Логические алгоритмы классификации

К. В. Воронцов, А.В. Зухба
vokov@forecsys.ru
a__l@mail.ru

февраль 2014

Содержание

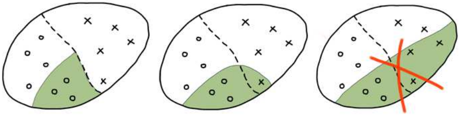
Логическая закономерность

$X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$ — обучающая выборка, $y_i = y(x_i)$.

Логическая закономерность (правило, rule) — это предикат $\varphi\colon X \to \{0,1\}$, удовлетворяющий двум требованиям:

1. *интерпретируемость*:
   1) φ записывается на естественном языке;
   2) φ зависит от небольшого числа признаков (1–7);
2. *информативность* относительно одного из классов $c \in Y$:
   $p(\varphi) = \#\bigl\{x_i\colon \varphi(x_i){=}1 \text{ и } y_i{=}c\bigr\} \to \max$;
   $n(\varphi) = \#\bigl\{x_i\colon \varphi(x_i){=}1 \text{ и } y_i{\neq}c\bigr\} \to \min$;

Если $\varphi(x) = 1$, то говорят «φ выделяет x» (φ covers x).



Требование интерпретируемости

1) $\varphi(x)$ записывается на естественном языке;
2) $\varphi(x)$ зависит от небольшого числа признаков (1–7);

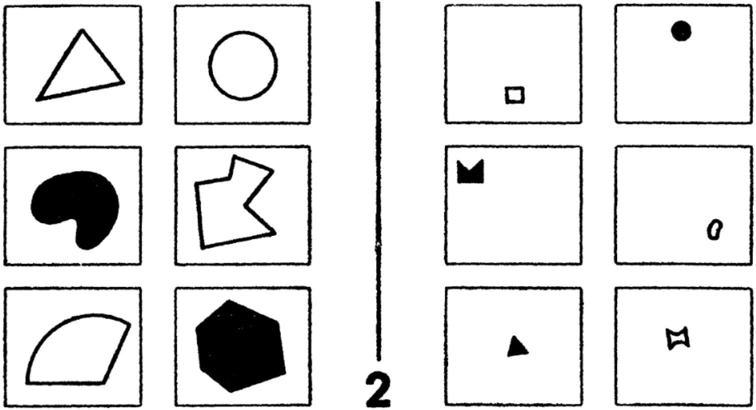
Пример (из области медицины)

Если *возраст* > 60 ***и*** *пациент ранее перенёс инфаркт,*
то *операцию не делать, риск отрицательного исхода 60%.*

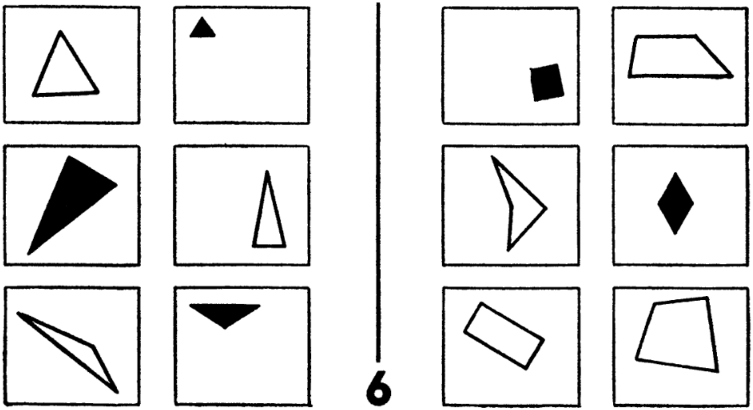
Пример (из области кредитного скоринга)

Если *в анкете указан домашний телефон*
и *зарплата* $> \$2000$ ***и*** *сумма кредита* $< \$5000$
то *кредит можно выдать, риск дефолта 5%.*

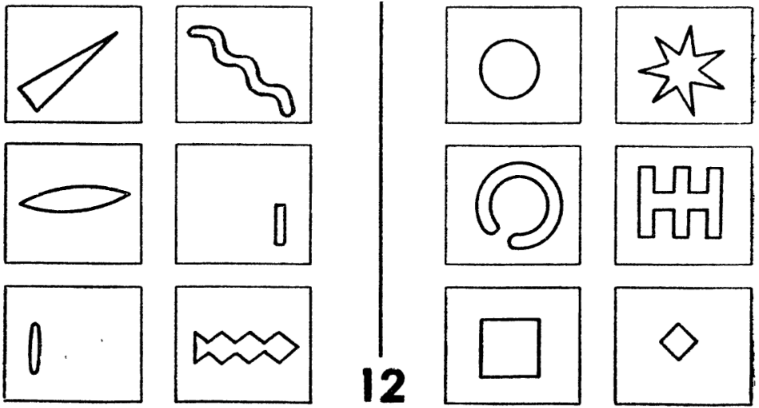
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



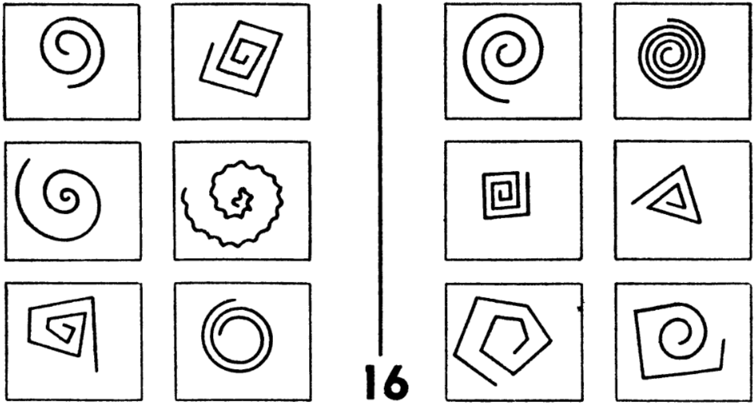
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



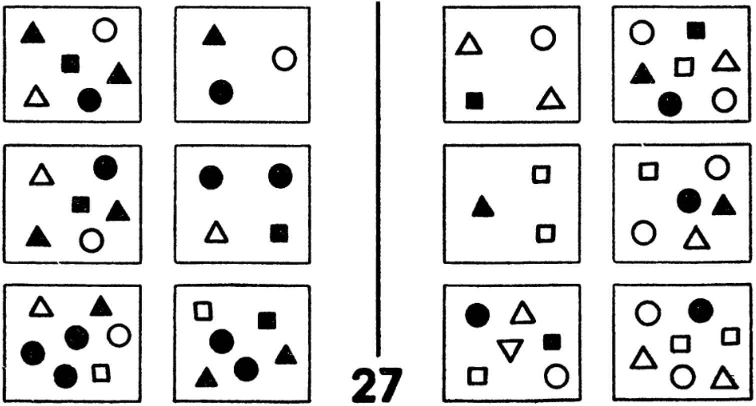
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



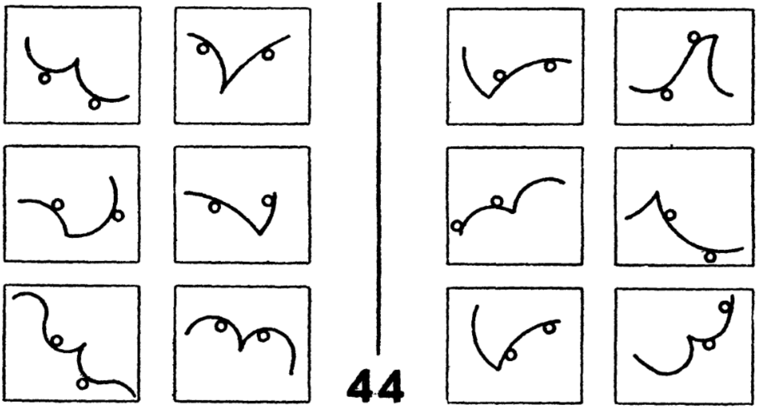
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



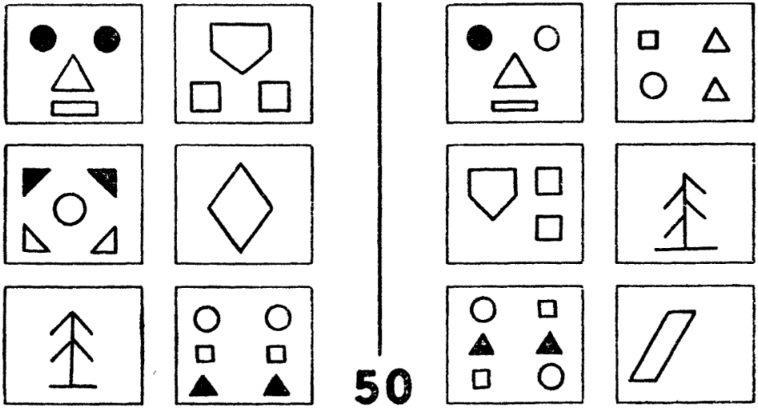
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



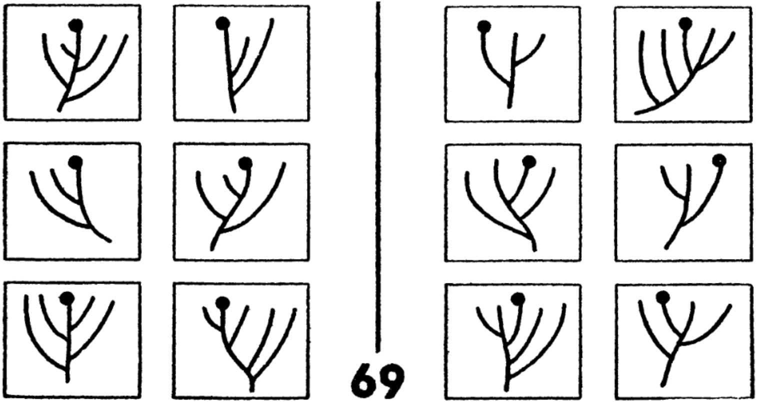
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Основные вопросы построения логических алгоритмов

1. Как изобретать признаки $f_1(x), \dots, f_n(x)$?
   — не наука, а искусство (размышления, озарения, эксперименты, консультации, мозговые штурмы,...)
2. Какого вида закономерности $\varphi(x)$ нам нужны?
   — простые формулы от малого числа признаков
3. Как определять информативность?
   — так, чтобы одновременно $p_c \to \max$, $n_c \to \min$
4. Как искать закономерности?
   — перебором подмножеств признаков
5. Как объединять закономерности в алгоритм?
   — любым классификатором ($\varphi(x)$ — это тоже признаки)

Закономерность — интерпретируемый высокоинформативный одноклассовый классификатор с отказами.

Часто используемые виды закономерностей

1. *Конъюнкция* пороговых условий (термов):

$$\varphi(x) = \bigwedge_{j \in J} \left[a_j \leqslant f_j(x) \leqslant b_j\right].$$

2. *Синдром* — когда выполнено не менее d термов из J, (при $d = |J|$ это конъюнкция, при $d = 1$ — дизъюнкция):

$$\varphi(x) = \left[\sum_{j \in J} \left[a_j \leqslant f_j(x) \leqslant b_j\right] \geqslant d\right],$$

Синдромы обнаруживаются во многих прикладных областях: в медицинской диагностике, в кредитном скоринге, в геологическом прогнозировании, и др.

Параметры J, a_j, b_j, d настраиваются по обучающей выборке.

Часто используемые виды закономерностей

3. *Полуплоскость* — линейная пороговая функция:

$$\varphi(x) = \Bigl[\sum_{j \in J} w_j f_j(x) \geqslant w_0 \Bigr].$$

4. *Шар* — пороговая функция близости:

$$\varphi(x) = \bigl[r(x, x_0) \leqslant w_0 \bigr],$$

АВО — алгоритмы вычисления оценок [Ю. И. Журавлёв, 1971]:

$$r(x, x_0) = \max_{j \in J} w_j \bigl| f_j(x) - f_j(x_0) \bigr|.$$

SCM — машины покрывающих множеств [M. Marchand, 2001]:

$$r(x, x_0) = \sum_{j \in J} w_j \bigl| f_j(x) - f_j(x_0) \bigr|^{\gamma}.$$

Параметры J, w_j, w_0, x_0 настраиваются по обучающей выборке путём оптимизации *критерия информативности*.

Логический (эвристический) критерий закономерности

$$\underbrace{\overbrace{\boxed{\begin{matrix}\varphi(x_i)=1\\ y_i=c\end{matrix}}}^{p_c(\varphi)}\boxed{\begin{matrix}\varphi(x_i)=0\\ y_i=c\end{matrix}}}_{P_c}\underbrace{\overbrace{\boxed{\begin{matrix}\varphi(x_i)=1\\ y_i\neq c\end{matrix}}}^{n_c(\varphi)}\boxed{\begin{matrix}\varphi(x_i)=0\\ y_i\neq c\end{matrix}}}_{N_c}$$

Определение

Предикат $\varphi(x)$ — *логическая* ε,δ*-закономерность класса* $c \in Y$

$$E_c(\varphi, X^\ell) = \frac{n_c(\varphi)}{p_c(\varphi) + n_c(\varphi)} \leqslant \varepsilon;$$

$$D_c(\varphi, X^\ell) = \frac{p_c(\varphi)}{\ell} \geqslant \delta.$$

Проблема: хотелось бы иметь один скалярный критерий.

Часто используемые критерии информативности

Проблема: надо сравнивать закономерности φ.
Как свернуть два критерия в один критерий информативности?

$$\begin{cases} p(\varphi) \to \max \\ n(\varphi) \to \min \end{cases} \quad \overset{?}{\Longrightarrow} \quad I(p,n) \to \max$$

Очевидные, но не всегда адекватные свёртки:

- $\dfrac{p}{p+n} \to \max$ (precision);
- $p - n \to \max$ (accuracy);
- $p - Cn \to \max$ (linear cost accuracy);
- $\dfrac{p}{P} - \dfrac{n}{N} \to \max$ (relative accuracy);

  $P = \#\{x_i\colon y_i{=}c\}$ — число «своих» во всей выборке;
  $N = \#\{x_i\colon y_i{\neq}c\}$ — число «чужих» во всей выборке.

Нетривиальность проблемы свёртки двух критериев

Пример.
Претенденты на звание «Критерий информативности»
при $P = 200$, $N = 100$ и различных p и n.

p	n	$p-n$	$p-5n$	$\frac{p}{P}-\frac{n}{N}$	$\frac{p}{n+1}$	I_c	IGain_c	$\sqrt{p}-\sqrt{n}$
50	0	**50**	50	0.25	50	22.65	23.70	7.07
100	50	**50**	−150	0	1.96	2.33	1.98	2.93
50	9	41	**5**	0.16	**5**	7.87	7.94	4.07
5	0	5	**5**	0.03	**5**	2.04	3.04	2.24
100	0	**100**	100	**0.5**	100	52.18	53.32	10.0
140	20	**120**	40	**0.5**	6.67	37.09	37.03	7.36

Часто используемые критерии информативности

Адекватные, но неочевидные критерии:

- энтропийный критерий информационного выигрыша:
$$h\left(\frac{P}{\ell}\right)-\frac{p+n}{\ell}h\left(\frac{p}{p+n}\right)-\frac{\ell-p-n}{\ell}h\left(\frac{P-p}{\ell-p-n}\right)\to\max,$$
где $h(q) = -q\log_2 q - (1-q)\log_2(1-q)$;
- критерий неслучайности FET (Fisher's Exact Test):
$C_P^p C_N^n / C_{P+N}^{p+n} \to \min$;
- критерий бустинга [Cohen, Singer, 1999]:
$\sqrt{p} - \sqrt{n} \to \max$
- нормированный критерий бустинга:
$\sqrt{p/P} - \sqrt{n/N} \to \max$;
- перестановочные тесты.

Статистический критерий информативности

Точный тест Фишера. Пусть X — в.п., выборка X^ℓ — i.i.d.
Гипотеза H_0: $y(x)$ и $\varphi(x)$ — независимые случайные величины.
Тогда вероятность реализации пары (p, n) описывается *гипергеометрическим распределением*:

$$\mathrm{P}(p,n) = \frac{C_P^p C_N^n}{C_{P+N}^{p+n}}, \quad 0 \leqslant p \leqslant P, \quad 0 \leqslant n \leqslant N,$$

где $C_N^n = \frac{N!}{n!(N-n)!}$ — биномиальные коэффициенты.

Определение

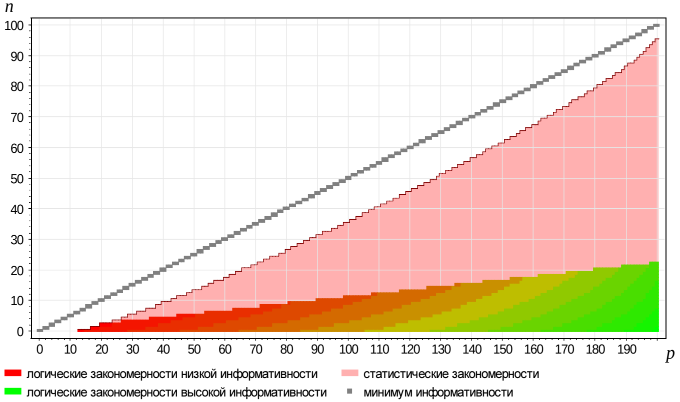
Информативность предиката $\varphi(x)$ относительно класса $c \in Y$:

$$I_c(\varphi, X^\ell) = -\ln \frac{C_{P_c}^{p_c(\varphi)} C_{N_c}^{n_c(\varphi)}}{C_{P_c+N_c}^{p_c(\varphi)+n_c(\varphi)}},$$

$I_c(\varphi, X^\ell) \geqslant I_0$ *— статистическая закономерность класса $c \in Y$.*

Соотношение логического и статистического критериев

Области логических ($\varepsilon = 0.1$) и статистических ($I_0 = 5$) закономерностей в координатах (p, n) при $P = 200$, $N = 100$.



Философский вопрос: закономерность == неслучайность?

Энтропийный критерий информативности

Пусть ω_0, ω_1 — 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:

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

Энтропия выборки X^ℓ, если исходы — это классы $y{=}c$, $y{\neq}c$:

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

Энтропия выборки X^ℓ после получения информации φ:

$$H(y|\varphi) = \frac{p+n}{\ell} h\left(\frac{p}{p+n}\right) + \frac{\ell-p-n}{\ell} h\left(\frac{P-p}{\ell-p-n}\right).$$

Информационный выигрыш (Information gain, IGain):

$$\mathrm{IGain}_c(\varphi, X^\ell) = H(y) - H(y|\varphi).$$

Соотношение статистического и энтропийного критериев

Определение

Предикат φ *— закономерность по энтропийному критерию, если* $\mathrm{IGain}_c(\varphi, X^\ell) > G_0$ *при некотором* G_0.

Теорема

Энтропийный критерий IGain_c *асимптотически эквивалентен статистическому* I_c:

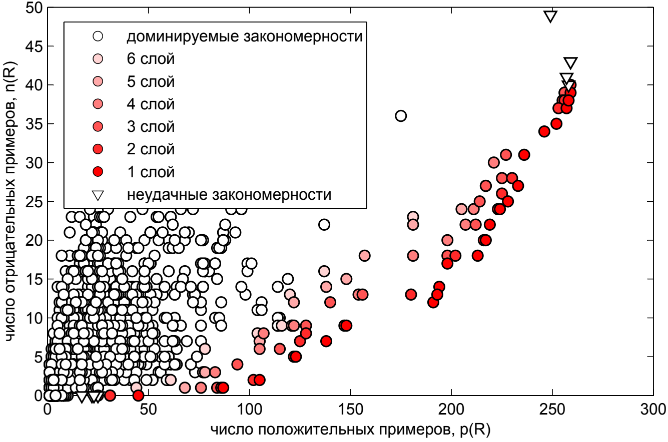
$$\mathrm{IGain}_c(\varphi, X^\ell) \to \frac{1}{\ell \ln 2} I_c(\varphi, X^\ell) \quad \text{при } \ell \to \infty.$$

Доказательство:

применить формулу Стирлинга к статистическому критерию.

Парето-критерий информативности в (p, n)-плоскости

Парето-фронт — множество недоминируемых закономерностей (точка φ недоминируема, если правее и ниже точек нет)

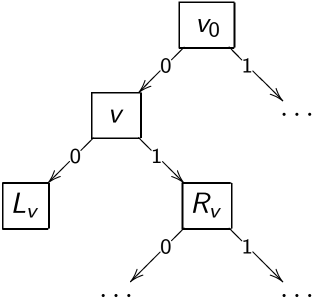


Определение бинарного решающего дерева

Бинарное решающее дерево — алгоритм классификации $a(x)$, задающийся бинарным деревом:

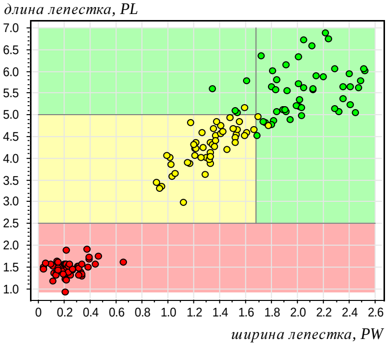
1) $\forall v \in V_{\text{внутр}} \;\to\;$ предикат $\beta_v : X \to \{0,1\}$, $\beta \in \mathscr{B}$
2) $\forall v \in V_{\text{лист}} \;\to\;$ имя класса $c_v \in Y$.

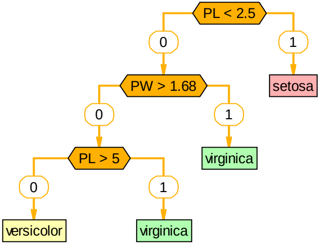
$v := v_0$;
пока $v \in V_{\text{внутр}}$
  если $\beta_v(x) = 1$ **то**
    переход вправо:
    $v := R_v$;
  иначе
    переход влево:
    $v := L_v$;
вернуть c_v.



Пример решающего дерева

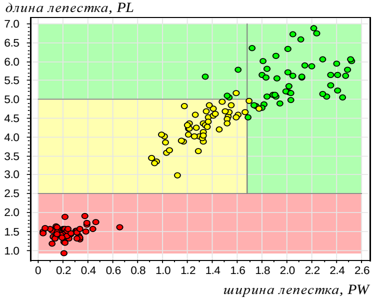
Задача Фишера о классификации цветков ириса на 3 класса, в выборке по 50 объектов каждого класса, 4 признака.

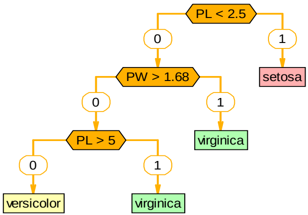




На графике: в осях двух самых информативных признаков (из 4) два класса разделились без ошибок, на третьем 3 ошибки.

Решающее дерево → покрывающий набор конъюнкций





$\varphi(x_i)=20$ y_i *red*	setosa $r_1(x) = [PL \leqslant 2.5]$
$\varphi(x_i)=20$ y_i *green*	virginica $r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$
$\varphi(x_i)=20$ y_i *green*	virginica $r_3(x) = [PL > 5] \wedge [PW \leqslant 1.68]$
$\varphi(x_i)=20$ y_i *yellow*	versicolor $r_4(x) = [PL > 2.5] \wedge [PL \leqslant 5] \wedge [PW < 1.68]$

Жадный алгоритм построения дерева ID3

1: **ПРОЦЕДУРА** LearnID3 $(U \subseteq X^\ell)$;

2: **если** все объекты из U лежат в одном классе $c \in Y$ **то**

3:     **вернуть** новый лист v, $c_v := c$;

4: найти предикат с максимальной информативностью:
$\beta := \arg\max\limits_{\beta \in \mathscr{B}} I(\beta, U)$;

5: разбить выборку на две части $U = U_0 \cup U_1$ по предикату β:
$U_0 := \bigl\{x \in U \colon \beta(x) = 0\bigr\}$;
$U_1 := \bigl\{x \in U \colon \beta(x) = 1\bigr\}$;

6: **если** $U_0 = \varnothing$ или $U_1 = \varnothing$ **то**

7:     **вернуть** новый лист v, $c_v :=$ Мажоритарный класс(U);

8: создать новую внутреннюю вершину v: $\beta_v := \beta$;
построить левое поддерево: $L_v :=$ LearnID3 (U_0);
построить правое поддерево: $R_v :=$ LearnID3 (U_1);

9: **вернуть** v;

Разновидности многоклассовых критериев ветвления

1. **Отделение одного класса** (слишком сильное ограничение):

$$I(\beta, X^\ell) = \max_{c \in Y} I_c(\beta, X^\ell).$$

2. **Многоклассовый энтропийный критерий:**

$$I(\beta, X^\ell) = \sum_{c \in Y} h\left(\frac{P_c}{\ell}\right) - \frac{p}{\ell} \sum_{c \in Y} h\left(\frac{p_c}{p}\right) - \frac{\ell - p}{\ell} \sum_{c \in Y} h\left(\frac{P_c - p_c}{\ell - p}\right),$$

где $P_c = \#\{x_i \colon y_i = c\}$, $\quad p = \#\{x_i \colon \beta(x_i) = 1\}$, $\quad h(z) \equiv -z \log_2 z$.

3. **Критерий Джини:**

$$I(\beta, X^\ell) = \#\bigl\{(x_i, x_j) \colon \beta(x_i) = \beta(x_j) \text{ и } y_i = y_j\bigr\}.$$

4. *D*-**критерий В.И.Донского:**

$$I(\beta, X^\ell) = \#\bigl\{(x_i, x_j) \colon \beta(x_i) \neq \beta(x_j) \text{ и } y_i \neq y_j\bigr\}.$$

Обработка пропусков

На стадии обучения:

- Если $\beta(x_i)$ не определено, то при вычислении $I(\beta, U)$ объект x_i исключается из выборки U.
- $q_v = \frac{|U_0|}{|U|}$ — оценка вероятности левой ветви, $\forall v \in V_{\text{внутр}}$.

На стадии классификации:

- $\hat{P}_v(y|x) = \beta_v(x)\hat{P}_{L_v}(y|x) + \bigl(1-\beta_v(x)\bigr)\hat{P}_{R_v}(y|x)$;
- $\beta_v(x)$ не определено $\Rightarrow$ *пропорциональное распределение*:
$$\hat{P}_v(y|x) = \begin{cases} [y = c_v], & v \in V_{\text{лист}}; \\ q_v\hat{P}_{L_v}(y|x) + (1-q_v)\hat{P}_{R_v}(y|x), & v \in V_{\text{внутр}}. \end{cases}$$
- Окончательное решение — наиболее вероятный класс:
$$y = \arg\max_{y \in Y} \hat{P}_{v_0}(y|x).$$

Решающие деревья ID3: достоинства и недостатки

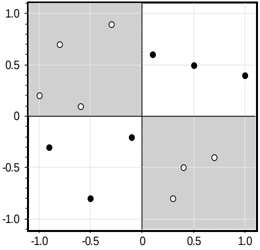
Достоинства:

- Интерпретируемость и простота классификации.
- Гибкость: можно варьировать множество $\mathscr{B}$.
- Допустимы разнотипные данные и данные с пропусками.
- Трудоёмкость линейна по длине выборки $O(|\mathscr{B}|h\ell)$.
- Не бывает отказов от классификации.

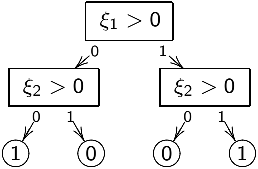
Недостатки:

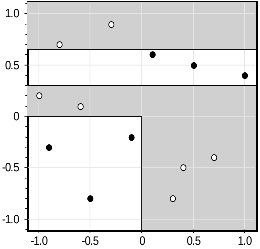
- Жадный ID3 переусложняет структуру дерева, и, как следствие, сильно переобучается.
- Фрагментация выборки: чем дальше v от корня, тем меньше статистическая надёжность выбора β_v, c_v.
- Высокая чувствительность к шуму, к составу выборки, к критерию информативности.

Жадный ID3 переусложняет структуру дерева

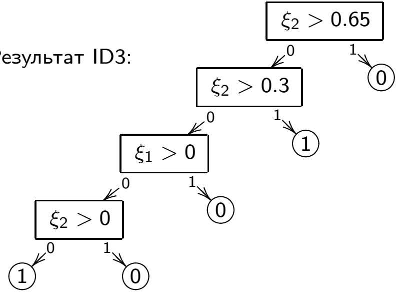


Оптимальное дерево для задачи XOR:





Результат ID3:



Редукция дерева («стрижка», pruning: C4.5, CART)

X^k — независимая контрольная выборка, $k \approx 0.5\ell$.

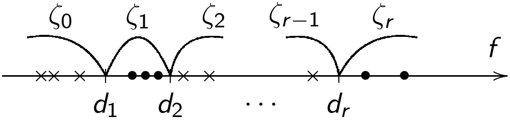
1: **для всех** $v \in V_{\text{внутр}}$
2: $S_v :=$ подмножество объектов X^k, дошедших до v;
3: **если** $S_v = \varnothing$ **то**
4: **вернуть** новый лист v, $c_v :=$ Мажоритарный класс(U);
5: число ошибок при классификации S_v четырьмя способами:
$r(v)$ — поддеревом, растущим из вершины v;
$r_L(v)$ — поддеревом левой дочерней вершины L_v;
$r_R(v)$ — поддеревом правой дочерней вершины R_v;
$r_c(v)$ — к классу $c \in Y$.
6: в зависимости от того, какое из них минимально:
сохранить поддерево v;
заменить поддерево v поддеревом L_v;
заменить поддерево v поддеревом R_v;
заменить поддерево v листом, $c_v := \arg\min_{c \in Y} r_c(v)$.

Вспомогательная задача бинаризации вещественного признака

Цель: сократить перебор предикатов вида $\left[\alpha \leqslant f(x) \leqslant \beta\right]$.

Дано: выборка значений вещественного признака $f(x_i)$, $x_i \in X^\ell$.
Найти: наилучшее (в каком-то смысле) разбиение области значений признака на относительно небольшое число зон:

$$\begin{aligned}
\zeta_0(x) &= [f(x) < d_1];\\
\zeta_s(x) &= [d_s \leqslant f(x) < d_{s+1}], \qquad s = 1, \ldots, r-1;\\
\zeta_r(x) &= [d_r \leqslant f(x)].
\end{aligned}$$



Способы разбиения области значений признака на зоны

1. Жадная максимизация информативности путём слияний
2. Разбиение на равномощные подвыборки
3. Разбиение по равномерной сетке
4. Объединение нескольких разбиений

Повышение интерпретируемости пороговых значений

Задача: на отрезке $[a, b]$ найти значение x с минимальным числом значащих цифр.
Если таких x несколько, выбрать $\arg\min\limits_x \left|\frac{1}{2}(a+b) - x\right|$.

Алгоритм разбиения области значений признака на зоны

Вход: выборка X^ℓ; класс $c \in Y$; параметры r и δ_0.
Выход: $D = \{d_1 < \dots < d_r\}$ — последовательность порогов;

$D := \varnothing$; упорядочить выборку X^ℓ по возрастанию $f(x_i)$;
для всех $i = 2, \dots, \ell$
  если $f(x_{i-1}) \neq f(x_i)$ и $[y_{i-1} = c] \neq [y_i = c]$ **то**
    добавить порог $\frac{1}{2}(f(x_{i-1}) + f(x_i))$ в конец D;
повторять
  для всех $d_i \in D$, $i = 1, \dots, |D| - 1$
    $\delta I_i := I_c(\zeta_{i-1} \vee \zeta_i \vee \zeta_{i+1}) - \max\{I_c(\zeta_{i-1}), I_c(\zeta_i), I_c(\zeta_{i+1})\}$;
  $i := \arg\max\limits_s \delta I_s$;
  если $\delta I_i > \delta_0$ **то**
    слить зоны ζ_{i-1}, ζ_i, ζ_{i+1}, удалив d_i и d_{i+1} из D;
пока $|D| > r + 1$.

Резюме в конце лекции

- Основные требования к логическим закономерностям:
  - интерпретируемость, информативность, различность.
- Преимущества решающих деревьев:
  - интерпретируемость,
  - допускаются разнотипные данные,
  - возможность обхода пропусков;
- Недостатки решающих деревьев:
  - переобучение,
  - фрагментация,
  - неустойчивость к шуму, составу выборки, критерию;
- Способы устранения этих недостатков:
  - редукция,
  - специальные виды деревьев ODT, ADT и др.
  - композиции (леса) деревьев — см. далее;