Линейные методы классификации

Паточенко Евгений

НИУ ВШЭ

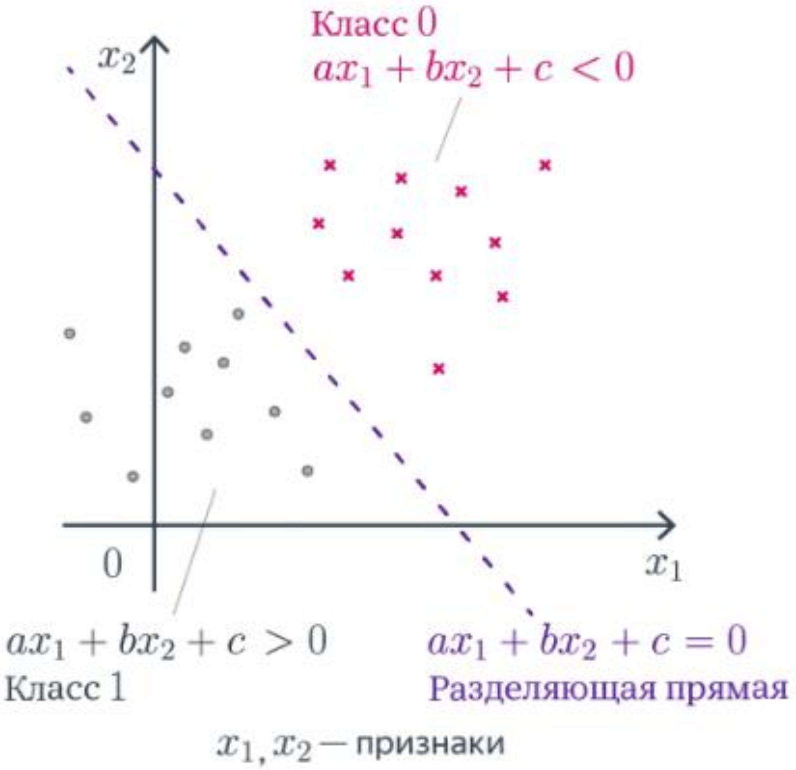
План занятия

- Линейные методы классификации (бинарная)
- Логистическая регрессия
- Метрики качества классификации

Линейные методы классификации (повтор)

Классификация

Модель машинного обучения, используемая для прогнозирования категориальной (дискретной) целевой переменной на основе одной или нескольких независимых переменных (признаков). Целевая переменная принимает конечное число классов или меток.



Может быть:

— Бинарной (классификация на два класса) $Y = \{0,1\}$

— Многоклассовой (классификация на M непересекающихся классов) $Y = \{1, \dots, M\}$

— Многоклассовой (классификация на M классов, которые могут пересекаться) $Y = \{0,1\}^M$

Линейные методы классификации (повтор)

Классификация

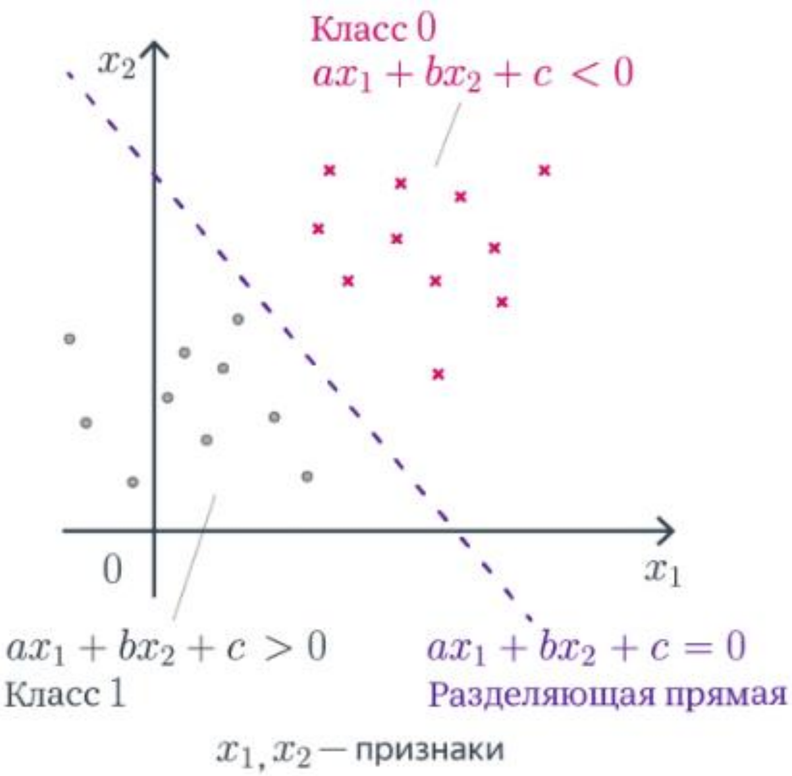
Модель машинного обучения, используемая для прогнозирования категориальной (дискретной) целевой переменной на основе одной или нескольких независимых переменных (признаков). Целевая переменная принимает конечное число классов или меток.

Может быть:

— Бинарной (классификация на два класса) $\boldsymbol{Y = \{0, 1\}}$

— Многоклассовой (классификация на M непересекающихся классов) $Y = \{1, \dots, M\}$

— Многоклассовой (классификация на M классов, которые могут пересекаться) $Y = \{0,1\}^M$



Бинарная классификация

Модель

$$f(x,\beta) = sign\left(\sum_{i=1}^{n} \beta_i x_i\right)$$

Бинарная классификация

Модель

$$f(x,\beta) = sign\left(\sum_{i=1}^{n} \beta_i x_i\right)$$

Если $\sum_{i=1}^{n} \beta_i x_i > 0, \text{то } sign(\sum_{i=1}^{n} \beta_i x_i) = +1$, объект относится к положительному классу

Если $\sum_{i=1}^{n} \beta_i x_i < 0, \text{то } sign(\sum_{i=1}^{n} \beta_i x_i) = -1$, объект относится к отрицательному классу

Бинарная классификация

Модель

$$f(x, \beta) = sign\left(\sum_{i=1}^{n} \beta_i x_i\right)$$

Если $\sum_{i=1}^{n} \beta_i x_i > 0$, то $sign(\sum_{i=1}^{n} \beta_i x_i) = +1$, объект относится к положительному классу

Если $\sum_{i=1}^{n} \beta_i x_i < 0$, то $sign(\sum_{i=1}^{n} \beta_i x_i) = -1$, объект относится к отрицательному классу

$\sum_{i=1}^{n} \beta_i x_i = 0$ — это уравнение разделяющей границы

Бинарная классификация

Модель

$$f(x, \beta) = sign\left(\sum_{i=1}^{n} \beta_i x_i\right)$$

Если $\sum_{i=1}^{n} \beta_i x_i > 0,$ то $sign(\sum_{i=1}^{n} \beta_i x_i) = +1$, объект относится к положительному классу

Если $\sum_{i=1}^{n} \beta_i x_i < 0,$ то $sign(\sum_{i=1}^{n} \beta_i x_i) = -1$, объект относится к отрицательному классу

$\sum_{i=1}^{n} \beta_i x_i = 0$ — это уравнение разделяющей границы

В двумерном случае разделяющая граница — это прямая

В многомерном — плоскость

Бинарная классификация

Обучение

При обучении классификатора мы минимизируем долю ошибок:

$$Q(f, X) = \frac{1}{n}\left(\sum_{i=1}^{n}[f(x_i) \neq y_i]\right) \to min,$$

где $[f(x_i) \neq y_i] = 1$, если предсказание на объекте неверное, иначе — 0

Бинарная классификация

Отступ (margin)

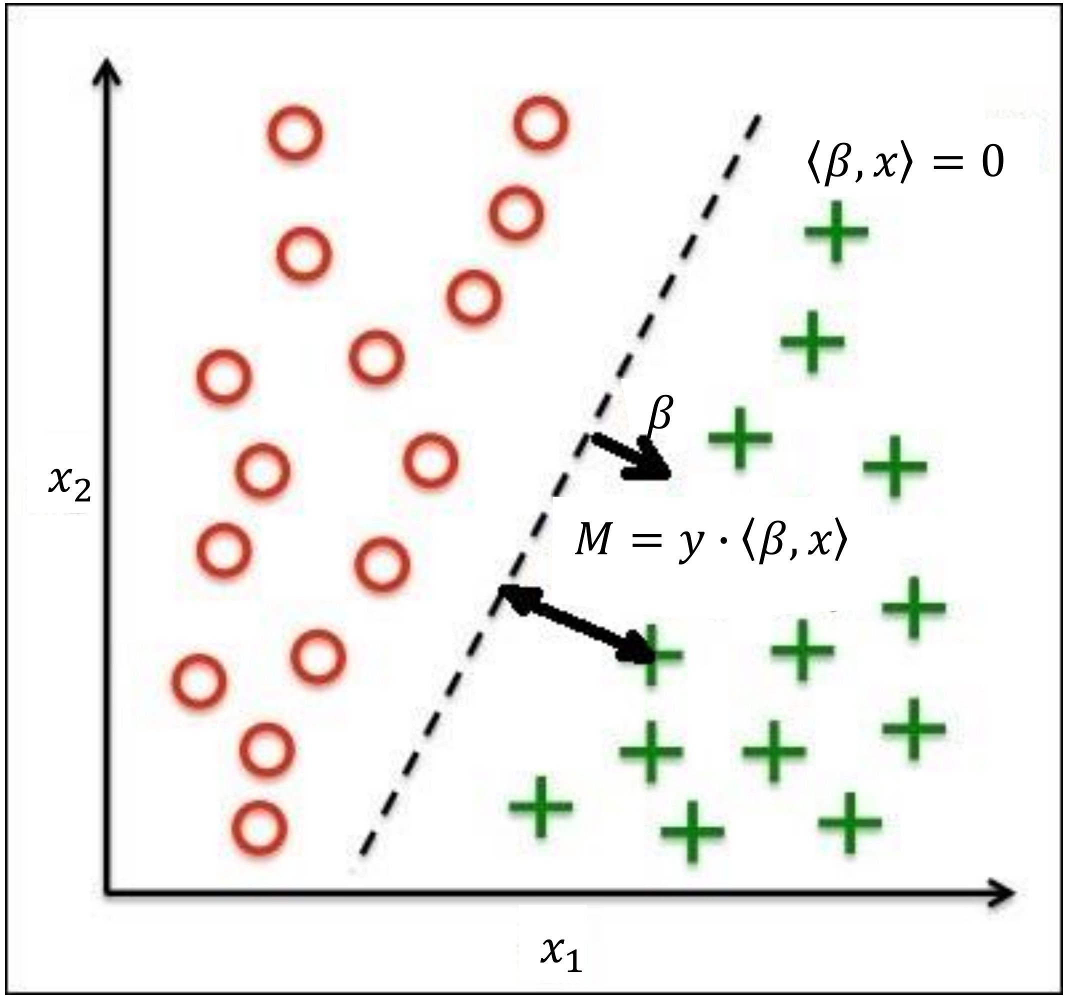
Отступ — степень уверенности классификатора в ответе

Обозначим отступ на i-ом объекте:

$$M_i = y_i \cdot (\beta, x_i)$$

Тогда решение задачи оптимизации при обучении классификатора эквивалентно решению задачи

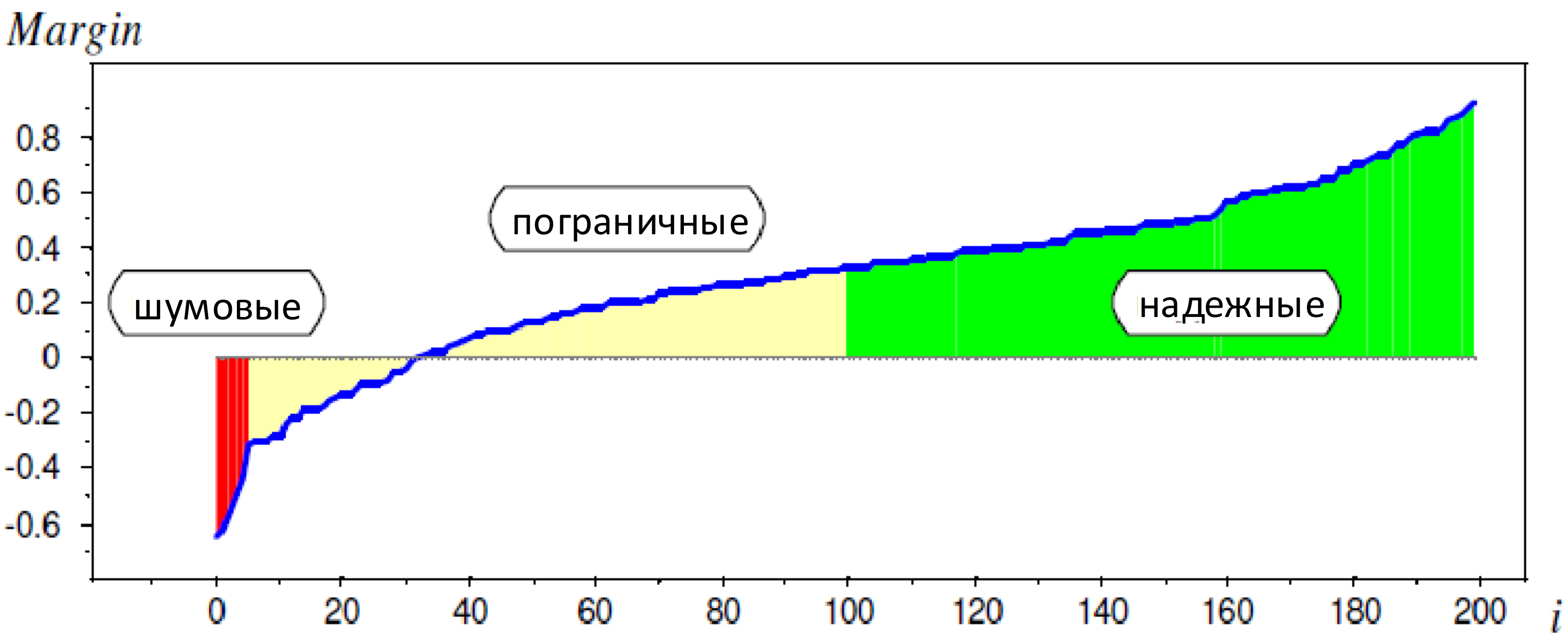
$$Q(f, X) = \frac{1}{n}\left(\sum_{i=1}^{n}[M_i < 0]\right) \to min$$



Бинарная классификация

Отступ

Чем ближе отступ к нулю, тем меньше уверенность алгоритма в ответе

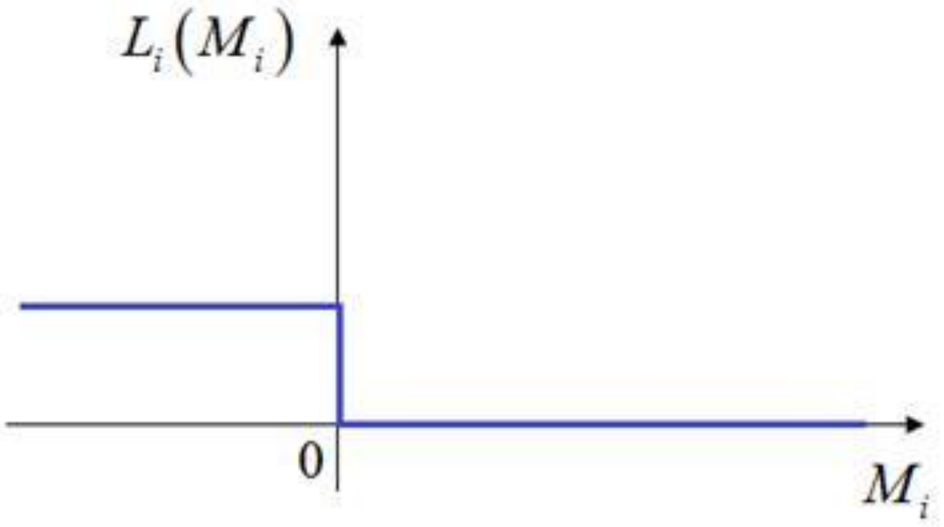


Бинарная классификация

Функция потерь

Функция, которую минимизируем при обучении: $Q(f, X) = \frac{1}{n}(\sum_{i=1}^{n}[M_i < 0]) \to min$

Такая функция потерь называется пороговой и она разрывна, что затрудняет процесс ее минимизации

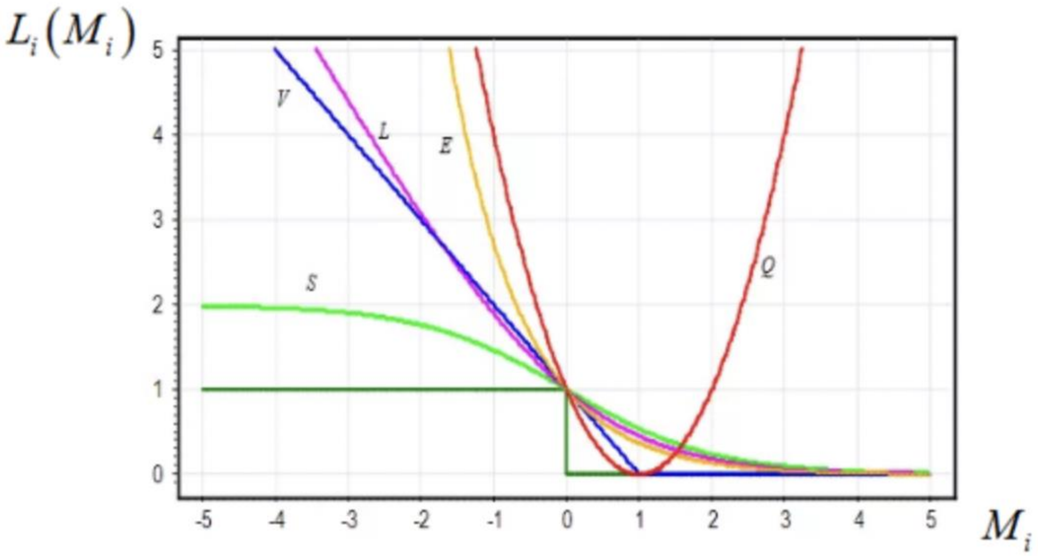


Бинарная классификация

Функция потерь

Для решения этой проблемы используются непрерывные или гладкие функции — верхнюю оценку пороговой функции. Они дифференцируемы и по значению больше либо равны исходной пороговой ⇒ автоматически минимизируют и пороговую.

Конкретную функцию выбирают в зависимости от задачи.



$V(M) = (1 - M)$ — кусочно-линейная

$\mathrm{H}(M) = (-M)$ — кусочно-линейная

$L(M) = \log\left(1 + e^{-M}\right)$ — логистическая

$Q(M) = (1 - M)^2$ — квадратичная

$S(M) = 2 \cdot (1 + e^{M})^{-1}$ — сигмоидная

$E(M) = e^{-M}$ — экспоненциальная

Логистическая регрессия

Определение

Логистическая регрессия — это линейный классификатор, который предсказывает вероятности классов, то есть:

$f(x, \beta)$ — это вероятность того, что $\mathrm{y} = +1$ на объекте x:

$$f(x, \beta) = P(y = +1 | x; \beta)$$

Логистическая регрессия

Определение

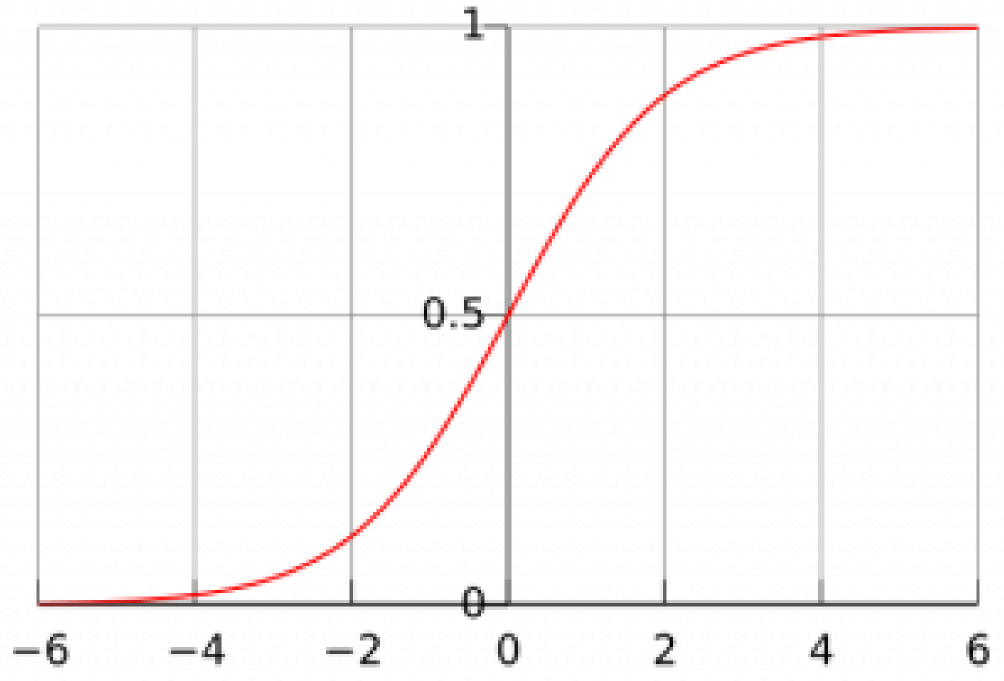
Логистическая регрессия — это линейный классификатор, который предсказывает вероятности классов, то есть:

$f(x, \beta)$ — это вероятность того, что y $= +1$ на объекте x:

$$f(x, \beta) = P(y = +1 | x; \beta)$$

Представляет собой сигмоиду:

$$f(x, \beta) = \sigma(\beta^T x) = \frac{1}{1 + e^{-\beta^T x}}$$

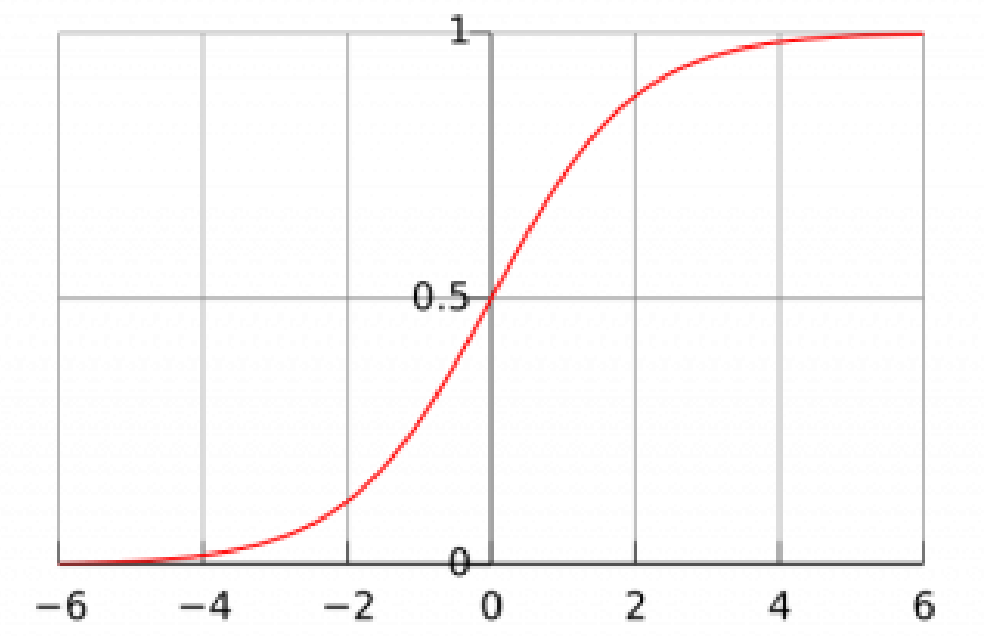


Напоминание

Сигмоида: $\sigma(z) = \frac{1}{1+e^{-z}}, \sigma(z) \in (0; 1)$

Логистическая регрессия

Почему сигмоида?



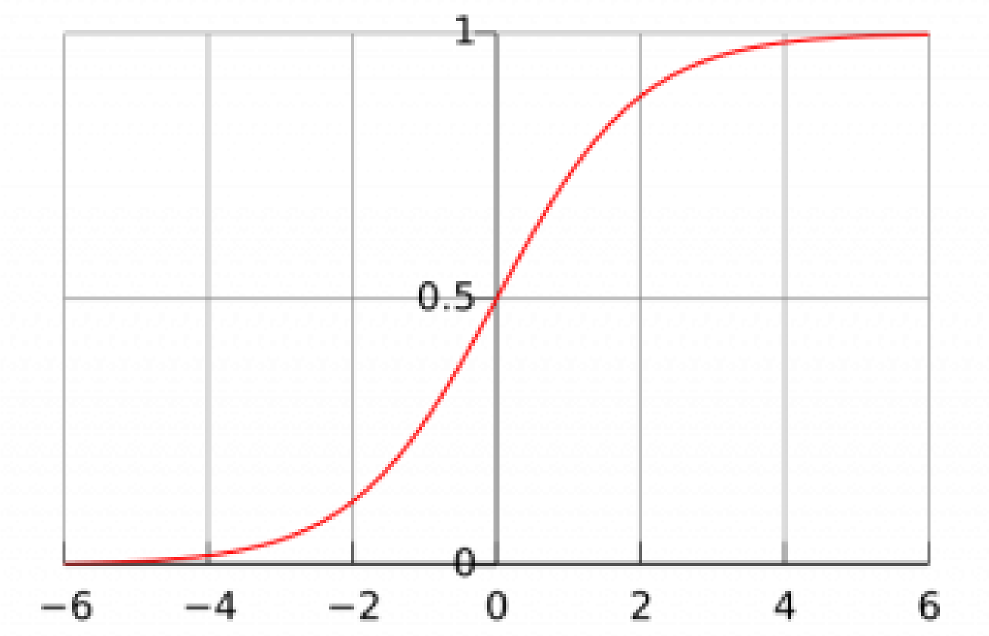
Напоминание

Сигмоида: $\sigma(z) = \frac{1}{1+e^{-z}}$

Логистическая регрессия

Почему сигмоида?

- Область допустимых значений $\sigma(z)$: $[0,1]$



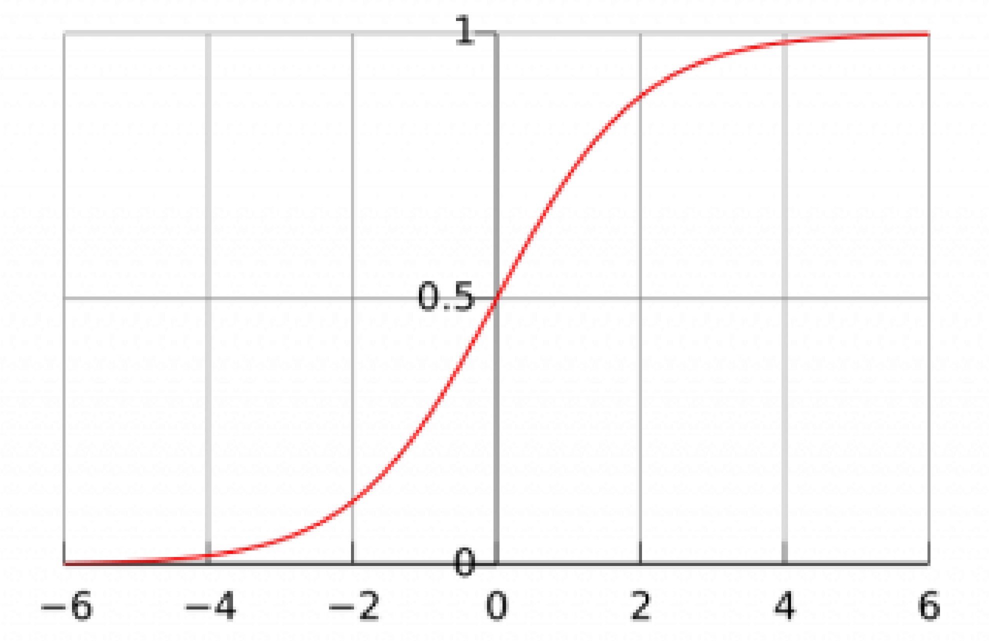
Напоминание

Сигмоида: $\sigma(z) = \frac{1}{1+e^{-z}}$

Логистическая регрессия

Почему сигмоида?

- Область допустимых значений $\sigma(z)$: $[0,1]$
- Интерпретируемая вероятность противоположного события $\sigma(-z) = 1 - \sigma(z)$



Напоминание

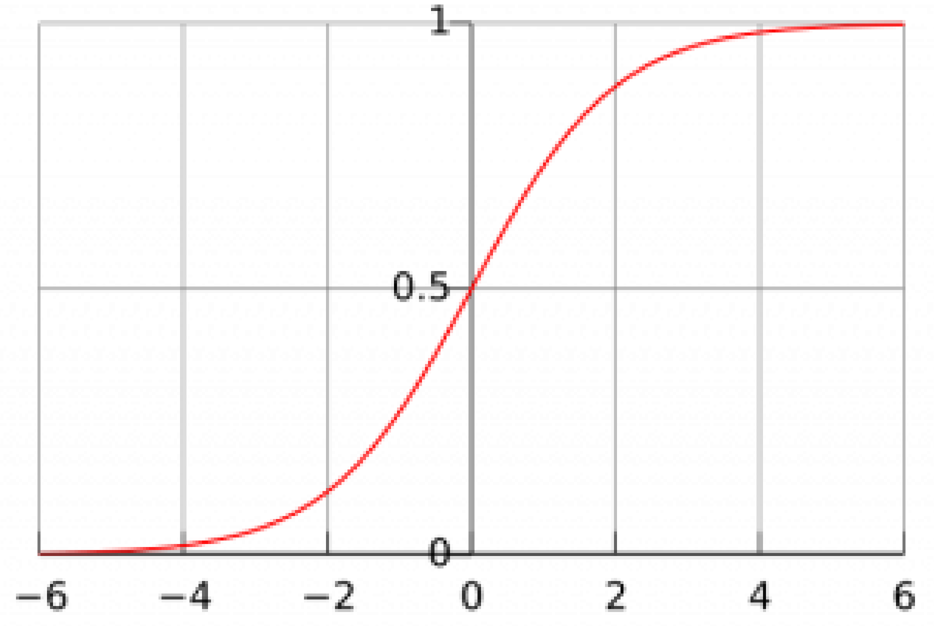
Сигмоида: $\sigma(z) = \frac{1}{1+e^{-z}}$

Логистическая регрессия

Почему сигмоида?

- Область допустимых значений $\sigma(z)$: $[0,1]$
- Интерпретируемая вероятность противоположного события $\sigma(-z) = 1 - \sigma(z)$
- Существование обратной функции:

$$z(s) = \ln\frac{s}{1-s}, \text{где s — значение сигмоиды}$$



Напоминание

Сигмоида: $\sigma(z) = \frac{1}{1+e^{-z}}$

Логистическая регрессия

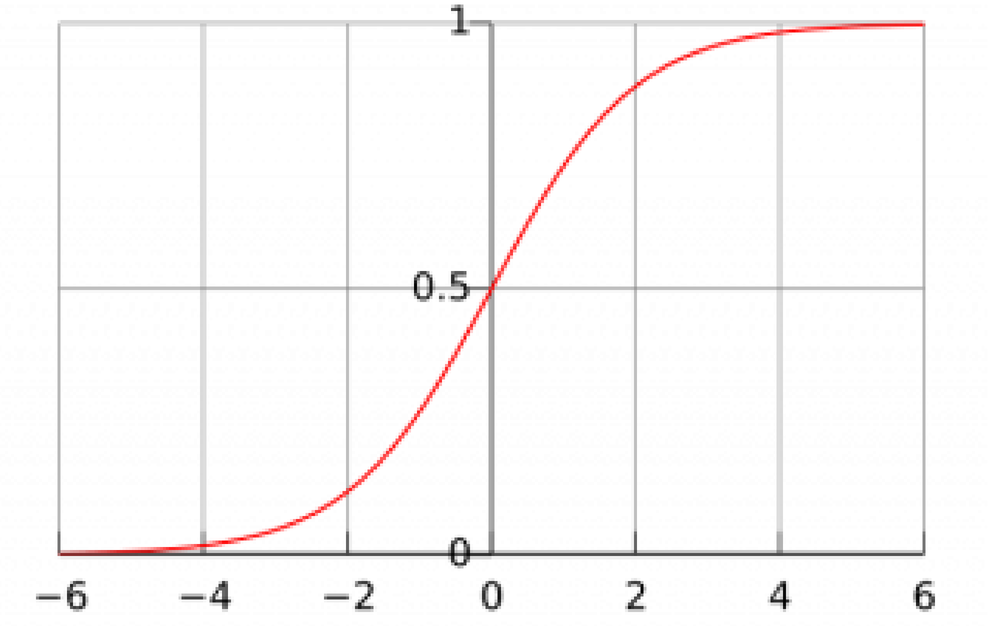
Почему сигмоида?

- Область допустимых значений $\sigma(z)$: $[0,1]$
- Интерпретируемая вероятность противоположного события $\sigma(-z) = 1 - \sigma(z)$
- Существование обратной функции:

$$z(s) = \ln\frac{s}{1-s}, \text{где s — значение сигмоиды}$$

- Удобно дифференцируема:

$$\sigma'^{(z)} = \sigma(z)(1 - \sigma(z))$$



Напоминание

Сигмоида: $\sigma(z) = \frac{1}{1+e^{-z}}$

Логистическая регрессия

Функция потерь

Квадратичную функцию для логистической регрессии использовать не получится:

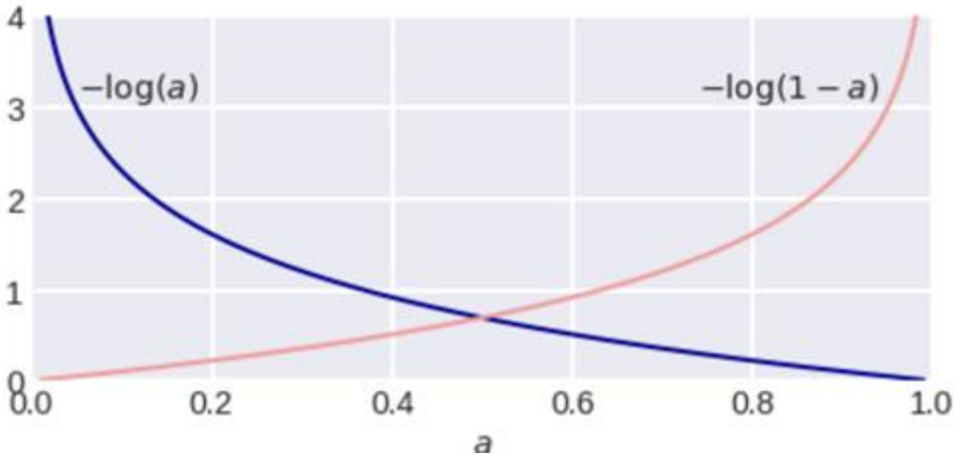
- Функция примет вид $Q(f, X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{1+e^{-\beta^T x}} - y\right)^2$, а это не выпуклая функция, то есть можем не попасть в глобальный минимум при оптимизации
- На совсем неправильных предсказаниях (например, вероятность 0% для положительного объекта), штраф будет маленьким: $(1-0)^2 = 1$

Логистическая регрессия

Функция потерь

Вместо квадратичной используется логистическая функция (log-loss):

$$logloss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot logf(x_i, \beta\,) + (1 - y_i)\log(1 - f(x_i, \beta\,))]$$



Логистическая ошибка
на одном объекте

Логистическая регрессия

Функция потерь

В отличие от квадратичной функции потерь, у логистической:

— если $f(x, \beta) = 1$ и $y = +1$ (алгоритм сделал верное предсказание), штраф $\mathrm{L}(f, y) = 0$

— если $f(x, \beta) \to 0,$ при $y = +1$ (т. е. предсказание неверное), штраф $\mathrm{L}(f, y) \to +\infty$

Логистическая регрессия

Константное решение

Пусть:

- a — ответ алгоритма
- n_1 — объектов положительного класса
- n_2 — объектов отрицательного класса
- $n_1 + n_2$ — всего объектов в выборке

Логистическая регрессия

Константное решение

Тогда функция потерь примет вид:

$$logloss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot \log\ a + (1 - y_i)\log(1 - a)]$$

Так как в выборке n_1 положительных и n_2 отрицательных объектов, функцию можно представить как:

$$logloss = -\frac{n_1}{n}\log a - \frac{n_2}{n}\log(1 - a)$$

Логистическая регрессия

Константное решение

Возьмем производную по a и приравняем к 0:

$$\frac{\partial logloss}{\partial a} = -\frac{n_1}{n}\frac{1}{a} + \frac{n_2}{n}\frac{1}{1-a} = 0$$

Выразим a:

$$a = \frac{n_1}{n}$$

Что в точности соответствует вероятности встретить объект положительного класса в выборке

Логистическая регрессия

Оптимизация logloss

Хотя logloss потенциально неограниченная функция, имеет смысл рассматривать значения лучше константного, поэтому принимается диапазон:

$$logloss \in \left[0, -\frac{n_1}{n}\log\frac{n_1}{n} - \frac{n_2}{n}\log\frac{n_2}{n}\right]$$

Метрики качества классификации

Accuracy

Accuracy — доля правильных ответов

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

Метрики качества классификации

Accuracy

Accuracy — доля правильных ответов

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

При сильном дисбалансе классов не отражает качество работы алгоритма

Метрики качества классификации

Accuracy

Пример: Нам необходимо обучить классификатор, который будет определять болен человек или здоров. В нашей обучающей выборке заболевание встречается у трех человек из тысячи. Чему будет равна accuracy **константного классификатора классификатора, который утверждает, что все здоровы?**

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

Метрики качества классификации

Accuracy

Пример: Нам необходимо обучить классификатор, который будет определять болен человек или здоров. В нашей обучающей выборке заболевание встречается у трех человек из тысячи. Чему будет равна accuracy **константного классификатора классификатора, который утверждает, что все здоровы?**

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

Accuracy составит 0,997

Можем ли мы считать, что модель хорошая?

Метрики качества классификации

Accuracy

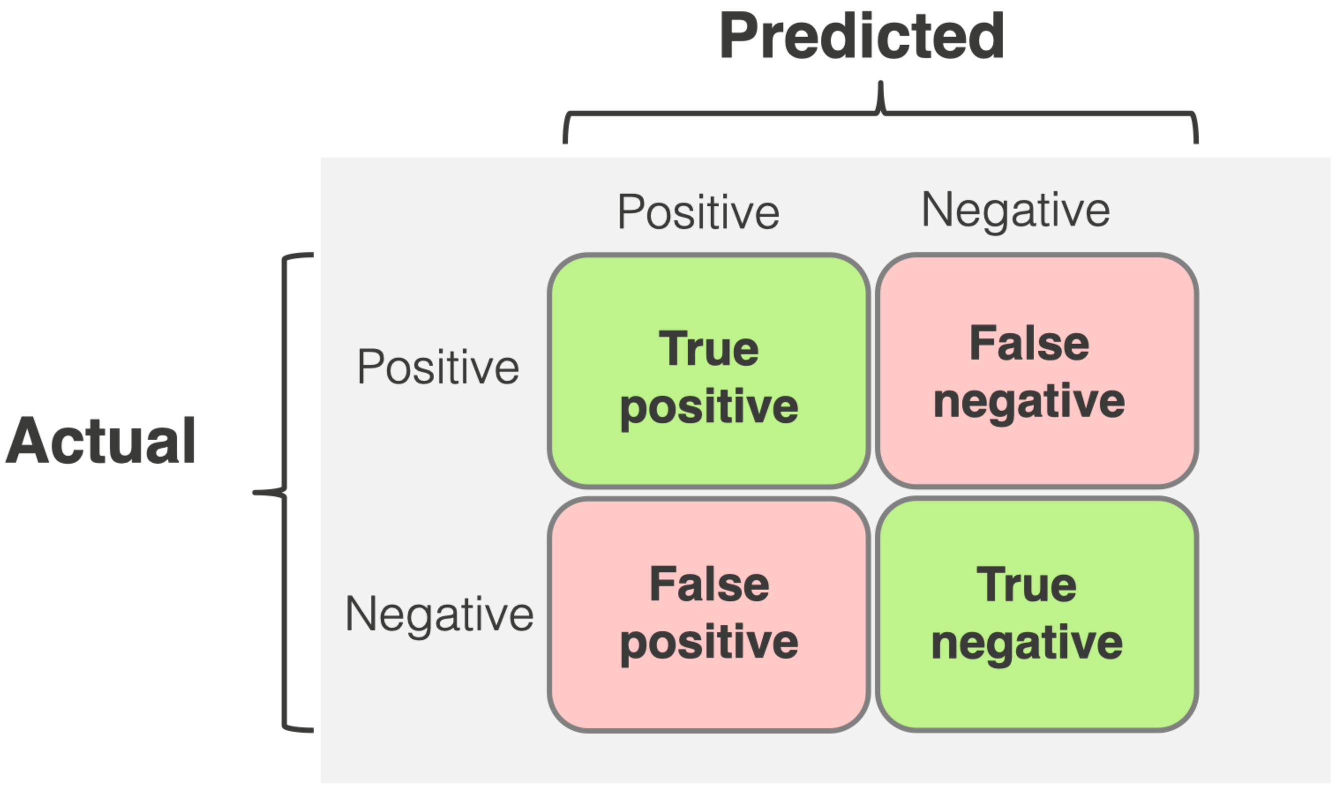
Пример: Нам необходимо обучить классификатор, который будет определять болен человек или здоров. В нашей обучающей выборке заболевание встречается у трех человек из тысячи. Чему будет равна accuracy **константного классификатора классификатора, который утверждает, что все здоровы?**

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

Accuracy составит 0,997

Метрики качества классификации

Матрица ошибок (confusion matrix)

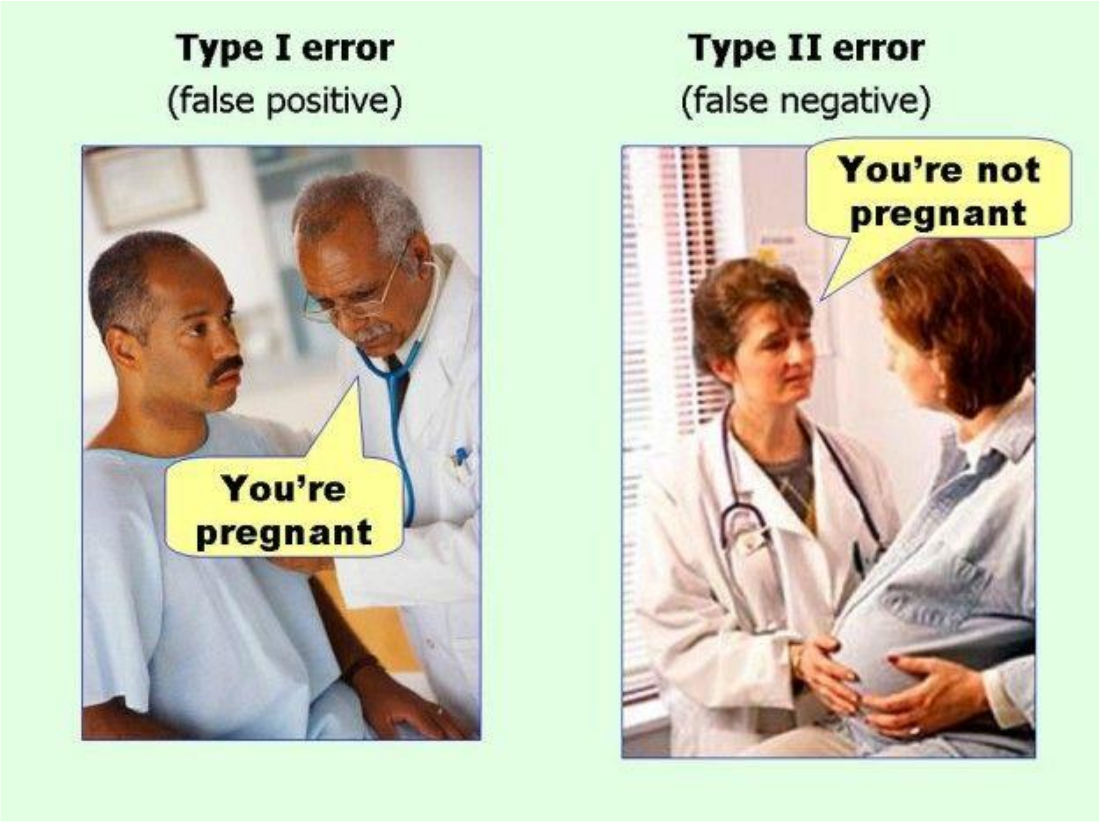


Метрики качества классификации

Матрица ошибок (confusion matrix)

Ошибка I рода: истинный класс отрицательный, предсказанный класс — положительный (False Positive)

Ошибка II рода: истинный класс положительный, предсказанный класс — отрицательный (False Negative)

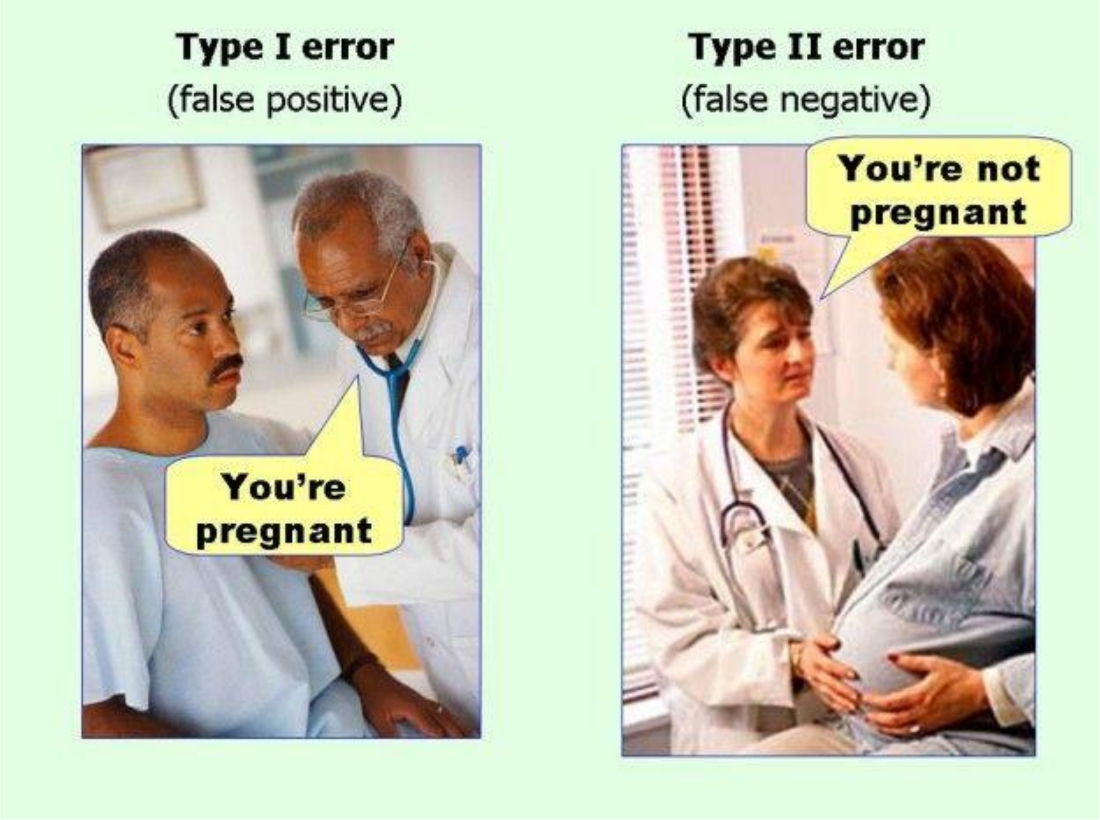


Метрики качества классификации

Матрица ошибок (confusion matrix)

Ошибка I рода: истинный класс отрицательный, предсказанный класс — положительный (False Positive)

Ошибка II рода: истинный класс положительный, предсказанный класс — отрицательный (False Negative)



Какая ошибка более критична?

Метрики качества классификации

Precision (точность)

Показывает, насколько можно

доверять классификатору при

$f(x) = +1$

$$precision(f, X) = \frac{TP}{TP + FP}$$

Метрики качества классификации

Precision (точность)

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

$$precision(f, X) = \frac{TP}{TP + FP}$$

Модель 1

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

$\text{Precision} = ?$

Модель 2

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

$\text{Precision} = ?$

Метрики качества классификации

Precision (точность)

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

$$precision(f, X) = \frac{TP}{TP + FP}$$

Модель 1

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

$\text{Precision} = 0{,}8$

Модель 2

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

$\text{Precision} = 0{,}96$

Метрики качества классификации

Recall (полнота)

Показывает, как много

объектов положительного

класса находит классификатор

$$\text{recall}(f, X) = \frac{TP}{TP+FN}$$

Метрики качества классификации

Recall (полнота)

Показывает, как много объектов положительного класса находит классификатор

$$\text{recall}(f, X) = \frac{TP}{TP + FN}$$

Модель 1

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

$\text{Precision} = 0{,}8$

$\text{Recall} = ?$

Модель 2

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

$\text{Precision} = 0{,}96$

$\text{Recall} = ?$

Метрики качества классификации

Recall (полнота)

Показывает, как много объектов положительного класса находит классификатор

$$\text{recall}(f, X) = \frac{TP}{TP+FN}$$

Модель 1

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

$\text{Precision} = 0{,}8$

$\text{Recall} = 0{,}8$

Модель 2

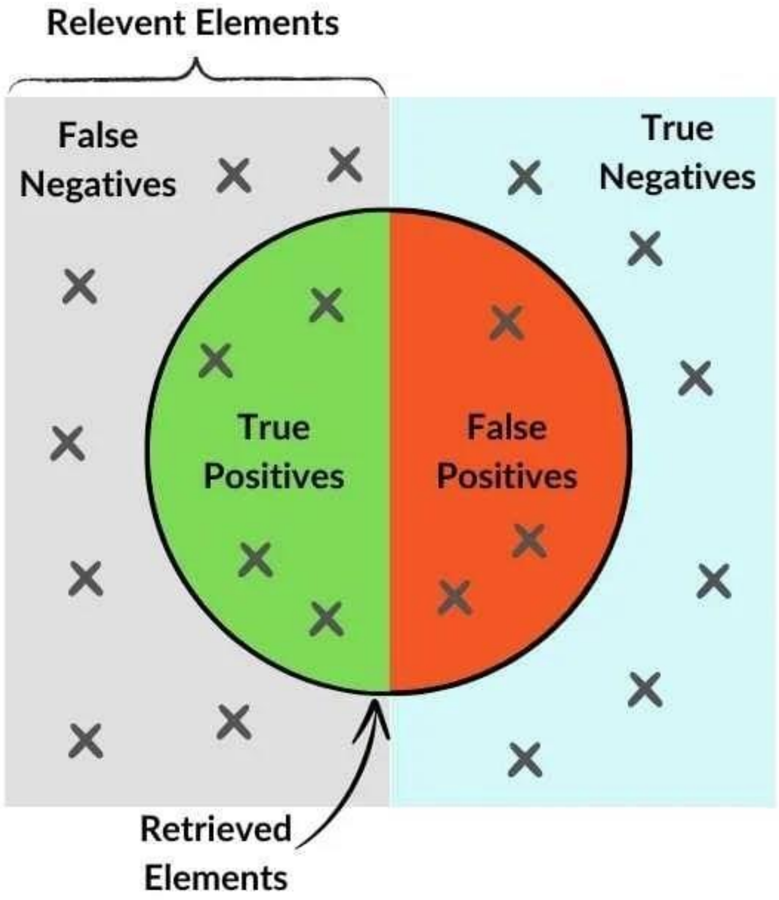
	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$a(x) = 1$ Получили кредит	48	2
$a(x) = -1$ Не получили кредит	52	98

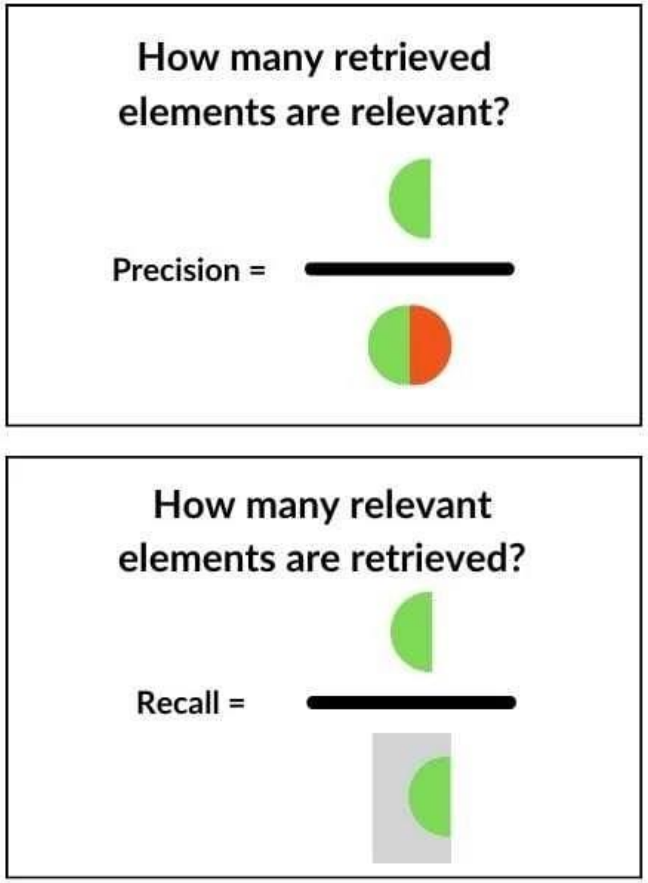
$\text{Precision} = 0{,}96$

$\text{Recall} = 0{,}48$

Метрики качества классификации

Precision vs recall





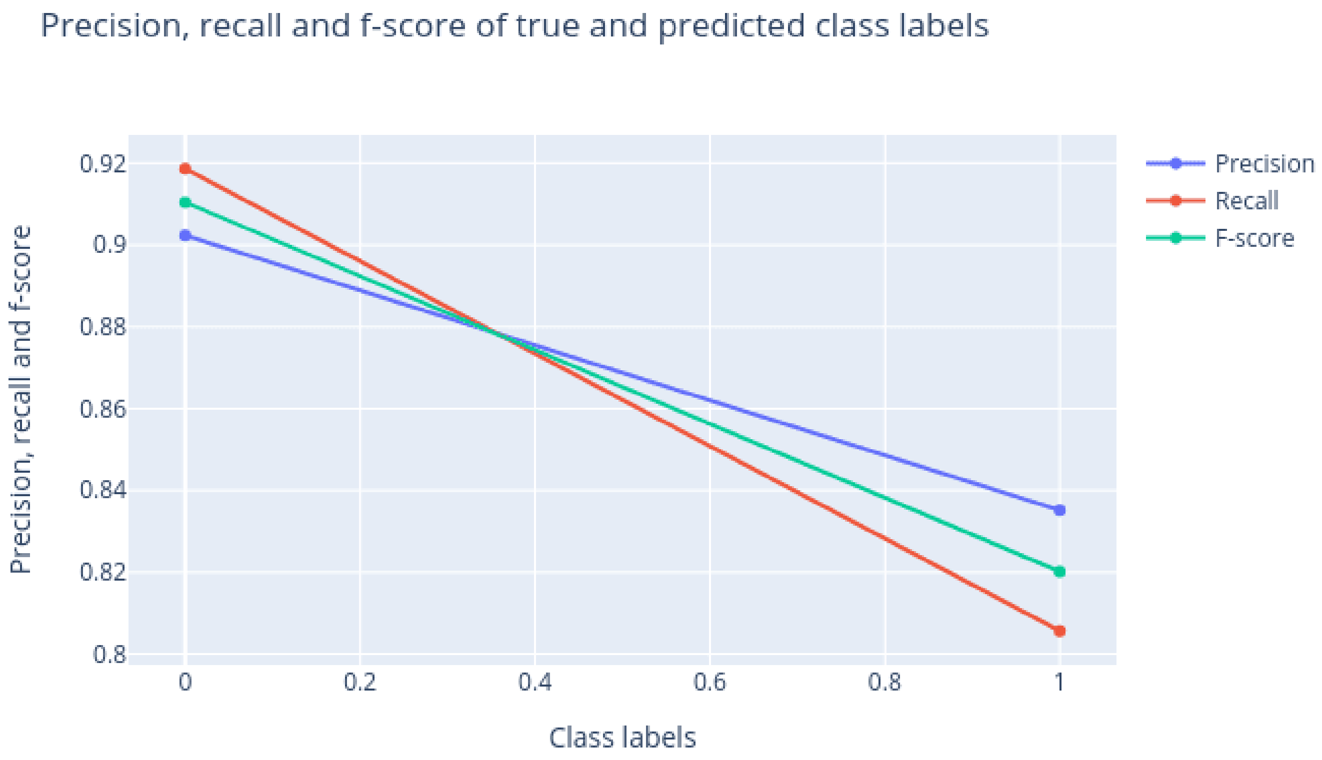
Метрики качества классификации

F-мера

Метрика, учитывающая и

точность, и полноту

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



Метрики качества классификации

Регулировка точности и полноты

Обозначим уверенность классификатора в том, что объект x относится к классу +1, за $p(x)$

Если $p(x) > 0{,}5$, мы относим объект к положительному классу, иначе — к отрицательному

Метрики качества классификации

Регулировка точности и полноты

Обозначим уверенность классификатора в том, что объект x относится к классу +1, за $p(x)$

Если $p(x) > 0{,}5$, мы относим объект к положительному классу, иначе — к отрицательному

Этот порог t можно менять в зависимости от задачи на любое число от 0 до 1

Метрики качества классификации

ROC-AUC

Метрика, которая помогает измерить качество всего семейства классификаторов независимо от выбранного порога

AUC — Area Under ROC Curve (площадь под ROC-кривой)

Метрики качества классификации

ROC-AUC

Метрика, которая помогает измерить качество всего семейства классификаторов независимо от выбранного порога

AUC — Area Under ROC Curve (площадь под ROC-кривой)

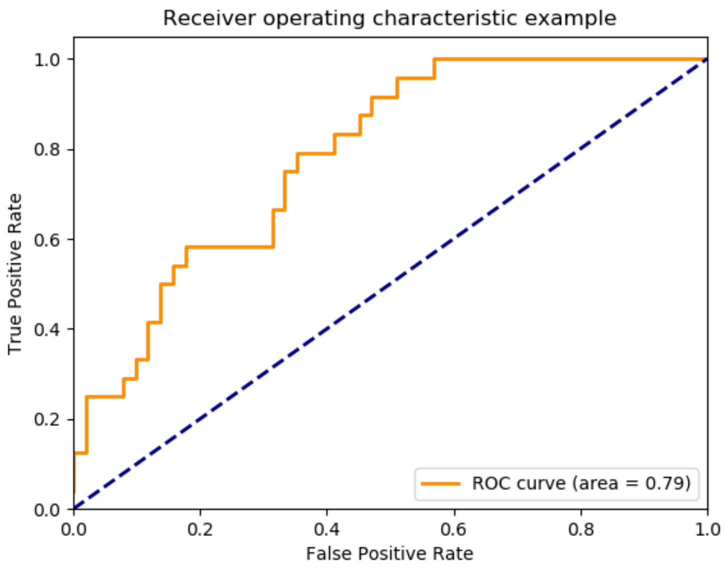
Для каждого значения порога t вычислим:

False Positive Rate $(FPR) = \frac{FP}{FP+TN}$ и True Positive Rate $(TPR) = \frac{TP}{TP+FN}$

Метрики качества классификации

ROC-AUC

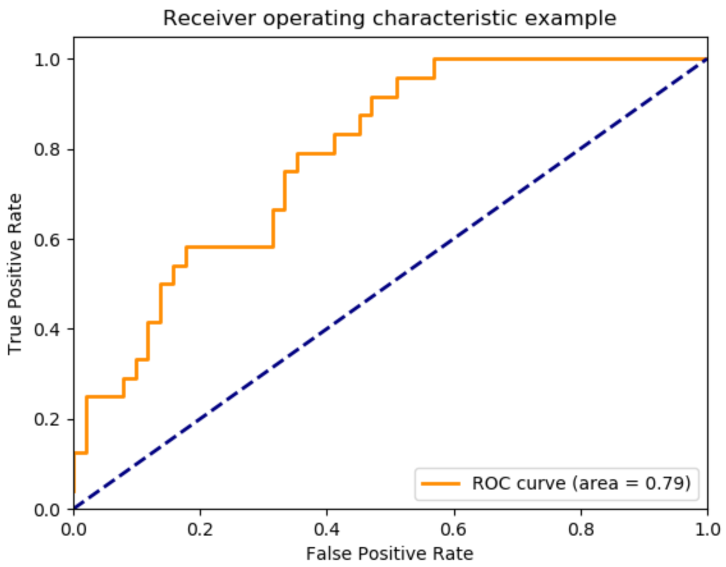
ROC-кривая — кривая, состоящая из точек с координатами (FPR, TPR) для всех возможных порогов



Метрики качества классификации

ROC-AUC

ROC-кривая — кривая, состоящая из точек с координатами (FPR, TPR) для всех возможных порогов

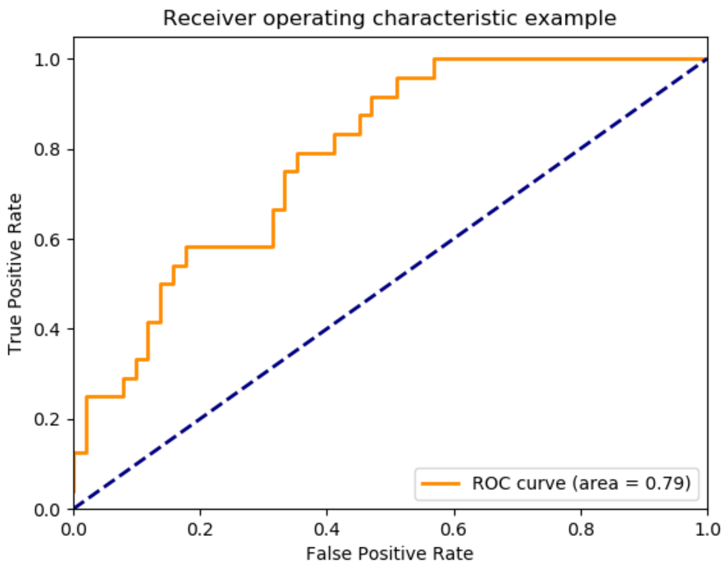


AUC ∈ [0; 1] — площадь под ROC-кривой

Метрики качества классификации

ROC-AUC

ROC-кривая — кривая, состоящая из точек с координатами (FPR, TPR) для всех возможных порогов



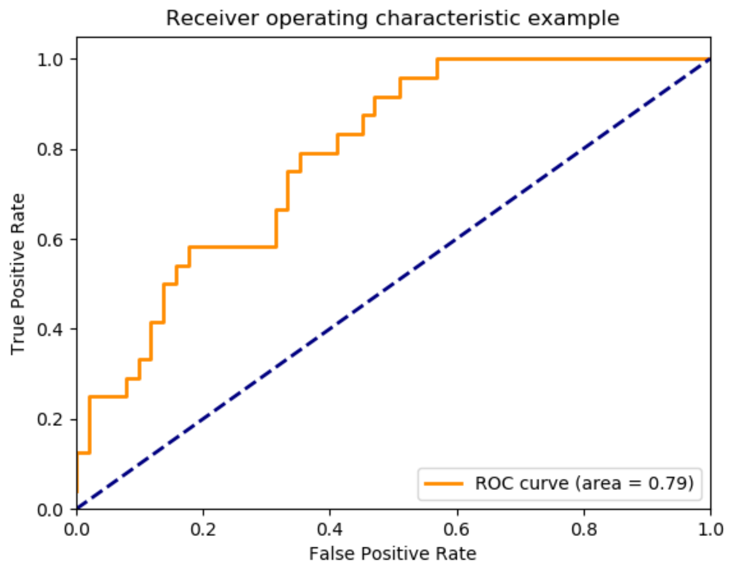
AUC ∈ [0; 1] — площадь под ROC-кривой

Чему равна AUC при идеальной классификации?

Метрики качества классификации

ROC-AUC

ROC-кривая — кривая, состоящая из точек с координатами (FPR, TPR) для всех возможных порогов



AUC ∈ [0; 1] — площадь под ROC-кривой

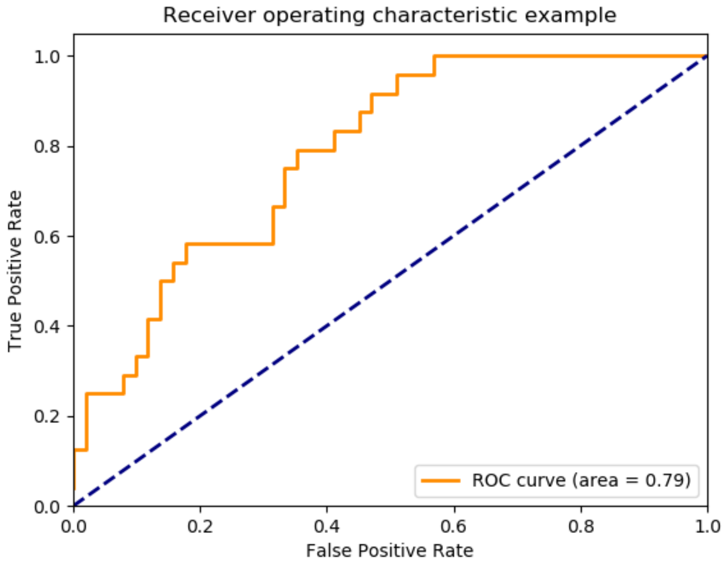
Чему равна AUC при идеальной классификации?

AUC = 1

Метрики качества классификации

ROC-AUC

ROC-кривая — кривая, состоящая из точек с координатами (FPR, TPR) для всех возможных порогов



AUC ∈ [0; 1] — площадь под ROC-кривой

Чему равна AUC при идеальной классификации?

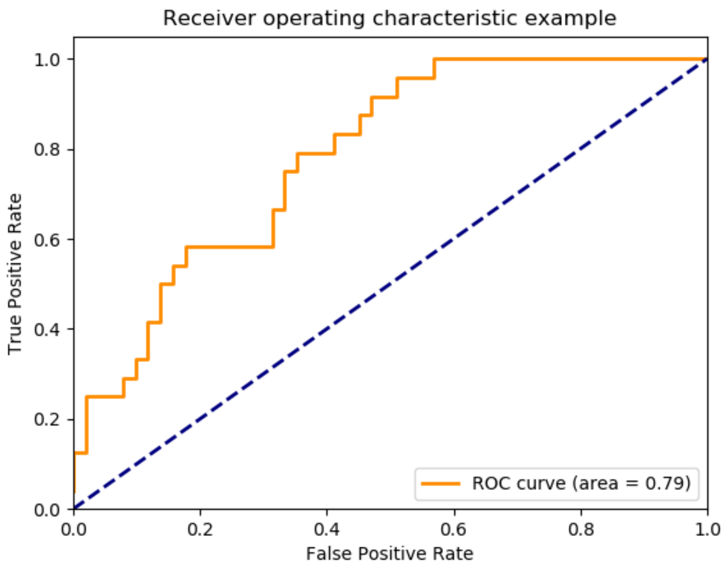
AUC = 1

Чему равна AUC при случайной классификации?

Метрики качества классификации

ROC-AUC

ROC-кривая — кривая, состоящая из точек с координатами (FPR, TPR) для всех возможных порогов



AUC ∈ [0; 1] — площадь под ROC-кривой

Чему равна AUC при идеальной классификации?

AUC = 1

Чему равна AUC при случайной классификации?

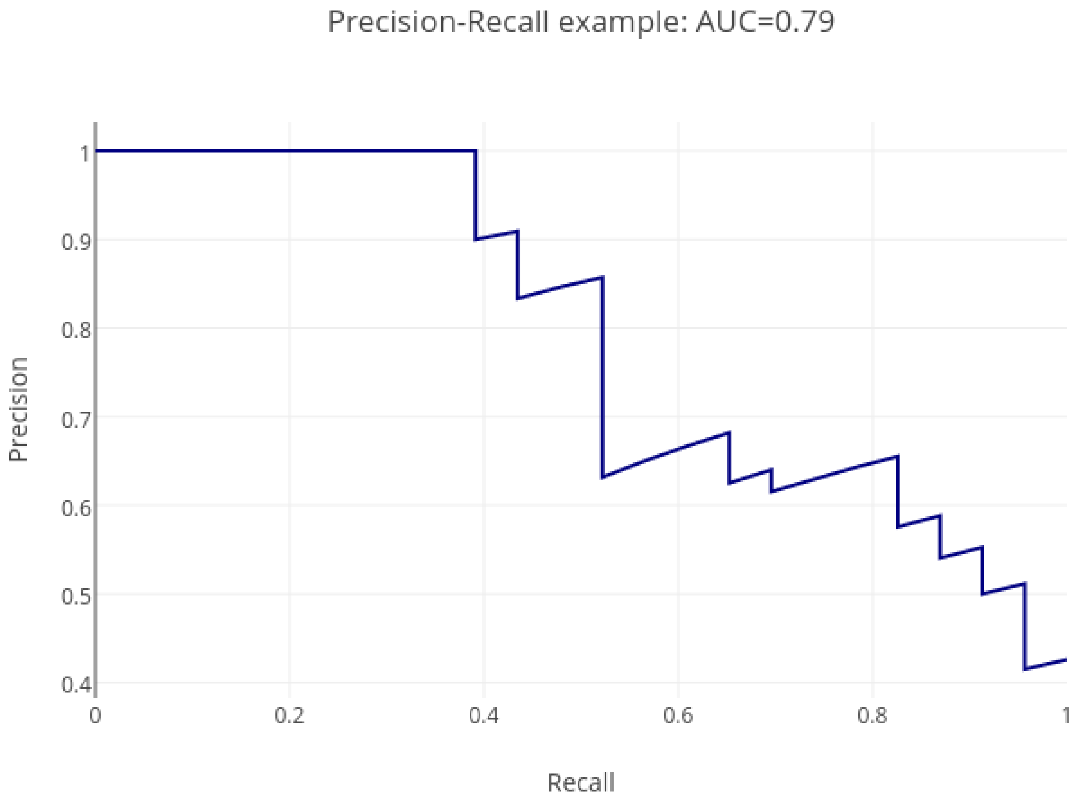
AUC = 0,5

Метрики качества классификации

PR-кривая

При малой доле объектов положительного класса ROC-AUC может давать неадекватно хороший результат.

Поэтому для задач с несбалансированными классами используют precision-recall-кривую



Метрики качества классификации

AUC-PR

Площадь под PR-кривой