Лекция 4:
Байесовский классификатор

Евгений Борисов

четверг, 11 октября 2018 г.

Классификатор: с чего все начинается?

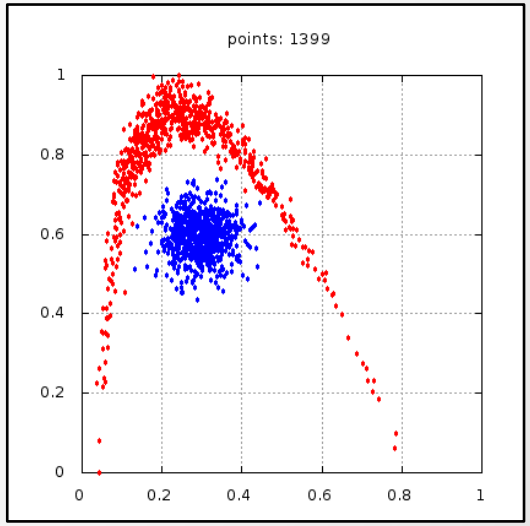
хорошие и плохие коты

извлекаем признаки

один кот — одна точка

→ [0.14, 12, ..., 345]

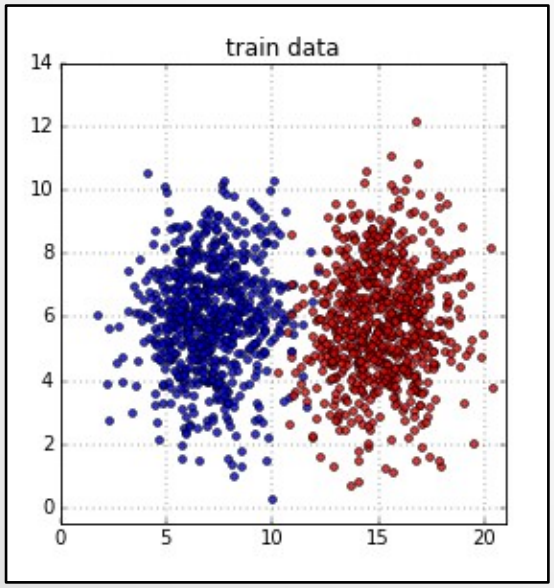
→ [78.0, 20, ..., 177]

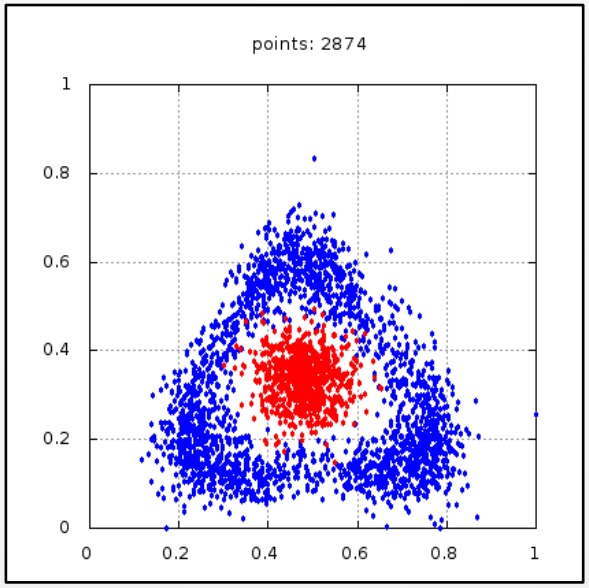


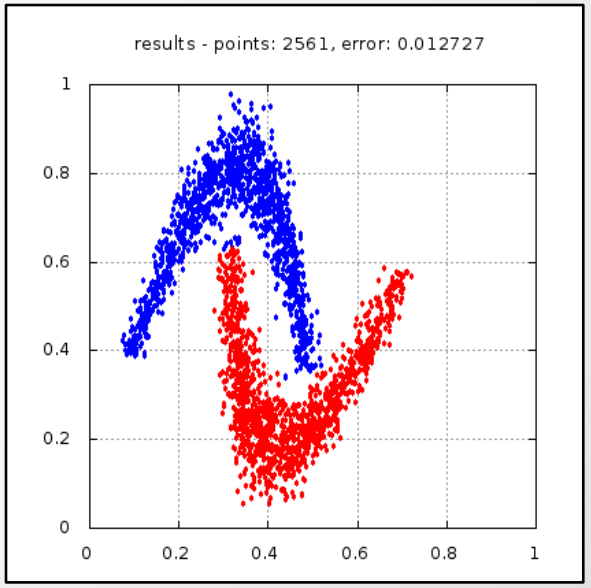
ML: классификация

разделения объектов на классы

Детектор котов:  → вектор-признак → есть/нет







Классификатор: о задаче

разделение данных на части (классы)
обучение «с учителем»

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

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

x - вектор-признак

y - метка класса

n - размер пространства признаков

m - количество примеров

Классификатор

формула Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Классификатор

X - объекты Y - ответы

X x Y - вероятностное пространство с плотностью p(x,y)

(x_i,y_i) - выборка

Задача: найти ф-цию (классификатор) a: X → Y с минимальной ошибкой

Совместная плотность: p(x,y) = p(x)P(y|x) = P(y)p(x|y)

P(y) - априорная вероятность класса y
p(x|y) - ф-ция правдоподобия класса y
P(y|x) - апостериорная вероятность класса y

принцип максимума апостериорной вероятности

$$a(x)=\underset{y\in Y}{argmax}\,P(y|x)=\underset{y\in Y}{argmax}\,P(y)\,p(x|y)$$

Классификатор: функционал среднего риска

a: X → Y - классификатор

$A_y = \{ x \in X \mid a(x) = y \}$, $y \in Y$ - разбиение X на части

Ошибка: объект *x* класса *y* попал в класс $s : A_s$, $s \neq y$

Вероятность ошибки: $$P(A_s, y) = \int_{A_s} p(x, y)\, dx$$

Потеря от ошибки: зададим $\lambda_{ys} \geq 0$ для всех пар $(y,s) \in Y \times Y$

Средний риск: мат.ожидание потери классификатора

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} P(A_s, y)$$

Классификатор: теорема

Теорема про оптимальный байесовский классификатор

пусть заданы:
- априорные вероятности классов P(y),
- плотности их распределений p(x|y)
- $\lambda_{ys} \geq 0$ потери от ошибки

тогда минимум среднего риска R(a) достигается классификатором

$$a(x) = \underset{s \in Y}{\operatorname{argmin}} \sum_{y \in Y} \lambda_{ys} P(y) p(x|y)$$

Дополнение:

если $\lambda_{yy}=0$ *и* $\lambda_{ys} = \lambda_y$ *для всех* $y,s \in Y$ *то*

$$a(x) = \underset{y \in Y}{\operatorname{argmax}} \lambda_y P(y) p(x|y)$$

Классификатор

$$a(x) = \underset{y \in Y}{\operatorname{argmax}} \lambda_y P(y) p(x|y)$$

λ_y - *потеря для объектов y*

P(y) - априорная вероятность класса y
(доля примеров класса y,
пропорция классов должна соответствовать)

p(x|y) - ф-ция правдоподобия класса y (плотность)

Классификатор: плотность

подходы к оценке плотности распределения:

- непараметрический
- параметрический
- смеси распределений

Классификатор: плотность

параметрический подход к оцениванию плотности

$$\hat{p}(x) = \varphi(x, \theta)$$

Классификатор: плотность

смеси распределений

$$\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi_j(x, \theta_j)$$

Классификатор: плотность

НЕпараметрический подход к оцениванию плотности

$$\hat{p}(x) = \sum_{j=1}^{m} \frac{1}{m V(h)} K\left(\frac{\rho(x, x_j)}{h}\right)$$

Классификатор: плотность

допущение (наивный Байес): признаки X - независимы друг от друга

тогда многомерную плотность
можно представить как произведение одномерных плотностей

$$p(x|y) = p_1(x_1|y) \ldots. p_n(x_n|y)$$

Классификатор: оценка плотности

дискретный случай (гистограмма) :

$$\hat{p}(x) = \frac{1}{m} \sum_{i=1}^{m} [x = x_i]$$

непрерывный случай:
эмпирическая оценка, окно ширины h

(доля объектов попавших в отрезок)

$$\hat{p}(x) = \frac{1}{2hm} \sum_{i=1}^{m} [|x - x_i| < h]$$

$$\hat{p}(x) = \frac{1}{mh} \sum_{i=1}^{m} \frac{1}{2} \left[\frac{|x - x_i|}{h} < 1 \right]$$

Классификатор: плотность

оценка Парзона-Розенблата

$$\hat{p}(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right)$$

K(r) - ядро

чётная ф-ция K(r)=K(-r)

нормированная $\int K(r)dr=1$

невозрастающая при r>0, неотрицательная ф-ция

Классификатор: плотность

оценка Парзона-Розенблата для класса y

K(r) - ядро

l_y - количество объектов y

ρ() - мера на X

V(h) - нормирующий множитель

$$\hat{p}(x|y) = \frac{1}{l_y V(h)} \sum_{i:y=y_i} K\left(\frac{\rho(x, x_i)}{h}\right)$$

Классификатор

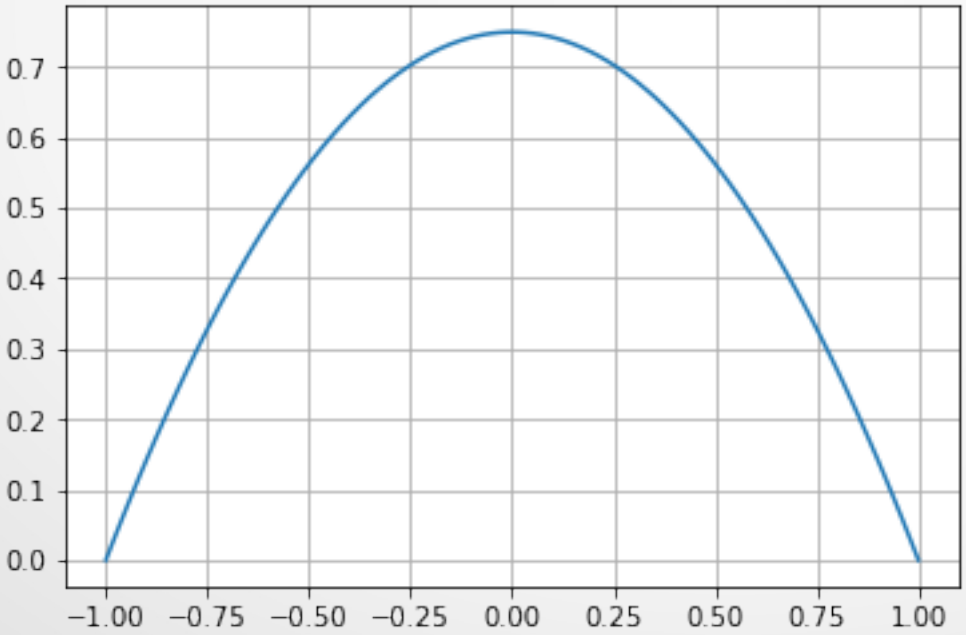
метод Парзоновского окна

$$a(x, X^l, h) = \underset{y \in Y}{argmax}\, \lambda_y \frac{P(y)}{l_y} \sum_{i: y = y_i} K\left(\frac{\rho(x, x_i)}{h}\right)$$

Классификатор

ядро Епанечникова

$$K(r)=\frac{3}{4}(1-r^2);|r|\leq 1$$

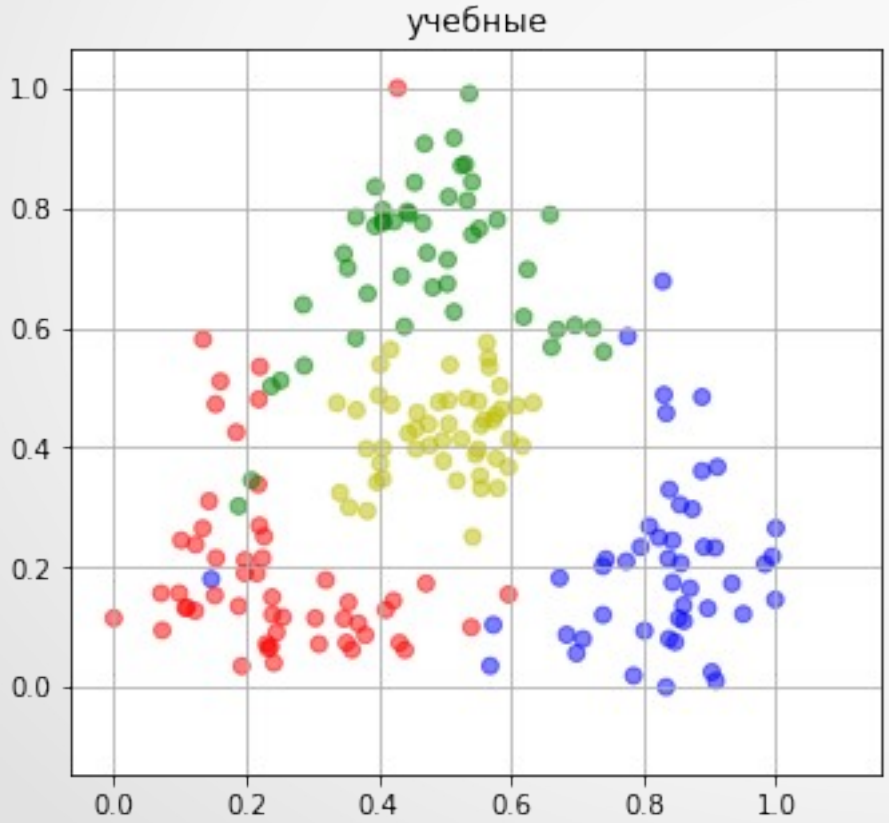


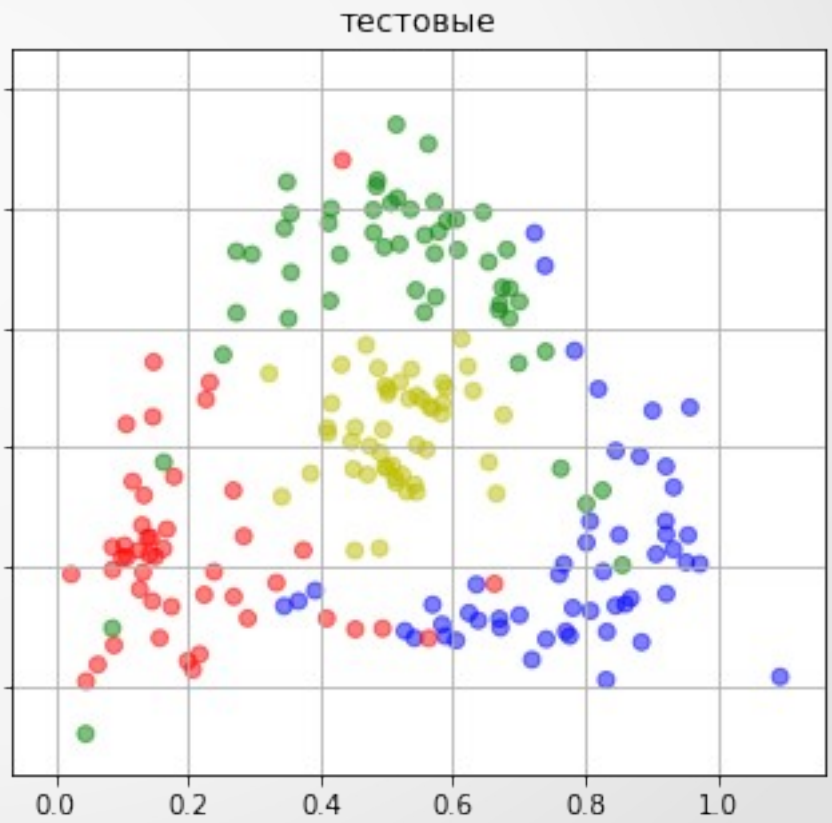
Классификатор

выбор оптимального размера окна
методом скользящего контроля (Leave One Out, LOO)

Классификатор: результат

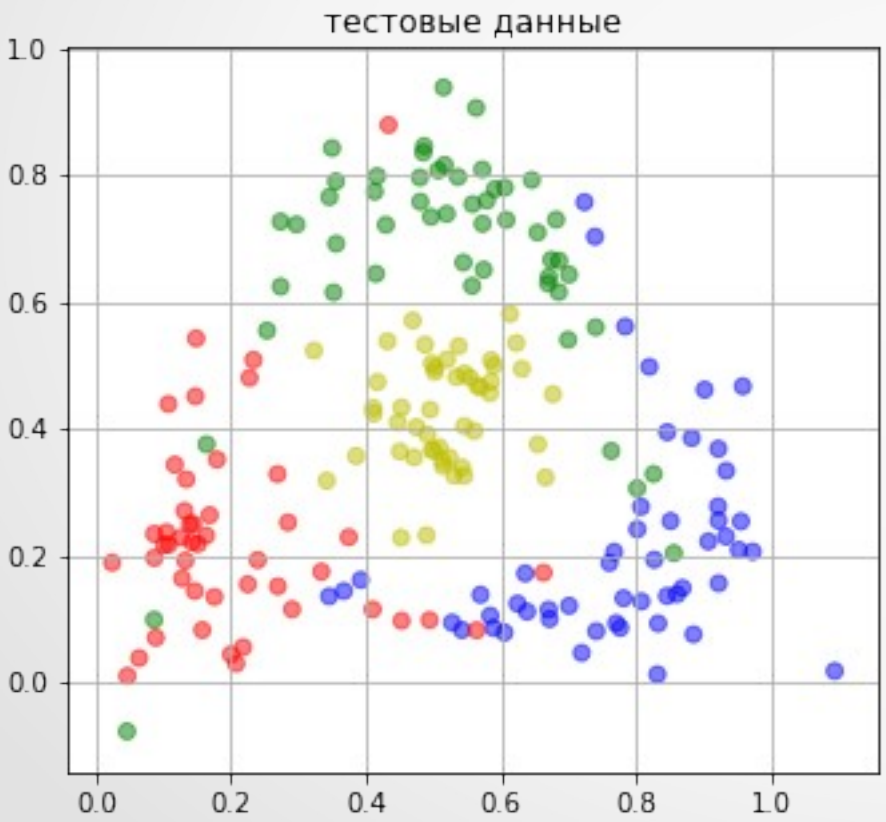
учебный набор

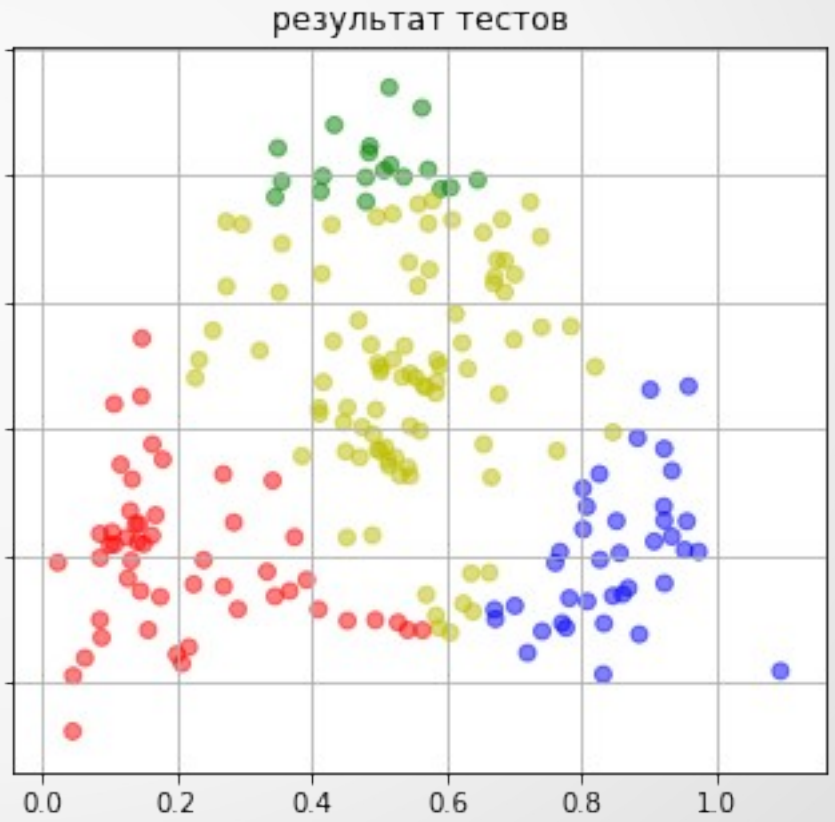




Классификатор: результат

результат теста



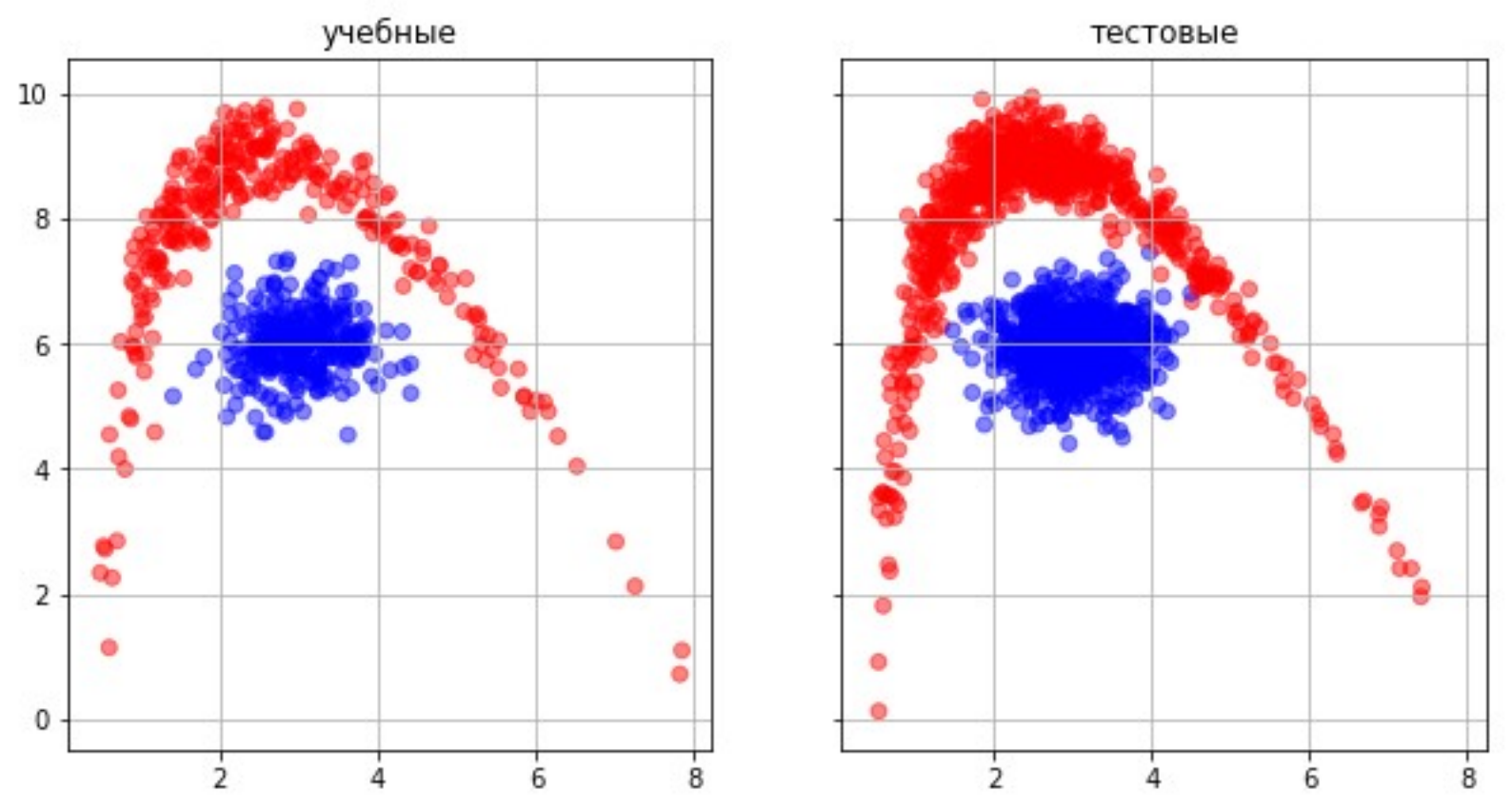


Классификатор: оценка результата 1

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

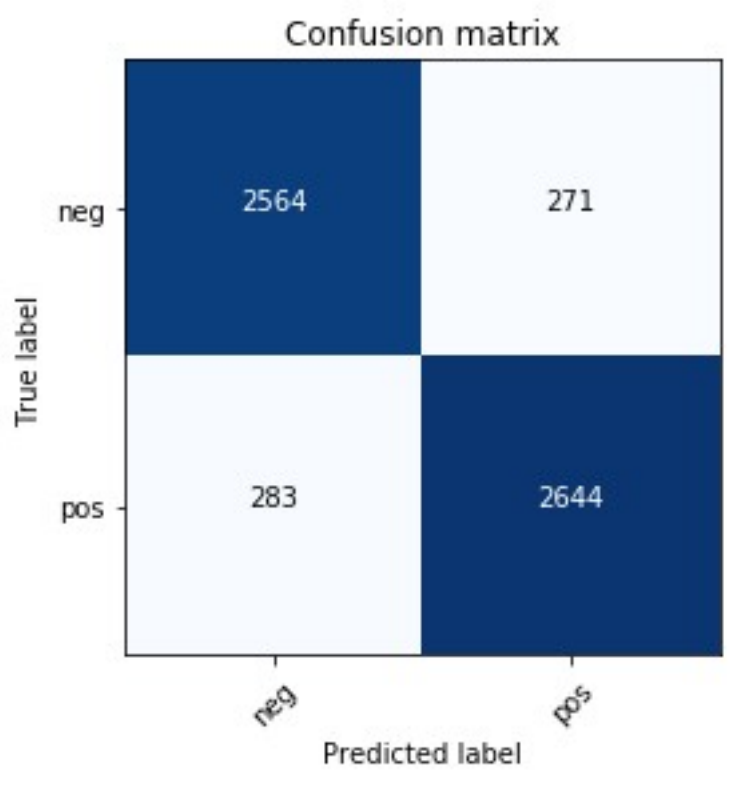
погрешность (accuracy)

правильные ответы / всего примеров

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

Классификатор: оценка результата 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

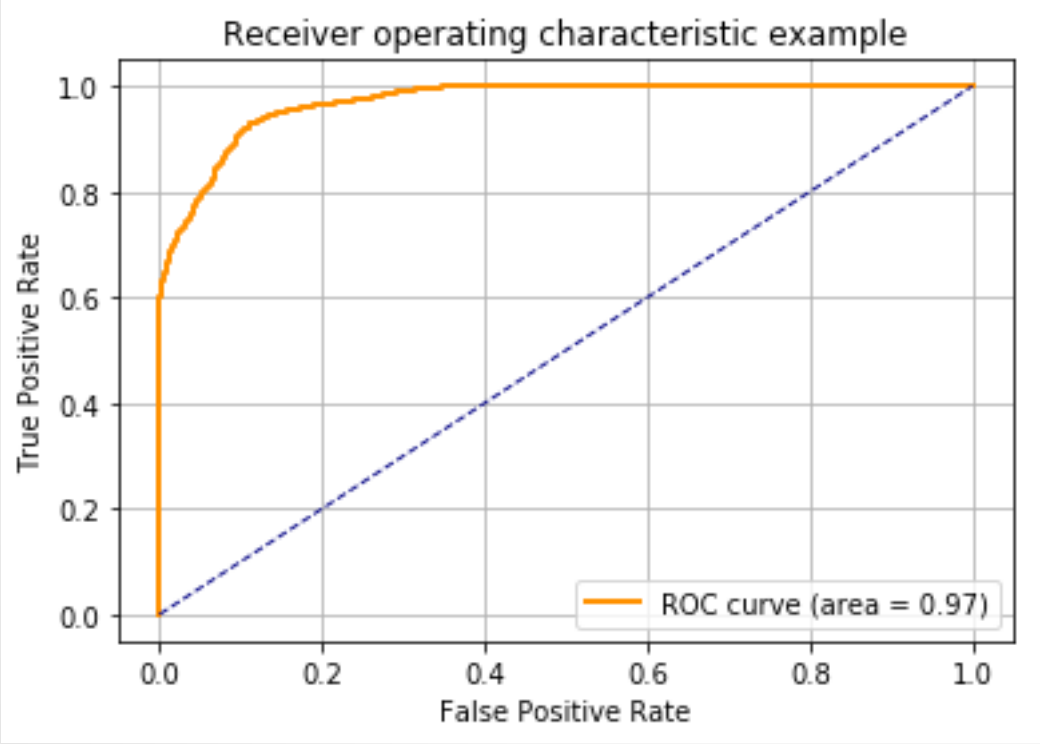
Классификатор: оценка результата 10

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. Байесовский классификатор.
http://mechanoid.kiev.ua/ml-bayes.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