Занятие 4
Линейные модели классификации.

Елена Кантонистова

elena.kantonistova@yandex.ru

ВШЭ, 2020

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ (НАПОМИНАНИЕ)

Обучающая выборка:

$x_1, x_2, \dots, x_n$ - объекты, $y_1, y_2, \dots, y_n$ - ответы (***любые числа***).

- Модель линейной регрессии:

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

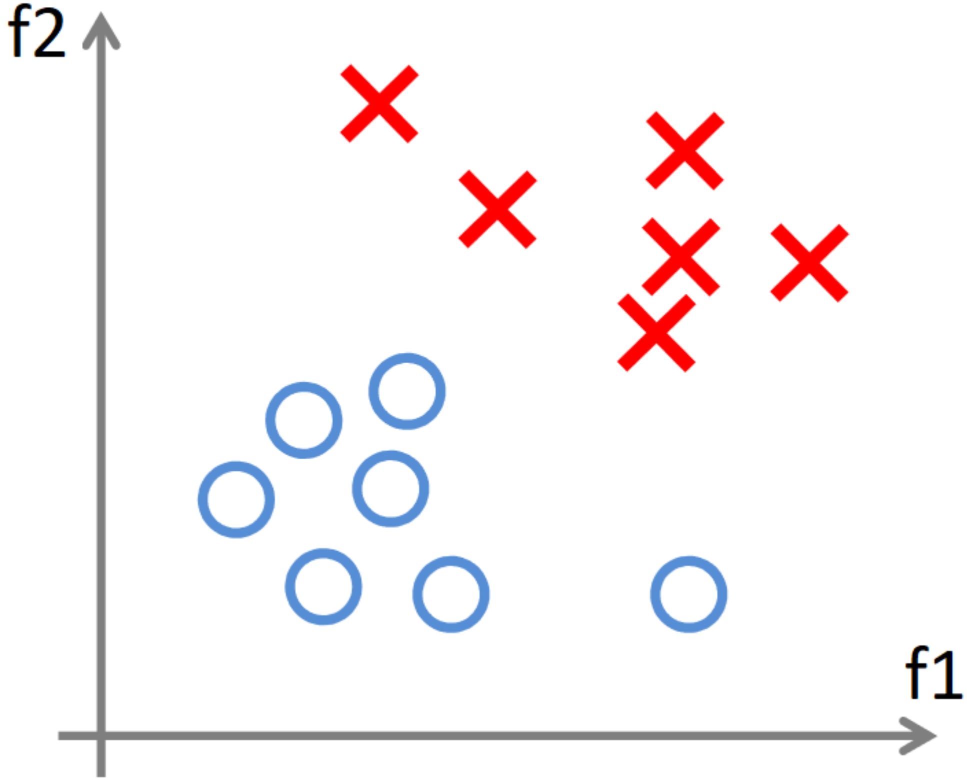
- Метод обучения – метод наименьших квадратов (***минимизируем разность между предсказанием и правильным ответом***):

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

БИНАРНАЯ КЛАССИФИКАЦИЯ

Обучающая выборка:

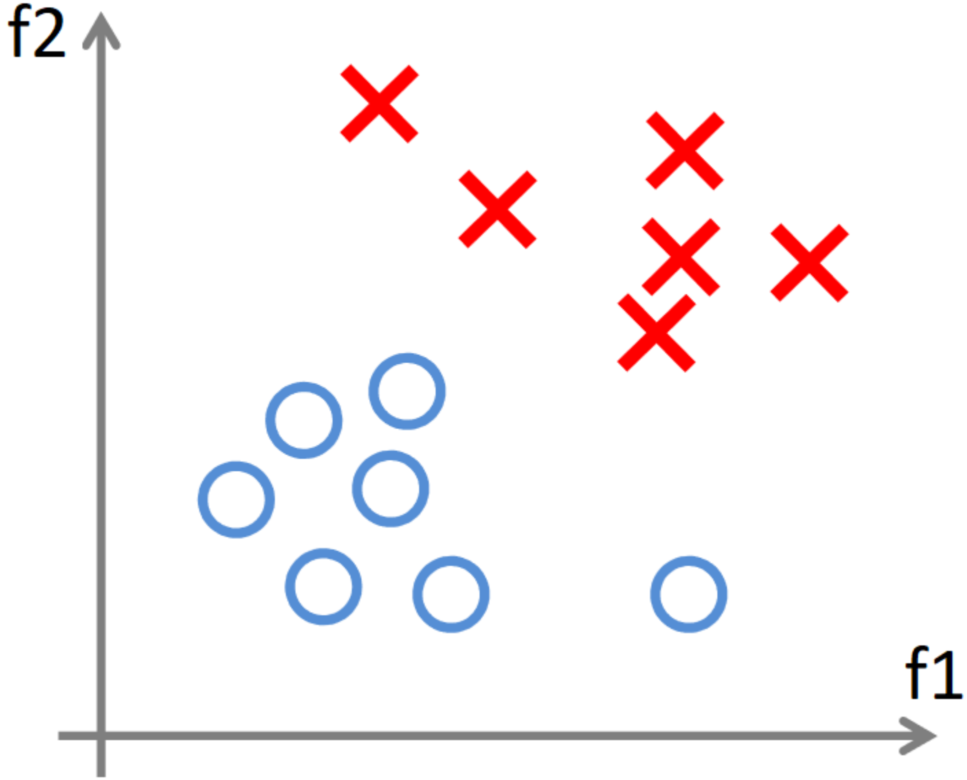
$x_1, x_2, \dots, x_n$ - объекты, $y_1, y_2, \dots, y_n$ - ответы (***+1 или -1***).



БИНАРНАЯ КЛАССИФИКАЦИЯ

Обучающая выборка:

$x_1, x_2, \ldots, x_n$ - объекты, $y_1, y_2, \ldots, y_n$ - ответы (***+1 или -1***).



Как выглядит модель линейного классификатора: $\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{w}) = ?$

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

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

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

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

- если $\sum_{j=1}^{n} w_j x_j > 0$, то $sign(\sum_{j=1}^{n} w_j x_j) = +1$, то есть объект отнесён к положительному классу
- если $\sum_{j=1}^{n} w_j x_j < 0$, то $sign(\sum_{j=1}^{n} w_j x_j) = -1$, то есть объект отнесён к отрицательному классу

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

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

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

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

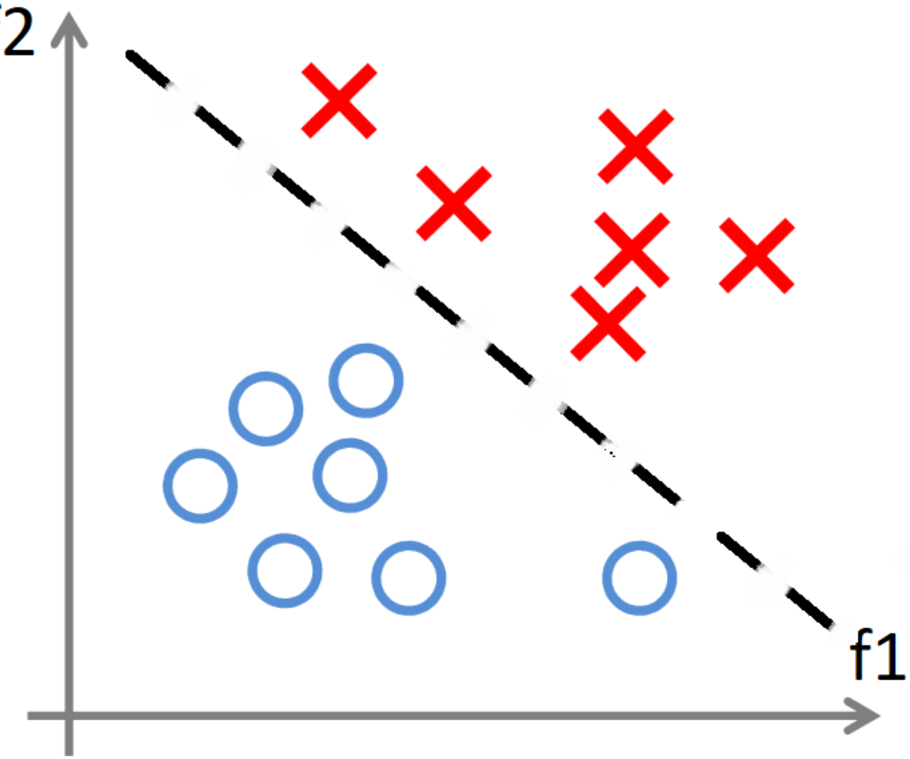
$$a(x,w) = sign\left(\sum_{j=1}^{n} w_j x_j\right)$$

Уравнение

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

– уравнение плоскости

(или прямой).



ОБУЧЕНИЕ КЛАССИФИКАТОРА

Как обучить линейный классификатор?

ОБУЧЕНИЕ КЛАССИФИКАТОРА

- Обучение - минимизация доли ошибок классификатора:

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[\boldsymbol{a}(\boldsymbol{x}_i) \neq \boldsymbol{y}_i] \rightarrow min\,,$$

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

Обучение классификатора

- Обучение - минимизация доли ошибок классификатора:

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[\boldsymbol{a}(\boldsymbol{x}_i) \neq \boldsymbol{y}_i] \to min \ \ (*),$$

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*.

ОБУЧЕНИЕ КЛАССИФИКАТОРА

- Обучение - минимизация доли ошибок классификатора:

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

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*.

Можно показать, что решение задачи (*) эквивалентно решению задачи

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

Отступ (margin)

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

ОТСТУП (MARGIN)

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

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

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

ОТСТУП (MARGIN)

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

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

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

$M = \mathrm{y} \cdot (w, x) < \mathbf{0}$.

Отступ (margin)

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

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

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

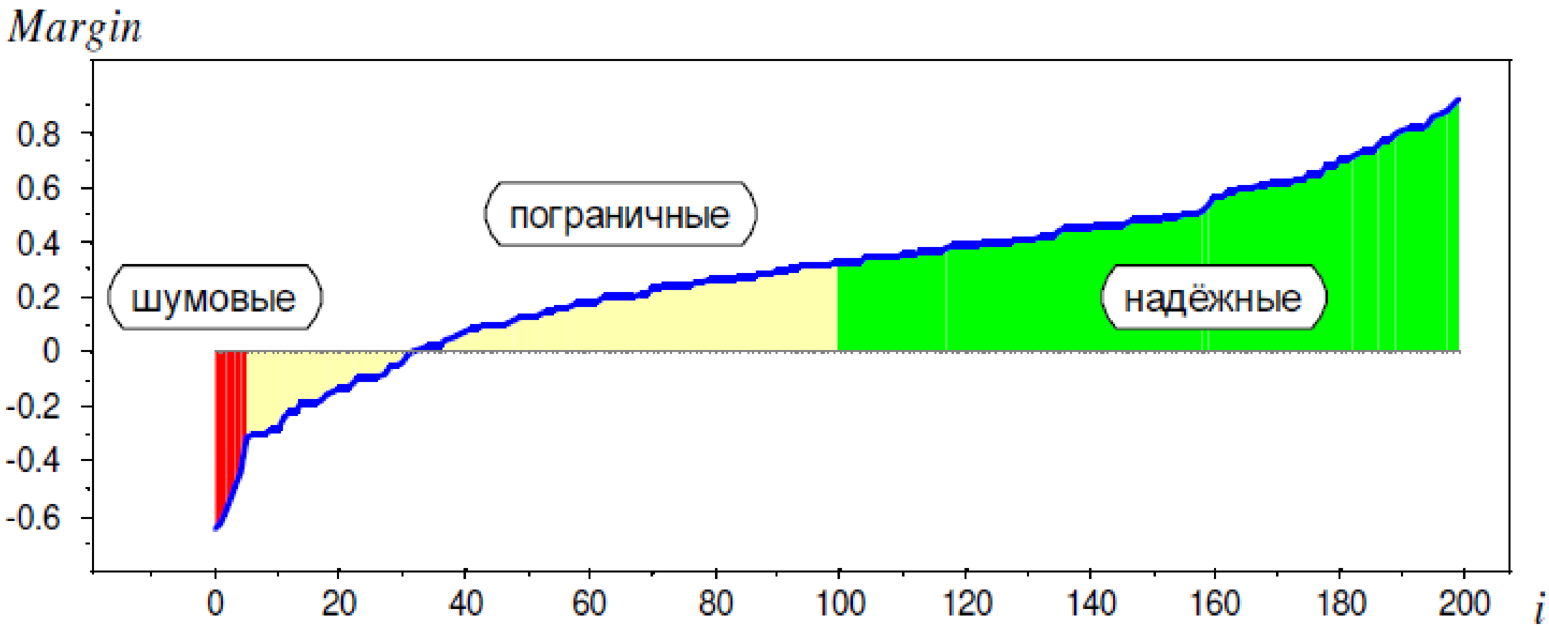
$M = \mathrm{y} \cdot (w, x) < \mathbf{0}$.

Случаи верной классификации:

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

ОТСТУП (MARGIN)

Ранжирование объектов по возрастанию отступа:



ВЕРХНИЕ ОЦЕНКИ ЭМПИРИЧЕСКОГО РИСКА

- $Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[\boldsymbol{M_i} < \mathbf{0}]$ – разрывная функция потерь, поэтому её минимум неудобно находить методом градиентного спуска (нельзя взять производную).

ВЕРХНИЕ ОЦЕНКИ ЭМПИРИЧЕСКОГО РИСКА

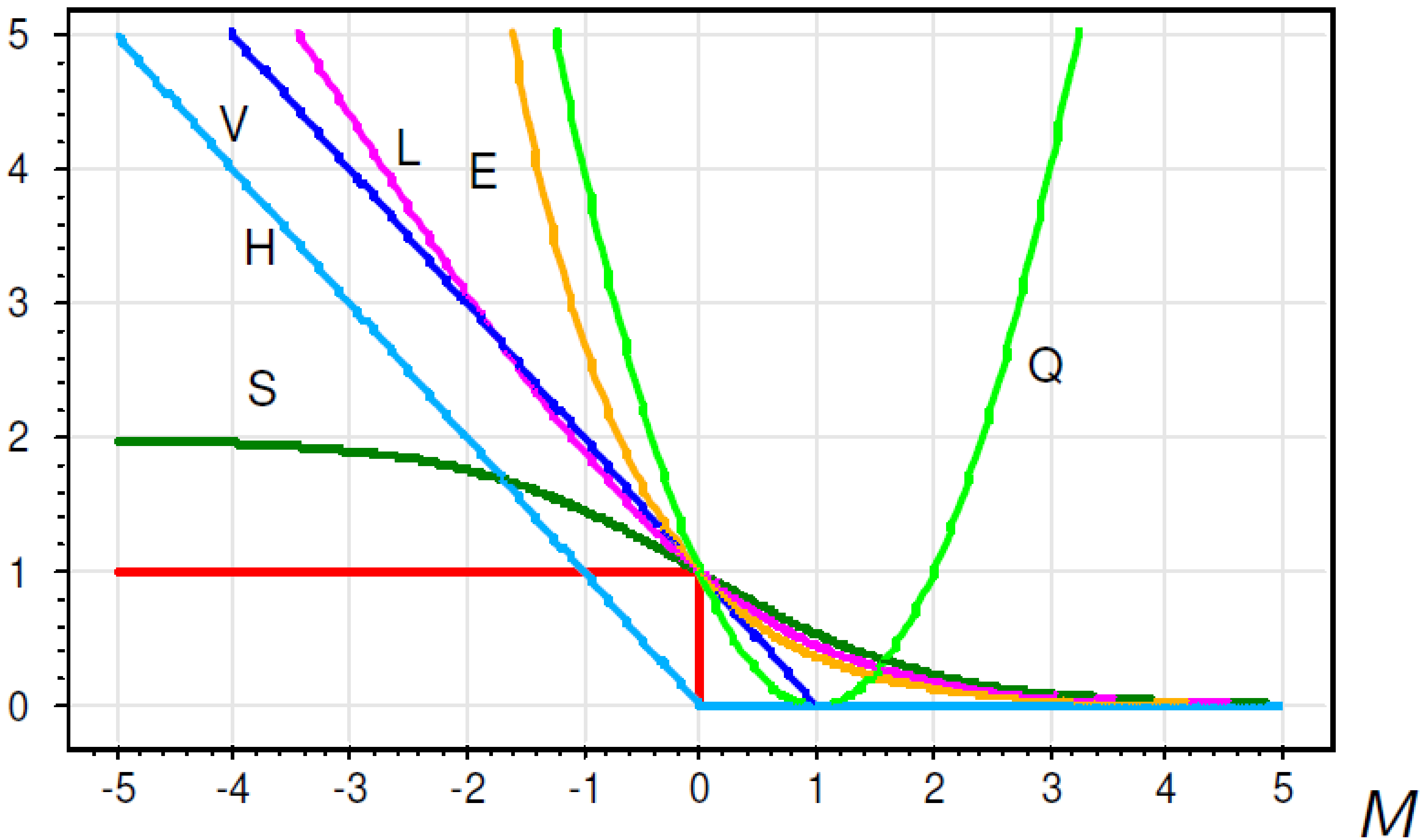
- $Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[\boldsymbol{M_i} < \mathbf{0}]$ – разрывная функция потерь, поэтому её минимум неудобно находить методом градиентного спуска (нельзя взять производную).
- Поэтому для решения задачи классификации *вместо разрывной функции $[M < 0]$ берут непрерывные или гладкие функции потерь $L(M)$*.

ФУНКЦИИ ПОТЕРЬ

Минимизируя различные функции потерь, получаем разные результаты. Поэтому разным функциям потерь соответствуют различные классификаторы.

- $L(M) = \log(1 + e^{-M})$ – логистическая функция потерь
- $\mathrm{V}(M) = (1 - \mathrm{M})_+ = \max(0{,}1 - M)$ – кусочно-линейная функция потерь (метод опорных векторов)
- $H(M) = (-\mathrm{M})_+ = \max(0, -M)$ – 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 $\boldsymbol{Q}$ происходит с помощью метода градиентного спуска:

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

МЕТРИКИ КАЧЕСТВА

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

- Accuracy – доля правильных ответов:

$$\boldsymbol{accuracy}(\boldsymbol{a}, \boldsymbol{X}) = \frac{\mathbf{1}}{\boldsymbol{l}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{l}} [\boldsymbol{a}(\boldsymbol{x}_{\boldsymbol{i}}) = \boldsymbol{y}_{\boldsymbol{i}}]$$

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

- Accuracy – доля правильных ответов:

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

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

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ: PRECISION, RECALL

- **Precision (точность):**

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

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
$\mathbf{a(x) = 1}$ Получили кредит	80	20
$\mathbf{a(x) = -1}$ 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
$\mathbf{a(x) = 1}$ Получили кредит	48	2
$\mathbf{a(x) = -1}$ 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 $a_1(x)$:

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

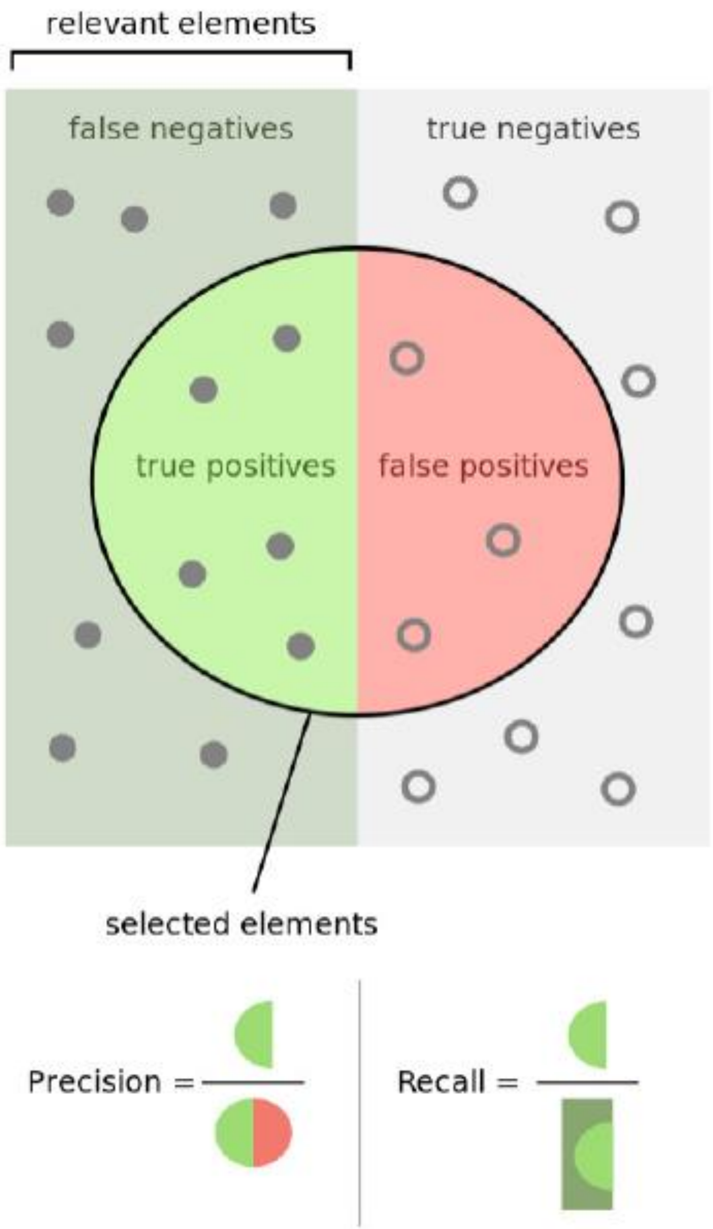
	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	80	20
$\mathbf{a(x) = -1}$ Не получили кредит	20	80

Модель $a_2(x)$:

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

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

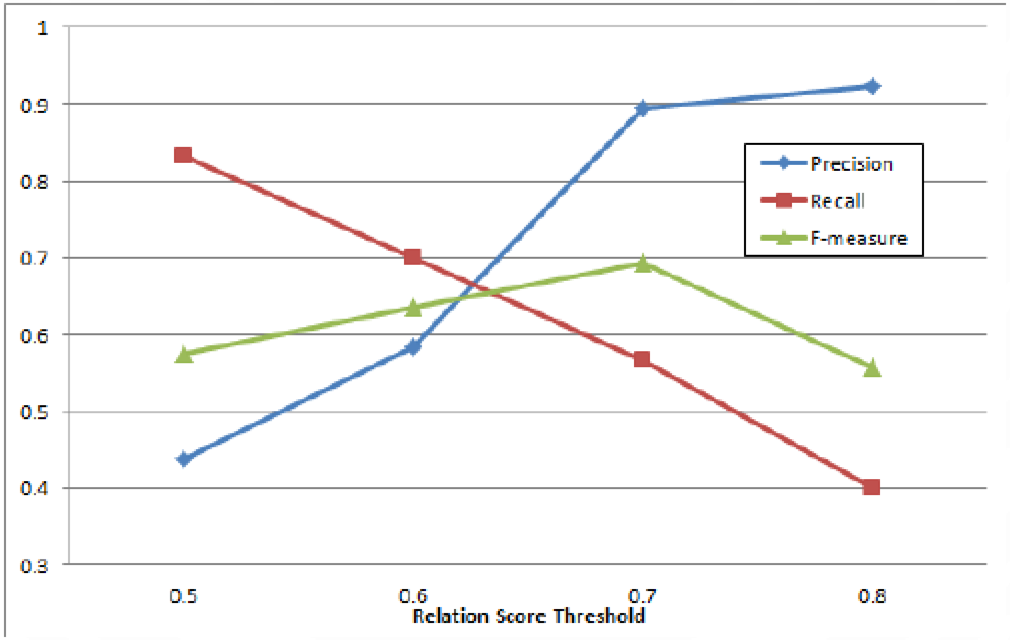
ТОЧНОСТЬ И ПОЛНОТА



F-МЕРА

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

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



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.

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:

- Например, если $t = 0$, то 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 возрастет (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

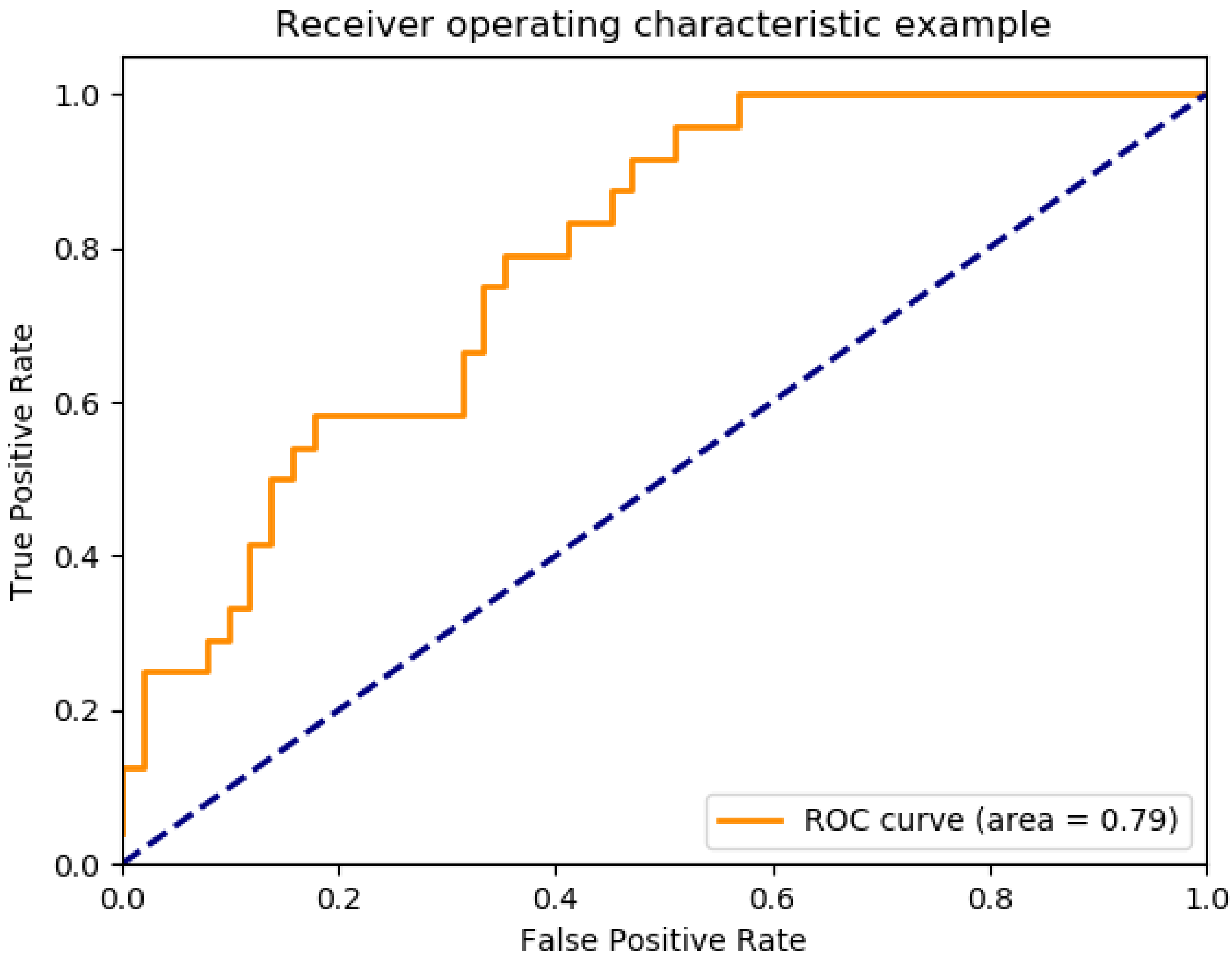
Для каждого значения порога t вычислим:

- False Positive Rate - долю неверно принятых объектов
- True Positive Rate -долю верно принятых объектов

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

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.

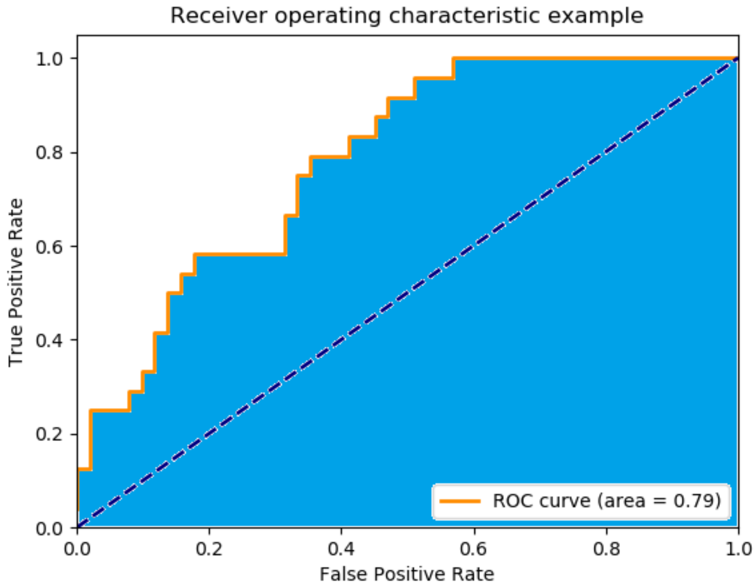
$$AUC \in [0; 1].$$

- $AUC = 1$ –

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

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

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

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

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

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:

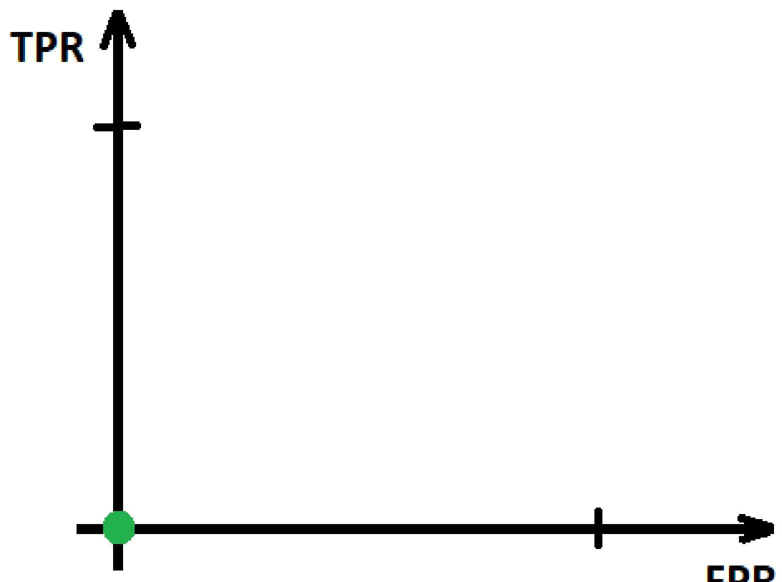
$(0.7, 0.4, 0.2, 0.1, 0.05)$

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

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

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

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



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

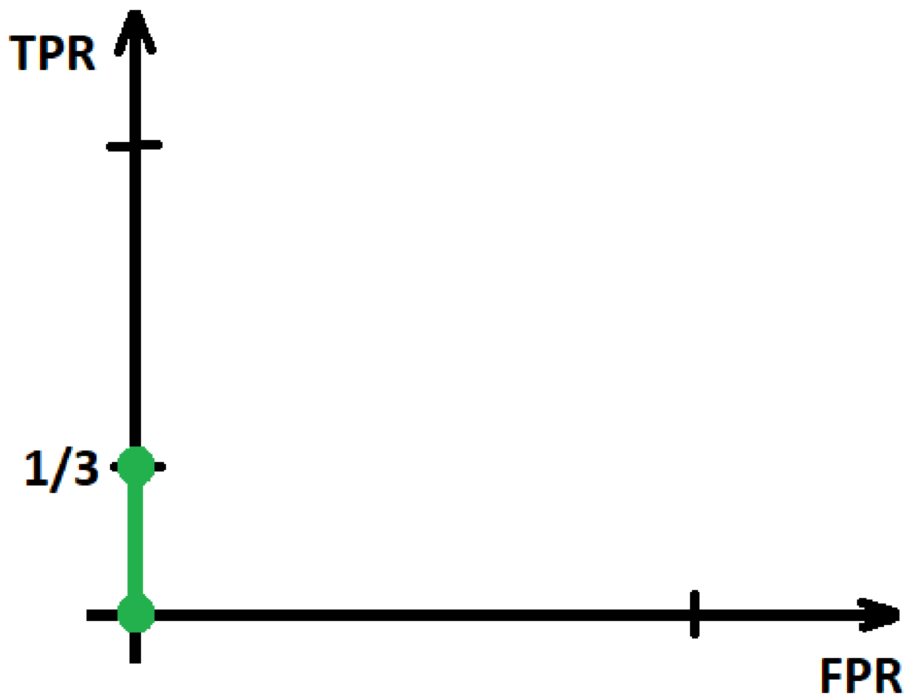
- Упорядочим объекты по

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	$+1$	$+1$

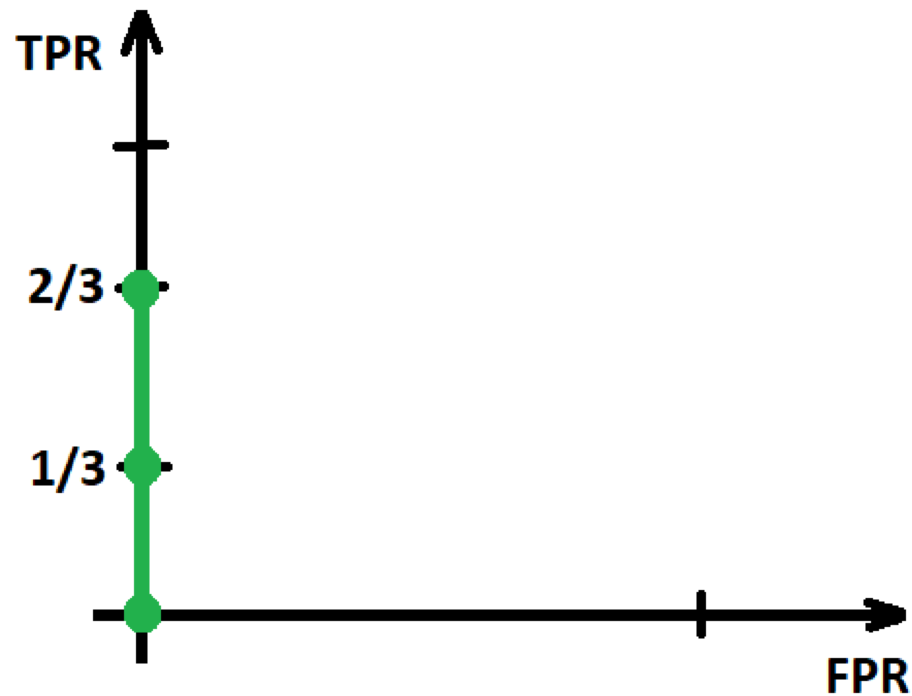
- Упорядочим объекты по

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

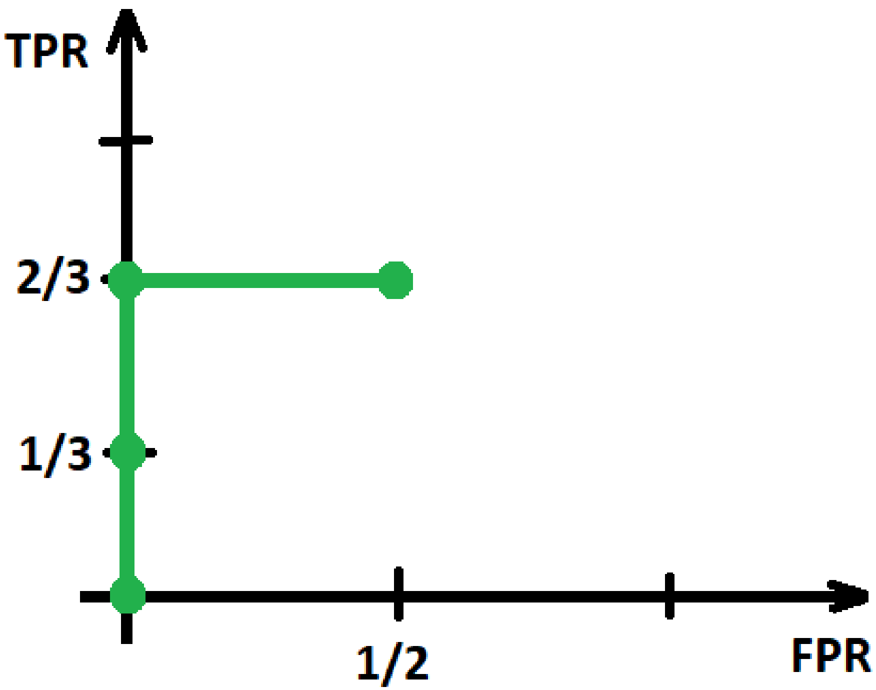
- Упорядочим объекты по

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

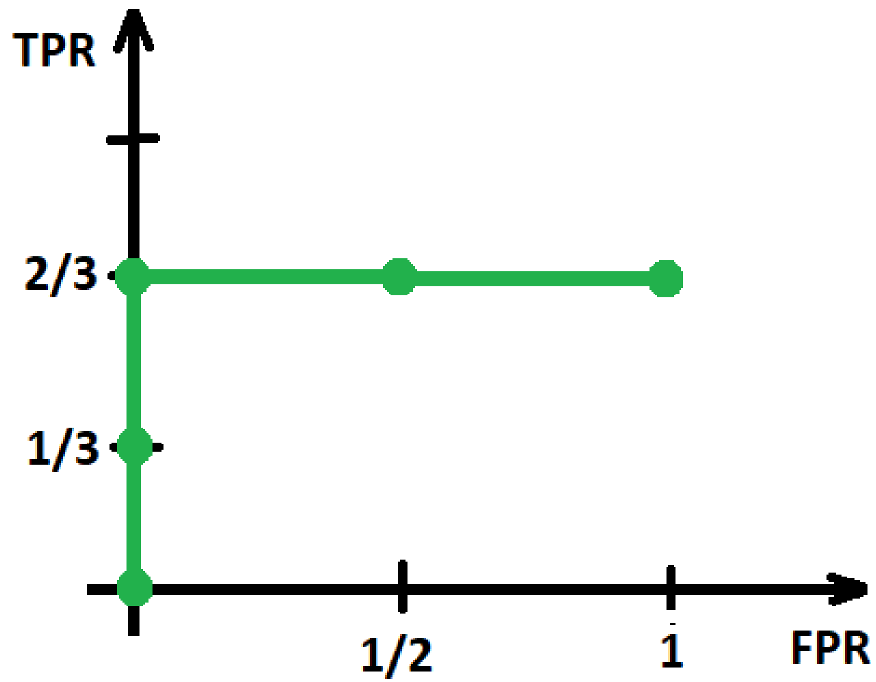
- Упорядочим объекты по

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	$+1$	$+1$

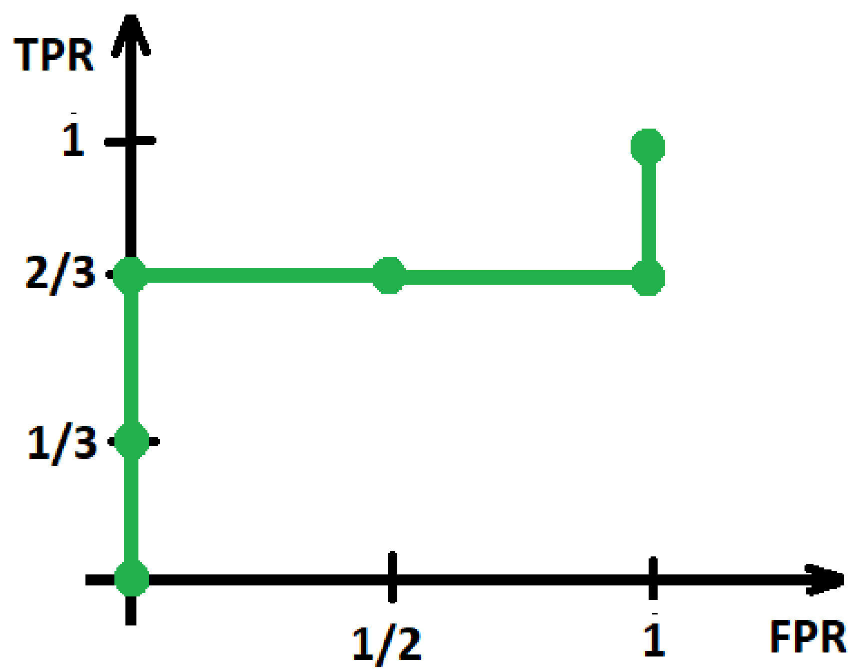
- Упорядочим объекты по

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

$$AUC = 2/3$$