Логические алгоритмы классификации

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

Видеолекции: http://shad.yandex.ru/lectures

25 февраля 2016

Содержание

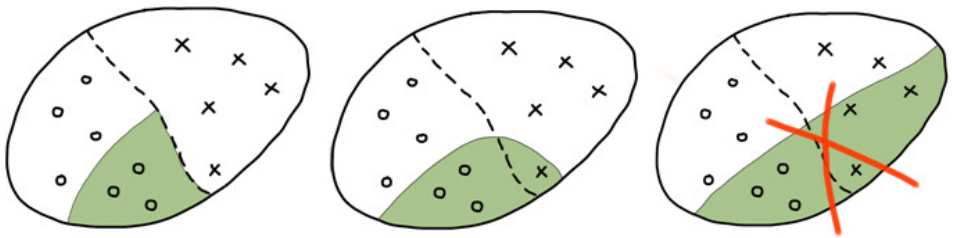
Логическая закономерность

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

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

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

Если $R(x) = 1$, то говорят «R выделяет x» (R covers x).



Требование интерпретируемости

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

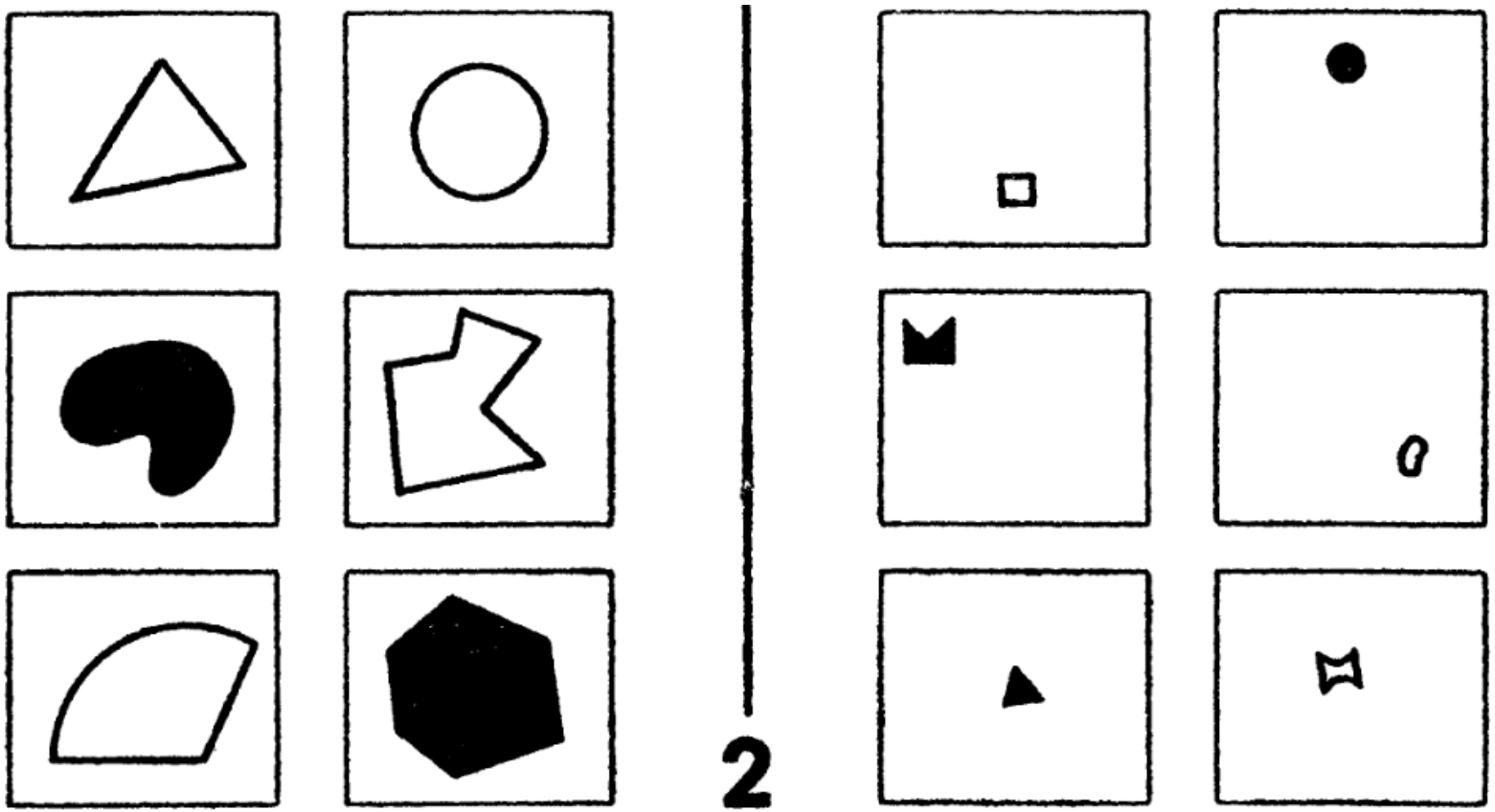
Пример (из области медицины)

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

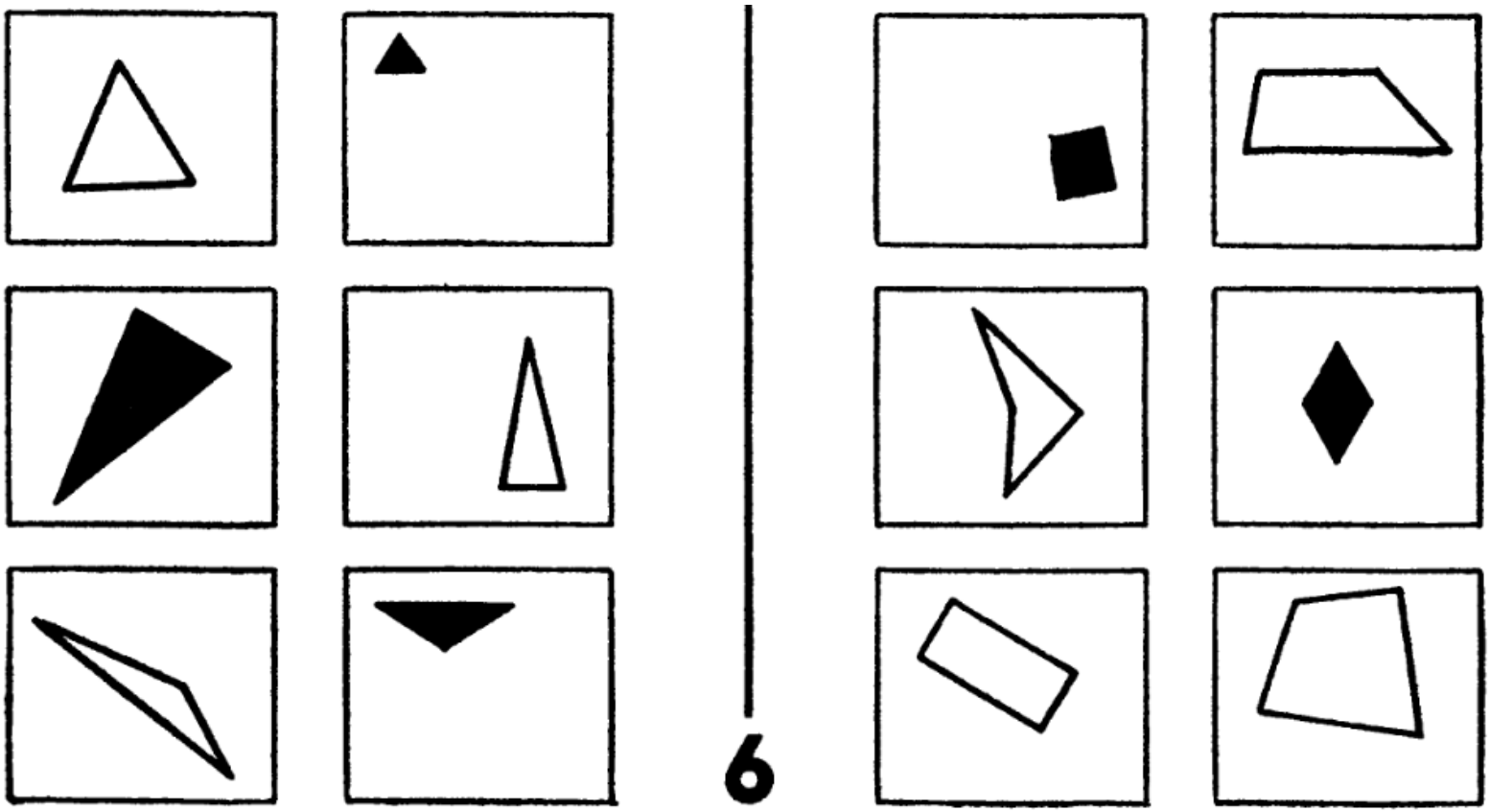
Пример (из области кредитного скоринга)

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

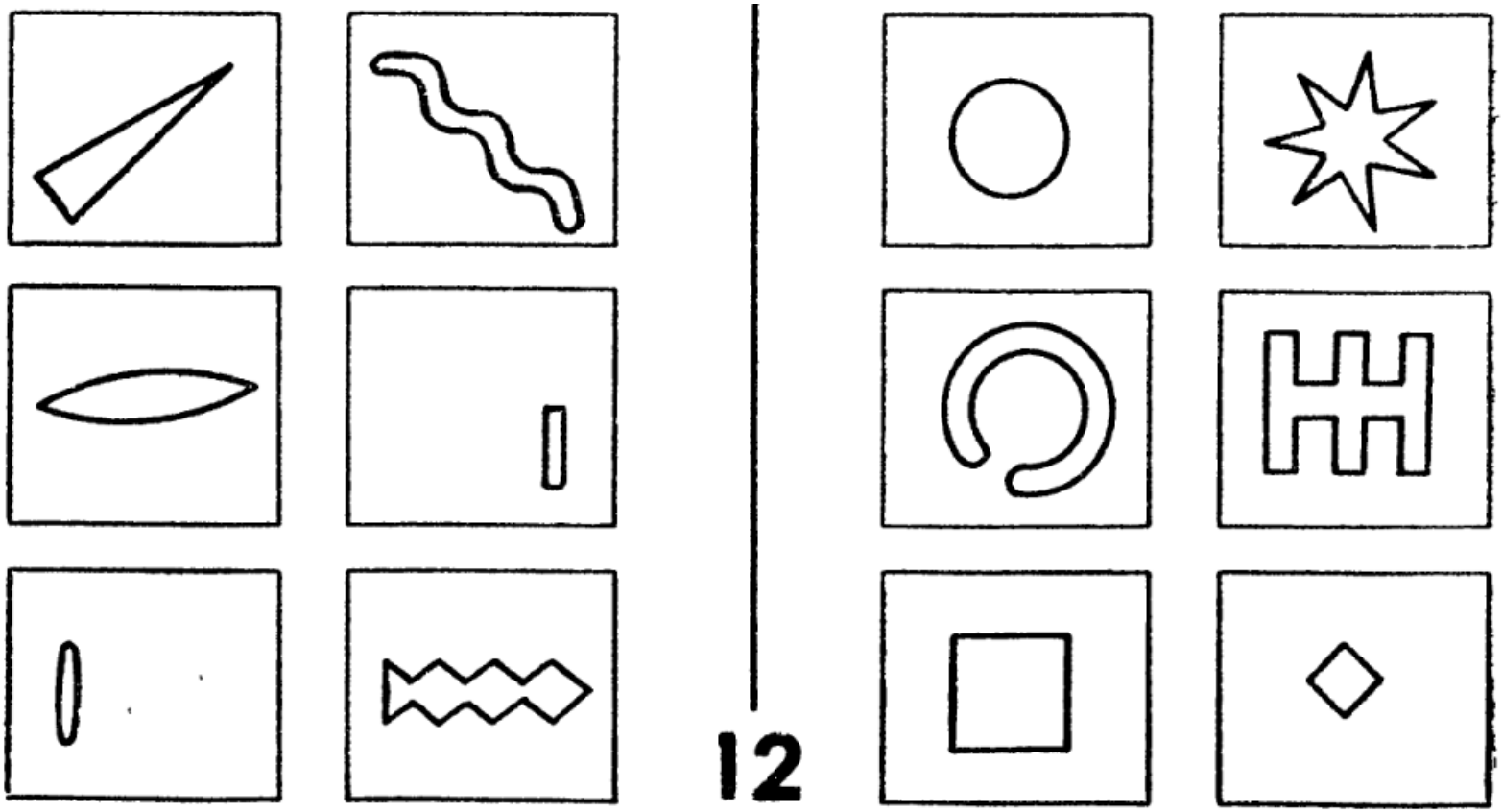
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



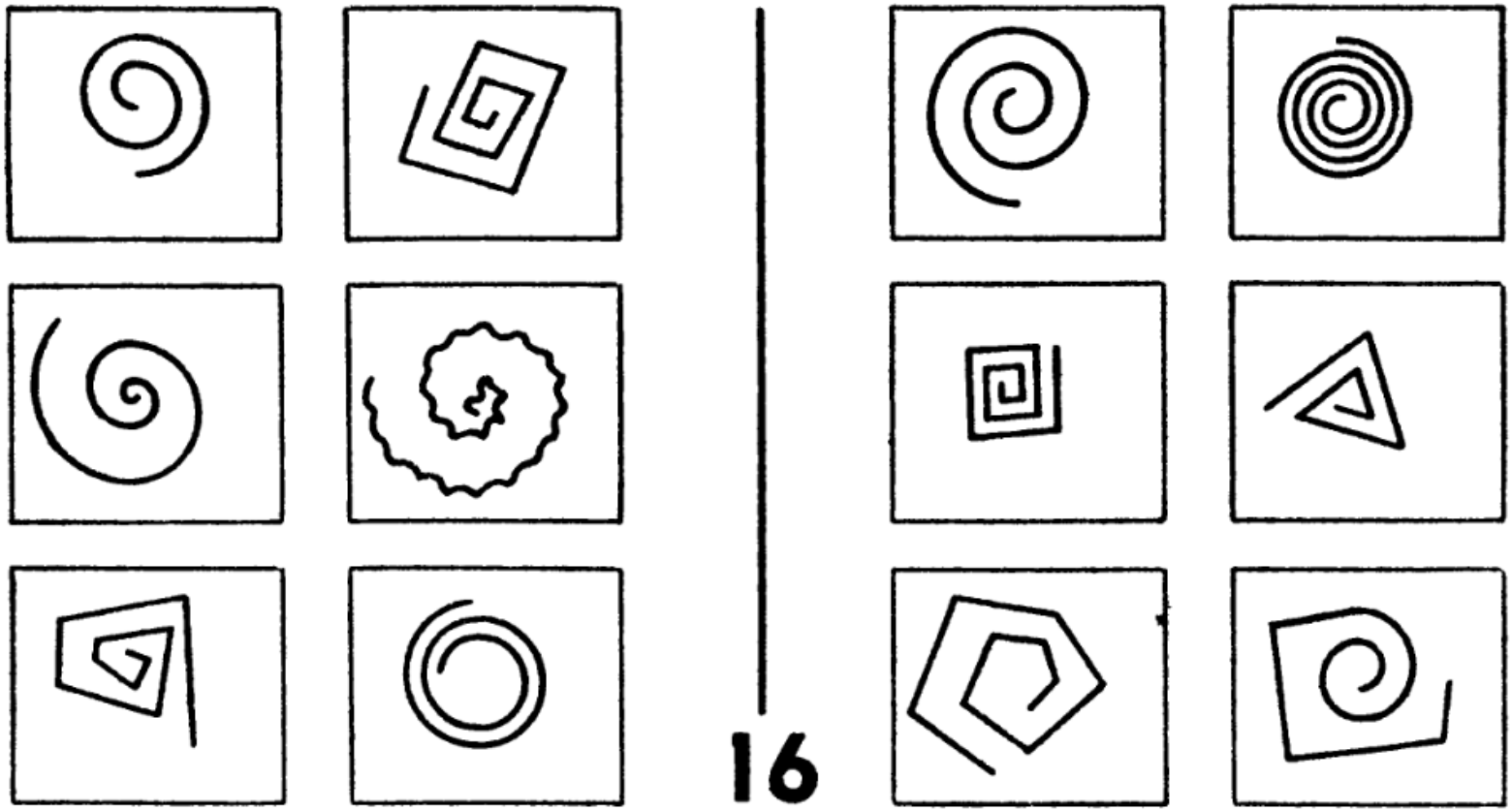
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



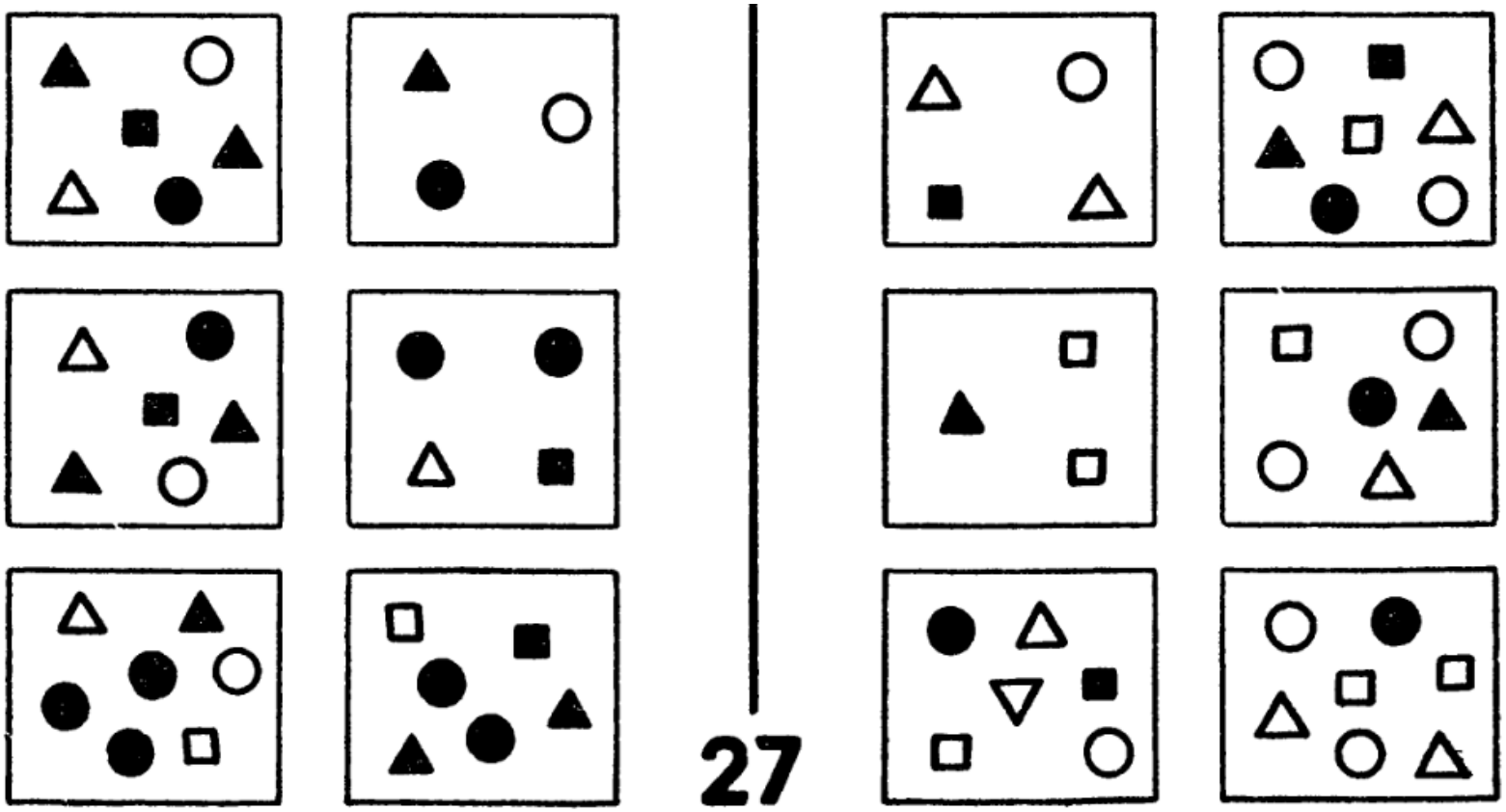
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



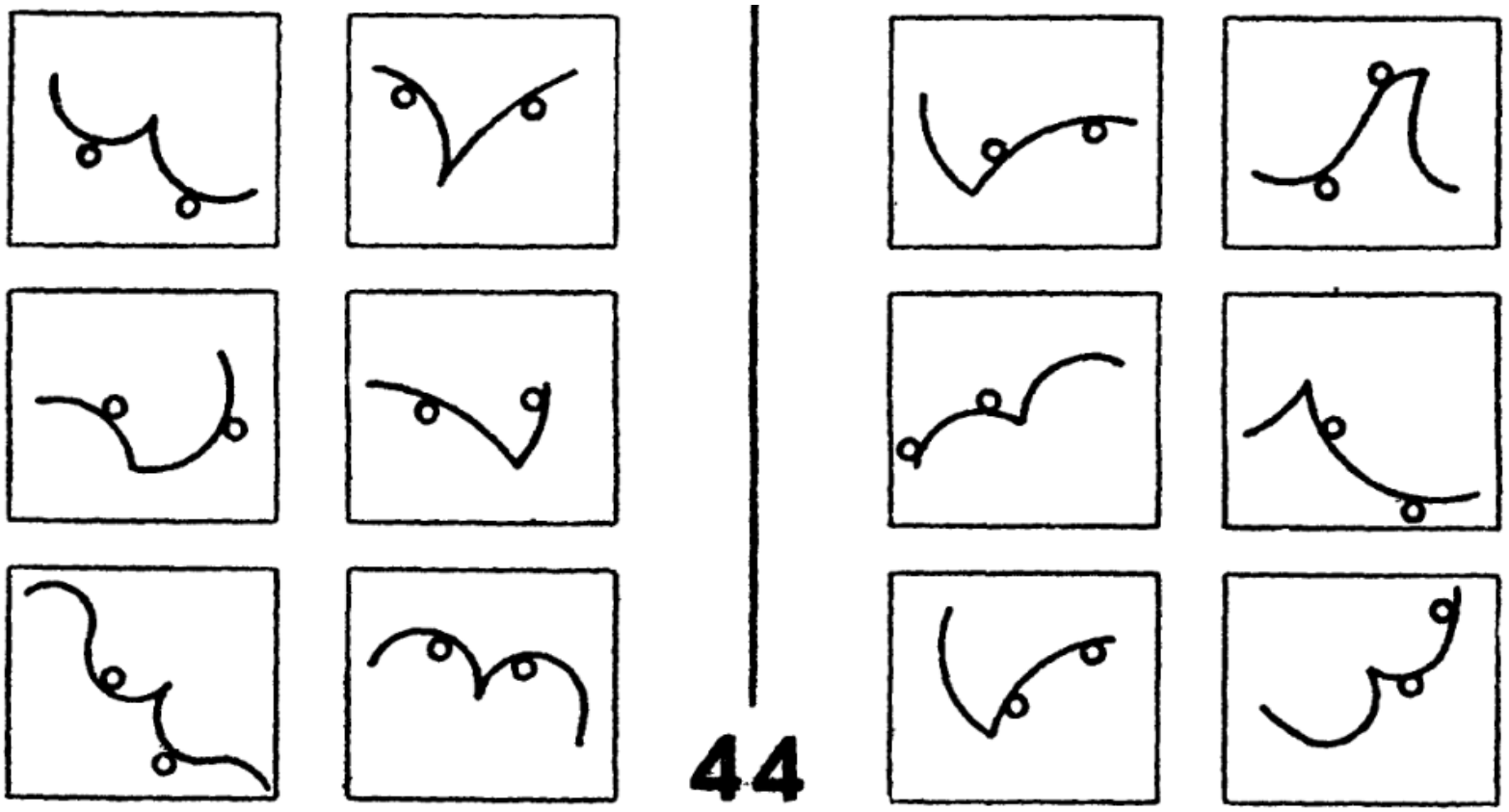
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



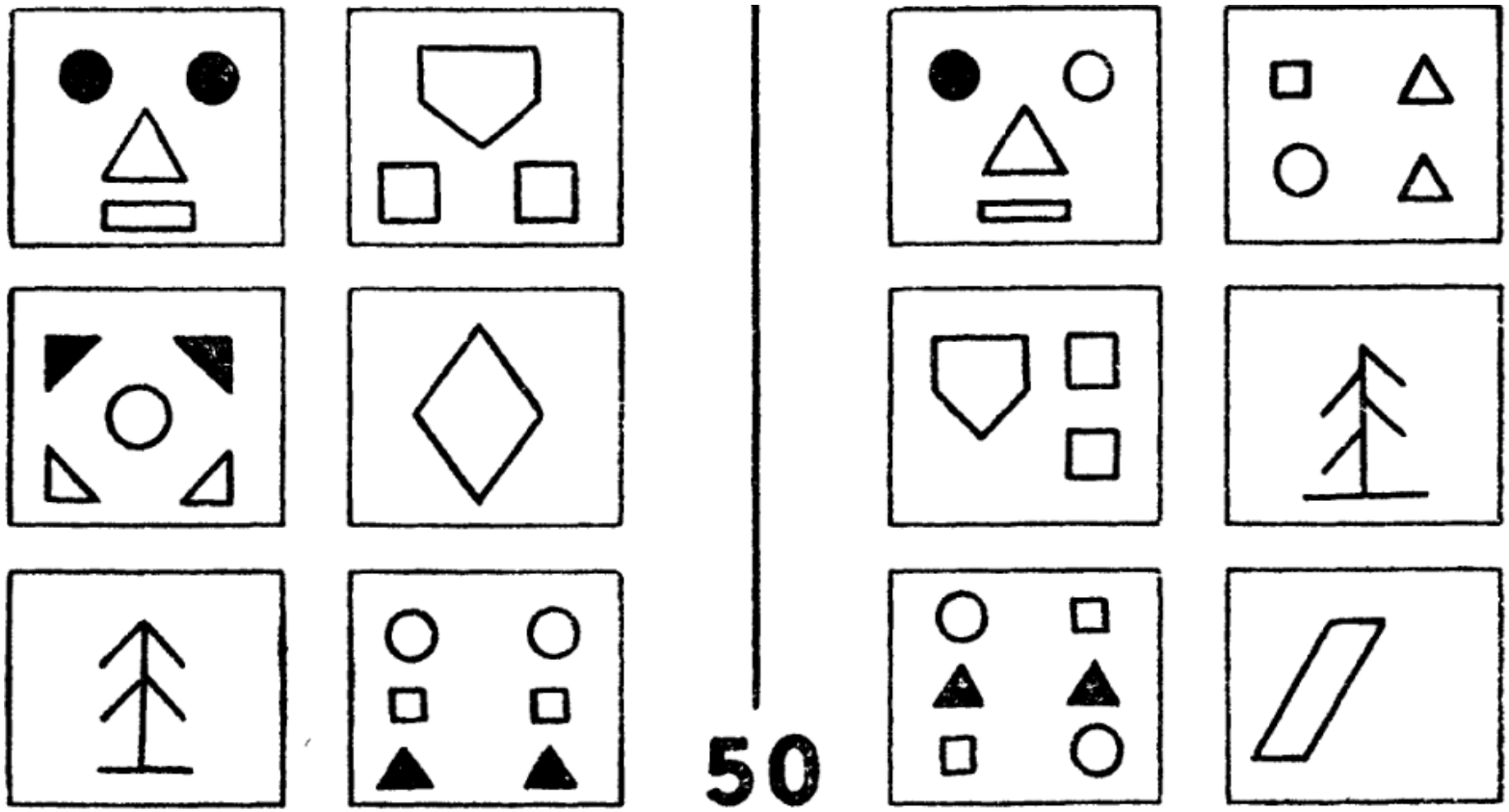
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



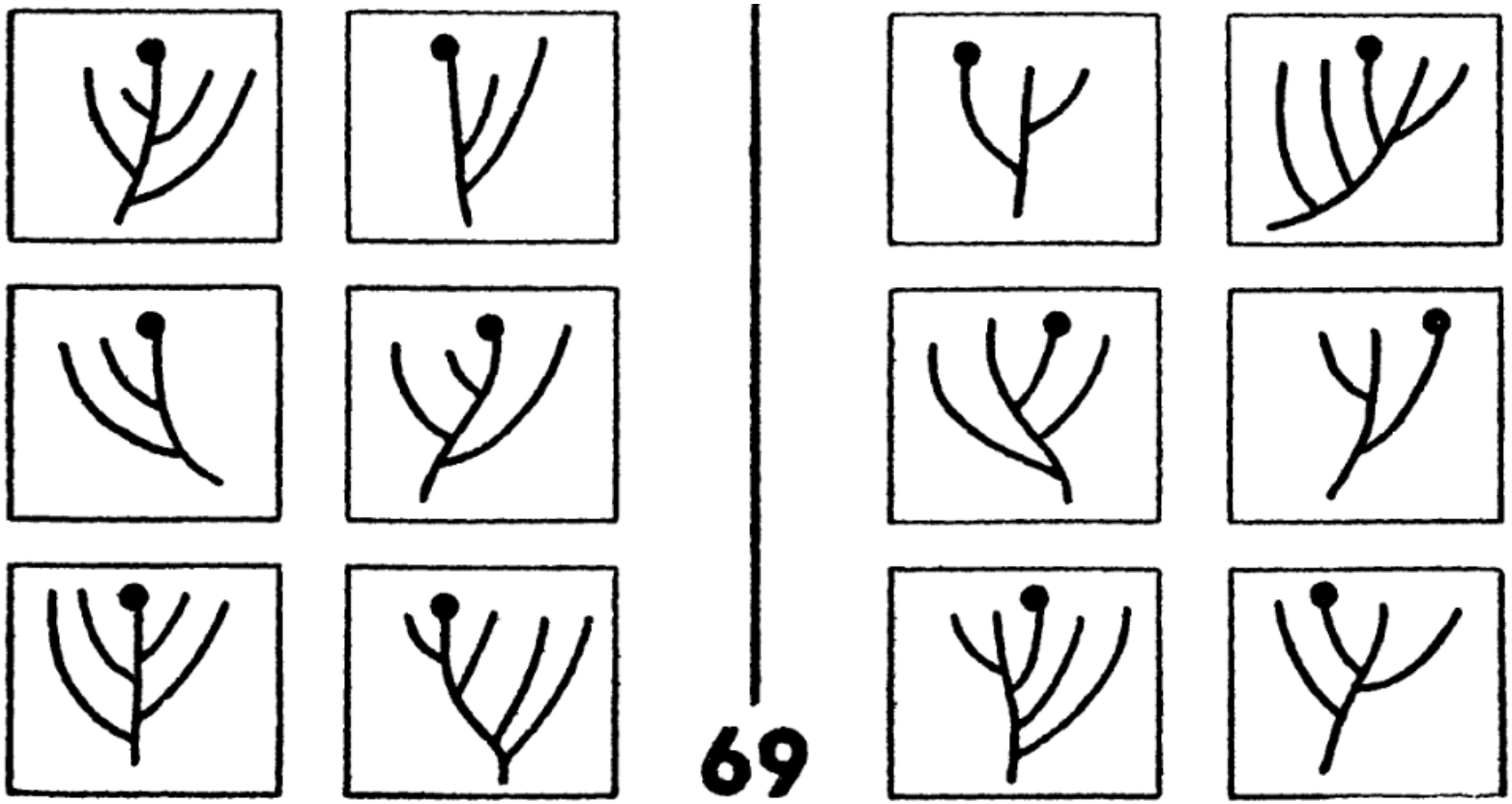
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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



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



Основные вопросы построения логических алгоритмов

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

Закономерность — интерпретируемый высокоинформативный одноклассовый классификатор с отказами.

Часто используемые виды закономерностей

1. Пороговое условие (решающий пень, decision stump):

$$R(x) = \big[f_j(x) \leqslant a_j\big] \text{ или } \big[a_j \leqslant f_j(x) \leqslant b_j\big].$$

2. *Конъюнкция* пороговых условий:

$$R(x) = \bigwedge_{j \in J} \big[a_j \leqslant f_j(x) \leqslant b_j\big].$$

3. *Синдром* — выполнение не менее d условий из $|J|$,
(при $d = |J|$ это конъюнкция, при $d = 1$ — дизъюнкция):

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

Параметры J, a_j, b_j, d настраиваются по обучающей выборке путём оптимизации *критерия информативности*.

Часто используемые виды закономерностей

4. *Полуплоскость* — линейная пороговая функция:

$$R(x) = \Big[\sum_{j\in J} w_j f_j(x) \geqslant w_0\Big].$$

5. *Шар* — пороговая функция близости:

$$R(x) = \big[\rho(x, x_0) \leqslant w_0\big],$$

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

$$\rho(x, x_0) = \max_{j\in J} w_j \big|f_j(x) - f_j(x_0)\big|.$$

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

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

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

Схема локального поиска информативных закономерностей

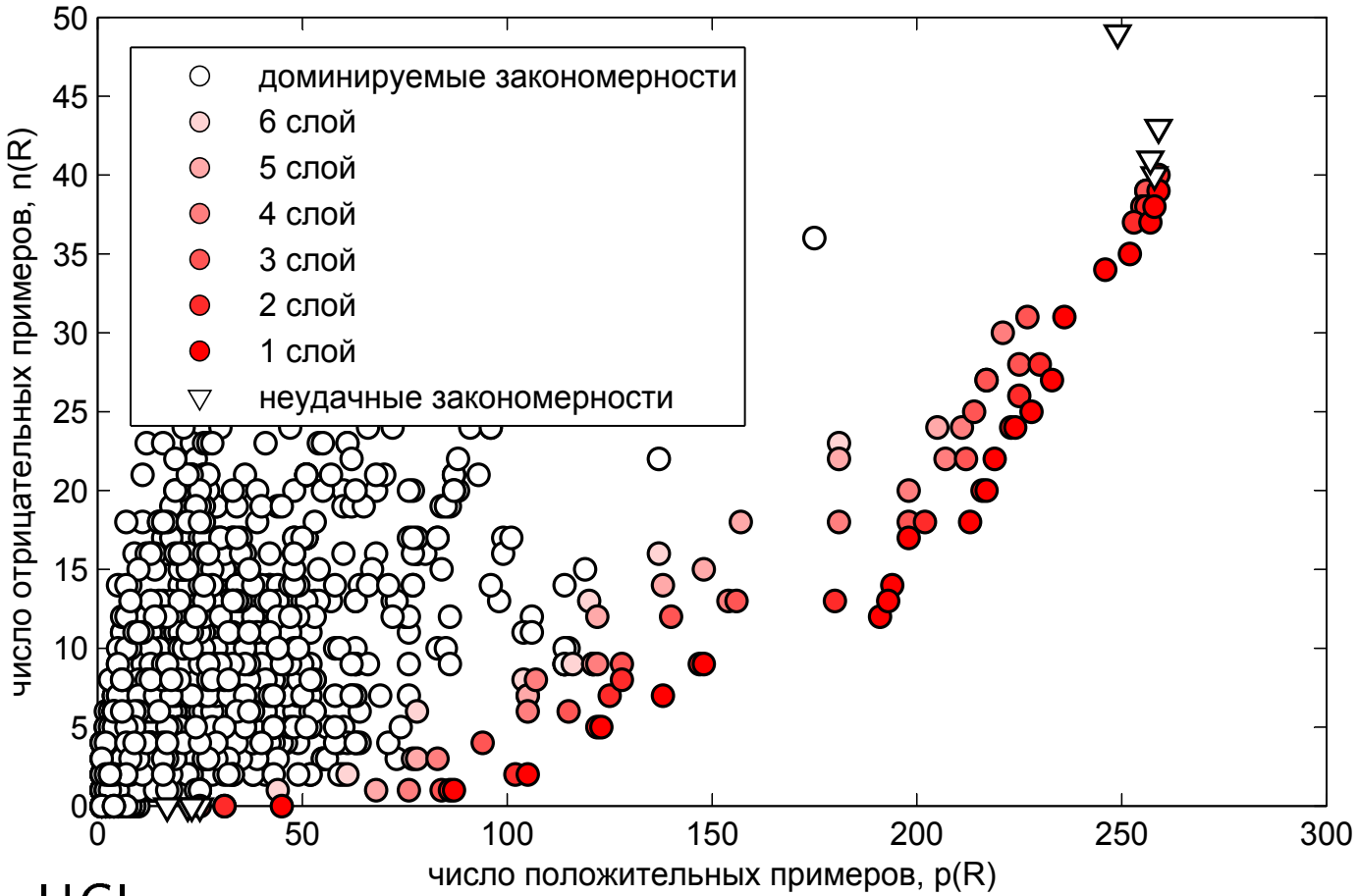
Вход: выборка X^ℓ;
Выход: множество закономерностей Z;

1: начальное множество правил Z;
2: **повторять**
3: $\quad Z' :=$ множество модификаций правил $R \in Z$;
4: $\quad$ удалить слишком похожие правила из $Z \cup Z'$;
5: $\quad$ оценить информативность всех правил $R \in Z'$;
6: $\quad Z :=$ наиболее информативные правила из $Z \cup Z'$;
7: **пока** правила продолжают улучшаться
8: **вернуть** Z.

Частные случаи:
— стохастический локальный поиск (SLS)
— генетические (эволюционные) алгоритмы
— метод ветвей и границ

Отбор закономерностей по информативности в (p, n)-плоскости

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



задача UCI:german

Определение решающего дерева (Decision Tree)

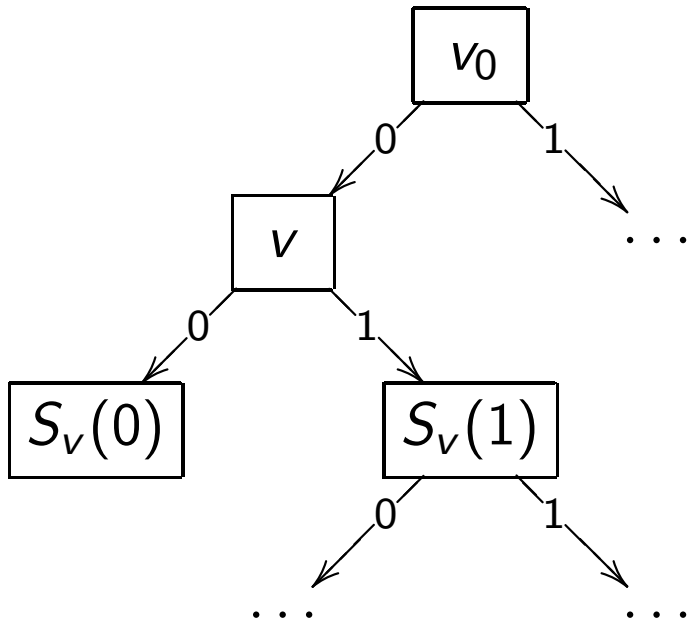
Решающее дерево — алгоритм классификации $a(x)$, задающийся *деревом* (связным ациклическим графом):

1) $V = V_{\text{внутр}} \sqcup V_{\text{лист}}$, $v_0 \in V$ — корень дерева;
2) $v \in V_{\text{внутр}}$: функции $f_v : X \to D_v$ и $S_v : D_v \to V$, $|D_v| < \infty$;
3) $v \in V_{\text{лист}}$: метка класса $y_v \in Y$.

1: $v := v_0$;
2: **пока** $v \in V_{\text{внутр}}$
3: $\quad v := S_v(f_v(x))$;
4: **вернуть** y_v;

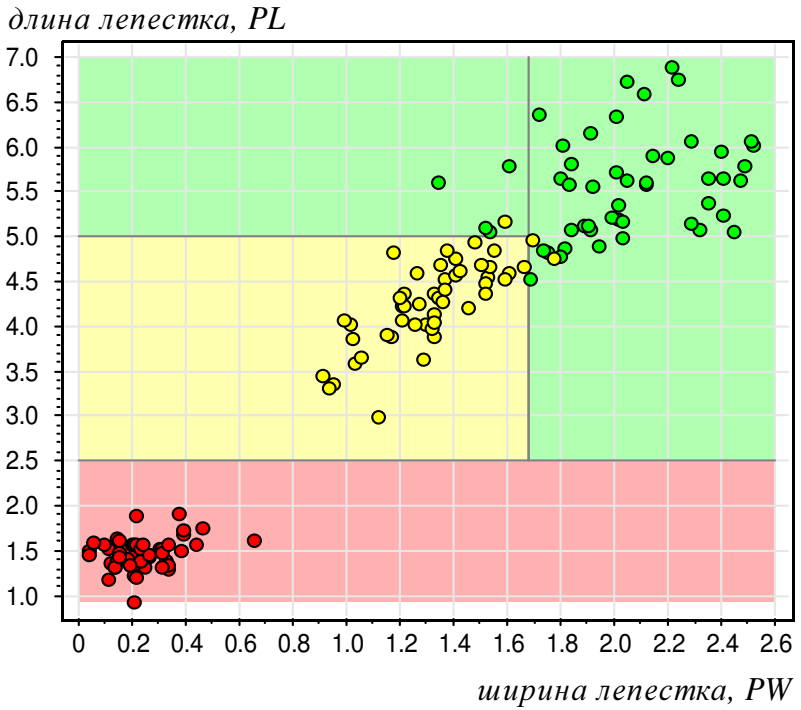
Частный случай: $D_v \equiv \{0, 1\}$ — бинарное решающее дерево

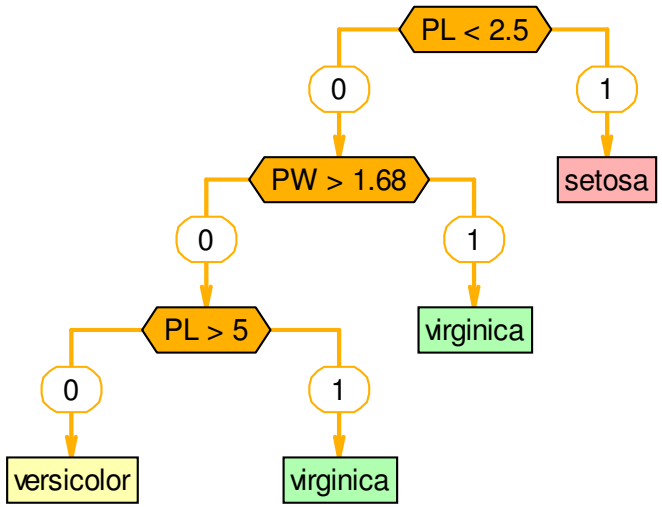
Пример: $f_v(x) = [f_j(x) \geqslant \theta_j]$



Пример решающего дерева

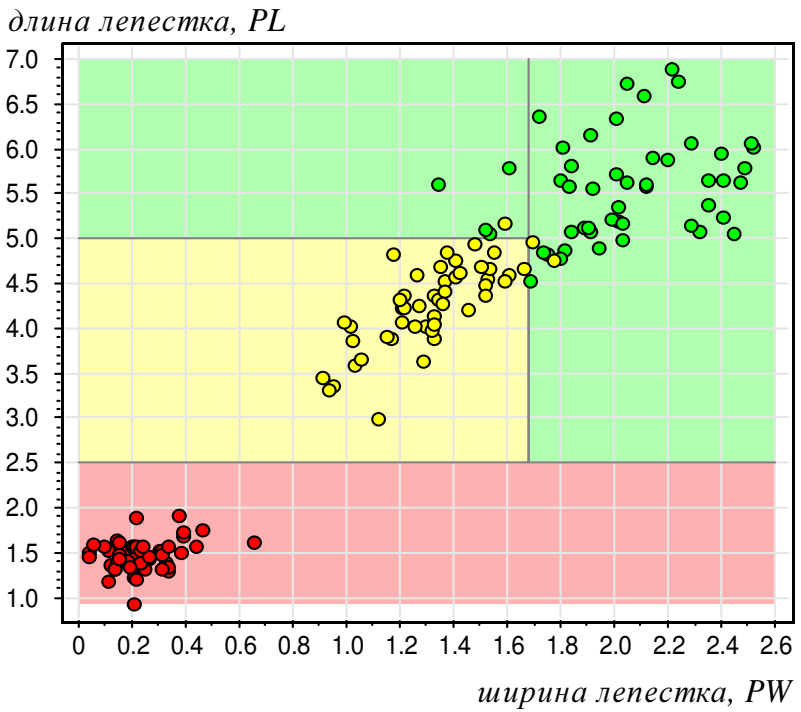
Задача Фишера о классификации цветков ириса на 3 класса, в выборке по 50 объектов каждого класса, 4 признака.

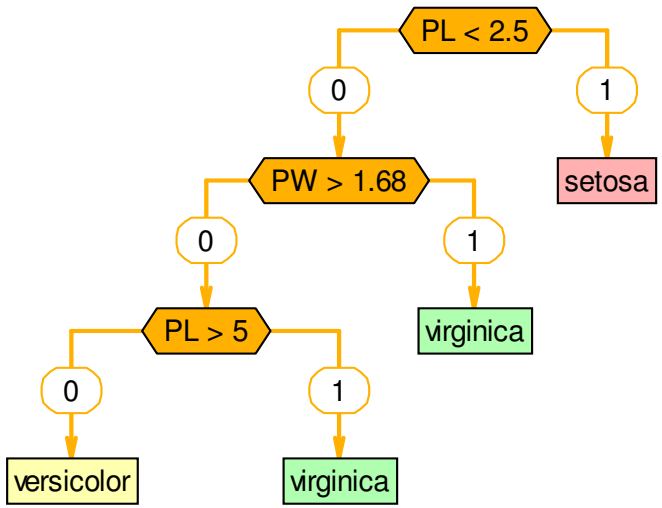




На графике: в осях двух самых информативных признаков (из 4) два класса разделились без ошибок, на третьем 3 ошибки.

Решающее дерево → покрывающий набор конъюнкций





setosa	$r_1(x) = [PL \leqslant 2.5]$
virginica	$r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$
virginica	$r_3(x) = [PL > 5] \wedge [PW \leqslant 1.68]$
versicolor	$r_4(x) = [PL > 2.5] \wedge [PL \leqslant 5] \wedge [PW < 1.68]$

Обучение решающего дерева: стратегия «разделяй и властвуй»

$v_0 := \text{TreeGrowing}\ (X^\ell)$;

ФУНКЦИЯ TreeGrowing $(U \subseteq X^\ell) \mapsto$ корень дерева v;
если StopCriterion (U) **то**
  вернуть новый лист v, взяв $y_v := \text{Major}\,(U)$;
найти признак, наиболее выгодный для ветвления дерева:
$f_v := \arg\max\limits_{f \in F} \text{Gain}\,(f, U)$;
если $\text{Gain}\,(f_v, U) < G_0$ **то**
  вернуть новый лист v, взяв $y_v := \text{Major}\,(U)$;
создать новую внутреннюю вершину v с функцией f_v;
$U_k := \{x \in U\colon f_v(x) = k\},\ \ k \in D_v$;
$S_v(k) := \text{TreeGrowing}\ (U_k),\ \ k \in D_v$;
вернуть v;

Мажоритарное правило: $\text{Major}\,(U) := \arg\max\limits_{y \in Y} P(y|U)$.

Вывод критерия ветвления

Оценка распределения вероятности классов в вершине $v \in V_{\text{внутр}}$:

$$P(y|x,U) = P(y|U) = \frac{1}{|U|}\sum_{x_i \in U}[y_i = y]$$

Принцип максимума (прологарифмированного) правдоподобия:

$$\sum_{x_i \in U} \log P(x_i, y_i) = \sum_{x_i \in U} \log P(y_i|x_i,U)\cancel{P(x_i)} \to \max$$

Обобщение: $\mathscr{L}(P)$ — функция потерь ($-\log P$, $-P$, $-P^2$ и т.п.):

$$\Phi(U) = \frac{1}{|U|}\sum_{x_i \in U}\mathscr{L}\big(P(y_i|x_i)\big) = \sum_{y \in Y}\underbrace{\frac{1}{|U|}\sum_{x_i \in U}[y_i = y]}_{P(y|U)}\mathscr{L}\big(P(y|U)\big) =$$

$$= \sum_{y \in Y} P(y|U)\mathscr{L}\big(P(y|U)\big) = \mathsf{E}_y\mathscr{L}\big(P(y|U)\big) \to \min$$

$\Phi(U)$ — *мера нечистоты* (impurity) распределения $P(y|U)$.

Вывод критерия ветвления

Оценка распределения вероятности классов при ветвлении вершины v по признаку f и разбиении выборки $U = \bigsqcup_{k \in D_v} U_k$:

$$P(y|x,U) = P(y|U_{f(x)}) = \sum_{k \in D_v} [f(x)=k]\, P(y|U_k)$$

Выигрыш от ветвления вершины v:

$$\text{Gain}\,(f,U) = \frac{1}{|U|} \sum_{x_i \in U} \Big(\mathscr{L}\big(P(y_i|U)\big) - \mathscr{L}\big(P(y_i|U_{f(x_i)})\big) \Big) =$$

... вводим суммирования $\sum_y [y_i = y]$ и $\sum_k [f(x_i)=k]$ ...

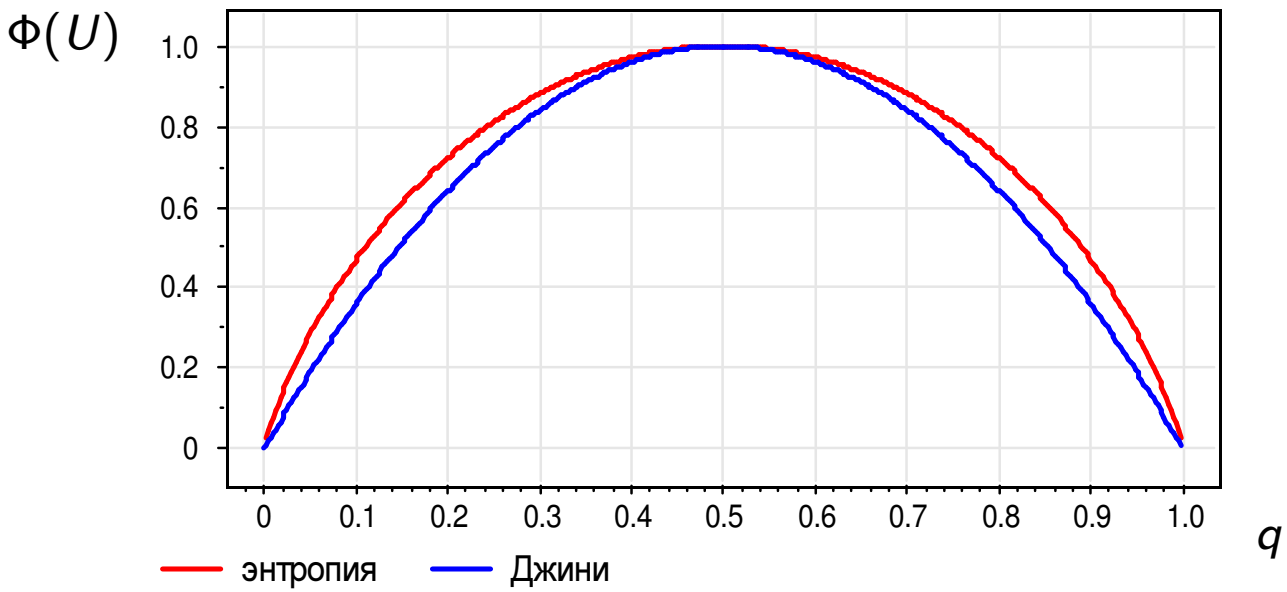
$$= \mathsf{E}_y \mathscr{L}\big(P(y|U)\big) - \sum_{k \in D_v} \frac{|U_k|}{|U|} \mathsf{E}_y \mathscr{L}\big(P(y|U_k)\big) =$$

$$= \Phi(U) - \sum_{k \in D_v} \frac{|U_k|}{|U|} \Phi(U_k) \to \max_{f \in F}$$

Критерий Джини и энтропийный критерий

Два класса, $Y = \{0, 1\}$, $P(y|U) = \begin{cases} q, & y=1 \\ 1-q, & y=0 \end{cases}$

- Если $\mathscr{L}(P) = -\log_2 P$, то
  $\Phi(U) = -q \log_2 q - (1-q)\log_2(1-q)$ — *энтропия* выборки.
- Если $\mathscr{L}(P) = 2(1-P)$, то
  $\Phi(U) = 4q(1-q)$ — Gini impurity.



Обработка пропусков

На стадии обучения:

- $f_v(x)$ не определено $\Rightarrow$ x_i исключается из U для $\text{Gain}\,(f_v, U)$
- $q_{vk} = \frac{|U_k|}{|U|}$ — оценка вероятности k-й ветви, $v \in V_{\text{внутр}}$
- $P(y|x,v) = \frac{1}{|U|} \sum\limits_{x_i \in U} [y_i = y]$ для всех $v \in V_{\text{лист}}$

На стадии классификации:

- $f_v(x)$ определено $\Rightarrow$ из дочерней $s = S_v(f_v(x))$ взять $P(y|x,v) = P(y|x,s)$.
- $f_v(x)$ не определено $\Rightarrow$ *пропорциональное распределение*: $P(y|x,v) = \sum\limits_{k \in D_v} q_{vk} P(y|x, S_v(k))$.
- Окончательное решение — наиболее вероятный класс: $a(x) = \arg\max\limits_{y \in Y} P(y|x, v_0)$.

Жадная нисходящая стратегия: достоинства и недостатки

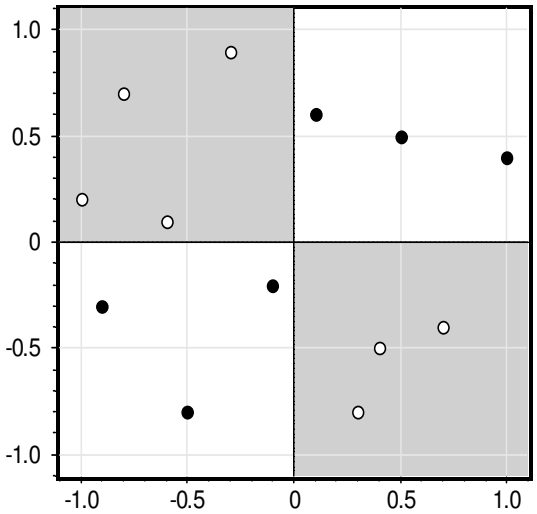
Достоинства:

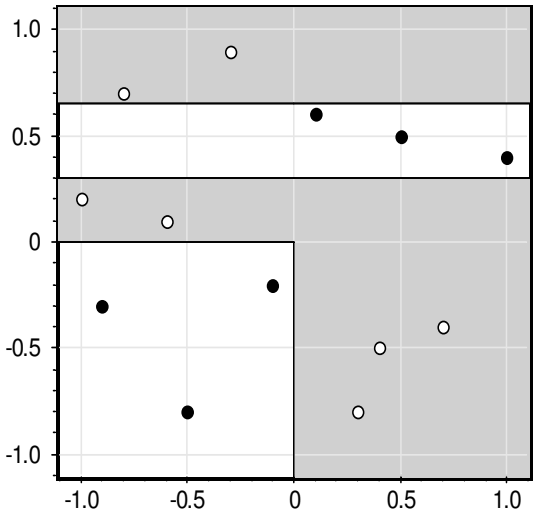
- Интерпретируемость и простота классификации.
- Гибкость: можно варьировать множество F.
- Допустимы разнотипные данные и данные с пропусками.
- Трудоёмкость линейна по длине выборки $O(|F|h\ell)$.
- Не бывает отказов от классификации.

Недостатки:

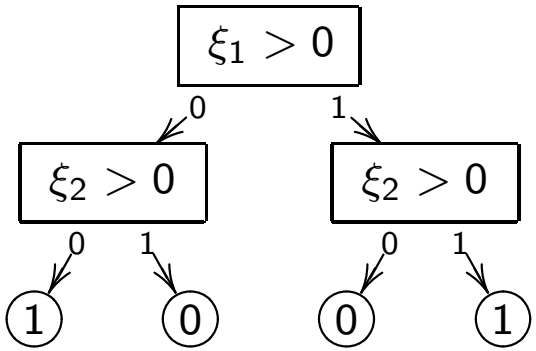
- Жадная стретегия переусложняет структуру дерева, и, как следствие, сильно переобучается.
- Фрагментация выборки: чем дальше v от корня, тем меньше статистическая надёжность выбора f_v, y_v.
- Высокая чувствительность к шуму, к составу выборки, к критерию информативности.

Жадная стратегия переусложняет структуру дерева

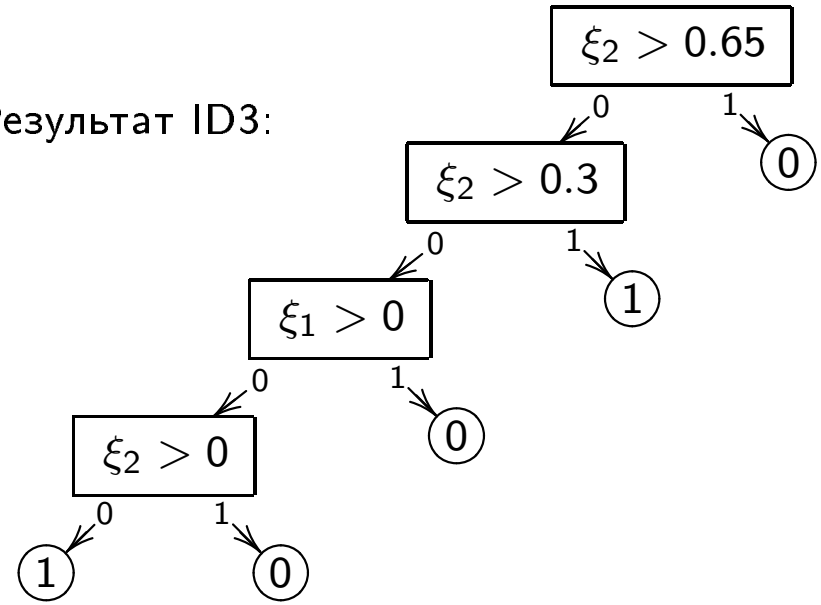




Оптимальное дерево для задачи XOR:



Результат ID3:



Усечение дерева (pruning)

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

1: **для всех** $v \in V_{\text{внутр}}$
2: $X_v^q :=$ подмножество объектов X^q, дошедших до v;
3: **если** $X_v^q = \varnothing$ **то**
4: **вернуть** новый лист v, $y_v := \text{Major}\,(U)$;
5: число ошибок при классификации X_v^q разными способами:
$\text{Err}(v)$ — поддеревом, растущим из вершины v;
$\text{Err}_k(v)$ — дочерним поддеревом $S_v(k)$, $k \in D_v$;
$\text{Err}_c(v)$ — классом $c \in Y$.
6: в зависимости от того, какое из них минимально:
сохранить поддерево v;
заменить поддерево v дочерним $S_v(k)$;
заменить поддерево v листом, $y_v := \arg\min\limits_{c \in Y} \text{Err}_c(v)$.

CART: деревья регрессии и классификации

Обобщение на случай регрессии: $Y = \mathbb{R}$, $y_v \in \mathbb{R}$.

U — множество объектов x_i, дошедших до вершины v

Критерий информативности — среднеквадратичная ошибка

$$\Phi(U) = \min_{y \in Y} \sum_{x_i \in U} (y - y_i)^2$$

Значения в терминальных вершинах — МНК-решение:

$$y_v = \frac{1}{|U|} \sum_{x_i \in U} y_i$$

Дерево регрессии $a(x)$ — это кусочно-постоянная функция.

CART: критерий Minimal Cost-Complexity Pruning

Среднеквадратичная ошибка со штрафом за сложность дерева

$$C_\alpha = \sum_{i=1}^{\ell} \big(a(x_i) - y_i\big)^2 + \alpha |V_{\text{лист}}| \to \min$$

При увеличении α дерево последовательно упрощается.
Причём последовательность вложенных деревьев единственна.

Из этой последовательности выбирается дерево с минимальной ошибкой на тестовой выборке (Hold-Out).

Для случая классификации используется аналогичная стратегия усечения, с критерием Джини.

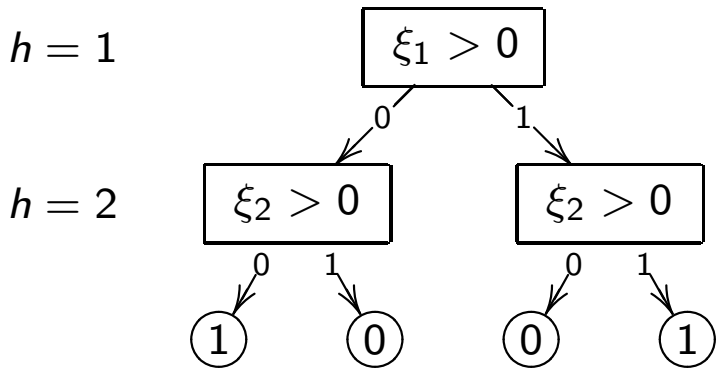
Небрежные решающие деревья (Oblivious Decision Tree, ODT)

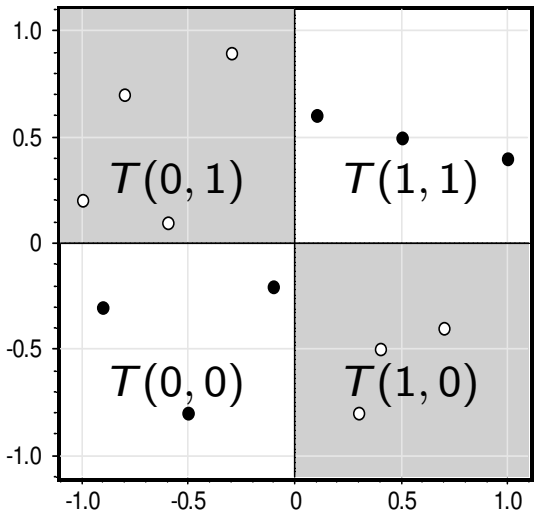
Решение проблемы фрагментации:
строится сбалансированное дерево глубины H, $D_v = \{0,1\}$;
для всех узлов уровня h условие ветвления $f_h(x)$ *одинаково*;
на уровне h ровно 2^{h-1} вершин; X делится на 2^H ячеек.

Классификатор задаётся *таблицей решений* $T\colon \{0,1\}^H \to Y$:

$$a(x) = T\big(f_1(x), \ldots, f_H(x)\big).$$

Пример: задача XOR, $H = 2$.





Алгоритм обучения ODT

Вход: выборка X^ℓ; множество признаков F; глубина дерева H;
Выход: признаки f_h, $h = 1, \ldots, H$; таблица $T\colon \{0,1\}^H \to Y$;

1: **для всех** $h = 1, \ldots, H$
2: найти предикат с максимальной информативностью:
$f_h := \arg\max\limits_{f \in F} \text{Gain}\,(f_1, \ldots, f_{h-1}, f)$;
3: классификация по мажоритарному правилу:
$T(\beta) := \text{Major}\,(U_{H\beta})$;

Выигрыш от ветвления на уровне h по всей выборке X^ℓ:

$$\text{Gain}\,(f_1, \ldots, f_h) = \Phi(X^\ell) - \sum_{\beta \in \{0,1\}^h} \frac{|U_{h\beta}|}{\ell}\, \Phi(U_{h\beta}),$$

$$U_{h\beta} = \left\{x_i \in X^\ell\colon f_s(x_i) = \beta_s,\ s = 1..h\right\},\quad \beta = (\beta_1, \ldots, \beta_h) \in \{0,1\}^h.$$

Случайный лес (Random Forest)

Голосование деревьев классификации, $Y = \{-1, +1\}$:

$$a(t) = \text{sign}\,\frac{1}{T}\sum_{t=1}^{T} b_t(x).$$

Голосование деревьев регрессии, $Y = \mathbb{R}$:

$$a(t) = \frac{1}{T}\sum_{t=1}^{T} b_t(x).$$

- каждое дерево $b_t(x)$ обучается по случайной выборке с повторениями
- в каждой вершине признак выбирается из случайного подмножества $\sqrt{n}$ признаков
- признаки и пороги выбираются по критерию Джини
- усечений (pruning) нет

Проблема оценивания информативности

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

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

Очевидные, но не всегда адекватные свёртки:

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

$P_c = \#\{x_i : y_i{=}c\}$ — число «своих» во всей выборке;
$N_c = \#\{x_i : y_i{\neq}c\}$ — число «чужих» во всей выборке.

Нетривиальность проблемы свёртки двух критериев

Пример:

при $P = 200$, $N = 100$ и различных p и n.

Простые эвристики не всегда адекватны:

p	n	$p-n$	$p-5n$	$\frac{p}{P}-\frac{n}{N}$	$\frac{p}{n+1}$	IStat$\cdot\ell$	IGain$\cdot\ell$	$\sqrt{p}-\sqrt{n}$
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

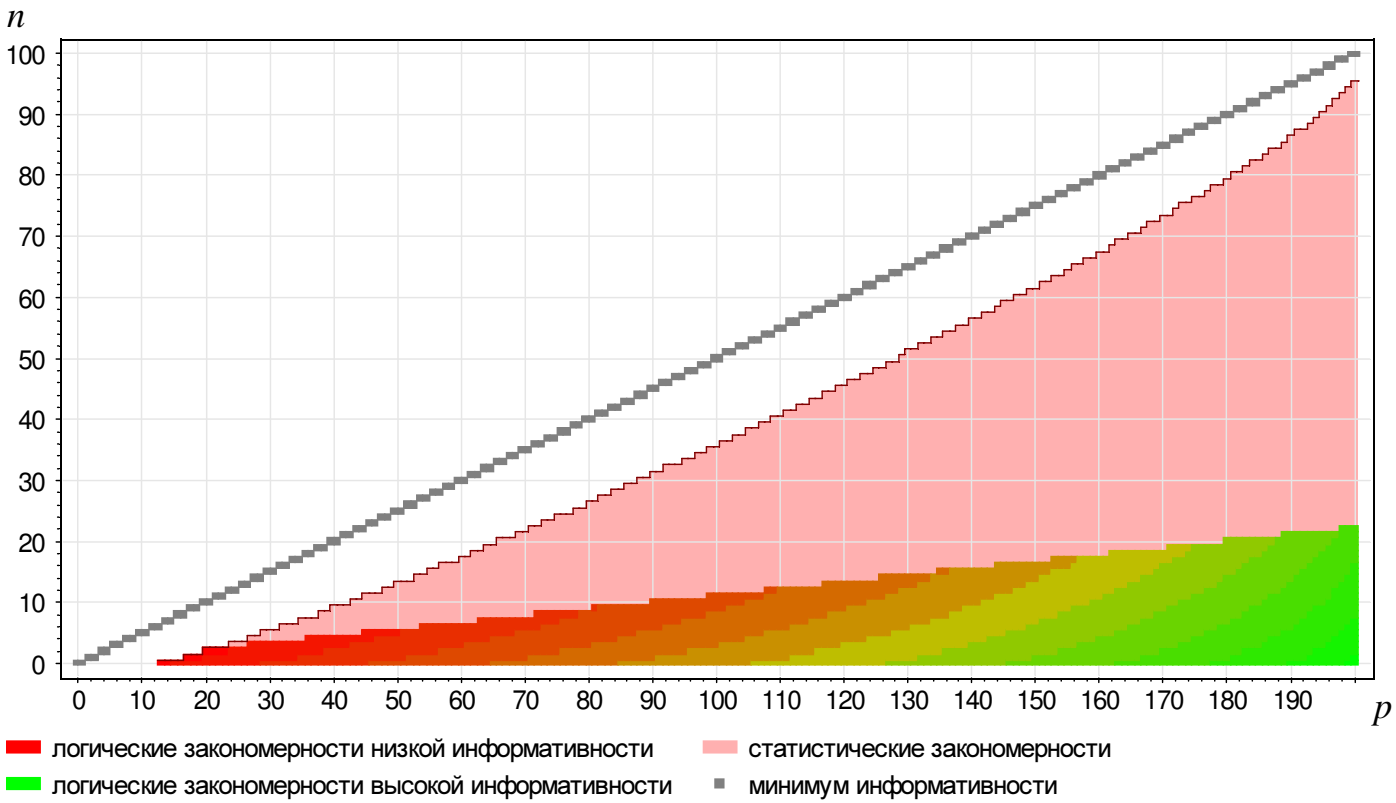
Часто используемые критерии информативности

Адекватные, но неочевидные критерии:

- энтропийный критерий прироста информации:
  $$\mathrm{IGain}(p,n) = h\left(\frac{P}{\ell}\right) - \frac{p+n}{\ell} h\left(\frac{p}{p+n}\right) - \frac{\ell-p-n}{\ell} h\left(\frac{P-p}{\ell-p-n}\right) \to \max,$$
  где $h(q) = -q\log_2 q - (1-q)\log_2(1-q)$
- критерий Джини (Gini impurity):
  $\mathrm{IGini}(p,n) = \mathrm{IGain}(p,n)$ при $h(q) = 4q(1-q)$
- точный статистический тест Фишера (Fisher's Exact Test):
  $$\mathrm{IStat}(p,n) = -\frac{1}{\ell}\log_2 \frac{C_P^p C_N^n}{C_{P+N}^{p+n}} \to \max$$
- критерий бустинга:
  $\sqrt{p} - \sqrt{n} \to \max$
- нормированный критерий бустинга:
  $\sqrt{p/P} - \sqrt{n/N} \to \max$

Где находятся закономерности в (p, n)-плоскости

Логические закономерности: $\frac{n}{p+n} \leqslant 0.1, \quad \frac{p}{P+N} \geqslant 0.05.$
Статистические закономерности: $\mathrm{IStat}(p, n) \geqslant 3.$



$P = 200$
$N = 100$

Вывод: неслучайность — ещё не значит закономерность.

Энтропийный критерий (вывод из теории информации)

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 $R(x_i)_{i=1}^{\ell}$:

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

Прирост информации (Information gain, IGain):

$$\mathrm{IGain}(p,n) = H(y) - H(y|R).$$

Статистический критерий информативности

Точный тест Фишера. Пусть X — в.п., выборка X^ℓ — i.i.d.
Гипотеза H_0: $y(x)$ и $R(x)$ — 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 $R(x)$ относительно класса $c \in Y$:

$$\mathrm{IStat}(p,n) = -\frac{1}{\ell} \log_2 \frac{C_P^p C_N^n}{C_{P+N}^{p+n}},$$

$\mathrm{IStat}(p,n) \geqslant I_0$ — *статистическая закономерность класса c.*

Соотношение статистического и энтропийного критериев

Определение

Предикат R — закономерность по энтропийному критерию, если $\mathrm{IGain}(p, n) > G_0$ *при некотором* G_0.

Теорема

Энтропийный критерий IGain *асимптотически эквивалентен статистическому* IStat*:*

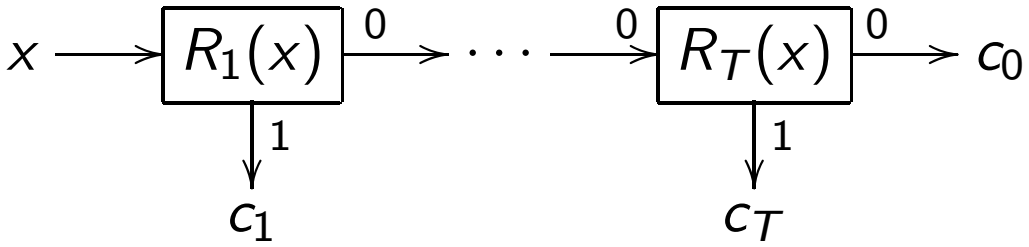
$$\mathrm{IStat}(p, n) \to \mathrm{IGain}(p, n) \quad \textit{при } \ell \to \infty.$$

Доказательство:

применить формулу Стирлинга к критерию IStat.

Определение решающего списка

Решающий список (Decision List, DL)
— алгоритм классификации $a\colon X \to Y$, который задаётся закономерностями $R_1(x), \ldots, R_T(x)$ классов $c_1, \ldots, c_T \in Y$:



1: **для всех** $t = 1, \ldots, T$
2: **если** $R_t(x) = 1$ **то**
3: **вернуть** c_t;
4: **вернуть** c_0 — отказ от классификации объекта x.

$$E(R_t, X^\ell) = \frac{n(R_t)}{n(R_t) + p(R_t)} \to \min \quad \text{— доля ошибок } R_t \text{ на } X^\ell$$

Жадный алгоритм построения решающего списка

Вход: выборка X^ℓ; семейство предикатов $\mathscr{B}$;
параметры: $T_{\max}$, $I_{\min}$, $E_{\max}$, ℓ_0;
Выход: решающий список $\{R_t, c_t\}_{t=1}^{T}$;

1: $U := X^\ell$;
2: **для всех** $t := 1, \ldots, T_{\max}$
3: выбрать класс c_t;
4: максимизация информативности $I(R, U)$ при ограничении на число ошибок $E(R, U)$:
$R_t := \underset{R \in \mathscr{B}:\ E(R,U) \leqslant E_{\max}}{\arg\max} I(R, U)$;
5: **если** $I(R_t, U) < I_{\min}$ **то выход**;
6: оставить объекты, не покрытые правилом R_t:
$U := \{x \in U : R_t(x) = 0\}$;
7: **если** $|U| \leqslant \ell_0$ **то выход**;

Замечания к алгоритму построения решающего списка

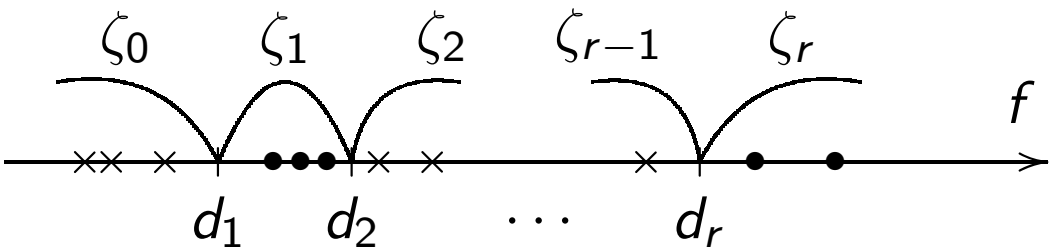
- Параметр $E_{\max}$ управляет сложностью списка:
  $E_{\max} \downarrow \quad \Rightarrow \quad p(R_t) \downarrow,\ T \uparrow.$
- Стратегии выбора класса c_t:
  1) все классы по очереди;
  2) на каждом шаге определяется оптимальный класс.
- Простой обход проблемы пропусков в данных.
- Другие названия:
  комитет с логикой старшинства (Majority Committee)
  голосование по старшинству (Majority Voting)
  машина покрывающих множеств (Set Covering Machine, SCM)
- **Недостаток:** низкое качество классификации

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:

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



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 $\arg\min_x \left|\frac{1}{2}(a+b) - x\right|$.

Алгоритм разбиения области значений признака на зоны

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

$D := \varnothing$; упорядочить выборку X^ℓ по возрастанию $f(x_i)$;
для всех $i = 2, \ldots, \ell$
  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 $d_i \in D$, $i = 1, \ldots, |D| - 1$
    $\delta I_i := I(\zeta_{i-1} \vee \zeta_i \vee \zeta_{i+1}) - \max\{I(\zeta_{i-1}), I(\zeta_i), I(\zeta_{i+1})\}$;
  $i := \arg\max_s \delta I_s$;
  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,
  - композиции (леса) деревьев.

Yandex MatrixNet = голосование (градиентный бустинг) над ODT.