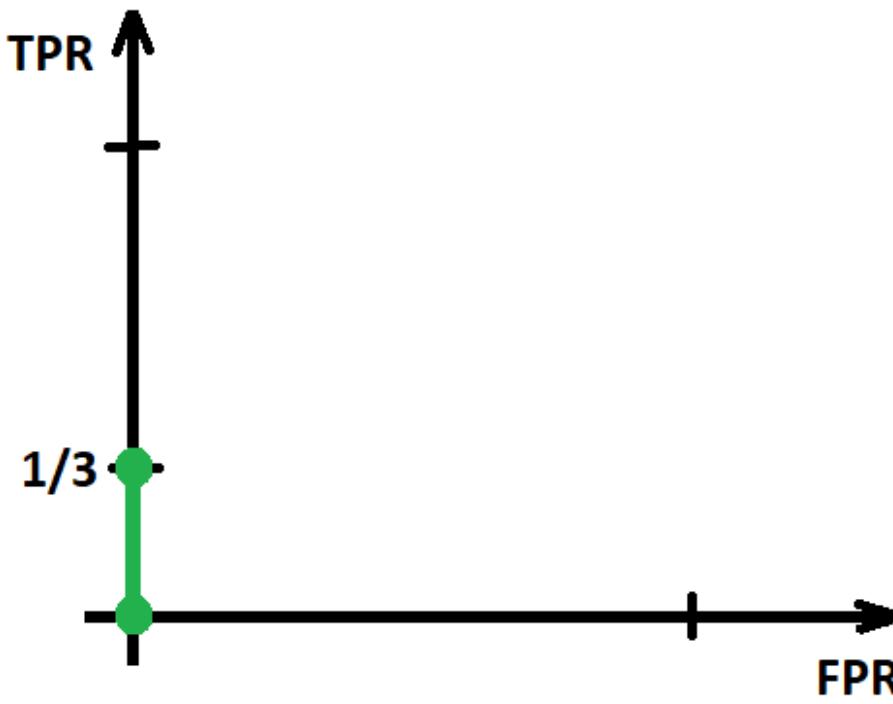
(0.7, 0.4, 0.2, 0.1, 0.05)

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:

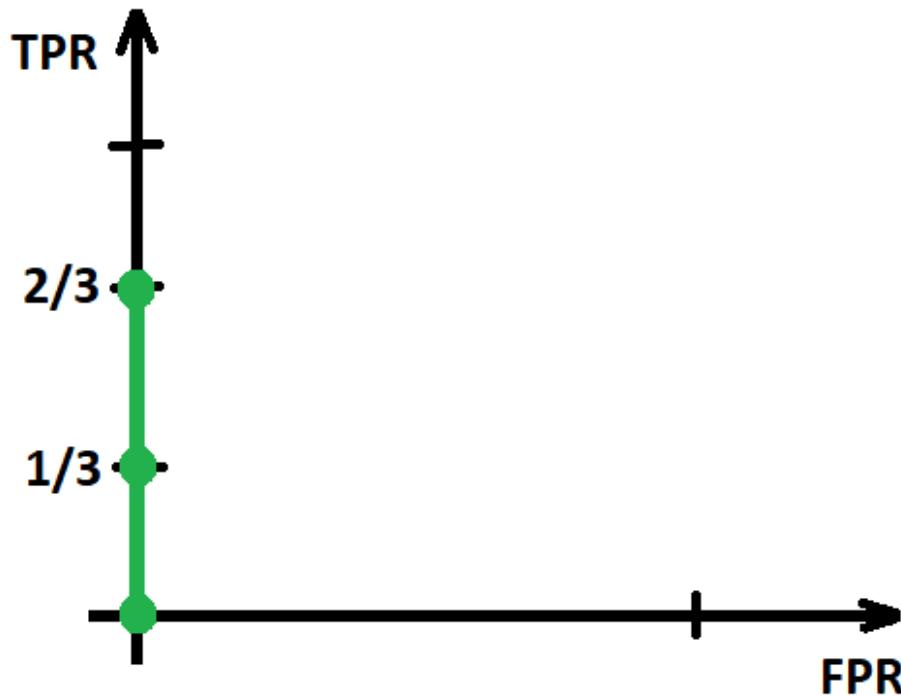
$$(0.7, 0.4, 0.2, 0.1, 0.05)$$

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:

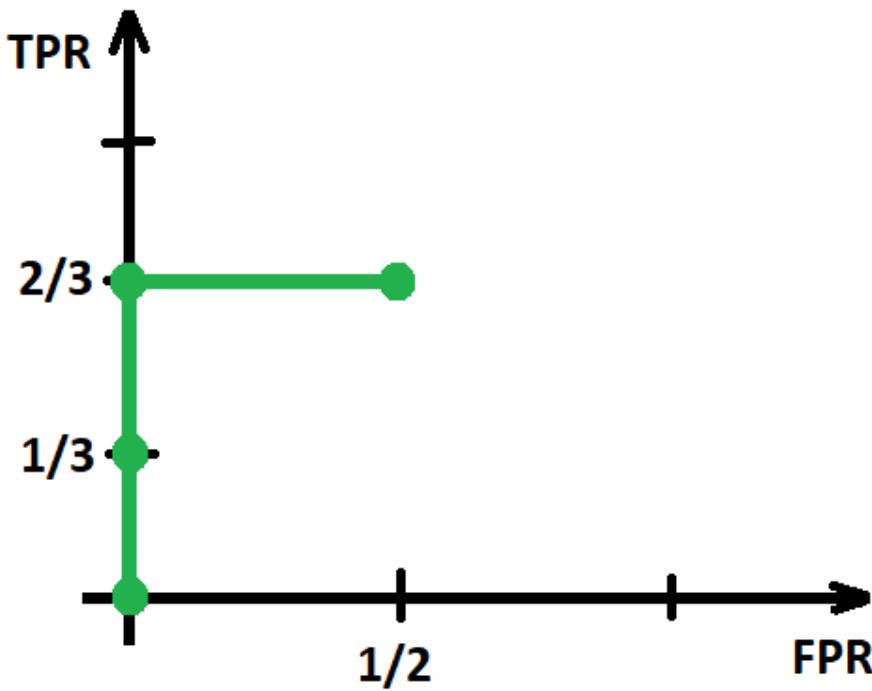
$b(x)$	0.2	0.4	0.1	0.7	0.05
y	-1	+1	-1	+1	+1

- Упорядочим объекты по убыванию предсказаний:
 $(0.7, 0.4, 0.2, 0.1, 0.05)$

**4 шаг: $t = 0.1$, 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:

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

- Упорядочим объекты по

убыванию предсказаний:

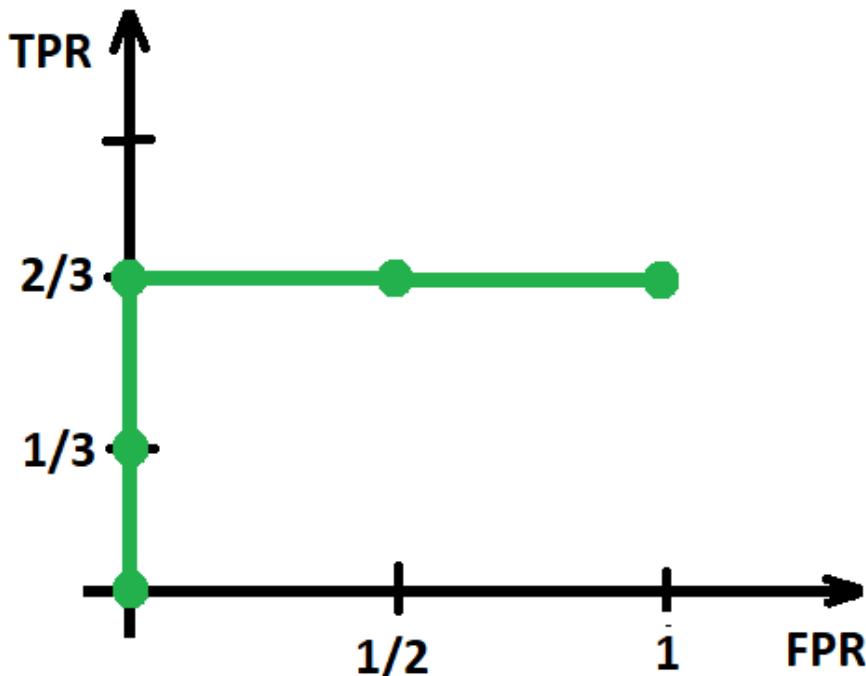
(0.7,0.4,0.2,0.1,0.05)

5 шаг: $t = 0.05$, 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:

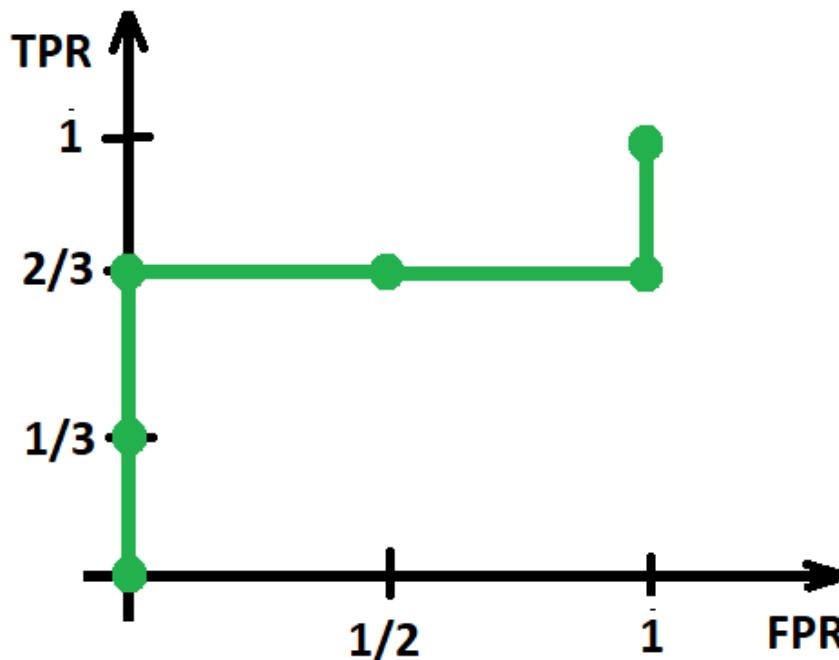
$b(x)$	0.2	0.4	0.1	0.7	0.05
y	-1	+1	-1	+1	+1

- Упорядочим объекты по убыванию предсказаний:
 $(0.7, 0.4, 0.2, 0.1, 0.05)$

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

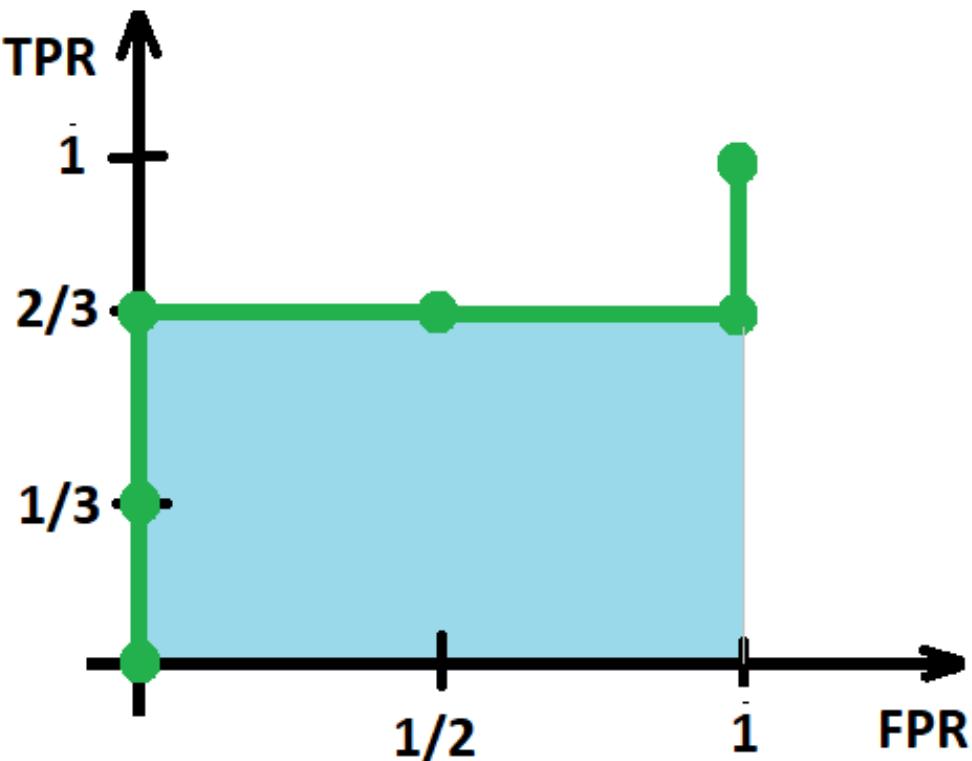
$$TPR = \frac{3}{3+0} = 1,$$

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



ПРИМЕР ПОСТРОЕНИЯ ROC-КРИВОЙ

$$AUC = 2/3$$

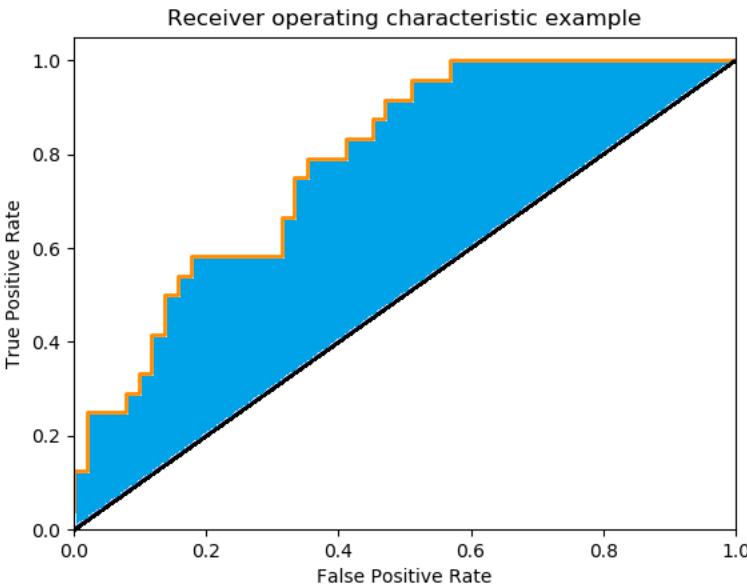


ИНДЕКС ДЖИНИ

Индекс Джини:

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

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

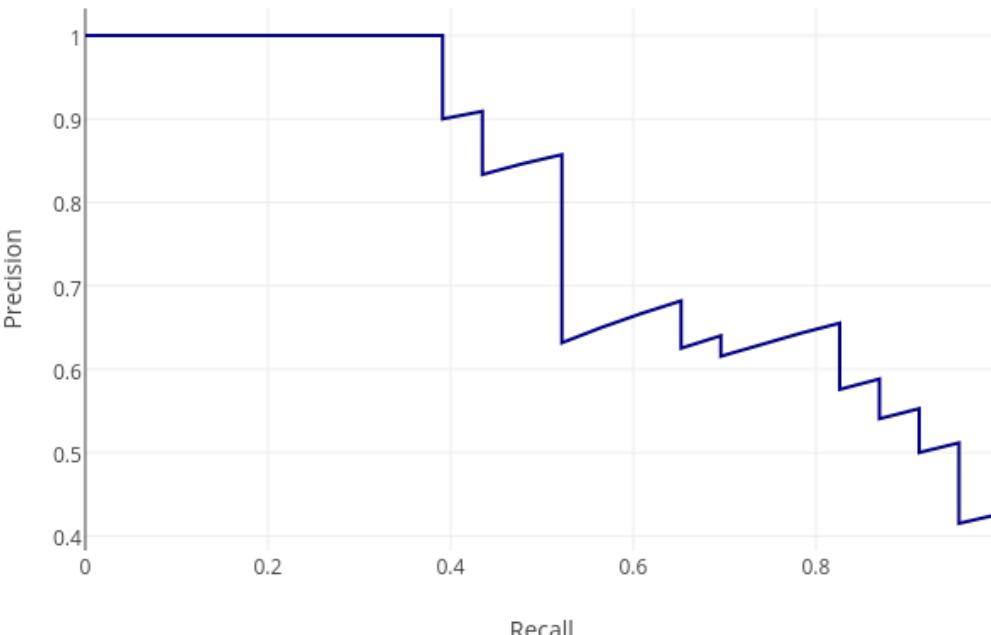


PRECISION-RECALL КРИВАЯ

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

Precision-Recall кривая:

Precision-Recall example: AUC=0.79



AUC-PR

AUC-PR – площадь под PR-кривой

Precision-Recall example: AUC=0.79

