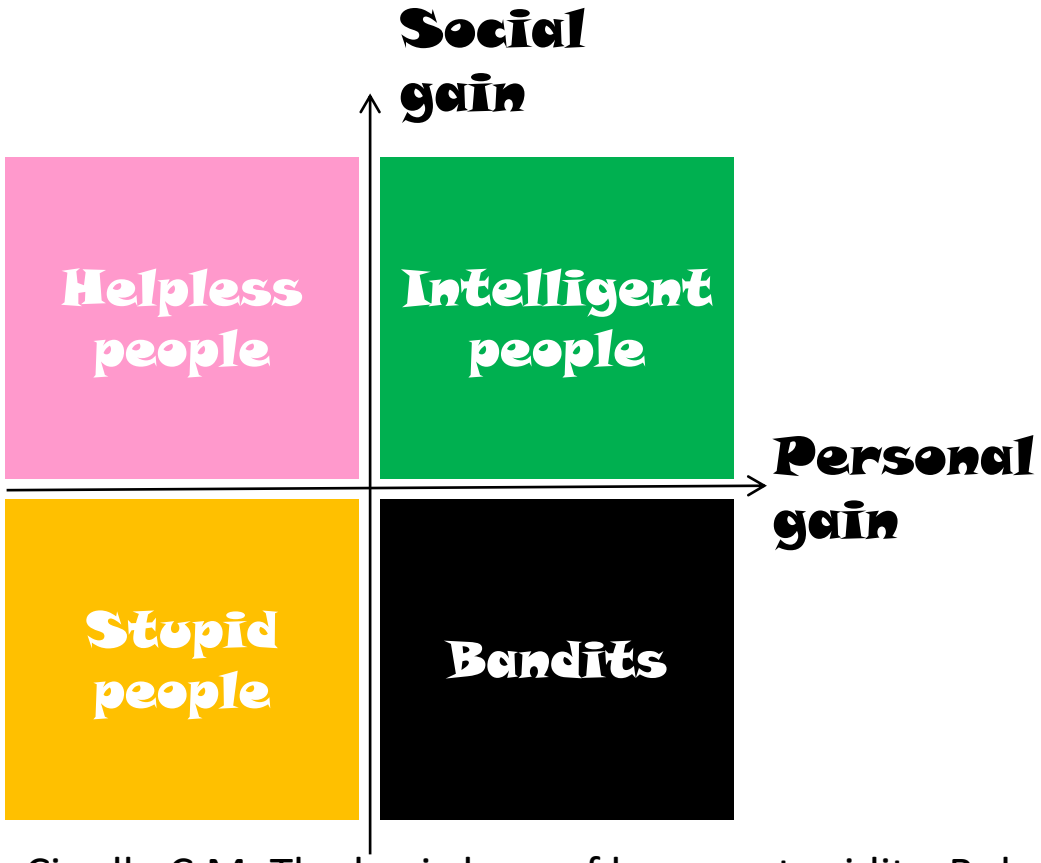
Задача классификации данных

Классификация – нить Ариадны в лабиринте природы.

Жорж Санд



Cipolla C.M. The basic laws of human stupidity. Bologna: il Muli, 2011

Содержание

- **Основные понятия**
- **Деревья решений**
- Байесовская классификация
- Классификация по ближайшим соседям
- Оценка качества классификации
- Ансамблевая классификация

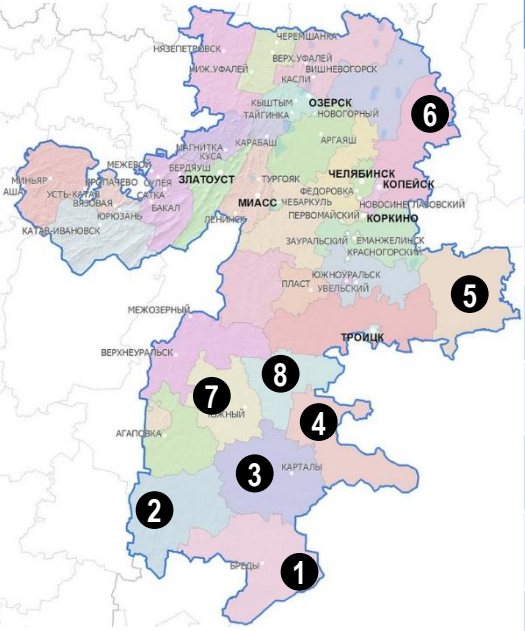
Задача классификации

- Построение формальной модели, которая распределяет объекты, имеющие одинаковую структуру по заранее известным группам (классам) в зависимости от похожести атрибутов объектов
- Основные задачи классификации
  - **Предсказание:** назначить корректный класс объекту, который предварительно не был рассмотрен
  - **Описание:** указать способ, с помощью которого можно отличать объекты различных классов

Атрибуты и метки класса

Приложение	Набор атрибутов	Метка класса
Кредитный скоринг клиентов	Пол, возраст и доход клиента, величина, процент и срок кредита	Надежный / ненадежный
Предсказание успеваемости студентов	Пол, курс, местный/приезжий, количество посещений, количество сданных заданий	Отлично / хорошо / удовлетворительно / неудовлетворительно
Выявление спама	Характеристики, полученные из заголовка и тела сообщения	Спам / не спам
Идентификация опухоли	Характеристики, полученные из снимков МРТ	Злокачественная / доброкачественная
Классификация галактик	Характеристики, полученные из снимков с телескопа	Эллиптическая / спиральная / нерегулярной формы

Пример: предсказание полезных ископаемых







#	$FeSO_4 \bullet 7H_2O$		$NH_4H_2PO_4$		$(Na,Ca)(Si,Al)_4O_8$		$SiO_2 \bullet nH_2O$		$LiAlSi_4O_{10}$		…	Мине-рал
	Нали чие	Плот ность	Нали чие	Плот ность	Нали чие	Плот ность	Нали чие	Плот ность	Нали чие	Плот ность		
1	Да	3.40	Нет	-	Да	7.80	Нет	-	Да	23.92		Железо
2	Нет	-	Да	7.22	Да	2.97	Да	5.97	Да	16.54		Медь
3	Да	4.67	Да	5.45	Да	5.43	Да	8.95	Да	28.49		Серебро
4	Нет	-	Да	3.12	Нет	-	Да	9.12	Нет	-		Цинк
5	Да	2.78	Да	0.18	Нет	-	Нет	-	Да	25.02		Железо
6	Да	1.02	Нет	-	Нет	-	Да	1.23	Да	2.12		НЕТ
7	Да	0.75	Нет	-	Нет	-	Да	3.10	Да	2.99		НЕТ
8	Нет	-	Да	0.36	Да	2.08	Нет	-	Нет	-		НЕТ

Пример: классификация позвоночных

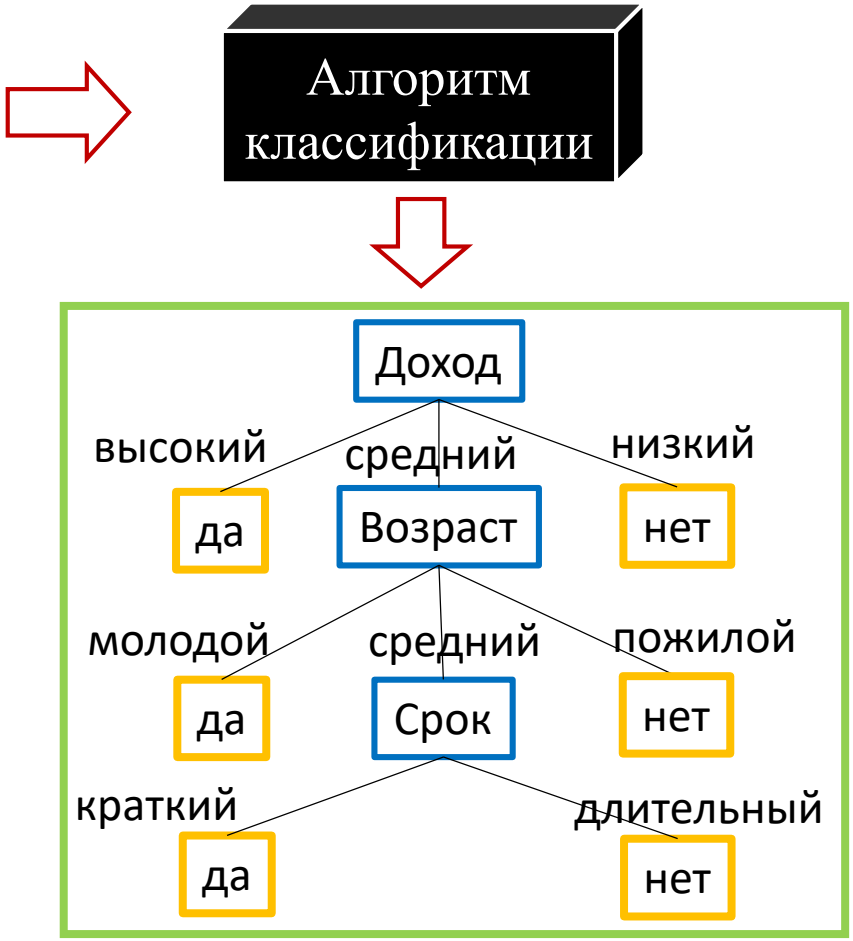
Vertebrate Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Vertebrate Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
gila monster	cold-blooded	scales	no	no	no	yes	yes	?

Процесс классификации: обучение (индукция)

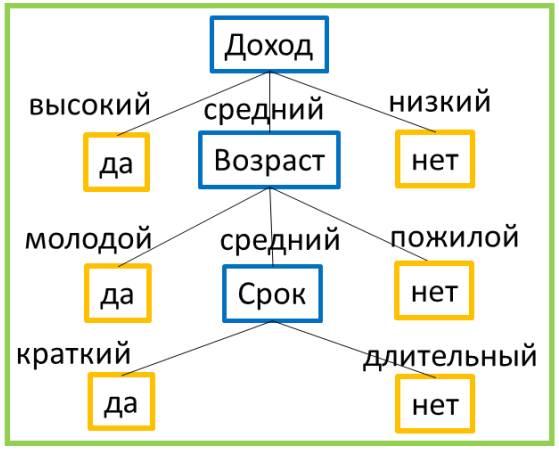
Обучающая выборка

ФИО	Доход	Возраст	Срок	Дать кредит
Бонд Дж.	средний	пожилой	краткий	нет
Бэннер Б.	высокий	пожилой	длительный	да
Кент К.	низкий	пожилой	краткий	нет
Паркер П.	низкий	молодой	длительный	нет
Скайуокер Э.	низкий	молодой	краткий	нет
Сойер Т.	высокий	молодой	длительный	да
Соло Х.	средний	средний	краткий	да
Старк Т.	высокий	средний	длительный	да
Джордан Х.	высокий	молодой	краткий	да
Хоулетт Дж.	средний	пожилой	длительный	нет



Модель классификации

Процесс классификации: оценка модели



Модель классификации

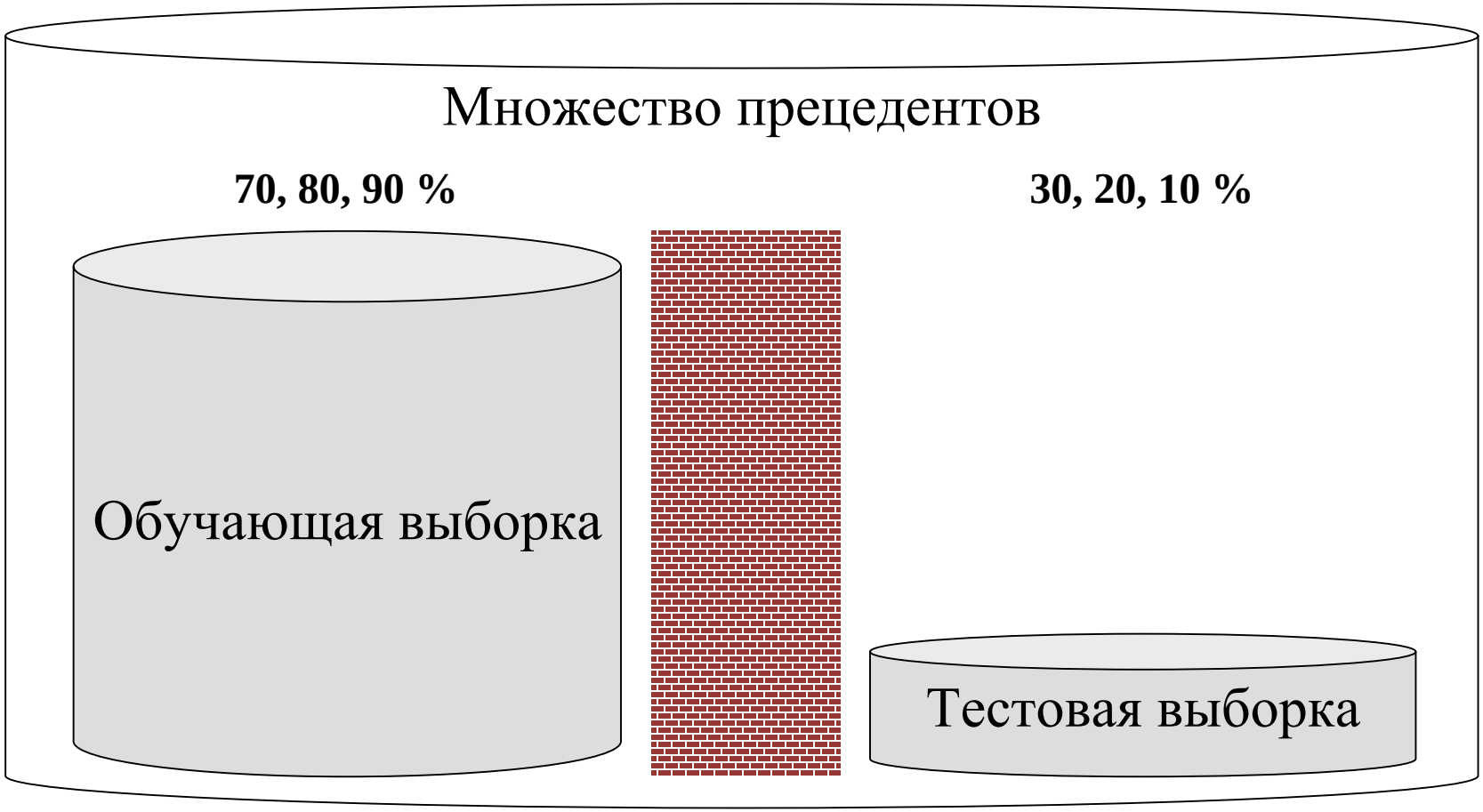
Тестовая выборка

ФИО	Доход	Возраст	Срок	Дать кредит
Вейдер Д.	низкий	средний	короткий	нет
Винду М.	низкий	молодой	длинный	нет
Дюррон К.	низкий	молодой	короткий	да
Органа Л.	высокий	средний	да	да
Хатт Дж.	средний	пожилой	длинный	нет

Дать кредит
нет
нет
нет
да
нет

Точность 80%

Процесс классификации: обучающая *vs.* тестовая выборки

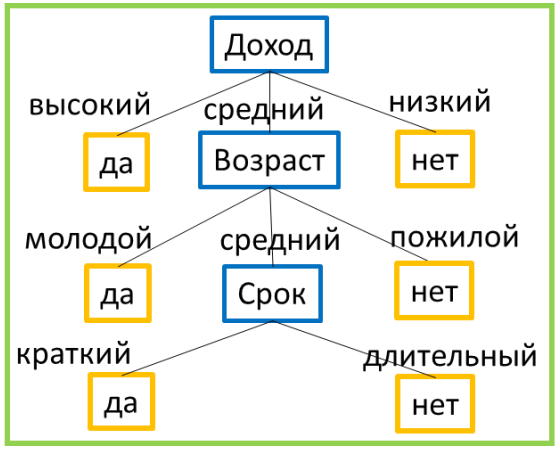


Оценка модели: матрица ошибок

		Реальный класс: **A**	Реальный класс: **B**
Предсказанный класс	**A**	TP	FP
	B	FN	TN

- $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
- $Error\ rate = \frac{FP+FN}{TP+FP+FN+TN}$

Процесс классификации: применение (дедукция)



Модель классификации

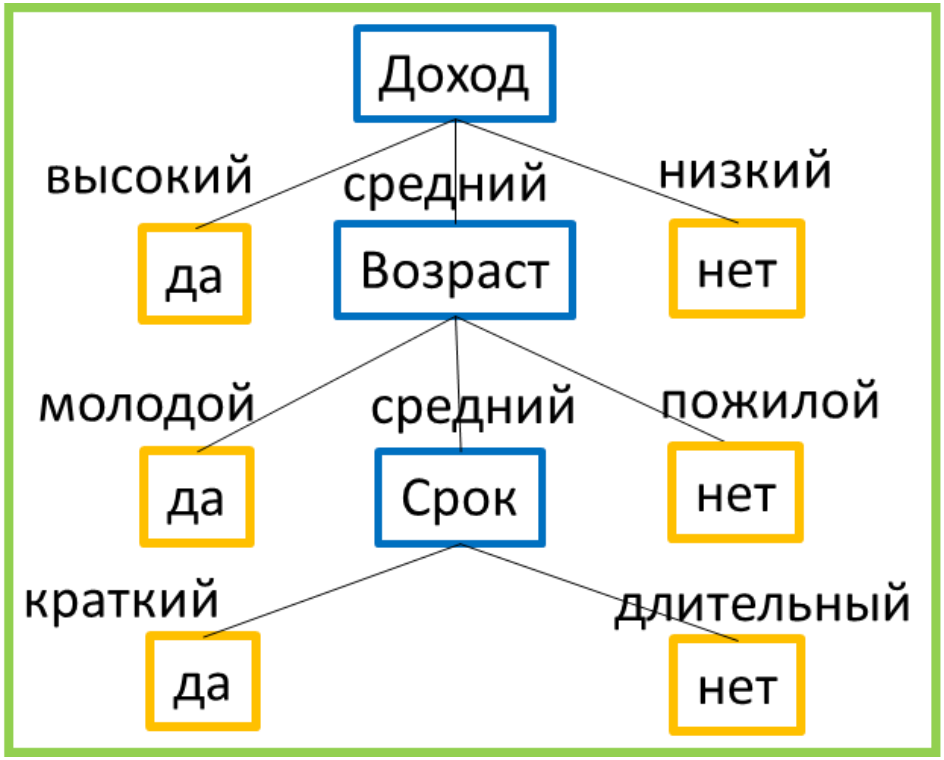
Неизвестные данные

ФИО	Доход	Возраст	Срок
Бендер О.С.	низкий	молодой	длительный
Воробьянинов И.М.	высокий	пожилой	длительный
Михельсон К.К.	высокий	пожилой	краткий
Востриков В.И.	низкий	средний	краткий
…			

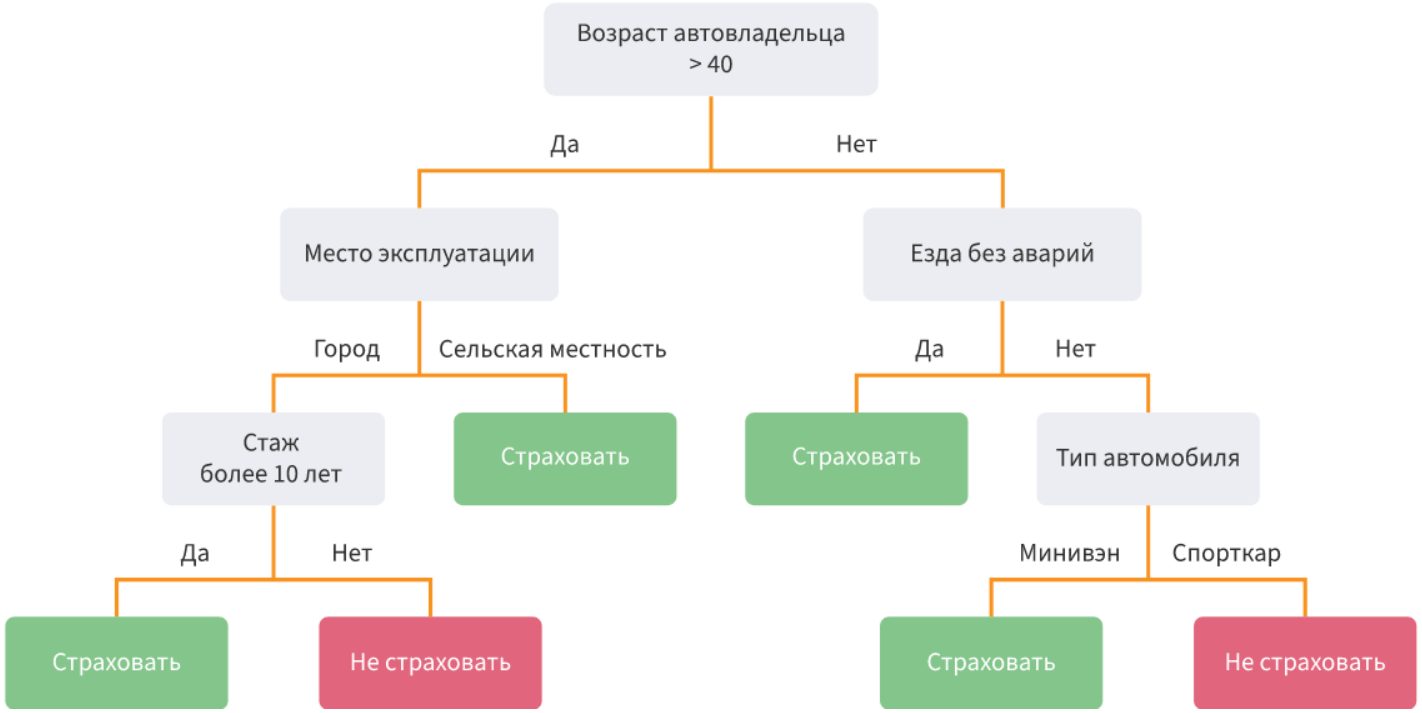
Дать кредит
нет
да
да
нет
…

Деревья решений

- Дерево решений – модель классификации в виде дерева, которое имеет
  - корневой узел и внутренние узлы: проверка условия на атрибут объекта
  - узлы-листья: метки классов
  - ребра: переходы по результату проверки



Деревья решений

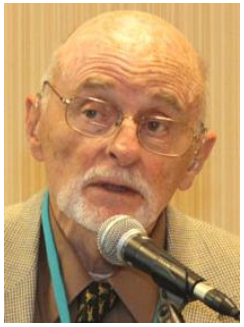


- Для применения не требуется компьютер
- Сферы применения: банковское дело, страхование, торговля, медицина, контроль качества продукции

Построение дерева решений (алгоритм Ханта)

NULL

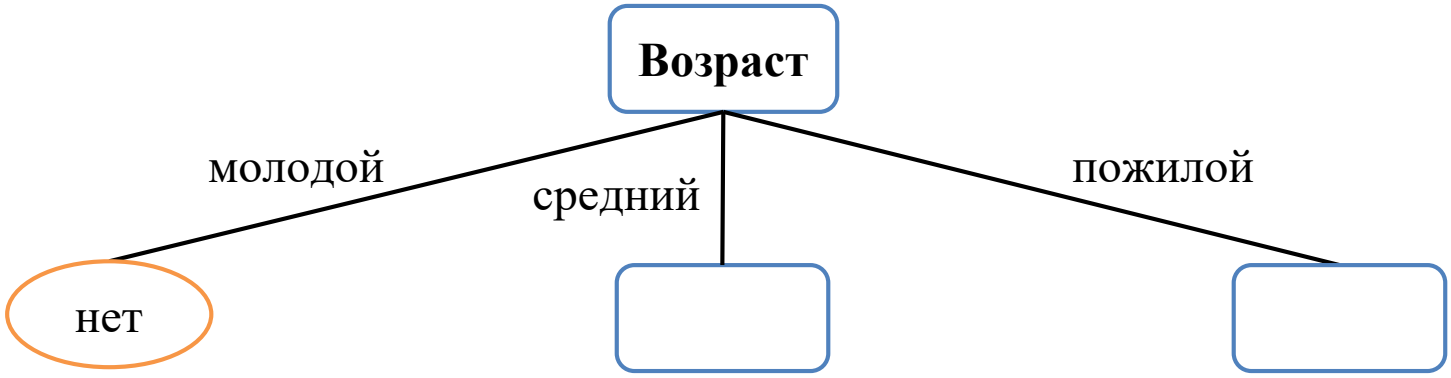
ФИО	Доход	Возраст	…	Дать кредит
Бонд Дж.	средний	пожилой		нет
Бэннер Б.	высокий	пожилой		да
Кент К.	низкий	пожилой		нет
Паркер П.	низкий	молодой		нет
Скайуокер Э.	низкий	молодой		нет
Сойер Т.	высокий	средний		да
Соло Х.	низкий	средний		нет
Старк Т.	высокий	средний		да



Earl Hunt
1933-2016

- Если все объекты из одного класса, то создать лист с меткой этого класса, иначе выбрать атрибут для разбиения

Построение дерева решений (алгоритм Ханта)



ФИО	Доход	Дать кредит
Паркер П.	низкий	нет
Скайуокер Э.	низкий	нет

ФИО	Доход	Дать кредит
Старк Т.	высокий	да
Соло Х.	низкий	нет
Сойер Т.	высокий	да

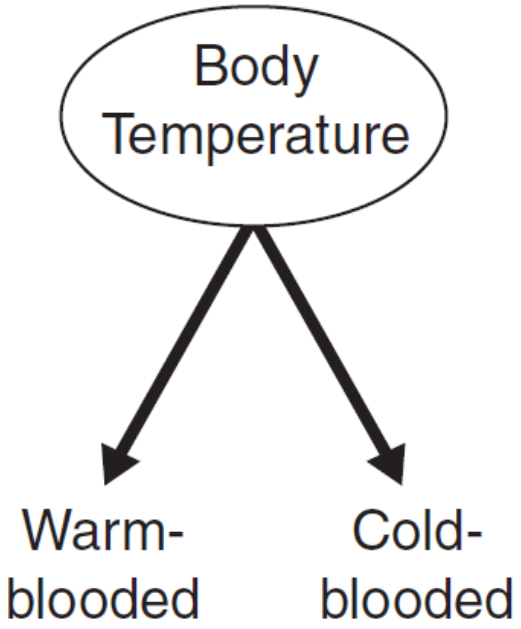
ФИО	Доход	Дать кредит
Кент К.	низкий	нет
Бонд Дж.	средний	нет
Бэннер Б.	высокий	да

- Разбить выборку в соответствии со значениями выбранного атрибута
- Рекурсивно построить поддеревья

Важные аспекты построения дерева решений

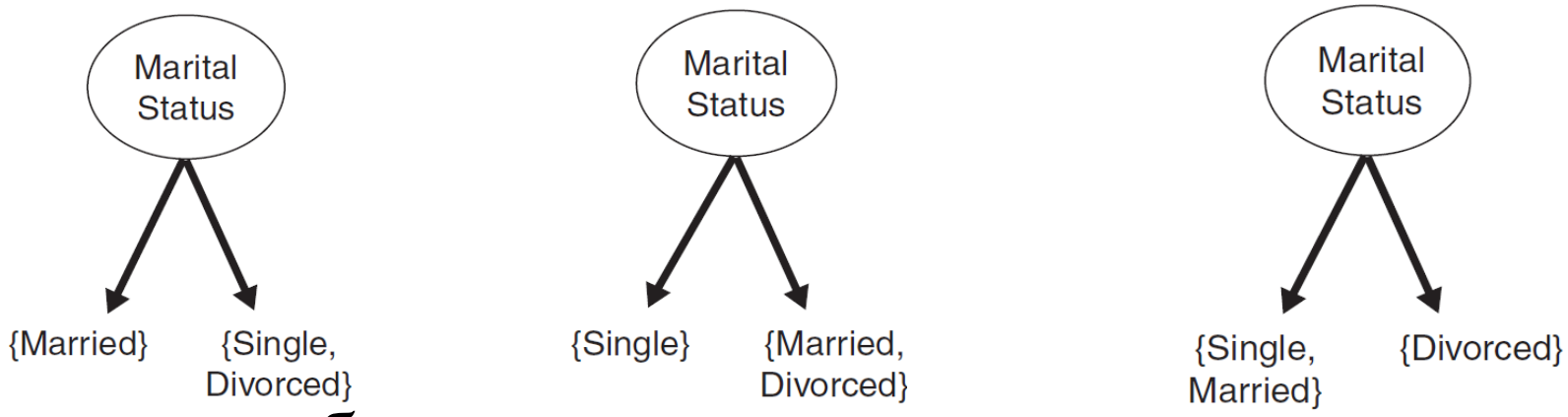
- Как задавать условие проверки атрибута?
  - Зависит от типа атрибута
- Как выбирать атрибут разбиения?
  - Нужен критерий выбора
- Когда останавливать построение дерева?
  - Когда все объекты выборки из одного класса либо имеют одинаковые значения атрибутов
  - Не всегда нужно и полезно, чтобы все объекты выборки были из одного класса (быстродействие, понятность дерева и др.)

Бинарные (булевы) атрибуты

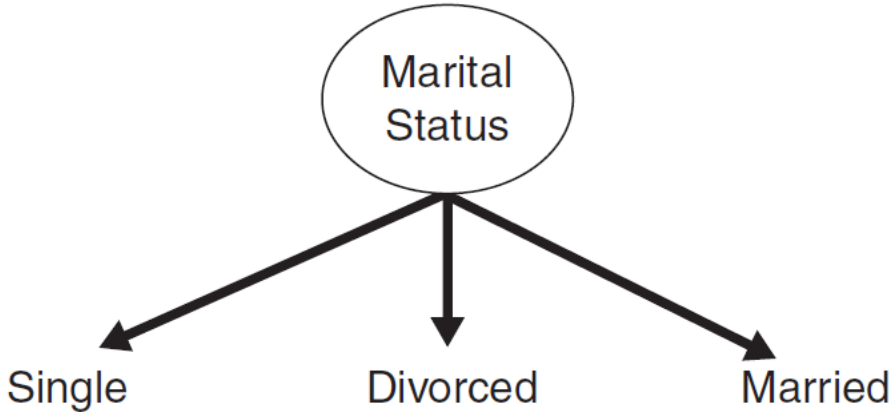


Перечислимые атрибуты

- Бинарное разбиение

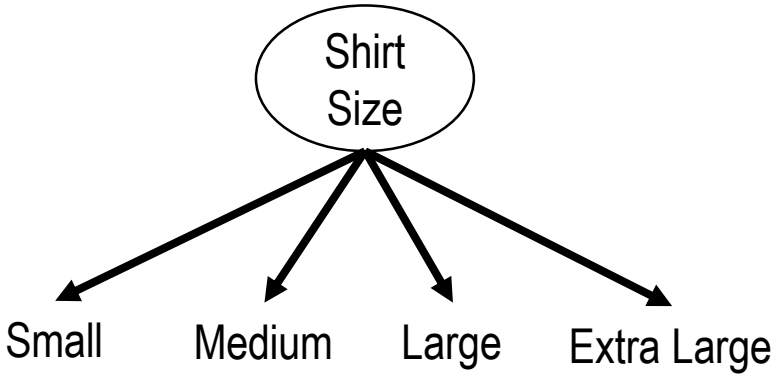


- n-арное разбиение

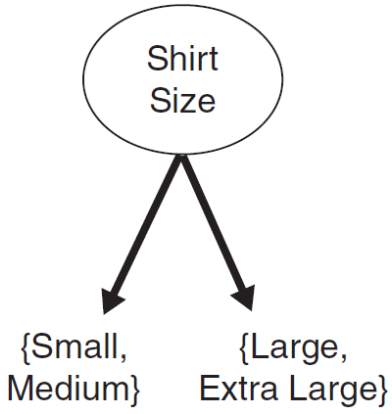


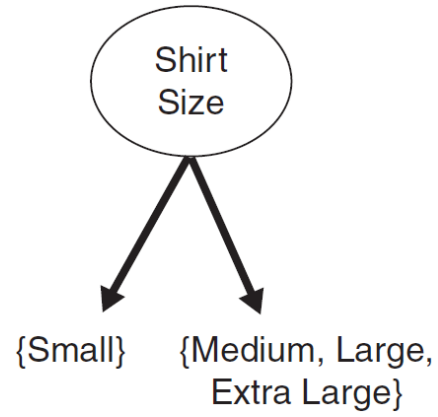
Порядковые атрибуты

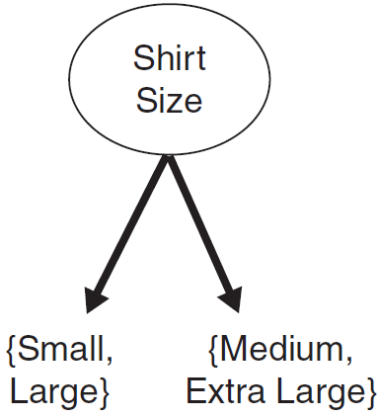
- n-арное разбиение



- Бинарное разбиение (с/без сохранением порядка)





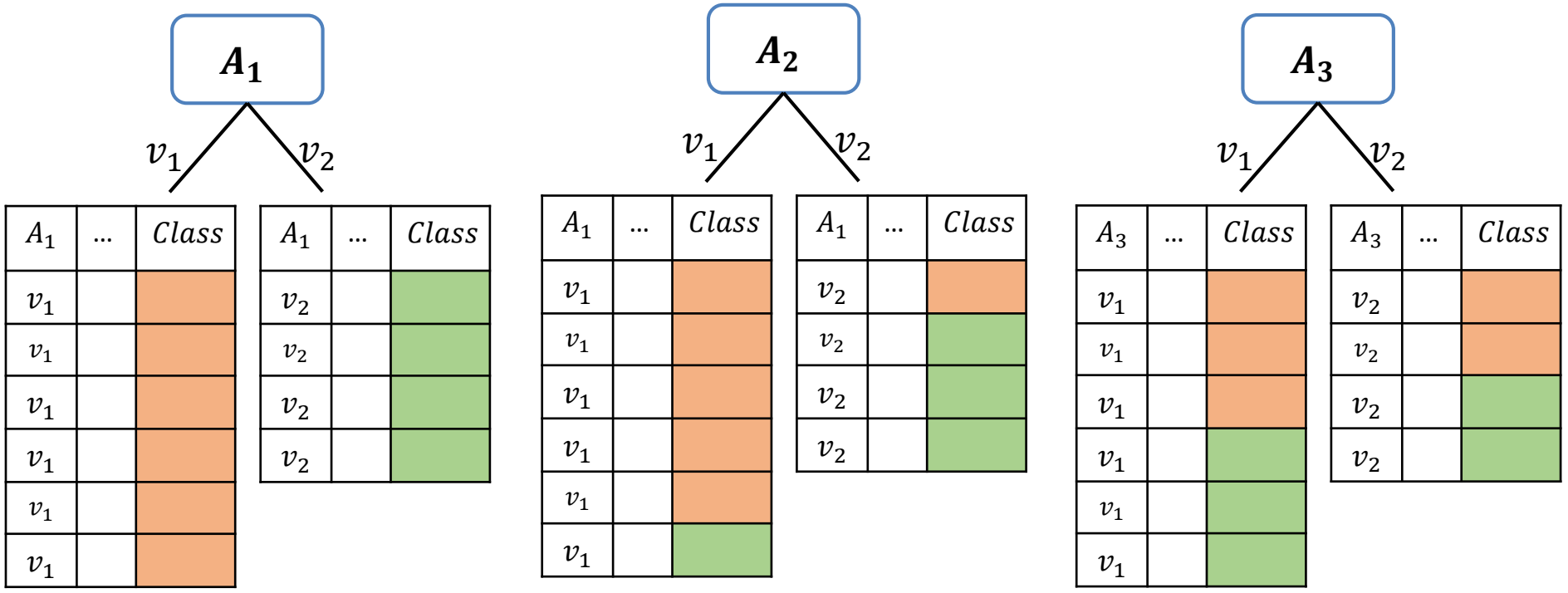


Непрерывные атрибуты

- Бинарное разбиение
  - $(Attr < value)\ or\ (Attr \geq value)$
  - рассмотреть все возможные $value$ и выбрать лучшее (большая трудоемкость)
- Разбиение на основе дискретизации
  - Гистограммы, кластеризация и др.
    - Пример: гистограммы равной ширины

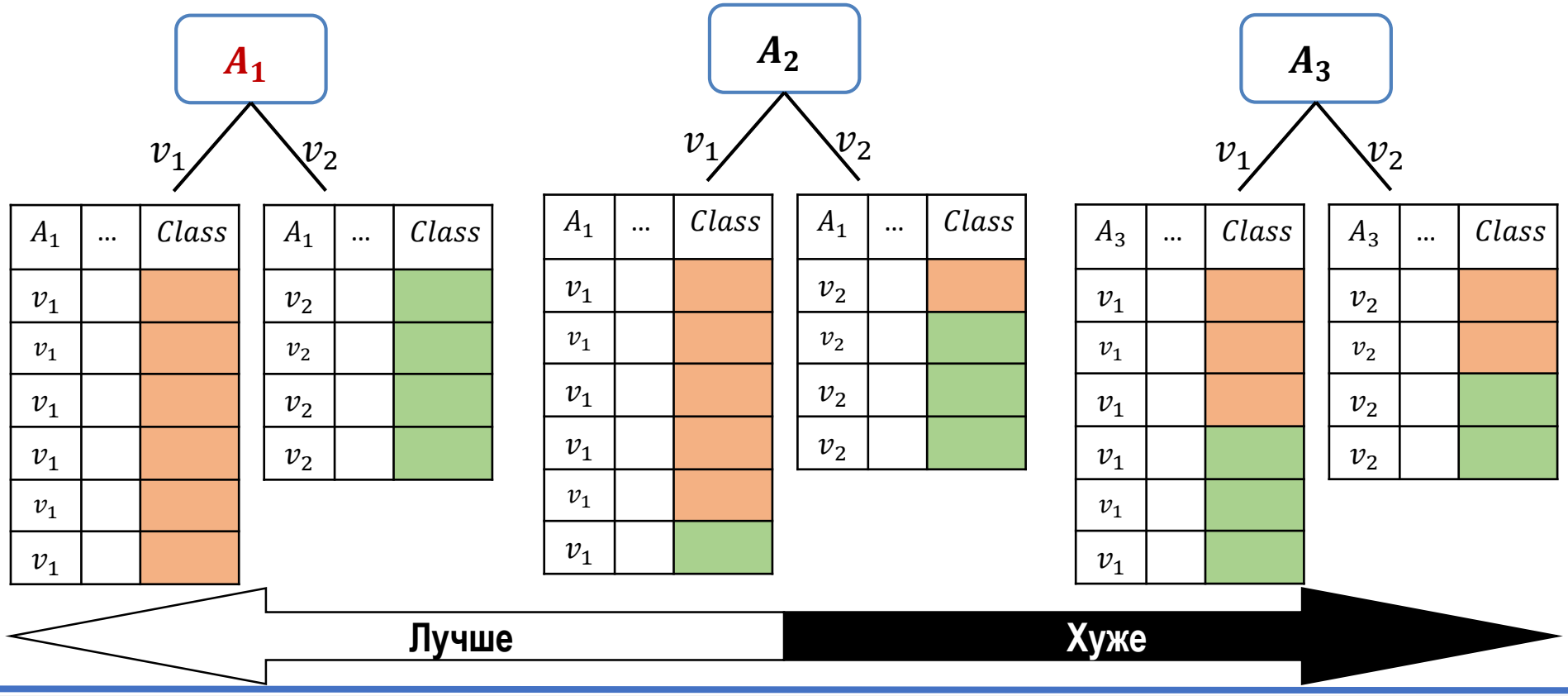
      $Width = \frac{\max\limits_{1 \leq i \leq n} value_i - \min\limits_{1 \leq i \leq n} value_i}{N}$, $N = 1 + \lfloor \log_2 n \rfloor$
  - Статическая (однократно перед построением дерева)
  - Динамическая (для каждого узла при построении)

Какой атрибут выбрать для разбиения?

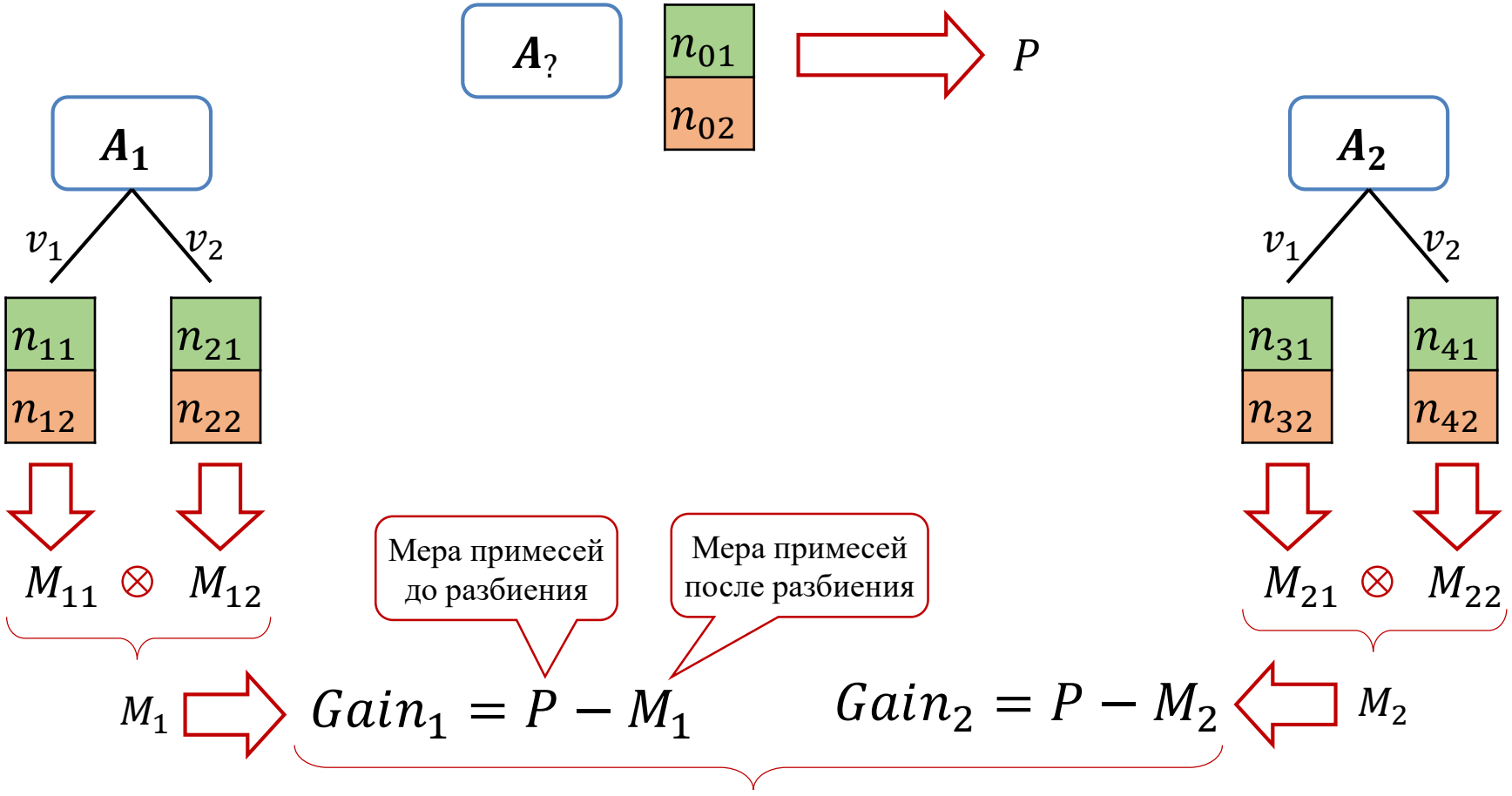


Выбор атрибута разбиения

- *Жадный подход*: выбрать атрибут, разбивающий выборку на подмножества с минимальной долей «примесей» (объектов иного класса)



Критерий выбора атрибута – прирост информации



Меры оценки доли примесей в узле дерева решений

- Индекс Джини

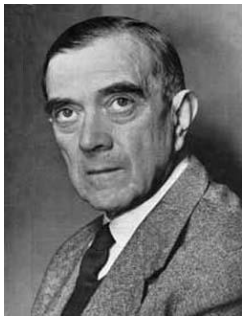
$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

- Энтропия

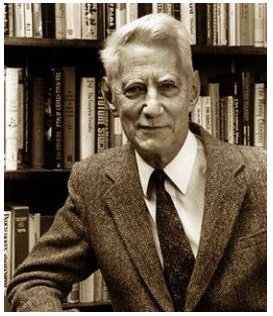
$$Entropy = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

- Ошибка классификации

$$Error = 1 - \max_{1 \le i \le n} p_i$$



Коррадо Джини
1884-1965



Клод Шеннон
1916-2001

n – количество классов

p_i – вероятность объектов i-го класса в выборке узла

Алгоритм CART (Classification and Regression Tree)

- $Gini = 1 - \sum_{i=1}^{n} p_i^2$
- $\max Gini = 1 - \frac{1}{n}$, когда объекты равномерно распределены по классам (наименее желательная ситуация)
- $\min Gini = 0$, когда объекты принадлежат одному классу (наиболее желательная ситуация)

Лео Брейман
1928-2005

Вычисление индекса Джини узла дерева решений

- $Gini = 1 - \sum_{i=1}^{n} p_i^2$

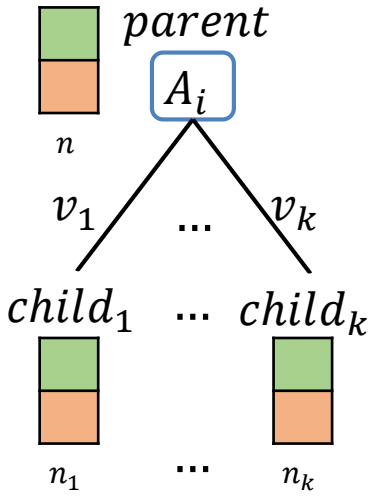
- 0
  6
  $Gini = 1 - \left(\frac{0}{6}\right)^2 - \left(\frac{6}{6}\right)^2 = 0$
- 1
  5
  $Gini = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.278$
- 2
  4
  $Gini = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = 0.444$

Вычисление прироста информации после разбиения узла по атрибуту

- $Gain(A) = Gini(parent) - Info(A)$

$$Gini(parent) = 1 - \sum_{i=1}^{n} p_i^2$$

$$Info(A) = \sum_{i=1}^{k} \frac{n_i}{n} Gini(child_i)$$



A – атрибут разбиения с k различными значениями

n – количество объектов в узле $parent$

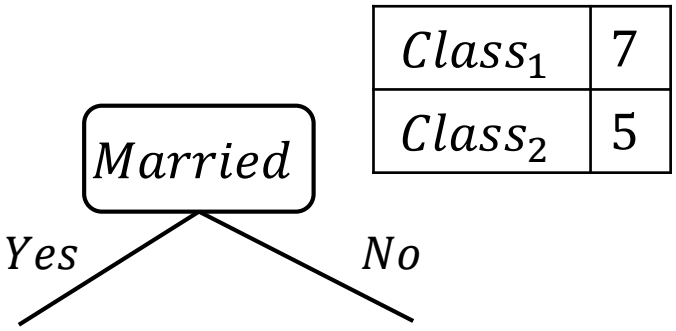
n_i – количество объектов в узле $child_i$

- Для разбиения выбирается атрибут

$$A_{split} = \arg\max_i Gain(A_i) = \arg\min_i Info(A_i)$$

Индекс Джини: бинарные атрибуты

- Выбор разбиения, которое имеет набольшие размер и чистоту



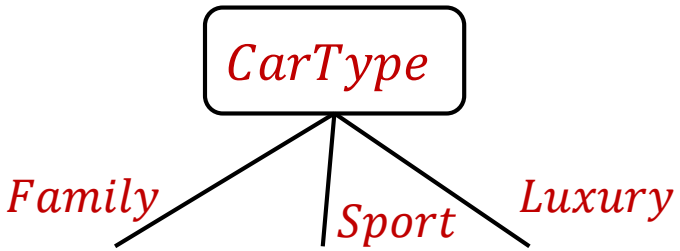
$Class_1$	7
$Class_2$	5

$Class_1$	5
$Class_2$	1

$Class_1$	2
$Class_2$	4

- $\boldsymbol{Gini(Married)} = 1 - \left(\frac{7}{12}\right)^2 - \left(\frac{5}{12}\right)^2 = \mathbf{0.486}$
- $\boldsymbol{Gini(Yes)} = 1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2 = \mathbf{0.278}$
- $\boldsymbol{Gini(No)} = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = \mathbf{0.444}$
- $\boldsymbol{Info(Married)} = \frac{6}{12} \cdot 0.278 + \frac{6}{12} \cdot 0.444 = \mathbf{0.361}$
- $\boldsymbol{Gain_{split}}(Married) = 0.486 - 0.361 = \mathbf{0.125}$

Индекс Джини: категориальные атрибуты



	Family	Sport	Luxury
$Class_1$	1	8	1
$Class_2$	3	0	7

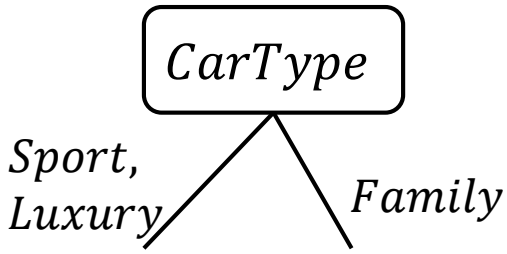
$$Gini(CarType) = 1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2 = 0.5$$

$$Gini(Family) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini(Sport) = 1 - \left(\frac{8}{8}\right)^2 - \left(\frac{0}{8}\right)^2 = 0$$

$$Gini(Luxury) = 1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2 = 0.219$$

$$Info = \frac{4}{20} \cdot 0.375 + \frac{8}{20} \cdot 0 + \frac{8}{20} \cdot 0.219 = 0.163$$

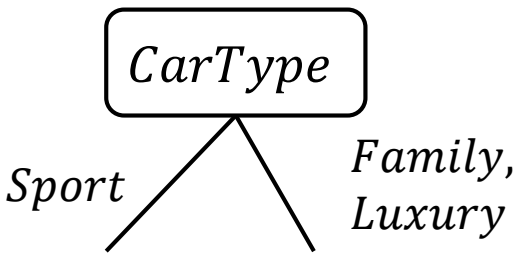


	Sport, Luxury	Family
$Class_1$	9	1
$Class_2$	7	3

$$Gini(Sport, Luxury) = 1 - \left(\frac{9}{16}\right)^2 - \left(\frac{7}{16}\right)^2 = 0.492$$

$$Gini(Family) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Info = \frac{16}{20} \cdot 0.492 + \frac{4}{20} \cdot 0.375 = 0.469$$



	Sport	Family, Luxury
$Class_1$	8	2
$Class_2$	0	10

$$Gini(Sport) = 1 - \left(\frac{8}{8}\right)^2 - \left(\frac{0}{8}\right)^2 = 0$$

$$Gini(Family, Luxury) = 1 - \left(\frac{2}{12}\right)^2 - \left(\frac{10}{12}\right)^2 = 0.375$$

$$Info = \frac{8}{20} \cdot 0 + \frac{12}{20} \cdot 0.375 = 0.167$$

Индекс Джини: вещественные атрибуты

- Сортировать значения вещественного атрибута
- Добавить точки разбиения (среднее соседних значений)
- Просматривать значения в линейном порядке, вычисляя $Gini$
- В качестве точки разбиения взять значение с min $Gini$

ID	...	*Income*	*Class*
1		125	*No*
2		100	*No*
3		70	*No*
4		120	*No*
5		95	*Yes*
6		60	*No*
7		220	*No*
8		85	*Yes*
9		75	*No*
10		90	*Yes*

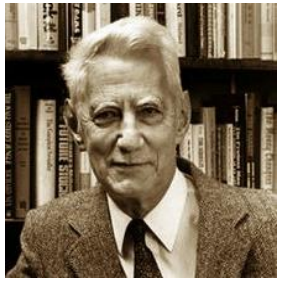
Class	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*	*No*	*No*
Income	60	70	75	85	90	95	100	120	125	220

split	**55**		**65**		**72**		**80**		**87**		**92**		**97**		**110**		**122**		**172**		**230**	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		**0.300**		0.343		0.375		0.400		0.420	

Энтропия: как подсчитать количество информации?

$$Entropy = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

n – количество объектов, p_i – частота i-го объекта



Клод Шеннон
1916-2001

to_be,_or_not_to_be,_that_is_the_question

i	Sym	num	p	-p*log2(p)
1	space	9	0.22	0.48
2	comma	2	0.05	0.21
3	a	1	0.02	0.13
4	b	2	0.05	0.21
5	e	4	0.10	0.33
6	h	2	0.05	0.21
7	i	2	0.05	0.21
8	n	2	0.05	0.21
9	o	5	0.12	0.37
10	q	1	0.02	0.13
11	r	1	0.02	0.13
12	s	2	0.05	0.21
13	t	7	0.17	0.44
14	u	1	0.02	0.13
		41		**3.41**

aaaaaaaaaaaaaaa…a

i	Sym	num	p	-p*log2(p)
1	a	41	1.00	0.00
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
		41		**0.00**

Алгоритм ID3 (Iterative Dichotomiser 3)

- $Entropy = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$
- $\max Entropy = \log_2 n$, когда объекты равномерно распределены по классам (наименее желательная ситуация)
- $\min Entropy = 0$, когда объекты принадлежат одному классу (наиболее желательная ситуация)

John Ross Quinlan

Вычисление энтропии узла дерева решений

- $Entropy = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$

$0 \cdot \log_2 0$ считается 0

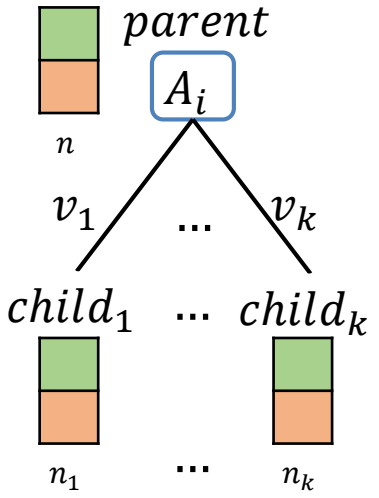
- 0 / 6 $Entropy = -\frac{0}{6} \cdot \log_2 0 - \frac{6}{6} \cdot \log_2 \frac{1}{1} = 0$
- 1 / 5 $Entropy = -\frac{1}{6} \cdot \log_2 \frac{1}{6} - \frac{5}{6} \cdot \log_2 \frac{5}{6} = 0.65$
- 2 / 4 $Entropy = -\frac{2}{6} \cdot \log_2 \frac{2}{6} - \frac{4}{6} \cdot \log_2 \frac{4}{6} = 0.92$

Вычисление прироста информации после разбиения узла по атрибуту

- $Gain(A) = Entropy(parent) - Info(A)$

$$Entropy(parent) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

$$Info(A) = \sum_{i=1}^{k} \frac{n_i}{n} Entropy(child_i)$$



A – атрибут разбиения с k различными значениями

n – количество объектов в узле $parent$

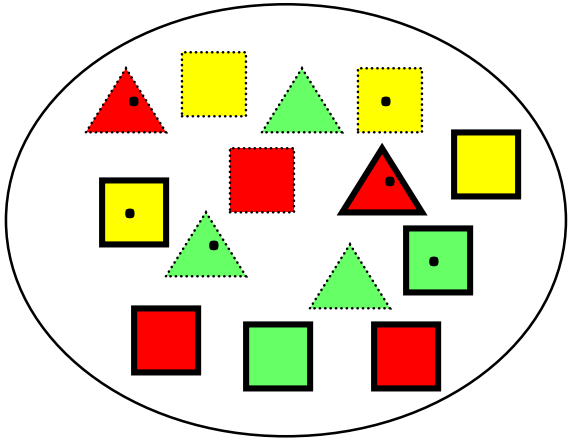
n_i – количество объектов в узле $child_i$

- Для разбиения выбирается атрибут

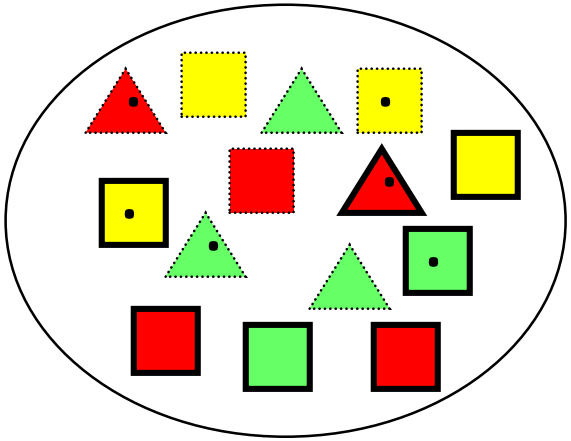
$$A_{split} = \arg\max_i Gain(A_i) = \arg\min_i Info(A_i)$$

Пример построения дерева решений по ID3

#	Attribute			Shape
	Color	Outline	Dot	
1	green	dashed	no	triange
2	green	dashed	yes	triange
3	yellow	dashed	no	square
4	red	dashed	no	square
5	red	solid	no	square
6	red	solid	yes	triange
7	green	solid	no	square
8	green	dashed	no	triange
9	yellow	solid	yes	square
10	red	solid	no	square
11	green	solid	yes	square
12	yellow	dashed	yes	square
13	yellow	solid	no	square
14	red	dashed	yes	triange



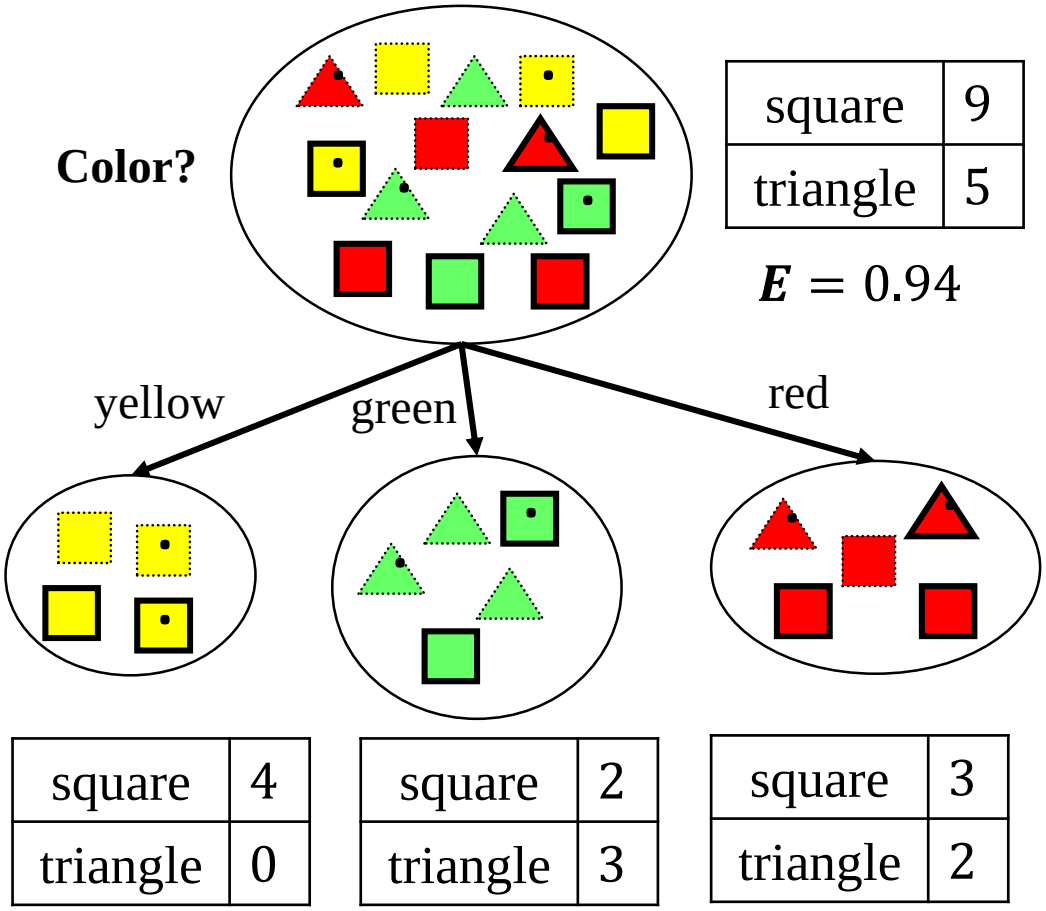
Пример: вычисление энтропии узла-родителя



square	9
triangle	5

- $\boldsymbol{Entropy} = -\frac{9}{14} \cdot \log_2 \frac{9}{14} - \frac{5}{14} \cdot \log_2 \frac{5}{14} = \mathbf{0.94}$

Пример: вычисление энтропии узлов-потомков



square	9
triangle	5

$\boldsymbol{E} = 0.94$

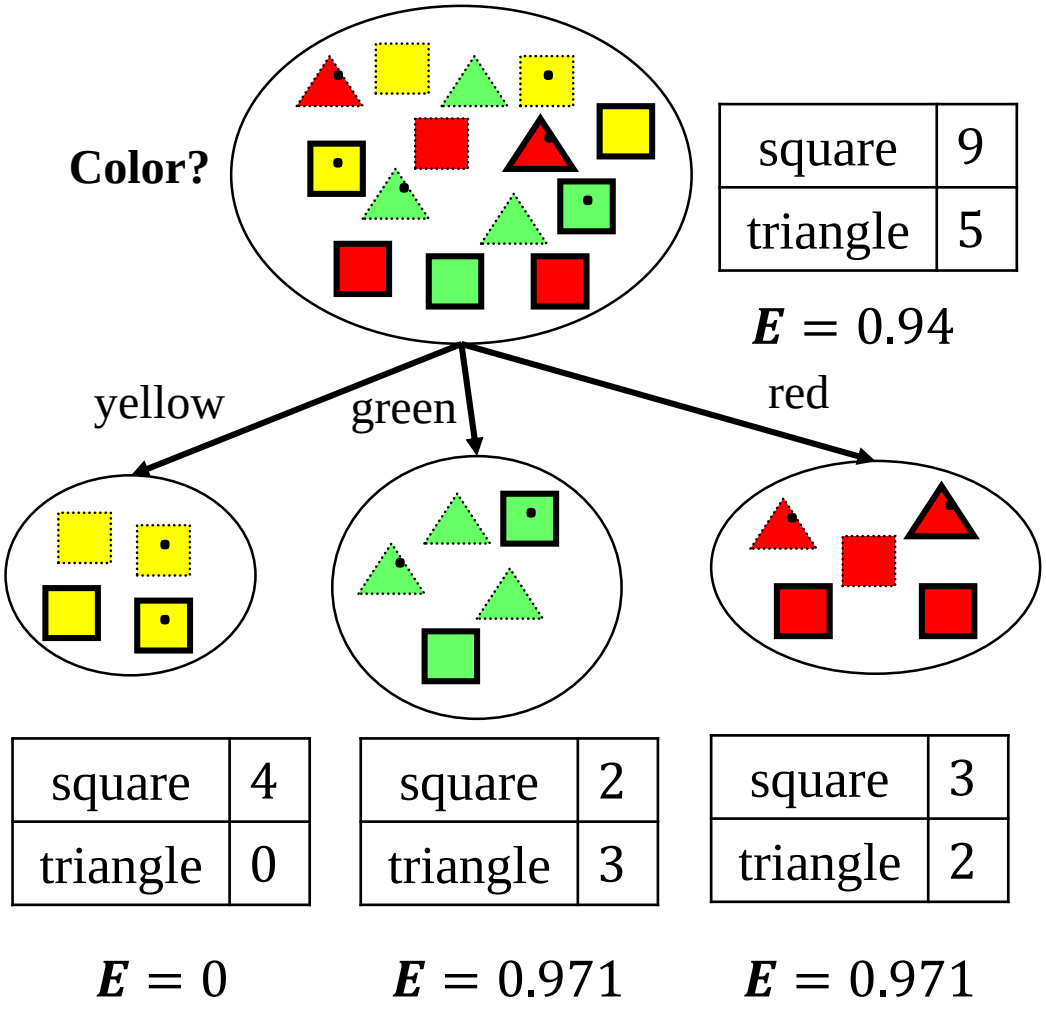
square	4
triangle	0

square	2
triangle	3

square	3
triangle	2

- $\boldsymbol{Entropy(yellow)} = 0$
- $\boldsymbol{Entropy(green)} = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = \mathbf{0.971}$
- $\boldsymbol{Entropy(red)} = -\frac{3}{5} \cdot \log_2 \frac{3}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5} = \mathbf{0.971}$

Пример: вычисление прироста информации



square	9
triangle	5

$E = 0.94$

square	4
triangle	0

$E = 0$

square	2
triangle	3

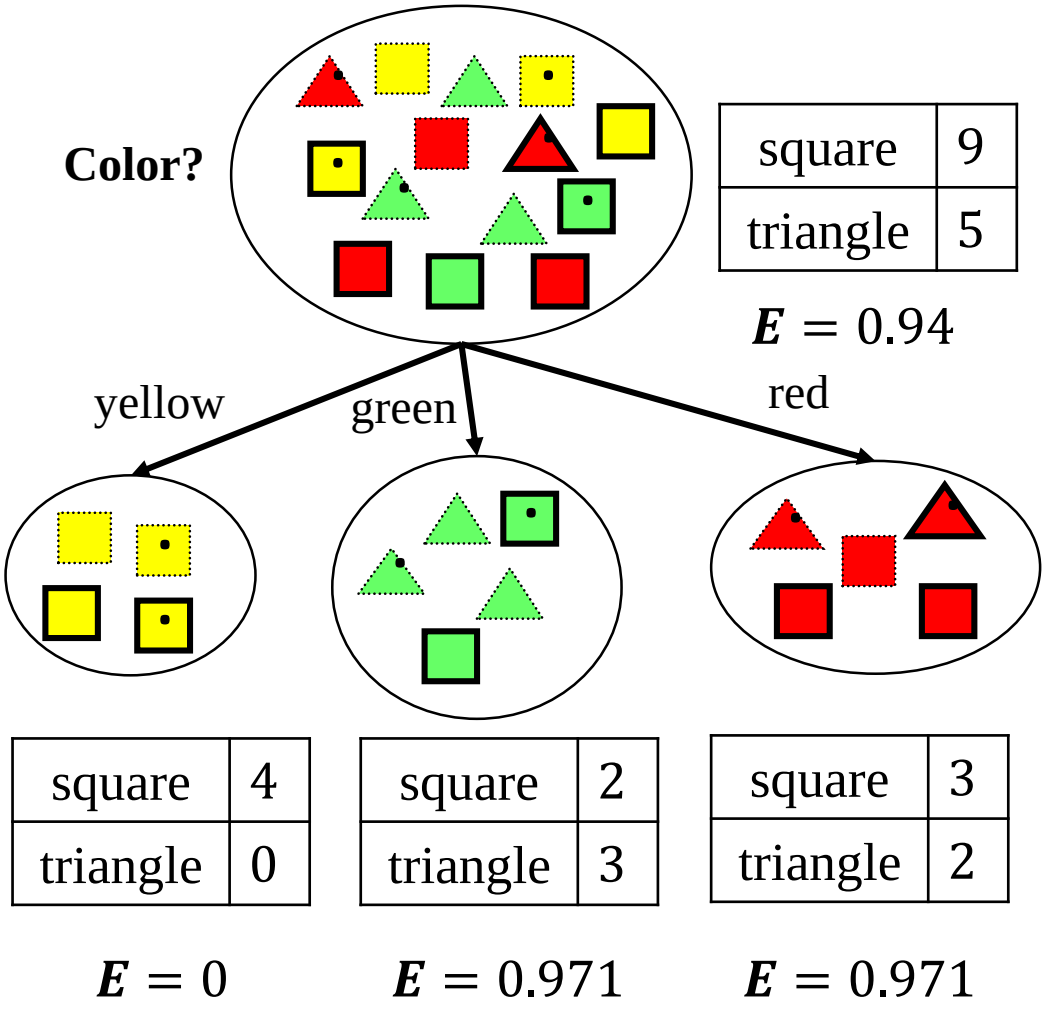
$E = 0.971$

square	3
triangle	2

$E = 0.971$

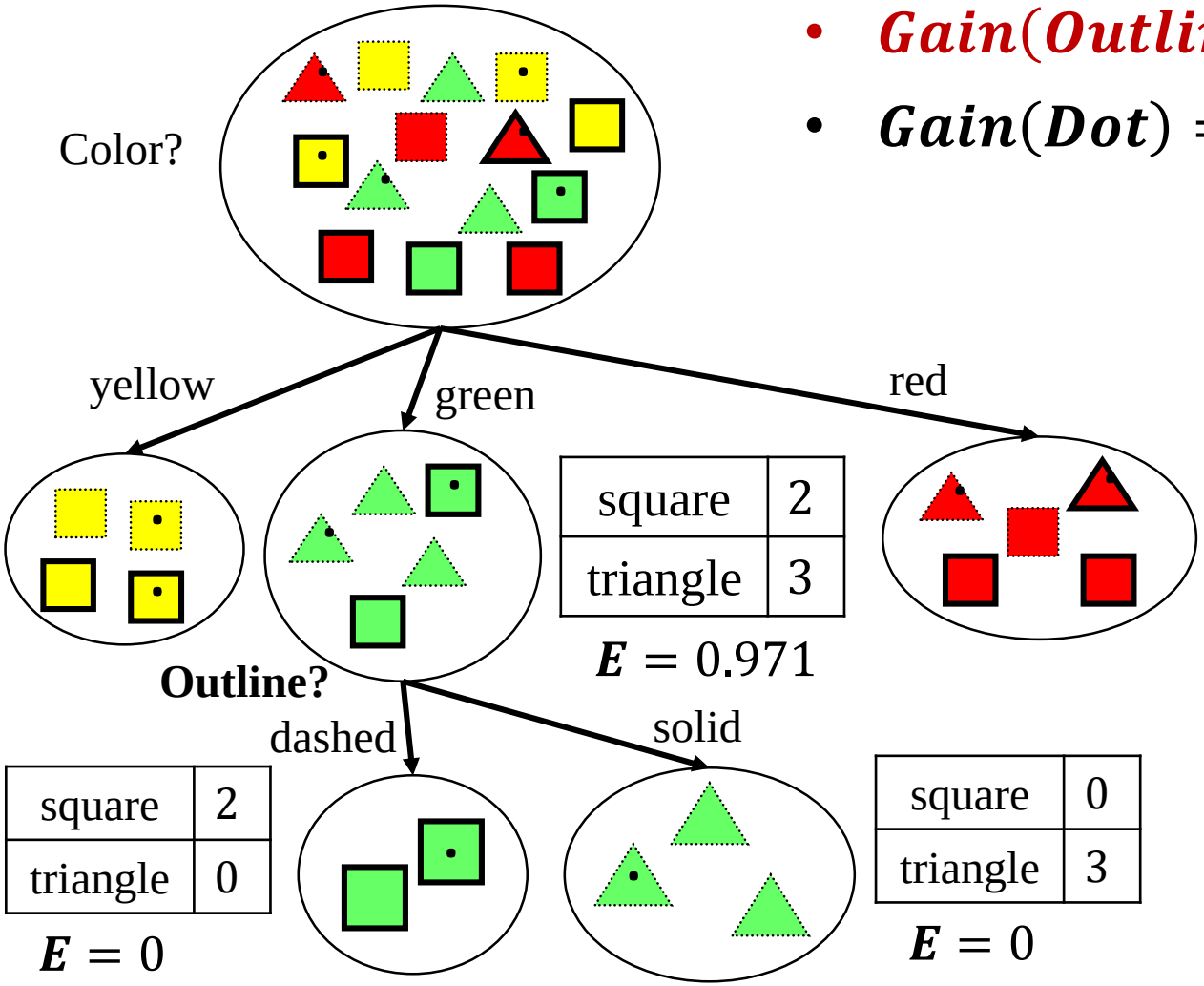
- $Info(Color) = \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0.971 + \frac{5}{14} \cdot 0.971 = \mathbf{0.694}$
- $Gain(Color) = 0.94 - 0.694 = \mathbf{0.246}$

Пример: выбор атрибута разбиения



- $Gain(Color) = 0.246$
- $Gain(Outline) = 0.151$
- $Gain(Dot) = 0.048$

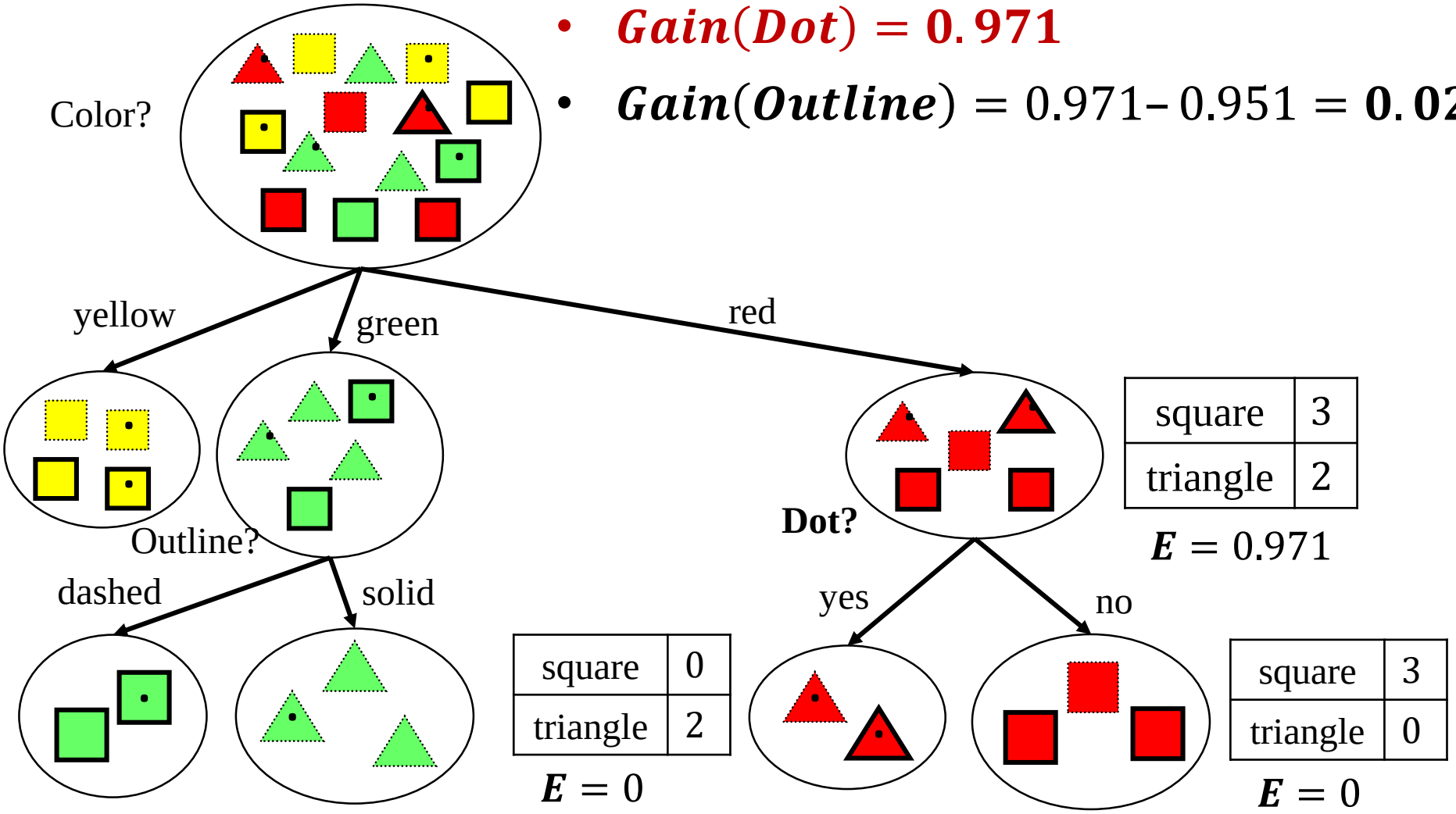
Пример: выбор атрибута разбиения



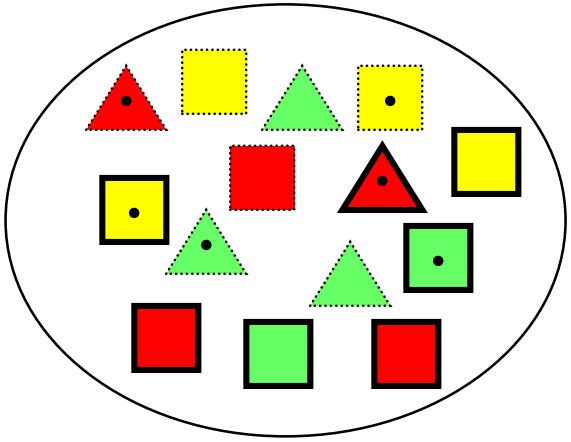
- $\boldsymbol{Gain(Outline) = 0.971}$
- $\boldsymbol{Gain(Dot)} = 0.971 - 0.951 = \mathbf{0.02}$

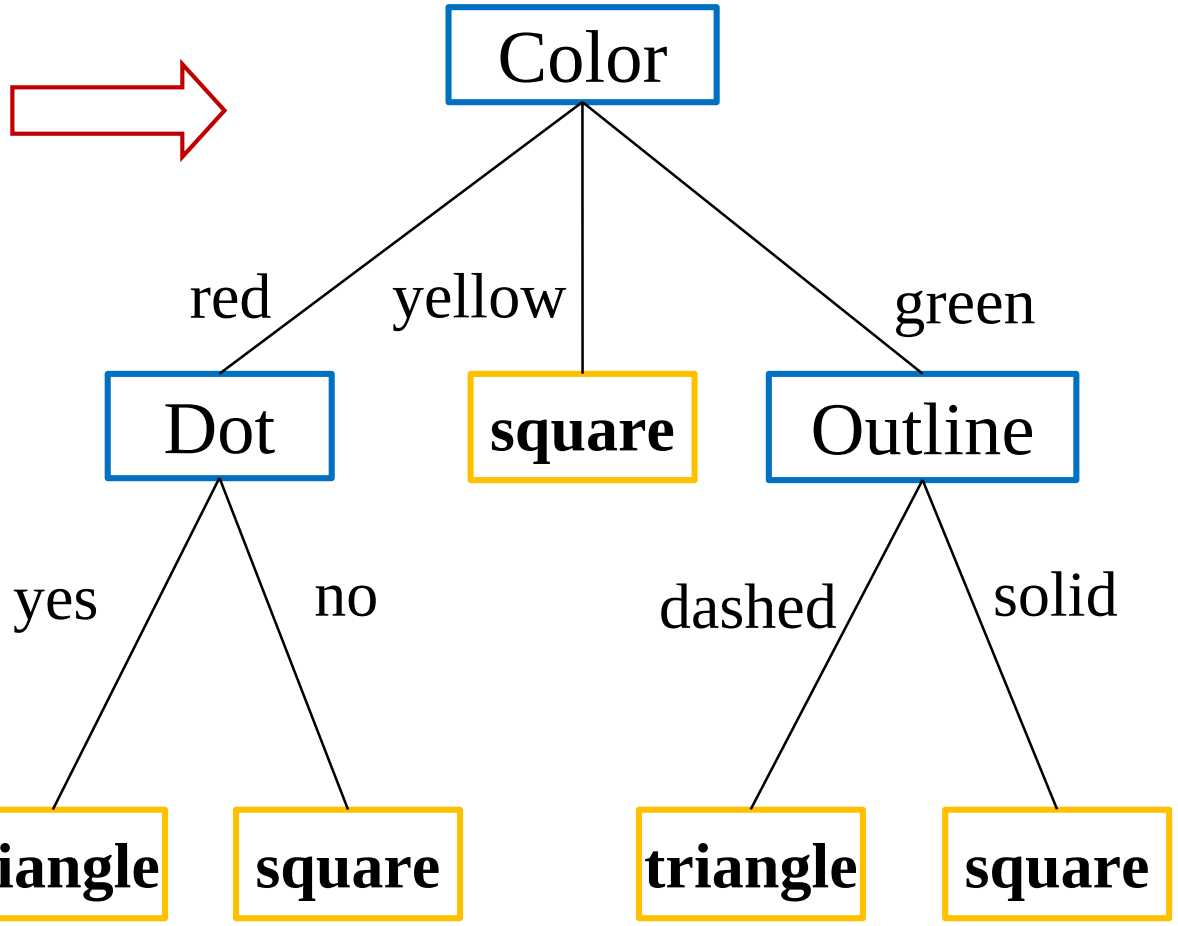
Пример: выбор атрибута разбиения

- $Gain(Dot) = 0.971$
- $Gain(Outline) = 0.971 - 0.951 = 0.02$

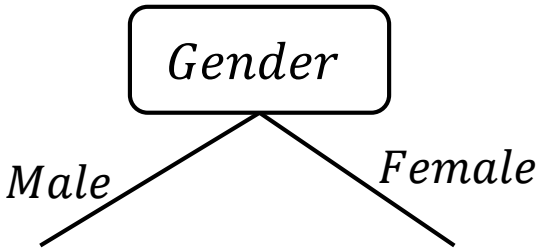


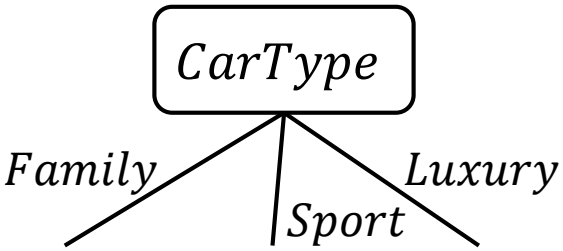
Пример: итоговое дерево решений

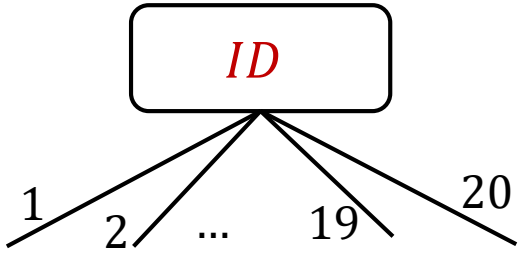




Проблема большого количества значений атрибута разбиения







	Male	Female
$Class_1$	6	4
$Class_2$	4	6

	Family	Sport	Luxury
$Class_1$	1	8	1
$Class_2$	3	0	7

	1	2	...	19	20
$Class_1$	1	0		0	1
$Class_2$	0	1		1	0

- $ID = \arg\min\{Info(Gender), Info(CarType), Info(ID)\}$, т.к. $Info(ID) = 0$
- Разбиение по атрибуту ID (уникальный идентификатор) бесполезно для классификации

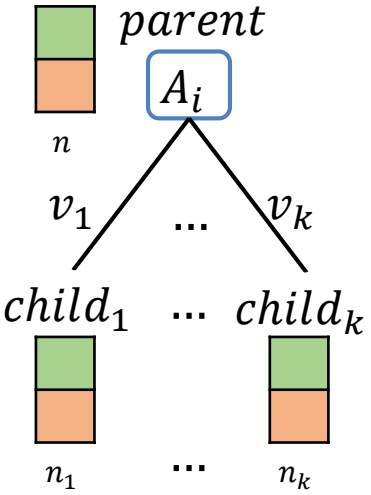
Доля прироста информации (алгоритм C4.5)

- $Gain_{ratio}(A) = \frac{Gain(A)}{Info_{split}(A)} = \frac{Entropy(parent) - Info(A)}{Info_{split}(A)}$

$$Entropy(parent) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

$$Info(A) = \sum_{i=1}^{k} \frac{n_i}{n} Entropy(child_i)$$

$$Info_{split}(A) = -\sum_{i=1}^{k} \frac{n_i}{n} \cdot \log_2 \frac{n_i}{n}$$

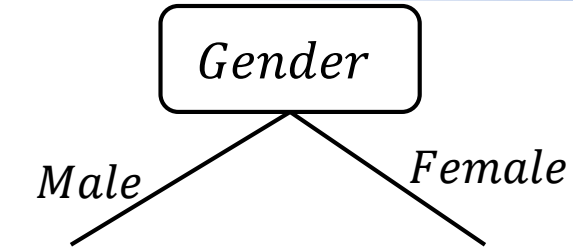


A – атрибут разбиения с k различными значениями
n – количество объектов в узле $parent$
n_i – количество объектов в узле $child_i$

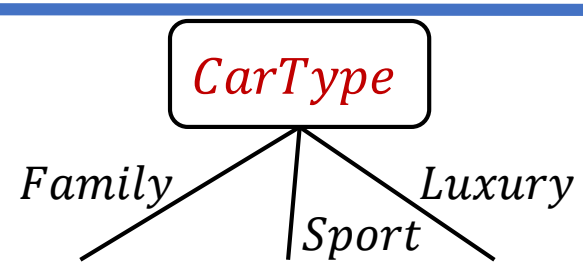
- Для разбиения выбирается атрибут

$$A_{split} = \arg\max_i Gain_{ratio}(A_i)$$

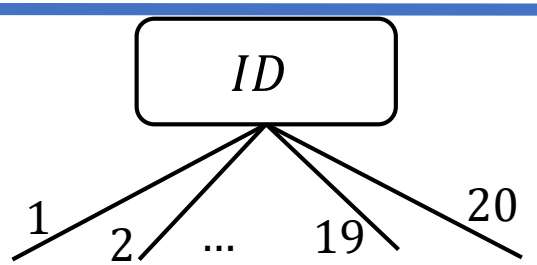
Доля прироста информации: выбор атрибута



	Male	Female
$Class_1$	6	4
$Class_2$	4	6



	Family	Sport	Luxury
$Class_1$	1	8	1
$Class_2$	3	0	7



	1	2	...	19	20
$Class_1$	1	0		0	1
$Class_2$	0	1		1	0

$$\boldsymbol{Entropy(Attr)} = -\frac{10}{20}\cdot\log_2\frac{10}{20} - \frac{10}{20}\cdot\log_2\frac{10}{20} = \mathbf{1}$$

$$\boldsymbol{Info(Gender)} = \frac{10}{20}\cdot\left(-\frac{6}{10}\cdot\log_2\frac{6}{10} - \frac{4}{10}\cdot\log_2\frac{4}{10}\right) + \frac{10}{20}\cdot\left(-\frac{4}{10}\cdot\log_2\frac{4}{10} - \frac{6}{10}\cdot\log_2\frac{6}{10}\right) = \mathbf{0.971}$$

$$\boldsymbol{Gain_{ratio}(Gender)} = \frac{1-0.971}{-\frac{10}{20}\cdot\log_2\frac{10}{20} - \frac{10}{20}\cdot\log_2\frac{10}{20}} = \frac{0.029}{1} = \mathbf{0.029}$$

$$\boldsymbol{Info(CarType)} = \frac{4}{20}\cdot\left(-\frac{1}{4}\cdot\log_2\frac{1}{4} - \frac{3}{4}\cdot\log_2\frac{3}{4}\right) + \frac{8}{20}\cdot\left(-\frac{1}{8}\cdot\log_2\frac{1}{8} - \frac{7}{8}\cdot\log_2\frac{7}{8}\right) = \frac{0.62}{1.52} = \mathbf{0.38}$$

$$\boldsymbol{Gain_{ratio}(CarType)} = \frac{1-0.38}{-\frac{4}{20}\cdot\log_2\frac{4}{20} - \frac{8}{20}\cdot\log_2\frac{8}{20} - \frac{8}{20}\cdot\log_2\frac{8}{20}} = \frac{0.62}{1.52} = \mathbf{0.41}$$

$$\boldsymbol{Info(ID)} = \frac{1}{20}\cdot\left(-\frac{1}{1}\cdot\log_2\frac{1}{1} - \frac{0}{1}\cdot\log_2\frac{0}{1}\right)\cdot 20 = \mathbf{0}$$

$$\boldsymbol{Gain_{ratio}(ID)} = \frac{1-0}{\left(-\frac{1}{20}\cdot\log_2\frac{1}{20}\right)\cdot 20} = \frac{1}{4.32} = \mathbf{0.23}$$

Сравнение

- Gain
  - тяготеет к атрибутам с большим количество значений
- Gain ratio
  - тяготеет к несбалансированным разбиениям, когда одна из частей существенно меньше других
- Gini
  - тяготеет к атрибутам с большим количество значений
  - имеются трудности, когда количество классов большое
  - тяготеет к равномощным разбиениям с равным количеством примесей

Мера ошибки неверной классификации

- $Error = 1 - \max\limits_{1 \le i \le n} p_i$
- $\max Error = 1 - \frac{1}{n}$, когда объекты равномерно распределены по классам (наименее желательная ситуация)
- $\min Error = 0$, когда объекты принадлежат одному классу (наиболее желательная ситуация)

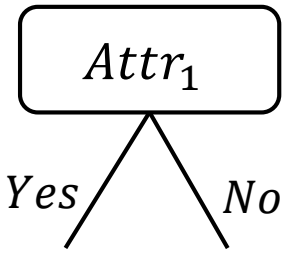
Вычисление ошибки классификации в узле дерева решений

- $Error = 1 - \max\limits_{1 \le i \le n} p_i$

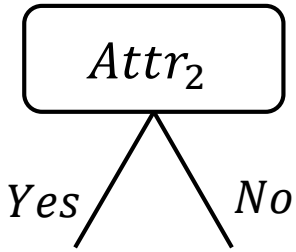
- 0 / 6 $Error = 1 - \max(\frac{0}{6}, \frac{6}{6}) = 0$
- 1 / 5 $Error = 1 - \max(\frac{1}{6}, \frac{5}{6}) = 0.166$
- 2 / 4 $Error = 1 - \max(\frac{2}{6}, \frac{4}{6}) = 0.333$

Ошибка классификации *vs.* Индекс Джини

	$Attr$
$Class_1$	7
$Class_2$	3
$Gini = 0.42$	
$Error = 0.3$	



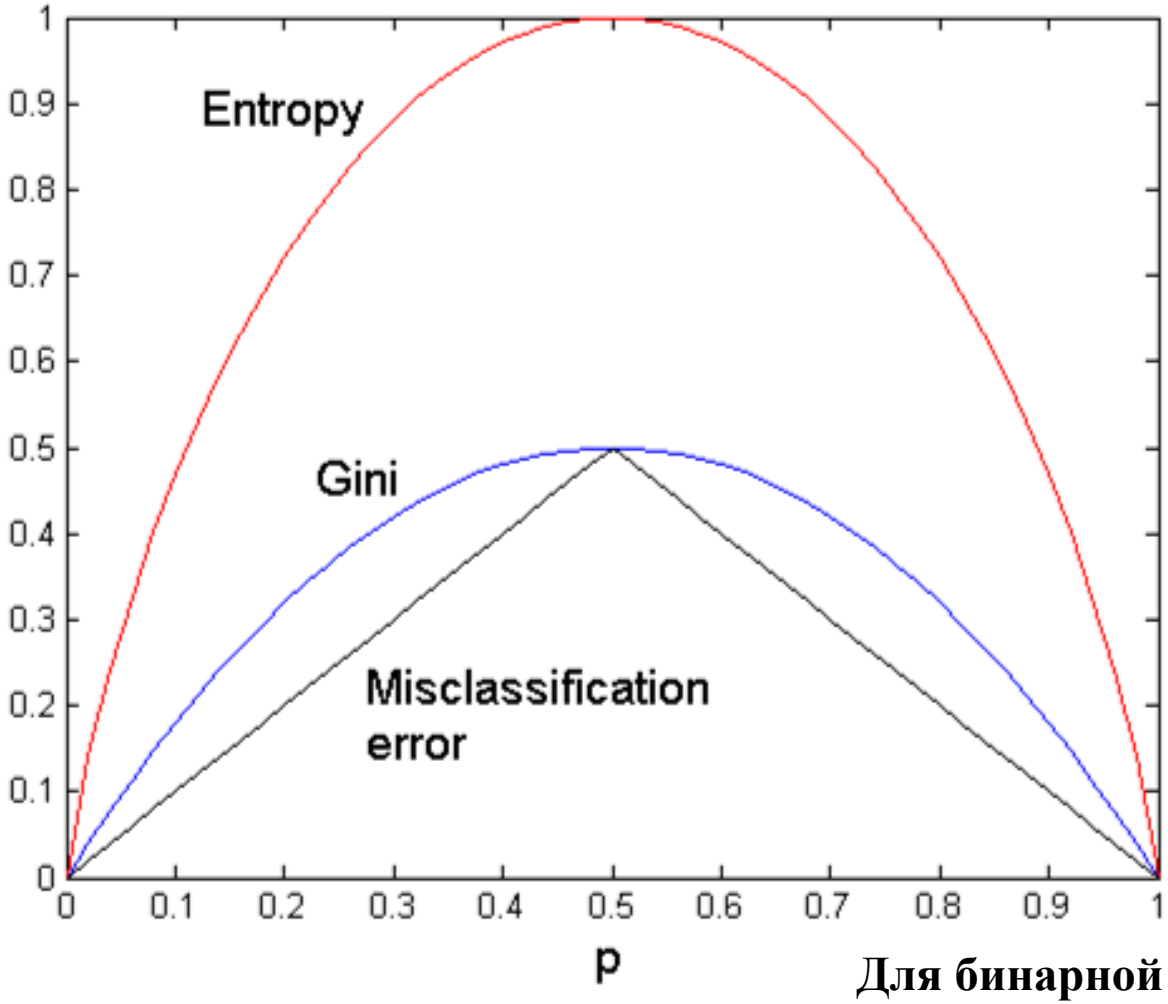
	Yes	No
$Class_1$	4	3
$Class_2$	3	0
$Gini = 0.342$		
$\boldsymbol{Error = 0.3}$		



	Yes	No
$Class_1$	3	4
$Class_2$	1	2
$\boldsymbol{Gini = 0.416}$		
$\boldsymbol{Error = 0.3}$		

- *Gini* улучшается, *Error* не изменяется

Сравнение мер оценки доли примесей



Для бинарной классификации

Алгоритм построения дерева решений

```
TreeGrowth (E, F)
 1: if stopping_cond(E,F) = true then
 2:    leaf = createNode().
 3:    leaf.label = Classify(E).
 4:    return leaf.
 5: else
 6:    root = createNode().
 7:    root.test_cond = find_best_split(E, F).
 8:    let V = {v|v is a possible outcome of root.test_cond }.
 9:    for each v ∈ V do
10:       E_v = {e | root.test_cond(e) = v and e ∈ E}.
11:       child = TreeGrowth(E_v, F).
12:       add child as descendent of root and label the edge (root → child) as v.
13:    end for
14: end if
15: return root.
```

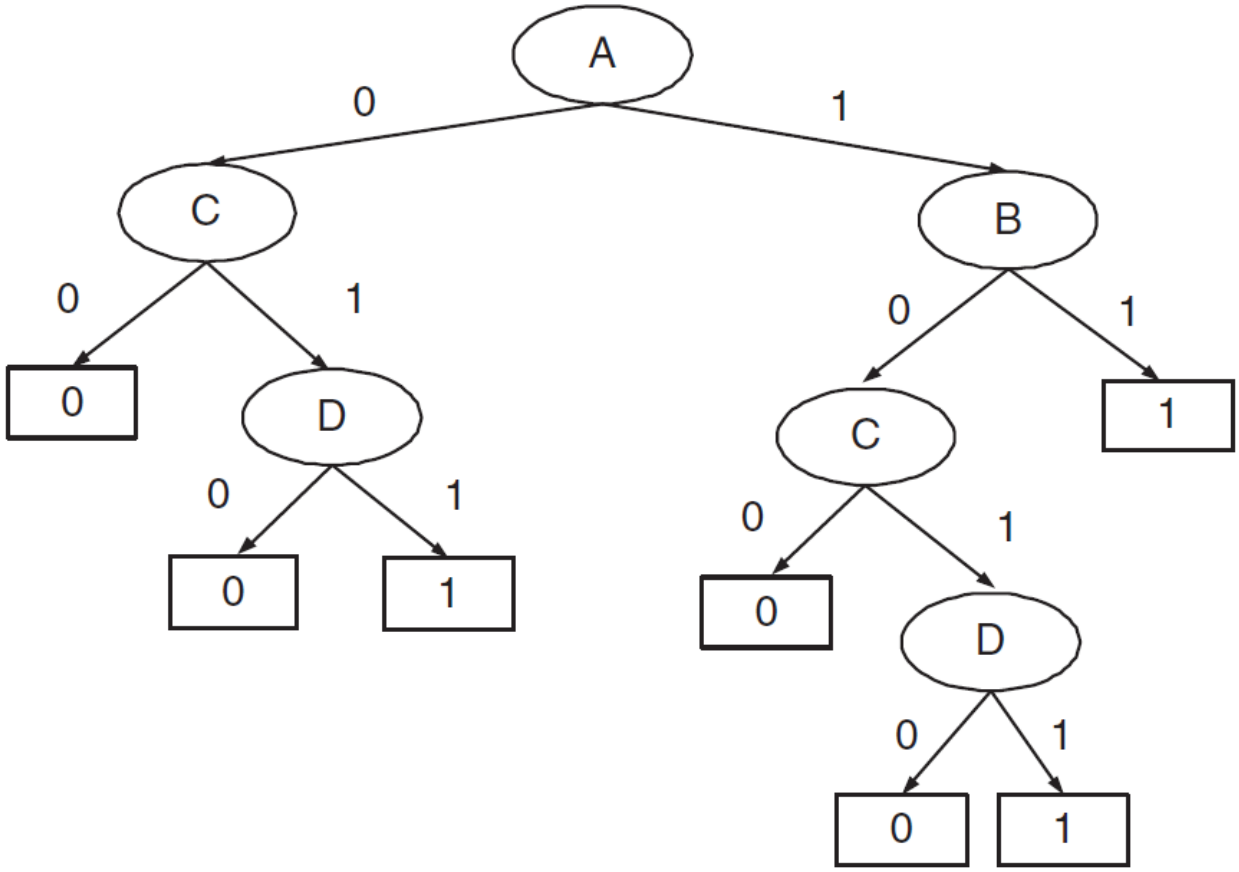
Деревья решений: плюсы и минусы

- Преимущества
  - Невысокая трудоемкость построения
  - Быстрая классификация ранее неизвестных объектов
  - Выразительность и простая интерпретация для небольших деревьев
  - Устойчивость к шуму в данных
  - Простая обработка избыточных и нерелевантных атрибутов (если атрибуты не взаимодействуют)
- Недостатки:
  - Пространство возможных деревьев решений экспоненциально велико
  - Жадные подходы часто не позволяют найти лучшее дерево
  - Не учитываются взаимодействия между атрибутами
  - Каждая граница решения включает только один атрибут

Выразительность

Таблица и дерево решений для булевой функции
$(A \wedge B) \vee (C \wedge \mathrm{D})$

A	B	C	D	class
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1

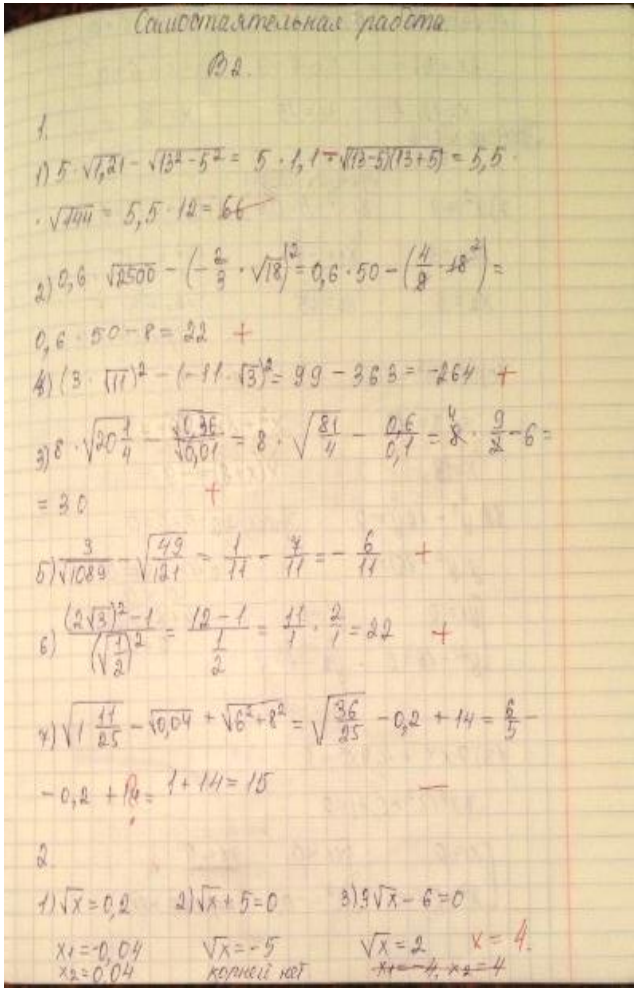


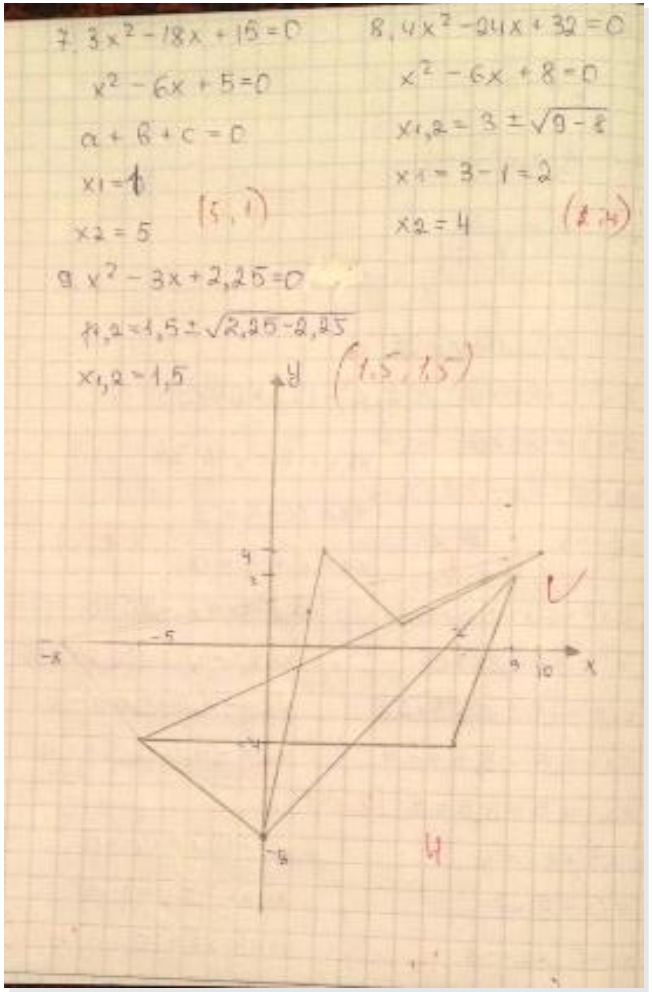
Обработка пропущенных значений

- В обучающей выборке
  - При подсчетах не учитывать объекты, в подвыборках каждого дочернего узла имеющие пропущенные значения в атрибуте разбиения
- В тестовой выборке
  - Предобработка: отбрасывание или/и восстановление (например, с помощью среднего или моды)
  - Без предобработки:
    - переход на основе др. атрибута, разбиение по которому даст разбиения, похожие на разбиения по атрибуту с пропущенным значением
    - ввести новый класс

Избыточные атрибуты

- Учитель
  + и –
  ~~Исправления~~
  ?
  ✔
  5+, 5, 5-, 4+,
  …
- Ученик
  ~~Исправления~~
  Замазывания

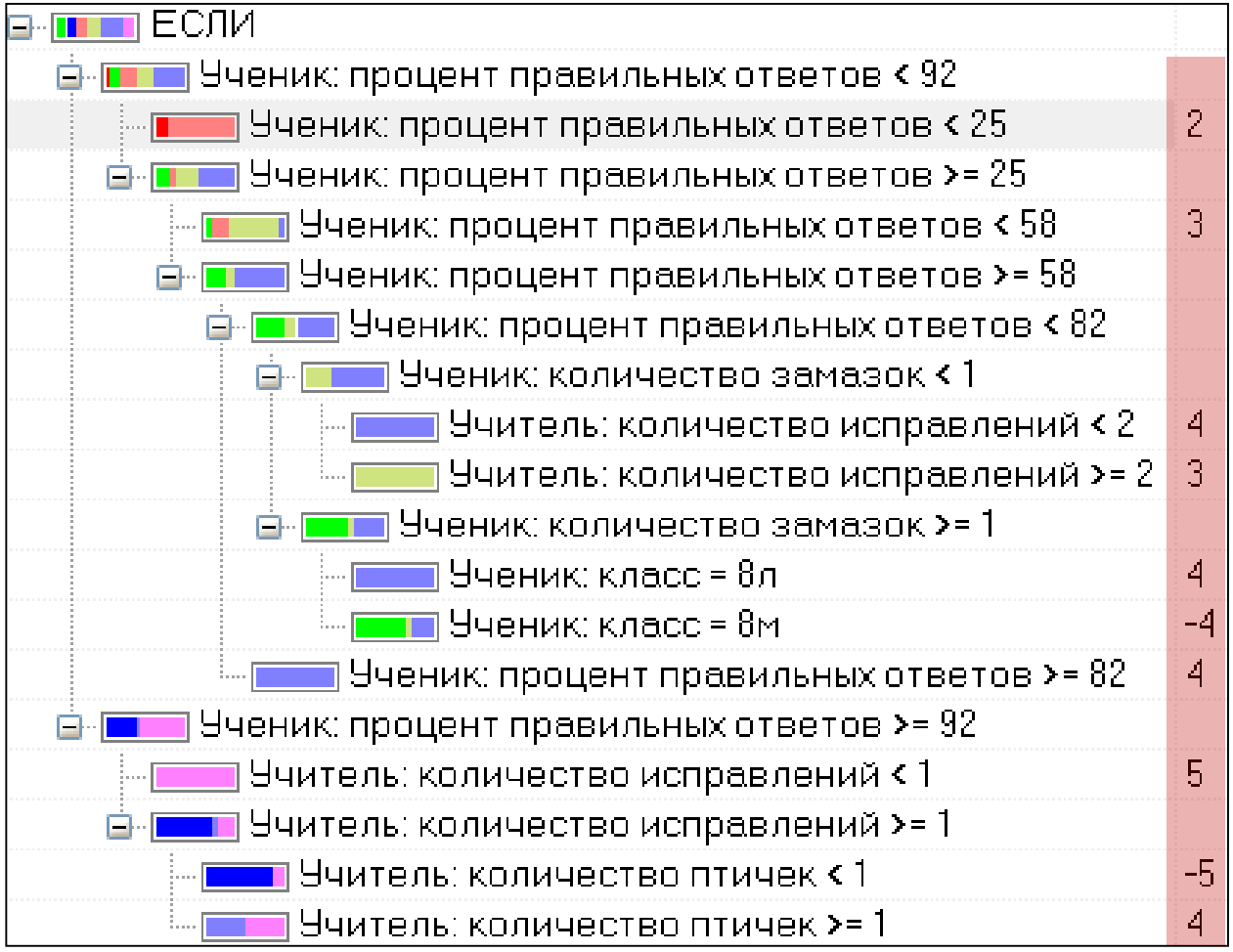




Избыточные атрибуты

№	Ученик		Проверка учителя				Выполнение работы		ОЦЕНКА
	Пол	Класс	% правильно выполненных заданий	К-во ✓	К-во ?	К-во испр-й	К-во испр-й	К-во замазок	
1	Ж	8м	67	0	0	1	3	3	4-
2	Ж	8м	60	0	2	0	5	9	3
3	Ж	8м	100	0	0	2	0	2	5-
4	Ж	8м	100	1	0	2	3	1	4
5	Ж	8м	100	1	0	2	0	5	5
6	М	8л	80	0	1	1	0	2	4
7	Ж	8м	50	1	0	6	0	1	4-
8	Ж	8л	20	0	0	2	2	1	3-
9	М	8м	40	1	2	0	3	4	2
10	М	8м	65	0	3	4	4	8	4-
…	…	…	…	…	…	…	…	…	…
70	М	8л	15	0	0	0	0	11	2

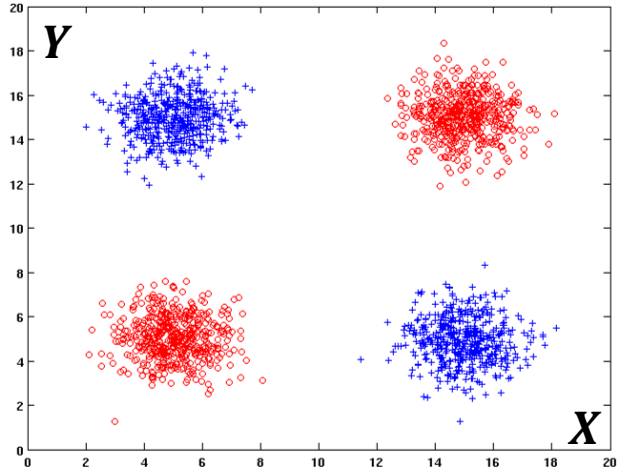
Избыточные атрибуты

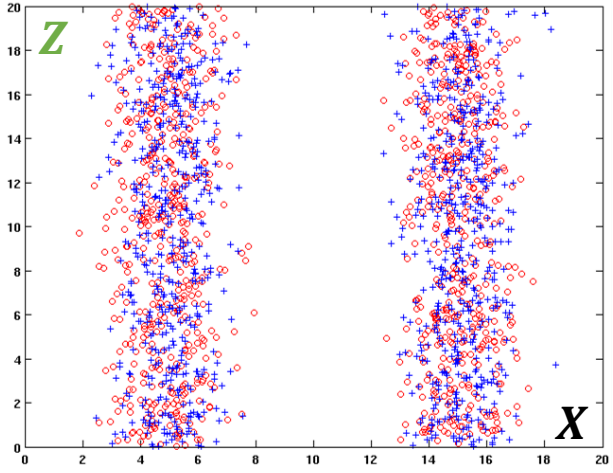


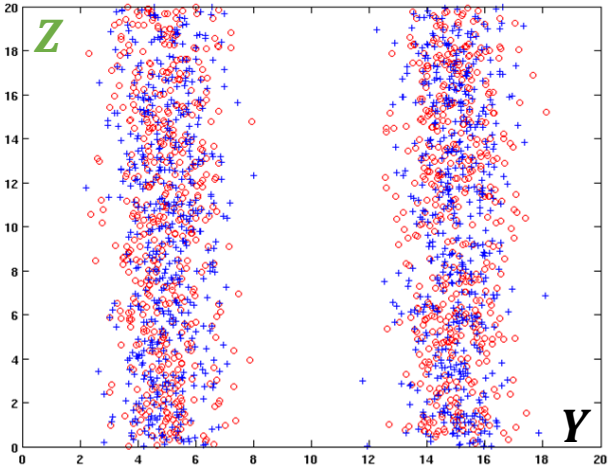
№	Ученик: Пол	Ученик: Класс	...	ОЦЕНКА
1	Ж	8м		4-
2	Ж	8м		3
3	Ж	8м		5-
4	Ж	8м		4
5	Ж	8м		5
6	М	8л		4
7	Ж	8м		4-
8	Ж	8л		3-
9	М	8м		2
10	М	8м		4-
...	...	...		...
70	М	8л		2

Взаимодействующие атрибуты

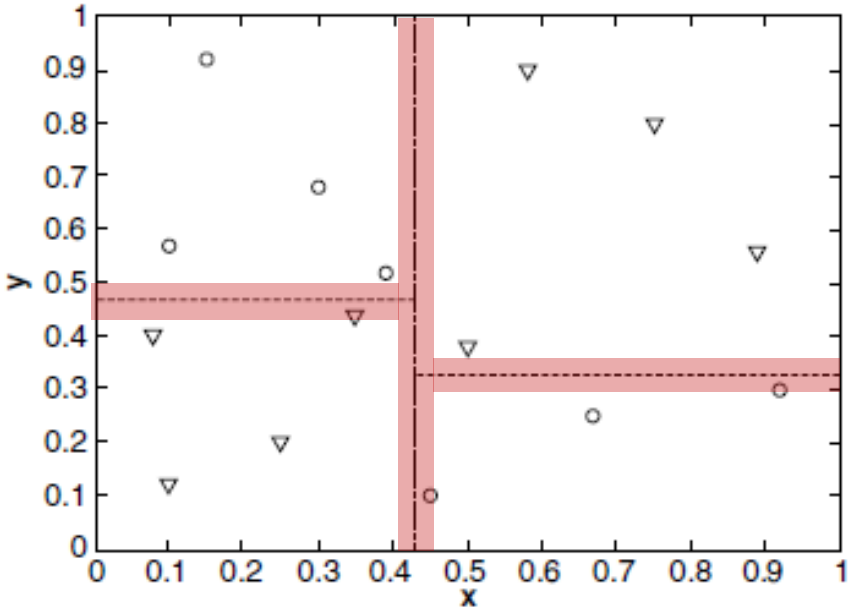
- Атрибуты считаются взаимодействующими, если они могут различать классы при совместном использовании (но не по отдельности)
- Жадный подход при выборе атрибута разбиения приводит к выбору не взаимодействующих, но менее полезных атрибутов, и, соответственно, к более сложным деревьям
- Пример
  - Набор данных: + 1000 шт., ○ 1000 шт.
  - Добавлен атрибут $\boldsymbol{Z}$: шум, в котором классы распределены равномерно
  - При $X \leq 10 \wedge Y \leq 10$ $Entropy(\boldsymbol{X}) = Entropy(\boldsymbol{Y}) = 0.99$, $Entropy(\boldsymbol{Z}) = 0.98$
  - Атрибут $\boldsymbol{Z}$ будет взят для разбиения

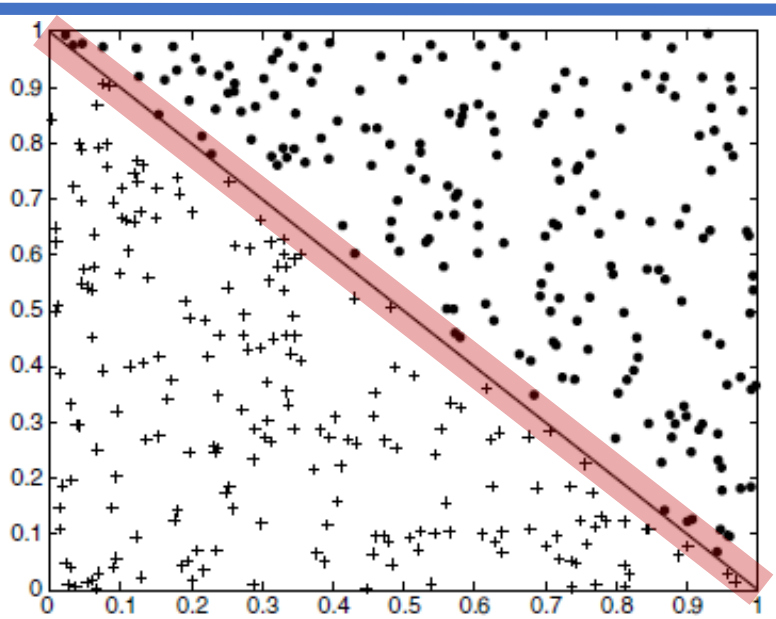




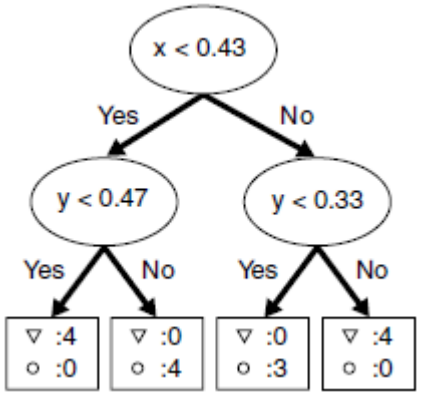


Прямолинейные разрезы





$x + y = 1$



?

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN: 978-0123814791
  - 8. Classification: Basic Concepts. 8.1. Basic Concepts, 8.2. Decision Tree Induction, pp. 327-350
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN: 978-0-13-312890-1
  - 3. Decision Tree Induction. 3.1 Basic Concepts, 3.2 General Framework for Classification, Decision Tree Classifier, pp. 113-146