

Лекция 9: логические методы

Евгений Борисов

логические методы

моделируем логику человеческих решений

интерпретируемость (для некоторых приложений это критично)

ЛОГИЧЕСКИЕ МЕТОДЫ

предикат - «простое» правило для выделения объектов

- предикат может быть описан естественным языком
- достаточно простая формула
- зависит от небольшого числа признаков

логические методы

примеры применения пороговых правил

если [возраст>60] **или** [ранее был инфаркт]
то операцию не делаем, риск неудачи > 60%

если [сумма<5000] **и** [зарплата>20000]
то кредит выдать, риск невозврата 5%

ЛОГИЧЕСКИЕ МЕТОДЫ

предикат - «простое» правило для выделения объектов

- предикат может быть описан естественным языком
- достаточно простая формула
- зависит от небольшого числа признаков

[длина > 10] и [ширина < 5] или [форма = квадрат]

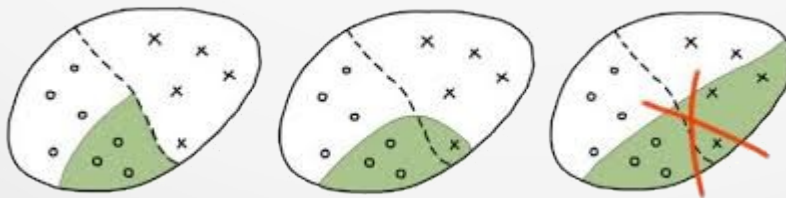
ЛОГИЧЕСКИЕ МЕТОДЫ

предикат - «простое» правило для выделения объектов

- предикат может быть описан естественным языком
- достаточно простая формула
- зависит от небольшого числа признаков

[длина > 10] и [ширина < 5] или [форма = квадрат]

- должен быть информативен, т.е. выделяет некоторое количество объектов одного класса



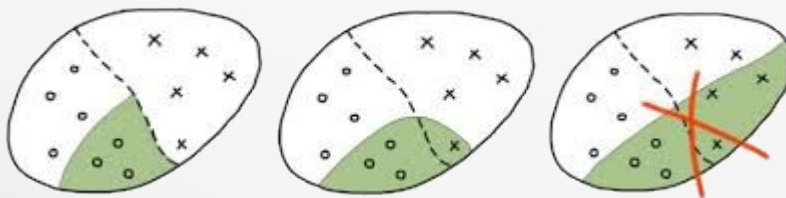
ЛОГИЧЕСКИЕ МЕТОДЫ

предикат - «простое» правило для выделения объектов

- предикат может быть описан естественным языком
- достаточно простая формула
- зависит от небольшого числа признаков

[длина > 10] и [ширина < 5] или [форма = квадрат]

- должен быть информативен, т.е. выделяет некоторое количество объектов одного класса

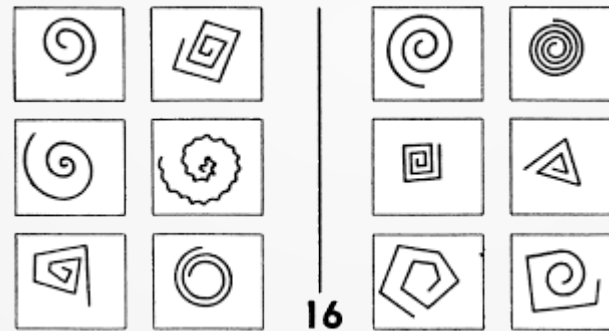


один предикат это маловато....

ЛОГИЧЕСКИЕ МЕТОДЫ

О ИНТУИТИВНОМ ПОНЯТИИ ЗАКОНОМЕРНОСТИ

тесты Бонгарда

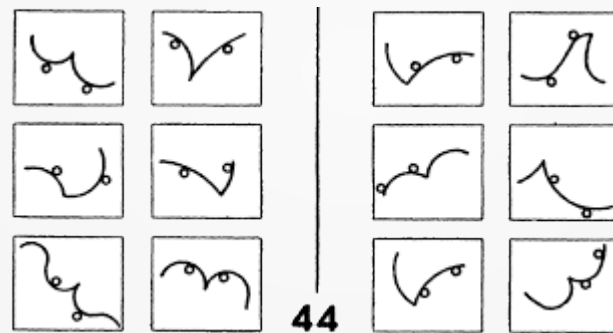


Бонгард М. М. Проблема узнавания.— М.: Физматгиз, 1967.

ЛОГИЧЕСКИЕ МЕТОДЫ

О ИНТУИТИВНОМ ПОНЯТИИ ЗАКОНОМЕРНОСТИ

тесты Бонгарда



Бонгард М. М. Проблема узнавания.— М.: Физматгиз, 1967.

ЛОГИЧЕСКИЕ МЕТОДЫ

закономерность - набор правил (предикатов)

- пороговое правило(decision stump) $R(x)=[a_i \leq f_i(x) < b_i]$

ЛОГИЧЕСКИЕ МЕТОДЫ

закономерность - набор правил (предикатов)

- пороговое правило(decision stump) $R(x)=[a_i \leq f_i(x) < b_i]$
- КОНЪЮНКЦИЯ $R(x)=\bigwedge_i [a_i \leq f_i(x) < b_i]$

ЛОГИЧЕСКИЕ МЕТОДЫ

закономерность - набор правил (предикатов)

- пороговое правило(decision stump) $R(x)=[a_i \leq f_i(x) < b_i]$
- КОНЪЮНКЦИЯ $R(x)=\bigwedge_i [a_i \leq f_i(x) < b_i]$
- синдром $R(x)=\left[\sum_i [a_i \leq f_i(x) < b_i] > d \right]$

ЛОГИЧЕСКИЕ МЕТОДЫ

закономерность - набор правил (предикатов)

- пороговое правило(decision stump) $R(x)=[a_i \leq f_i(x) < b_i]$
- конъюнкция $R(x)=\bigwedge_i [a_i \leq f_i(x) < b_i]$
- синдром $R(x)=\left[\sum_i [a_i \leq f_i(x) < b_i] > d \right]$
- полуплоскость $R(x)=\left[\sum_i w_i \cdot f_i(x) \geq w_0 \right]$

ЛОГИЧЕСКИЕ МЕТОДЫ

закономерность - набор правил (предикатов)

- пороговое правило(decision stump) $R(x)=[a_i \leq f_i(x) < b_i]$
- конъюнкция $R(x)=\bigwedge_i [a_i \leq f_i(x) < b_i]$
- синдром $R(x)=\left[\sum_i [a_i \leq f_i(x) < b_i] > d\right]$
- полуплоскость $R(x)=\left[\sum_i w_i \cdot f_i(x) \geq w_0\right]$
- шар $R(x)=[\rho(x_0, x) \leq w_0]$

ЛОГИЧЕСКИЕ МЕТОДЫ

задача: нужно отбирать «хорошие» закономерности

вопрос: как их оценивать

ЛОГИЧЕСКИЕ МЕТОДЫ

введём понятие информативности

как определить информативность предиката ?

предикат выделил объекты

r - количество позитивных

n - количество негативных

ЛОГИЧЕСКИЕ МЕТОДЫ

введём понятие информативности

как определить информативность предиката ?

предикат выделил объекты

p - количество позитивных

n - количество негативных

«простые» эвристики

p	n	$p-n$	$p-5n$	$\frac{p}{p}-\frac{n}{N}$	$\frac{p}{n+1}$
50	0	50	50	0.25	50
100	50	50	-150	0	1.96
50	9	41	5	0.16	5
5	0	5	5	0.03	5
100	0	100	100	0.5	100
140	20	120	40	0.5	6.67

ЛОГИЧЕСКИЕ МЕТОДЫ

информативность - энтропийный критерий

два исхода с вероятностями q и $1-q$

количество информации: $I_0 = -\log_2(q)$ $I_1 = -\log_2(1-q)$

энтропия - математическое ожидание количества информации

$$h(q) = -q \cdot \log_2(q) - (1-q) \cdot \log_2(1-q)$$

энтропия выборки **S** :

исходы q это принадлежность к классу **y**

$$H(y) = h\left(\frac{P}{S}\right)$$

S - количество объектов в выборке
 P - количество объектов класса **y**
(позитивных) в выборке

ЛОГИЧЕСКИЕ МЕТОДЫ

информативность - энтропийный критерий

S - количество объектов в выборке

P - количество объектов класса y (позитивных) в выборке

$$H(y) = h\left(\frac{P}{S}\right) \quad \text{энтропия выборки } S$$

предикат R выделил в S объекты

p - количество позитивных

n - количество негативных

$$H(y|R) = \frac{(p+n)}{S} \cdot h\left(\frac{p}{p+n}\right) + \frac{s-p-n}{S} \cdot h\left(\frac{P-p}{S-p-n}\right) \quad \text{энтропия выборки } S \text{ после}$$

получения информации R

информационный выигрыш (Information gain)

$$iGain(y, R) = H(y) - H(y|R)$$

ЛОГИЧЕСКИЕ МЕТОДЫ

P - количество объектов класса **y** (позитивных) в выборке

N - количество объектов класса не **y** (негативных) в выборке

S - количество объектов в выборке ($S = P+N$)

предикат **R** выделил в S объекты

p - количество позитивных

n - количество негативных

информативность

точный статистический тест Фишера

$$iStat(y, R) = \frac{-1}{S} \log_2 \left(\frac{C_P^p \cdot C_N^n}{C_S^{p+n}} \right)$$

ЛОГИЧЕСКИЕ МЕТОДЫ

предикат **R** выделил в **S** объекты

p - количество позитивных

n - количество негативных

q_c - априорная вероятность класса c, выделенного предикатом R

информативность

неопределенность Джини (Gini impurity)

$$Gini(y, R) = \sum_c q_c \cdot (1 - q_c) = \frac{p}{p+n} \cdot \left(1 - \frac{p}{p+n}\right) + \frac{n}{p+n} \cdot \left(1 - \frac{n}{p+n}\right)$$

ЛОГИЧЕСКИЕ МЕТОДЫ

основные вопросы построения логического классификатора

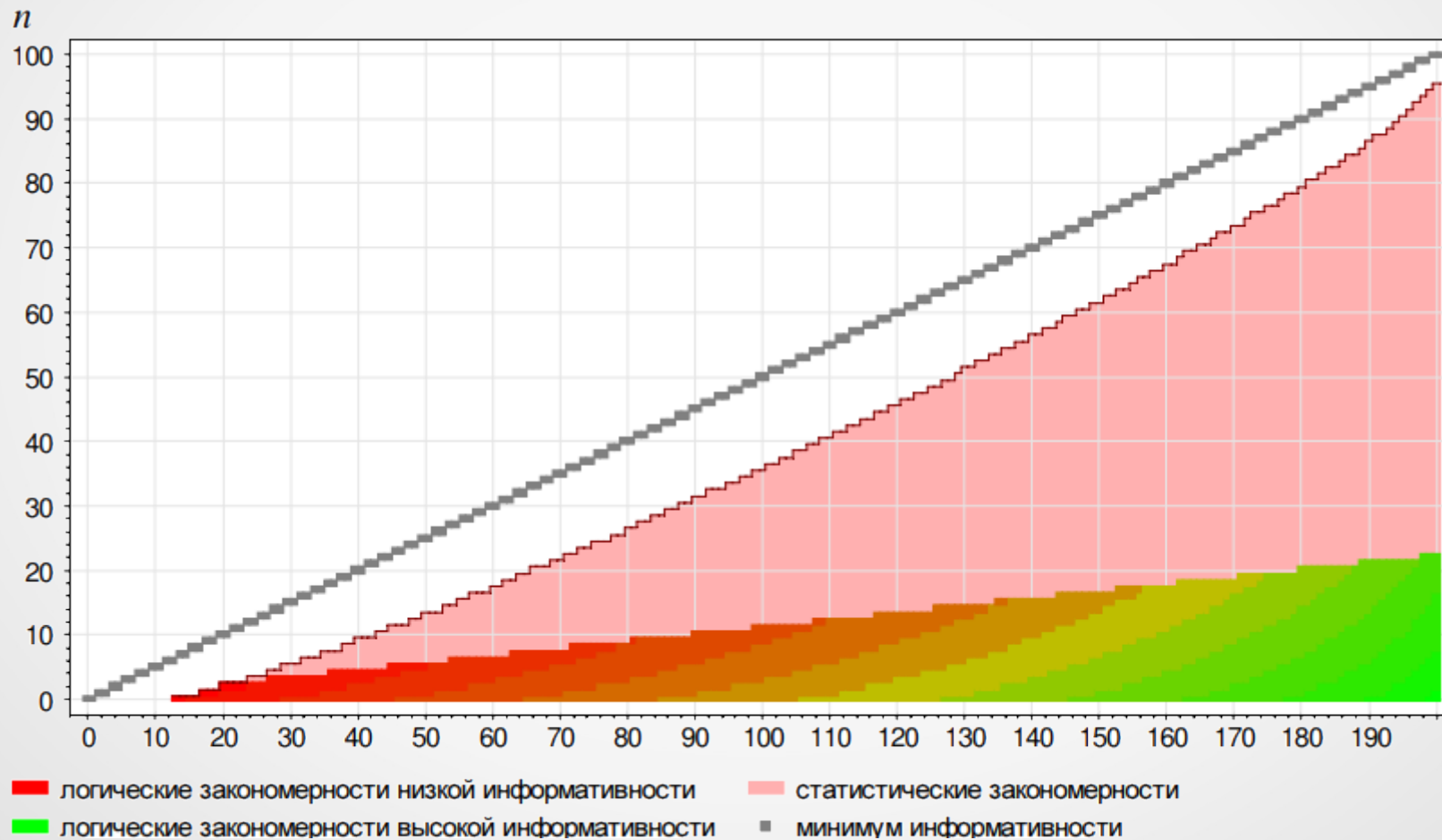
- как извлекать признаки
не наука, но творчество
- какого вида закономерности нужны
простые, малое количество признаков
- как определить информативность
iGain, Gini ...
- как искать закономерности
ограниченный перебор (rule induction)
- как объединить закономерности в алгоритм

ЛОГИЧЕСКИЕ МЕТОДЫ

как искать закономерности

P=200

N=100



неслучайность это ещё не закономерность

ЛОГИЧЕСКИЕ МЕТОДЫ

как объединить закономерности в алгоритм:

решающее дерево

рекурсивное разделение данных на две части

строим простой предикат -
ищем признак **i** и порог **b** для него

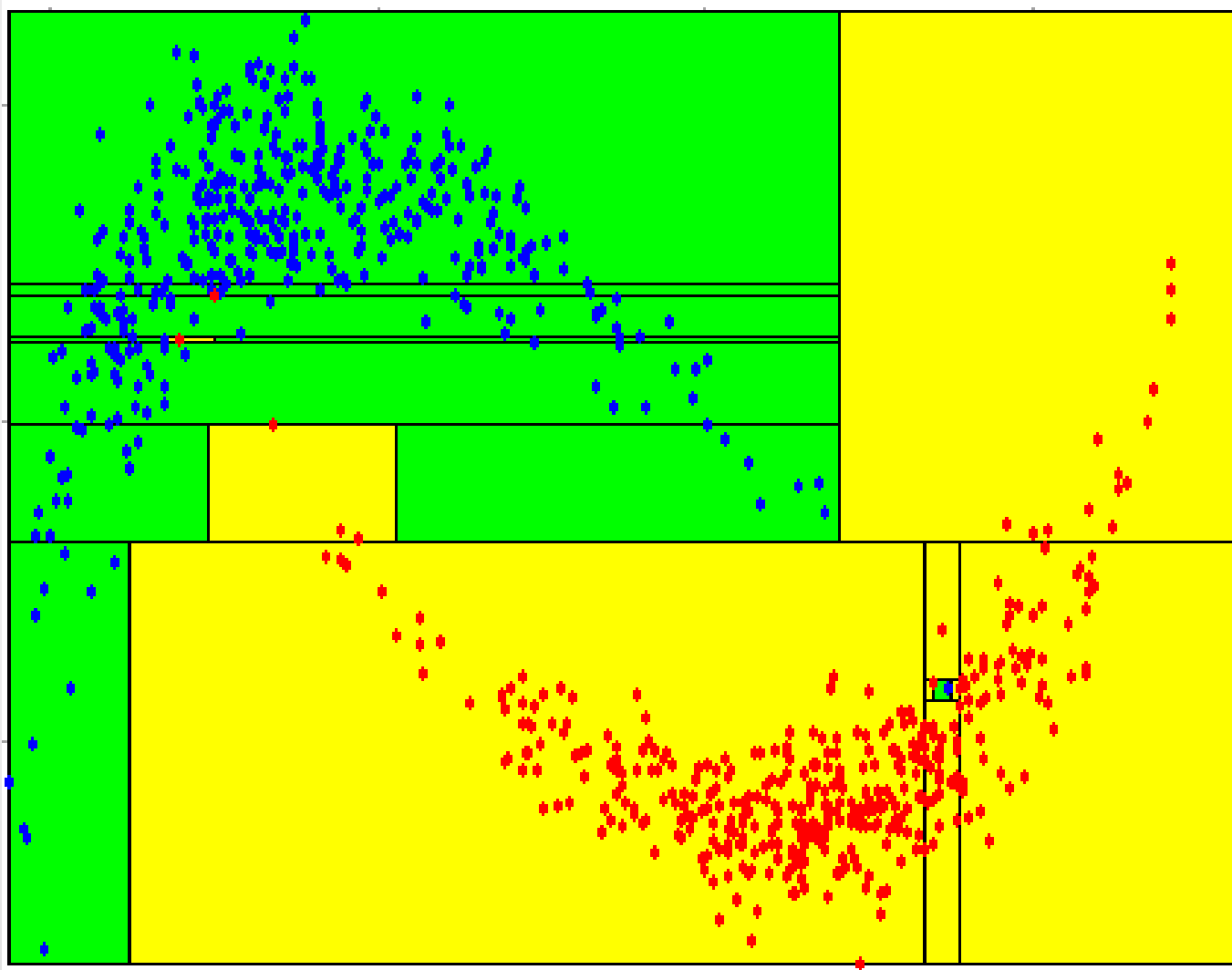
максимизируем информативность

$$\max_{i,b} (iGain(y, [X_i > b]))$$

$$\min(X_i) < b < \max(X_i)$$

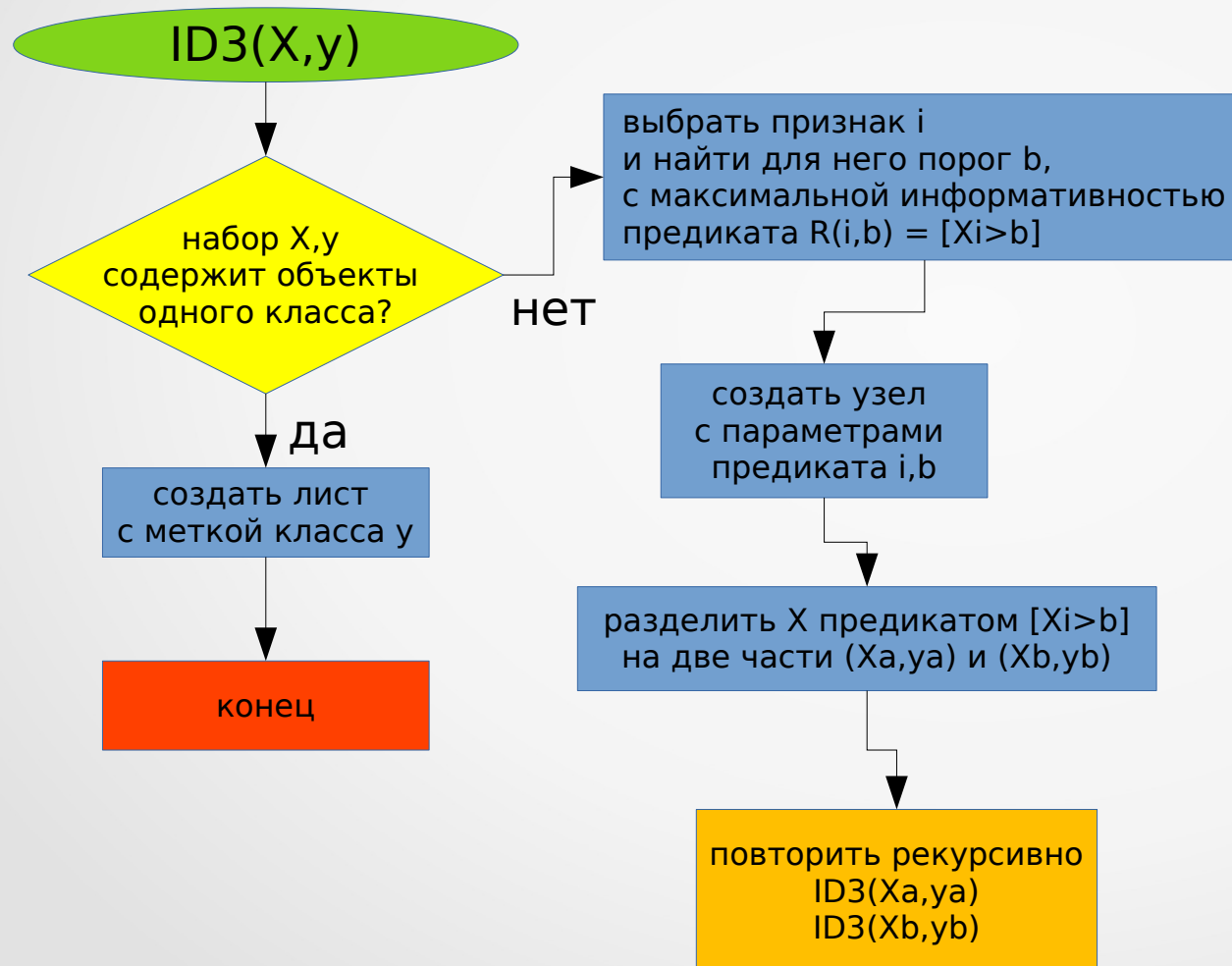
ЛОГИЧЕСКИЕ МЕТОДЫ

разделение набора объектов решающим деревом



ЛОГИЧЕСКИЕ МЕТОДЫ

как объединить закономерности в алгоритм:
решающее дерево, алгоритм ID3



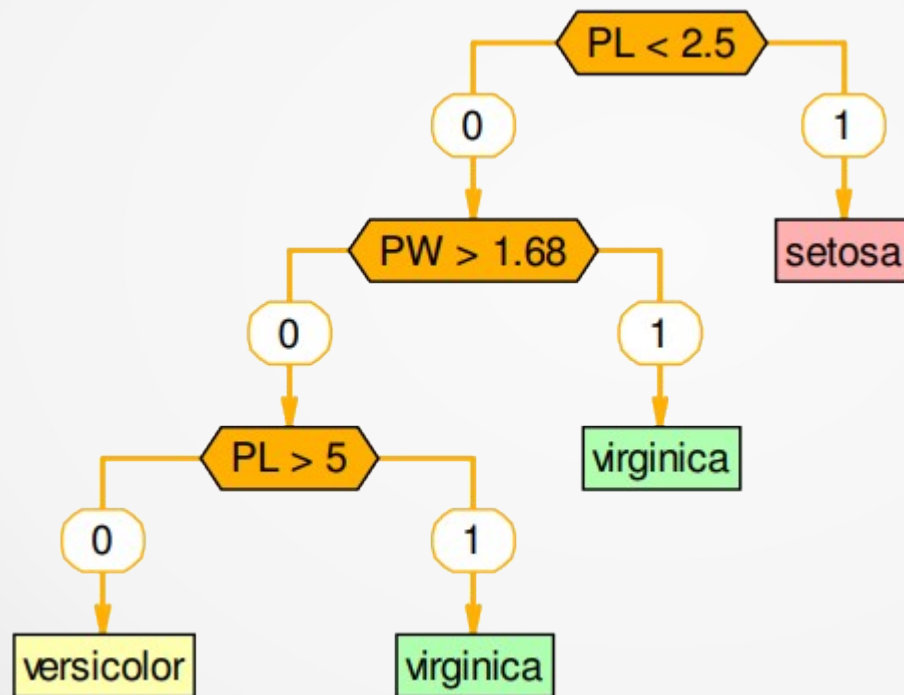
$$\max_{i,b} (iGain(y, [X_i > b]))$$

$$\min(X_i) < b < \max(X_i)$$

рекурсивное
разделение
данных на две
части

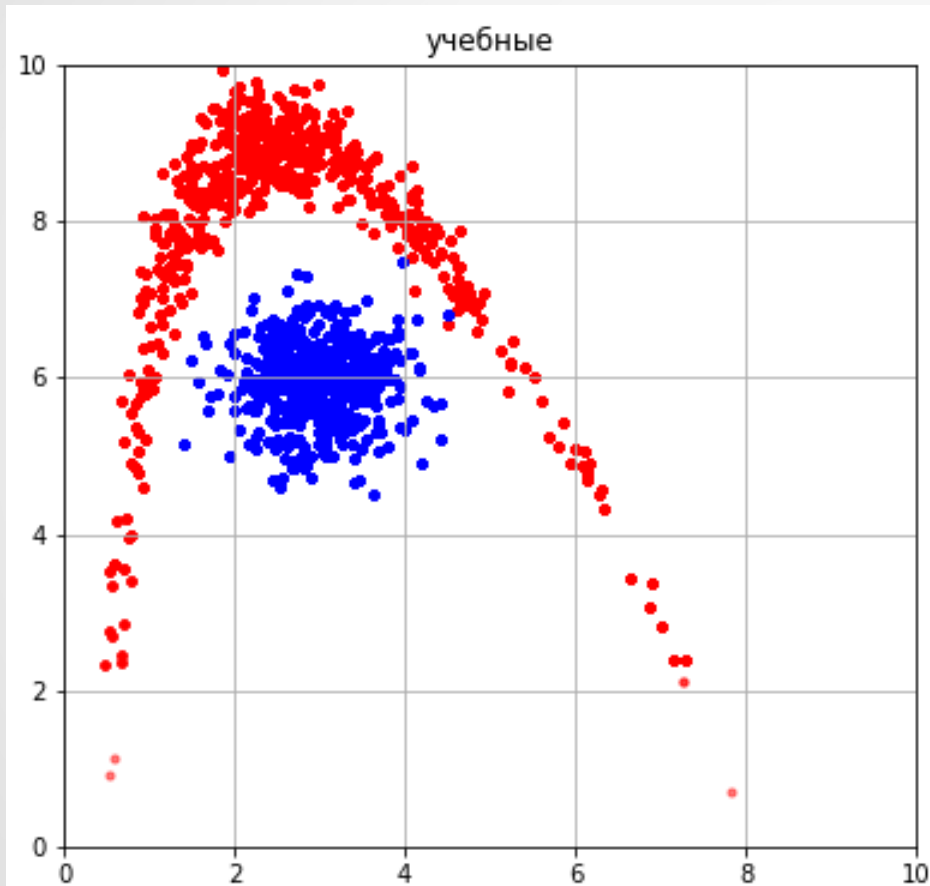
ЛОГИЧЕСКИЕ МЕТОДЫ

пример дерева для набора iris



ЛОГИЧЕСКИЕ МЕТОДЫ

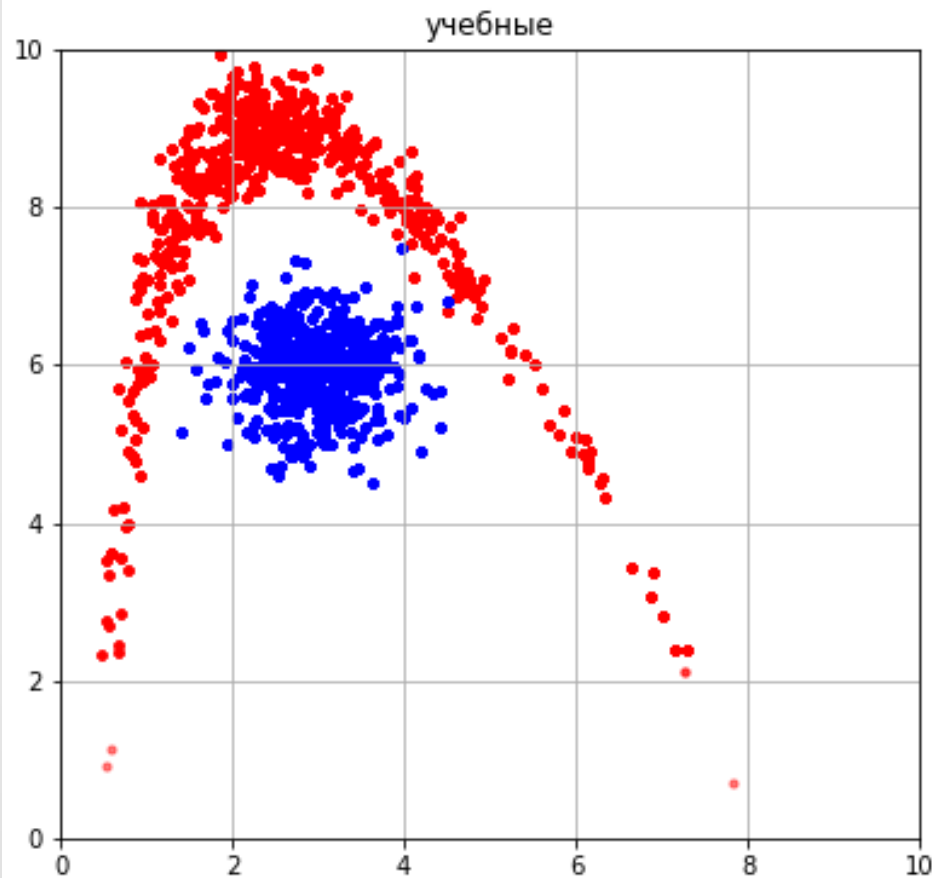
результат работы решающего дерева



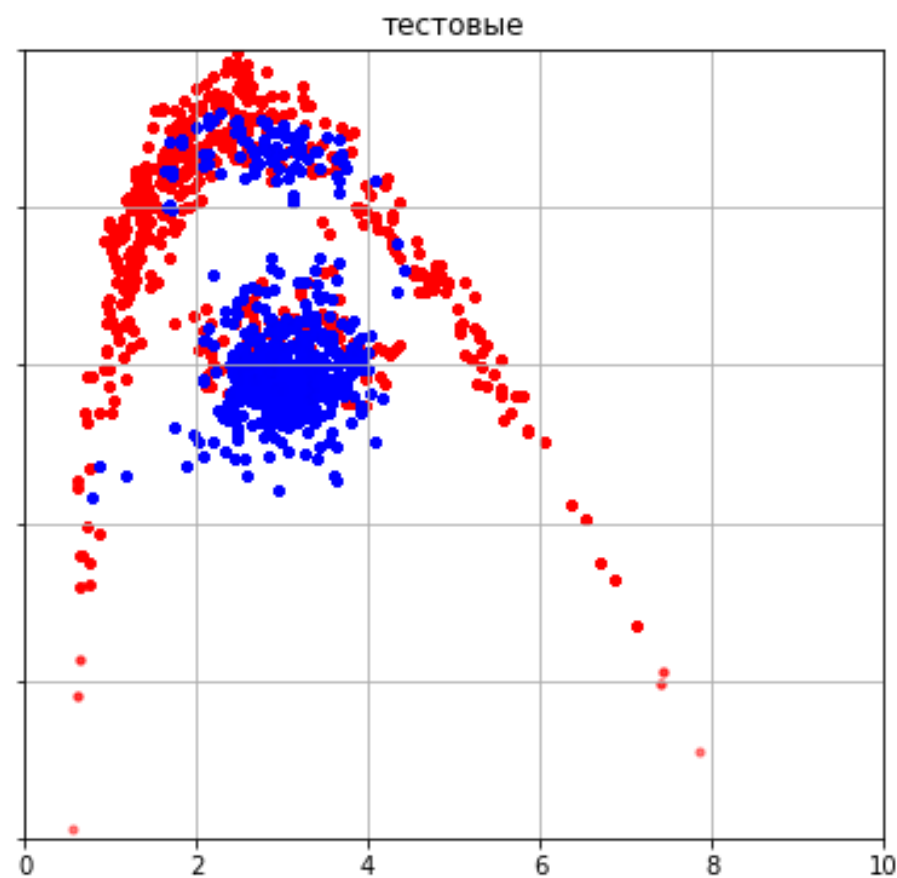
на учебном наборе - 100% точность

ЛОГИЧЕСКИЕ МЕТОДЫ

результат работы решающего дерева



на учебном наборе - 100% точность



на тесте - переобучение

ЛОГИЧЕСКИЕ МЕТОДЫ

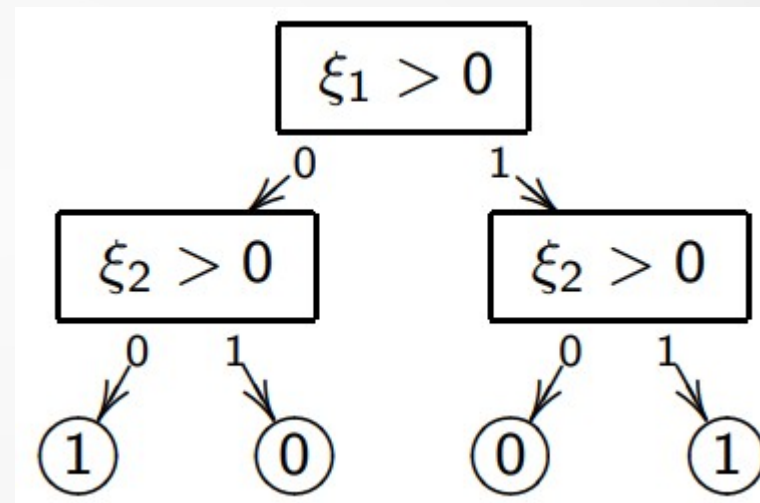
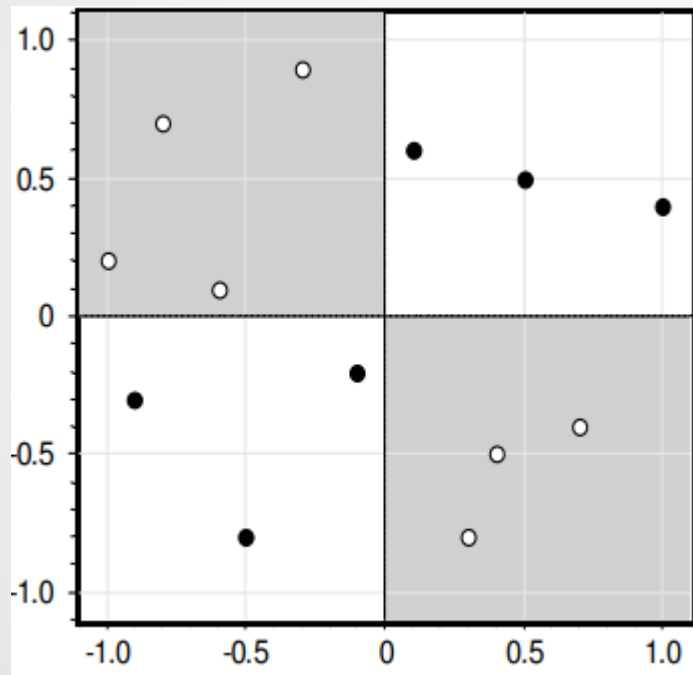
решающее дерево

достоинство: интерпретируемость результата

недостаток: переобучение, неустойчивы к шуму

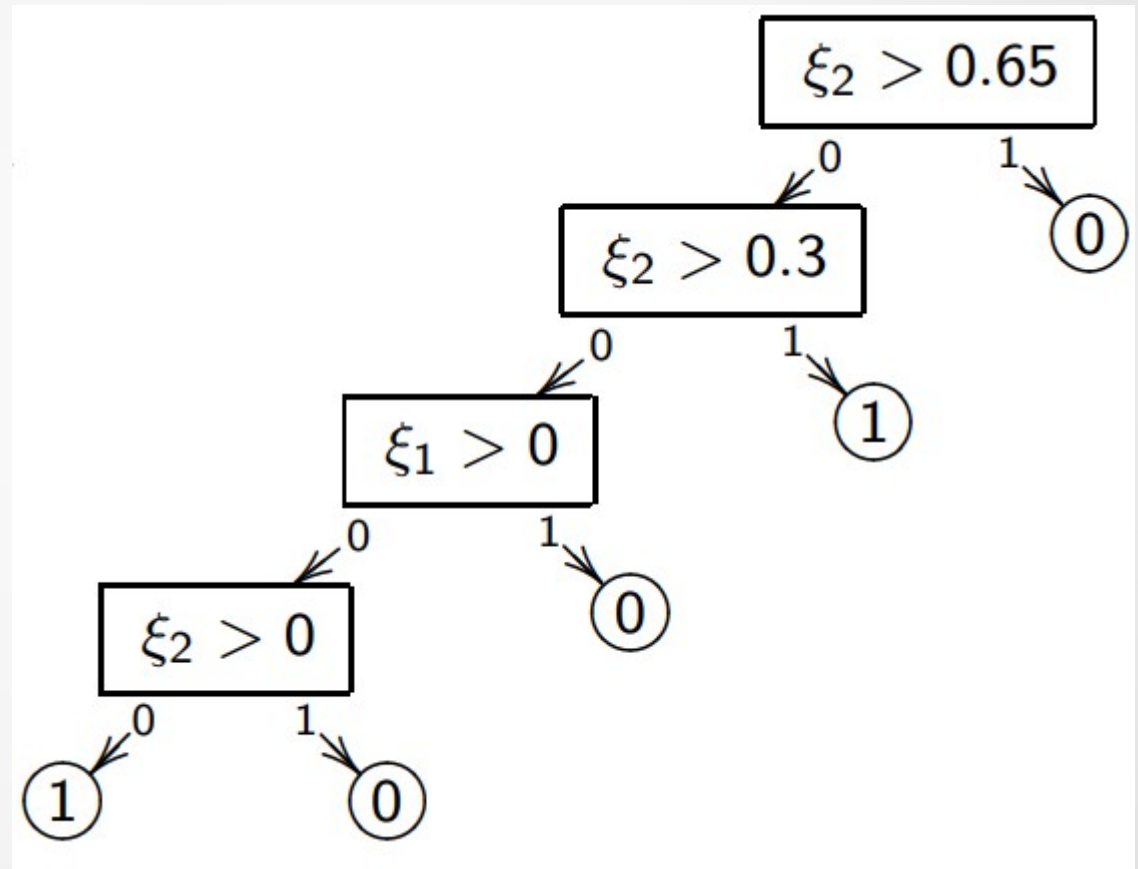
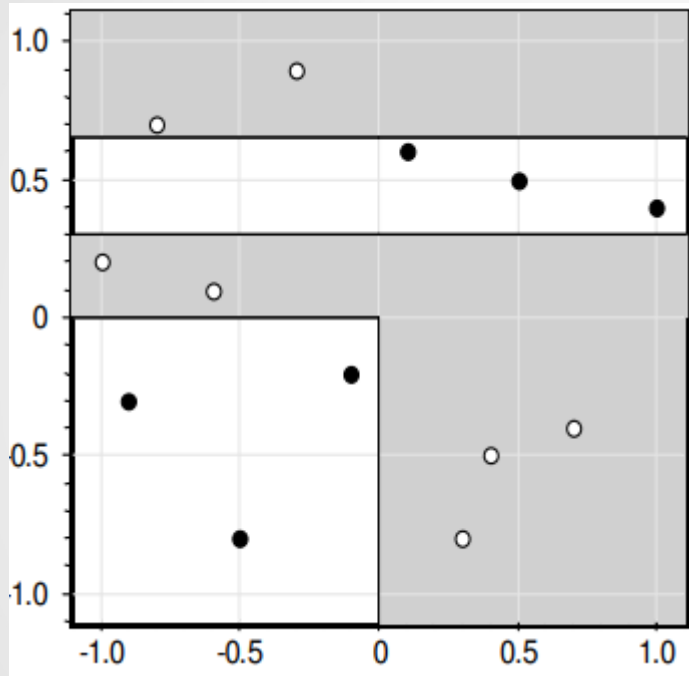
ЛОГИЧЕСКИЕ МЕТОДЫ

задача XOR : оптимальное дерево



ЛОГИЧЕСКИЕ МЕТОДЫ

задача XOR : результат «жадной» стратегии для дерева



ЛОГИЧЕСКИЕ МЕТОДЫ

pruning - обрезка решающего дерева

pre-pruning - критерий раннего останова.

если информативность меньше порога или глубина велика
то прекращаем ветвление

post-pruning - пост-редукция.

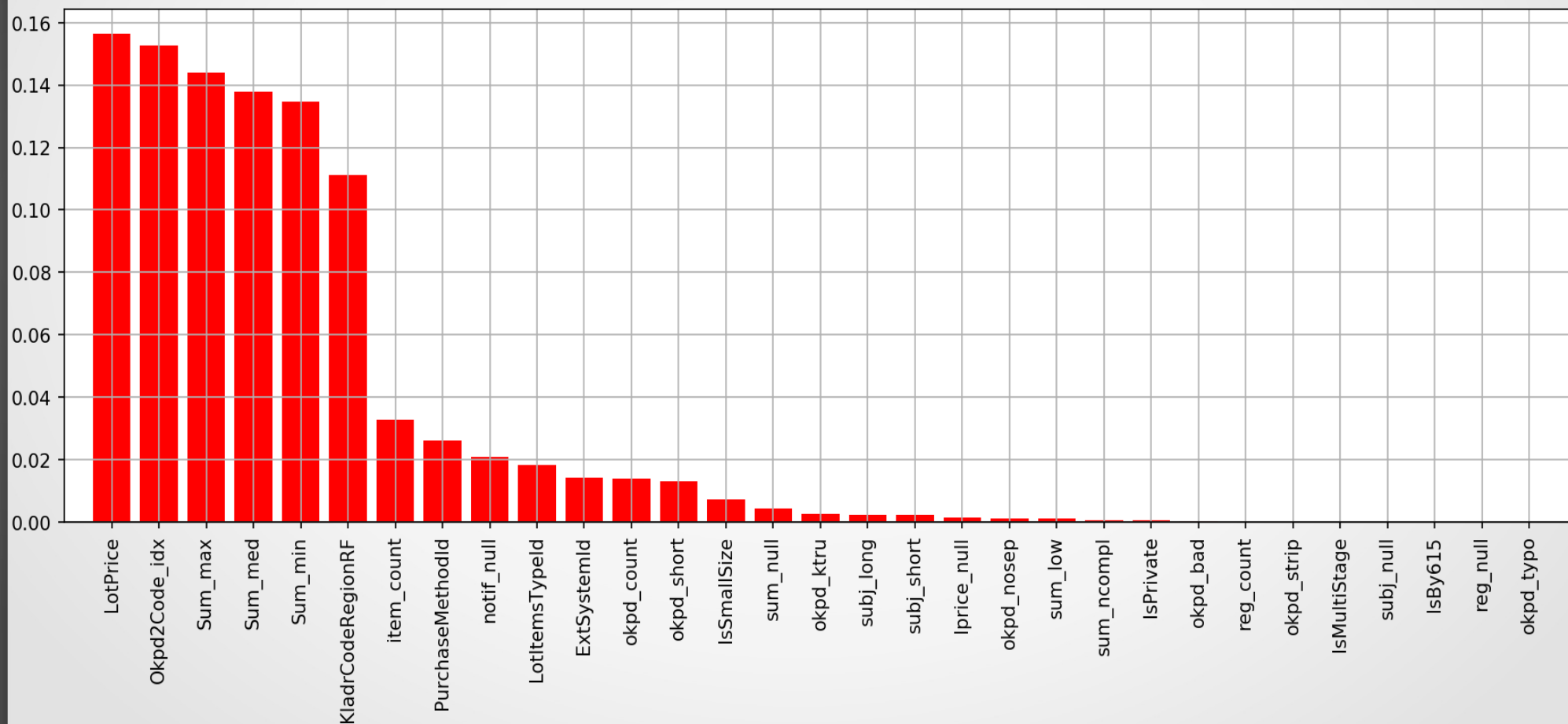
простматриваем все внутренние вершины дерева

проверяем их качество на тестовой выборке,

заменяем листом, где качество после разделения ухудшается

ЛОГИЧЕСКИЕ МЕТОДЫ

Оценка важности признаков (feature importances)



ЛОГИЧЕСКИЕ МЕТОДЫ

Оценка важности признаков (feature importances)

$$I_t = \frac{N_t}{N} \cdot \left(G_t - \frac{N_{tR}}{N_t} \cdot G_R - \frac{N_{tL}}{N_t} \cdot G_L \right)$$

G_t - неопределенность Джини (Gini impurity) в узле t

N - всего объектов учебной выборки,

N_t - количество объектов в узле t ,

G_L - неопределенность Джини для левой ветки

N_{tL} - количество объектов после разделения в узле t слева,

G_R - неопределенность Джини для правой ветки

N_{tR} - количество объектов после разделения в узле t справа,

ЛОГИЧЕСКИЕ МЕТОДЫ: ЛИТЕРАТУРА

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

- К.В. Воронцов Логические алгоритмы классификации. - курс "Машинное обучение" ШАД Яндекс 2014
- Е.С.Борисов Классификатор на основе решающего дерева.
<http://mechanoid.kiev.ua/ml-dtree.html>

ЛОГИЧЕСКИЕ МЕТОДЫ



Вопросы ?

ЛОГИЧЕСКИЕ МЕТОДЫ: ПРАКТИКА

ИСТОЧНИКИ ДАННЫХ ДЛЯ ЭКСПЕРИМЕНТОВ



sklearn.datasets
UCI Repository
kaggle



ЗАДАНИЕ

- посчитать число узлов и листьев
- pre-pruning (ограничить глубину дерева)