Лекция 4: классификатор логистическая регрессия

Евгений Борисов

Классификатор: о задаче

разделение данных на части (классы)

Учебный набор: [объект, ответ]

Задача: классификатор

объект → вектор-признак → результат

Обучение: минимизация ошибки

ошибка = результат - правильный ответ

Критерий остановки:

достигнут порог значения ошибки,
и/или порог количества циклов

Классификатор: данные 1

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} & y^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} & y^{(2)} \\ \vdots & \vdots & & \vdots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} & y^{(m)} \end{bmatrix}$$

x - вектор-признак

y - метка класса

n - размер пространства признаков

m - количество примеров

Классификатор: данные 2

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} ; \quad X = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & \dots & x_n^{(2)} \\ & \vdots & & \\ 1 & x_1^{(m)} & \dots & x_n^{(m)} \end{bmatrix} ; \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

x - вектор-признак

y - метка класса

n - размер пространства признаков

m - количество примеров

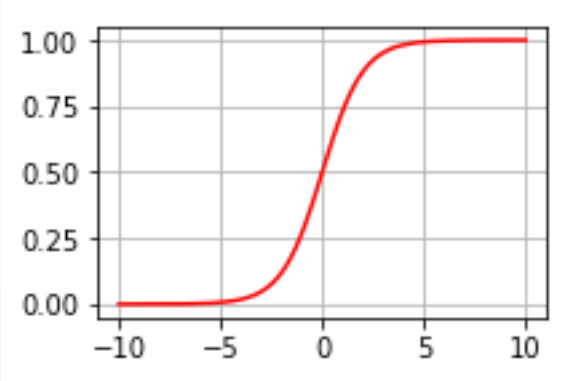
θ - параметры

Классификатор: классификатор 1

$$z(x) = \theta^T x = \theta_0 + \theta_1 x_1 + \ldots \theta_n x_n$$

ф-ция сигмоид

$$h_\theta(z) = \frac{1}{1 + e^{-z}}$$

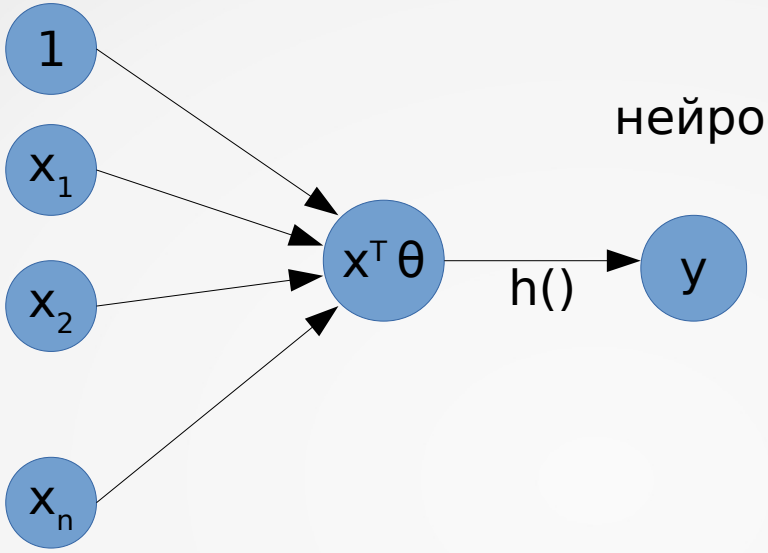


x - вектор-признак

n - размер пространства признаков

θ - параметры

Классификатор: классификатор 2



нейросеть из одного нейрона

x - вектор-признак

n - размер пространства признаков

θ - параметры

h - ф-ция сигмоид

y - выход

Классификатор: ф-ция потери

перекрёстная энтропия (cross entropy)

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} \cdot log(h_\theta(x^{(i)})) + (1 - y^{(i)})log(1 - h_\theta(x^{(i)}))\right]$$

Энтропия — это то, как много информации вам не известно о системе.

y - номер класса объекта x

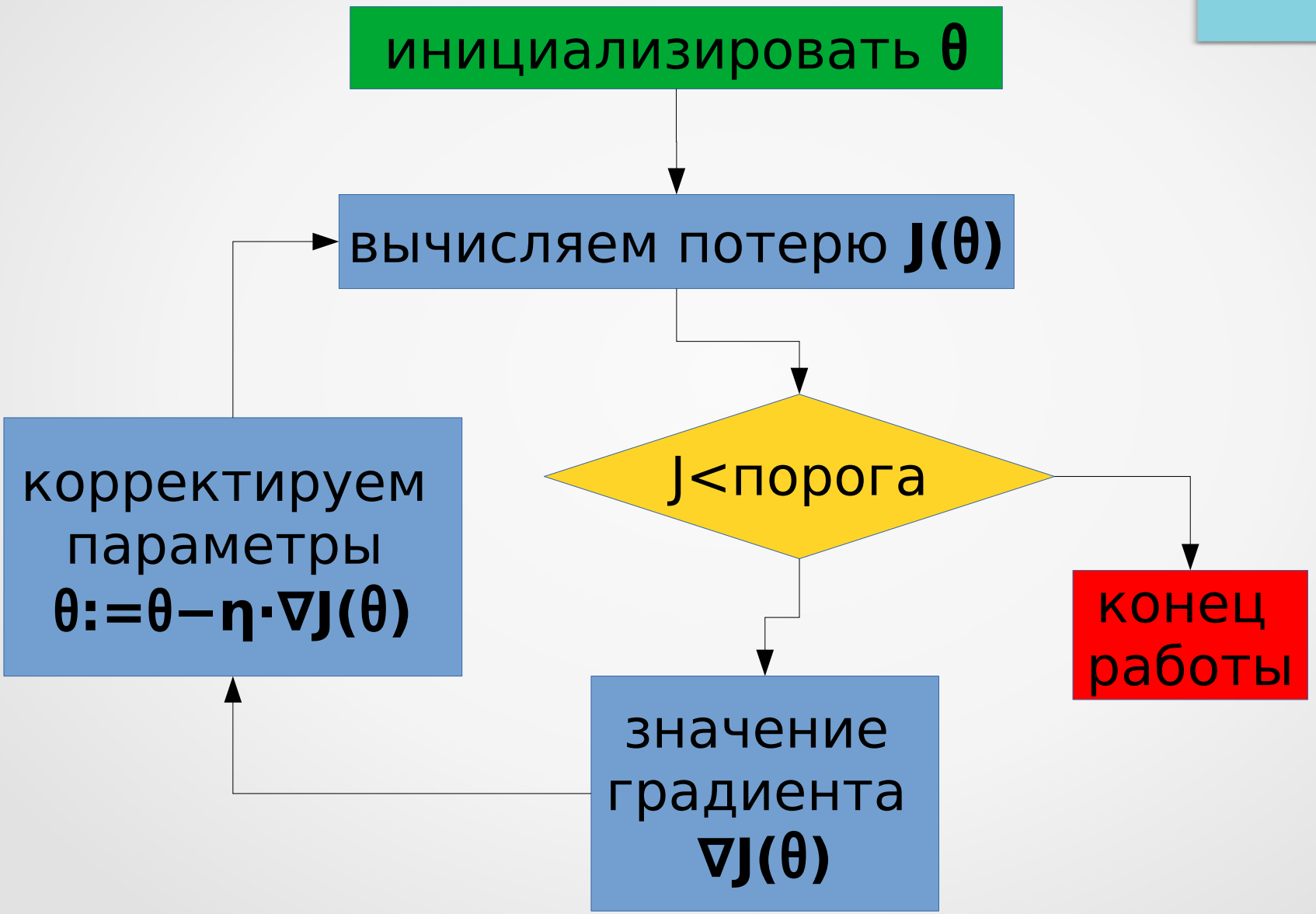
o = h($x^T \theta$) - ответ класстфикатора

if (y==1) log(o) else log(1-o)

задача оптимизации

$$\min_{\theta} J(\theta)$$

Классификатор: метод градиентного спуска

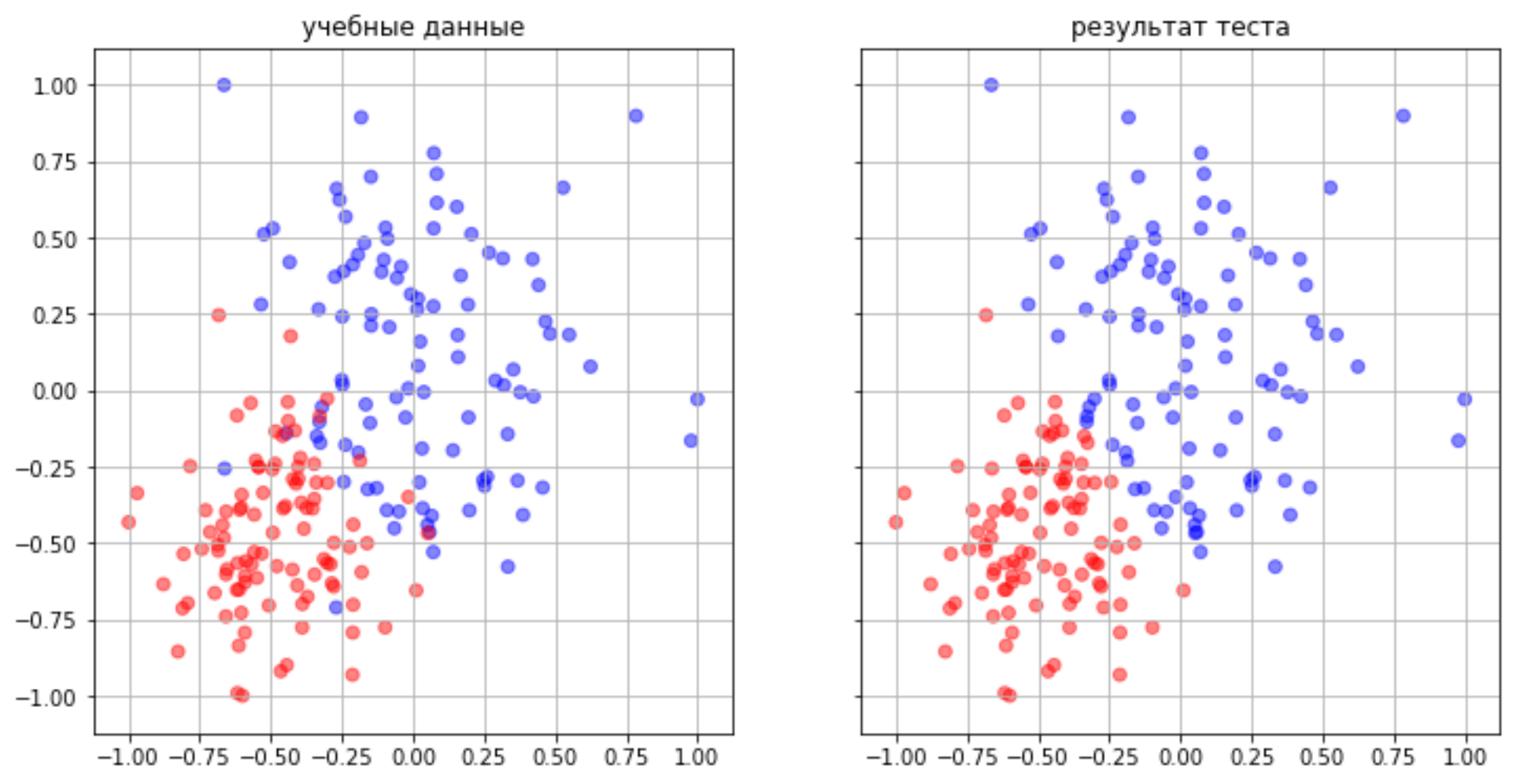


Классификатор: градиент и параметры

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

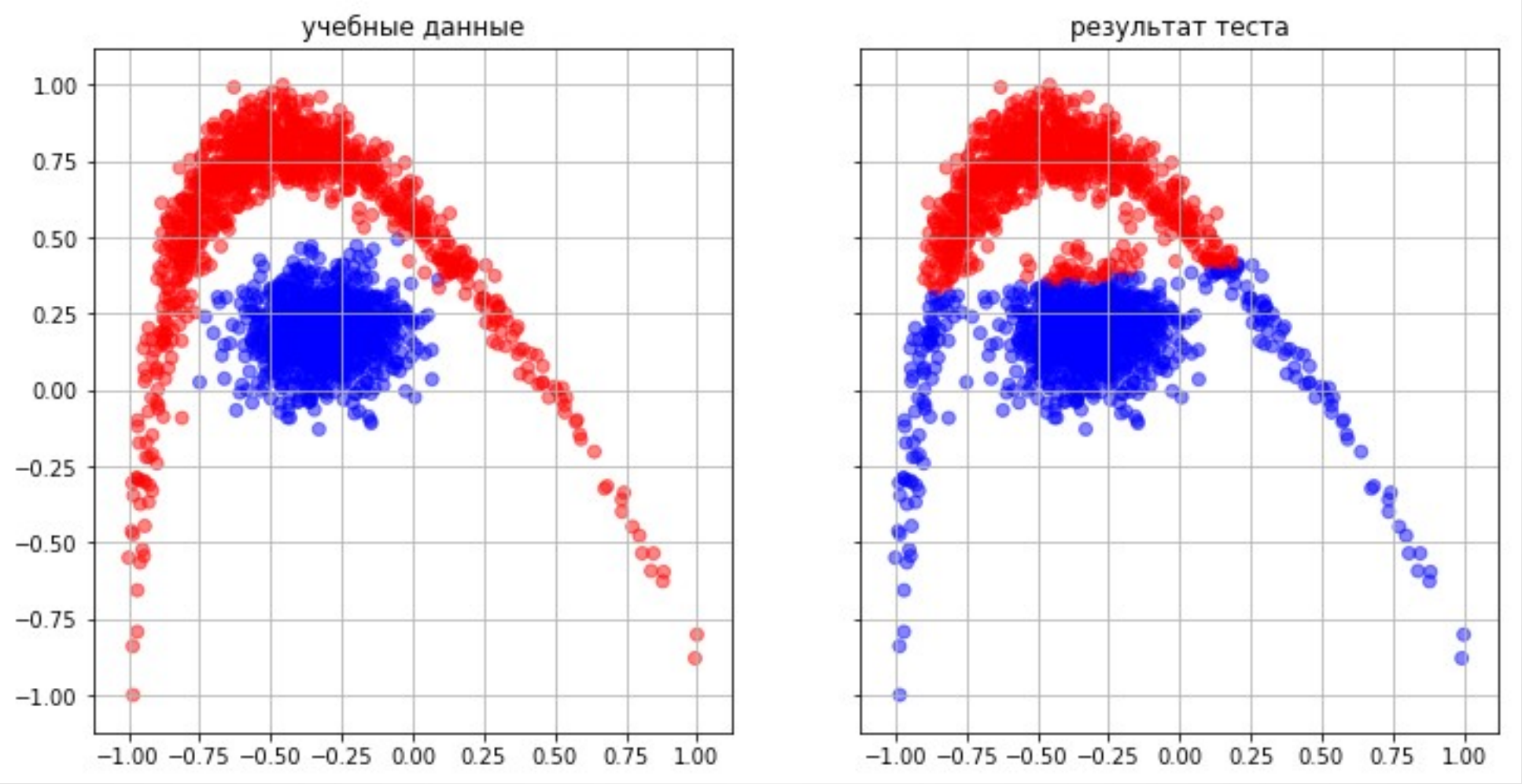
Классификатор: результат

линейный классификатор и линейно разделимые данные



Классификатор: результат 2

линейный классификатор и линейно **не**разделимые данные



Классификатор: нелинейный

$$z(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 \ldots + \theta_{24} x_1^4 x_2^4$$

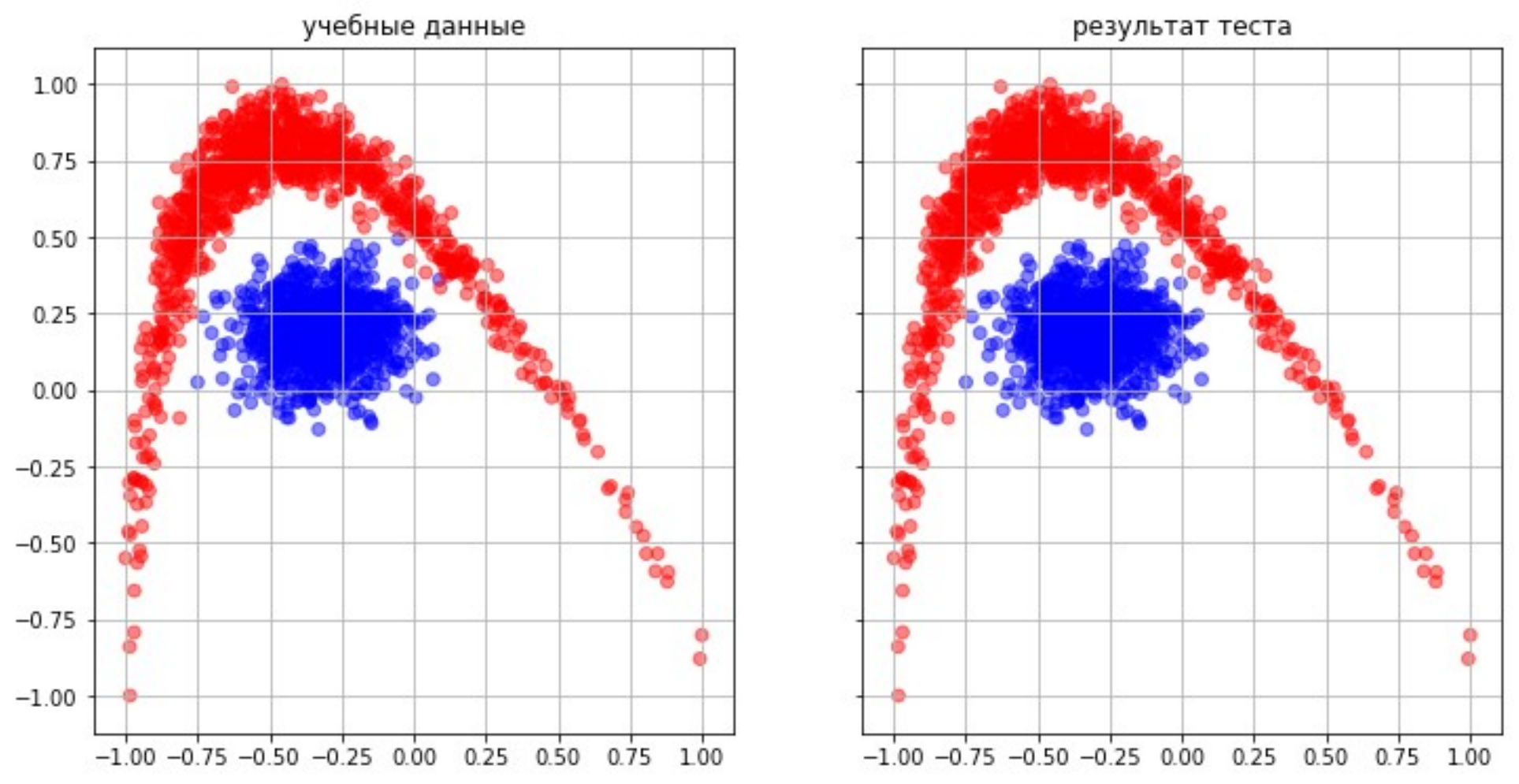
строим полином степени k на признаках X

повышаем размерность пространства признаков

увеличиваем число параметров θ

Классификатор: результат 3

нелинейный классификатор и линейно **не**разделимые данные



Классификатор: порог скора

$p = h(x^T \theta)$ - оценка (score) классификатора

[p > b] - ответ класстфикатора, т. е. номер класса 0 или 1

b - порог скора
x - вектор-признак объекта
θ - параметры классификатора
h - ф-ция сигмоид

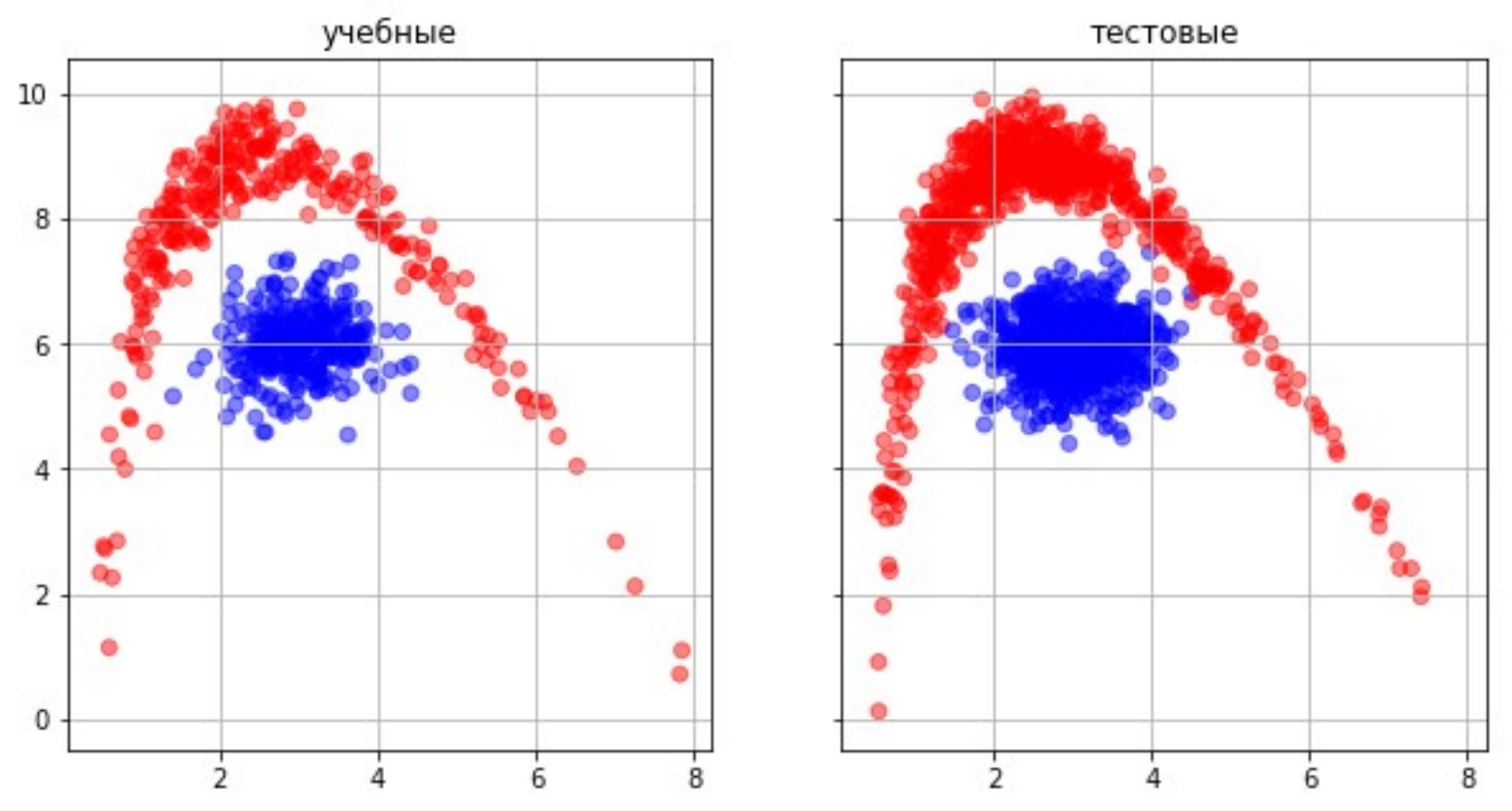
пример: $0 \le p \le 1 \quad , \; b = 0.5$

Классификатор: оценка результата 1

разделяем набор данных
- учебный
- тестовый

недообучение (underfitting)
большая ошибка на учебном наборе

переобучение (overfitting)
малая ошибка на учебном наборе
большая ошибка на тестовом наборе



Классификатор: оценка результата 2

метрики качества на тестовом наборе

- погрешность (accuracy)
- матрица ошибок (confusion matrix)
- точность (precision)
- полнота (recall)
- F-мера
- ROC/AUC

Классификатор: оценка результата 3

погрешность (accuracy)

правильные ответы / всего примеров

оценка для сбалансированного набора, т.е. количество примеров в классах +- одинаковое

Классификатор: оценка результата 4

Пример:

тест на болезнь «зеленуху» имеет вероятность ошибки 0.1 (как позитивной, так и негативной)

зеленухой болеет 10% населения.

Какая вероятность того,
что человек болен зеленухой,
если у него позитивный результат теста?

Классификатор: оценка результата 5

Подсказка: *формула Байеса*

$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

Классификатор: оценка результата 6

Подсказка: *формула Байеса*

$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

Ответ: *(0.9*0.1) / (0.9*0.1 + 0.1*0.9) = 0.5*

Классификатор: оценка результата 7

Подсказка: *формула Байеса*

$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

Ответ: *(0.9*0.1) / (0.9*0.1 + 0.1*0.9) = 0.5*

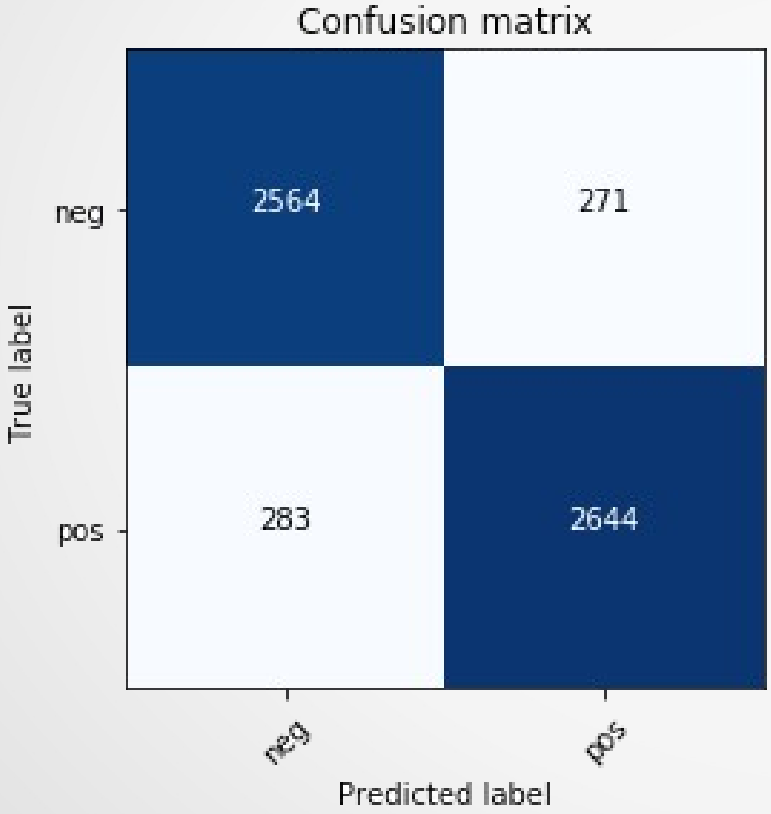
Замечание о корректности этого результата:
анализы проводят, когда есть подозрение на какую-то болезнь,

поэтому вероятность болезни надо вычислять по этой «подозрительной» группе.

это понижает требования к точности прибора...

Классификатор: оценка результата 8

матрица ошибок (confusion matrix)



два класса — четыре группы

- TP истинно положительные
- TN истинно отрицательные
- FP ложно положительные
- FN ложно отрицательные

Классификатор: оценка результата 9

точность (precision)

TP / (TP + FP)

(метрики для отдельного класса)

доля объектов действительно принадлежащих данному классу относительно всех объектов, которые классификатор отнес к этому классу

полнота (recall)

TP / (TP + FN)

доля объектов, найденных классификатором, относительно всех объектов этого класса

F-мера

*(precision*recall) / (precision+recall)*

усреднение точности и полноты

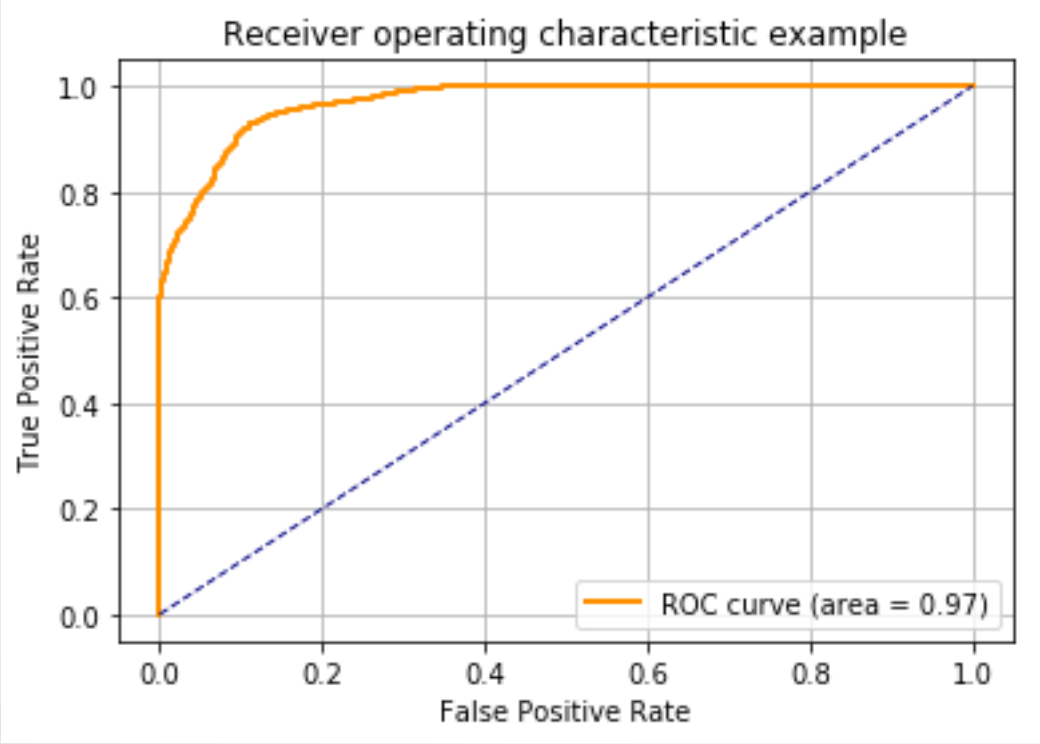
Классификатор: оценка результата 10

Пример *classification_report*

	precision	recall	f1-score	support
0	0.90	0.90	0.90	2835
1	0.91	0.90	0.91	2927
avg / total	0.90	0.90	0.90	5762

Классификатор: оценка результата 11

ROC - receiver operating characteristic,
рабочая характеристика приёмника



TPR=TP/(TP+FN)
полнота(recall), доля объектов, найденных классификатором, относительно всех объектов этого класса

FPR=FP/(FP+TN)
доля объектов negative класса алгоритм предсказал неверно

ROC - показывает зависимость полноты **TPR**
от доли ложно-негативных **FPR**
при изменении порога скора

AUC - area under ROC curve,
площадь под ROC-кривой
характеристика качества классификации

Классификатор: литература

Борисов Е.С. Классификатор на основе логистической регрессии.

http://mechanoid.kiev.ua/ml-regression-class.html

git clone https://github.com/mechanoid5/ml_lectorium.git

Классификатор: почти последний слайд...

Вопросы ?

Классификатор: практика

источники данных для экспериментов

sklearn.datasets

UCI Repository

kaggle