

Логические алгоритмы классификации

Воронцов Константин Вячеславович
vokov@forecsys.ru

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>

27 февраля 2014

Содержание

1 Понятия закономерности и информативности

- Понятие закономерности
- Интерпретируемость
- Информативность

2 Решающие деревья

- Алгоритм ID3
- Небрежные решающие деревья — ODT
- Бинаризация данных

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 $R: X \rightarrow \{0, 1\}$, удовлетворяющий двум требованиям:

① *интерпретируемость*:

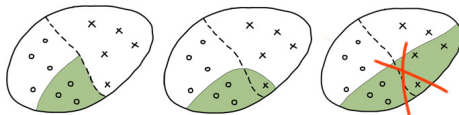
- 1) R записывается на естественном языке;
- 2) R зависит от небольшого числа признаков (1–7);

② *информативность* относительно одного из классов $c \in Y$:

$$p(R) = \#\{x_i: R(x_i)=1 \text{ и } y_i=c\} \rightarrow \max;$$

$$n(R) = \#\{x_i: R(x_i)=1 \text{ и } y_i \neq c\} \rightarrow \min;$$

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);

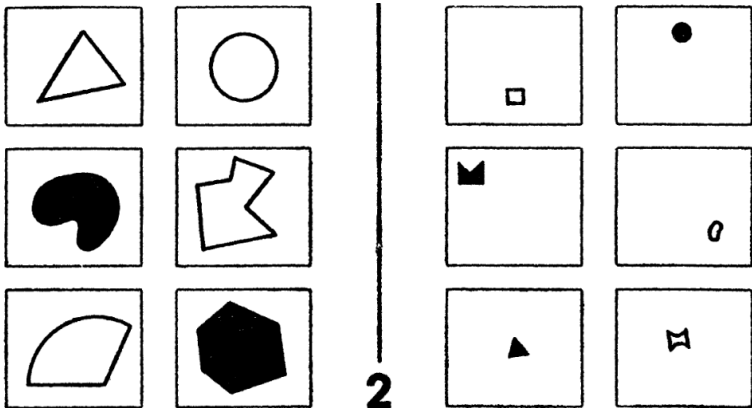
Пример (из области медицины)

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

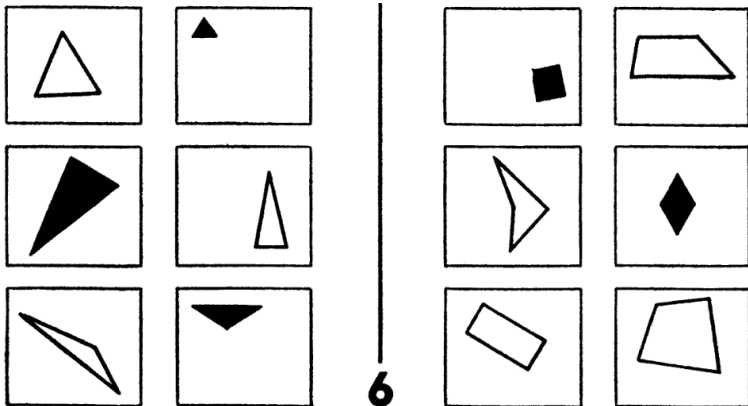
Пример (из области кредитного скоринга)

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

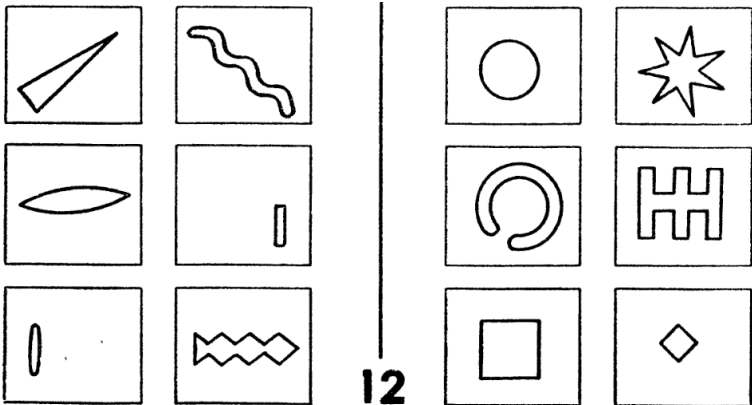
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



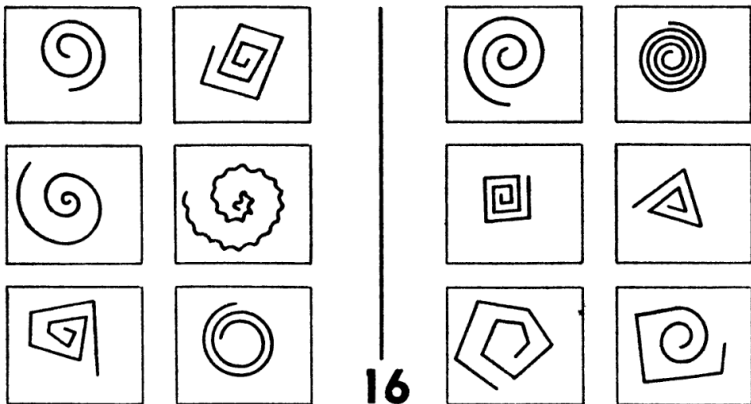
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



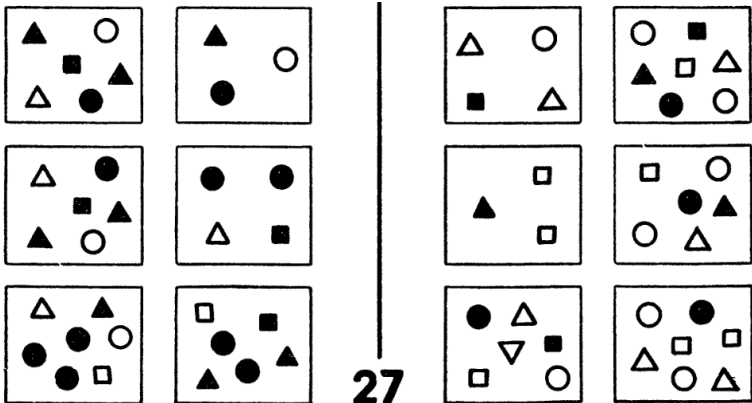
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



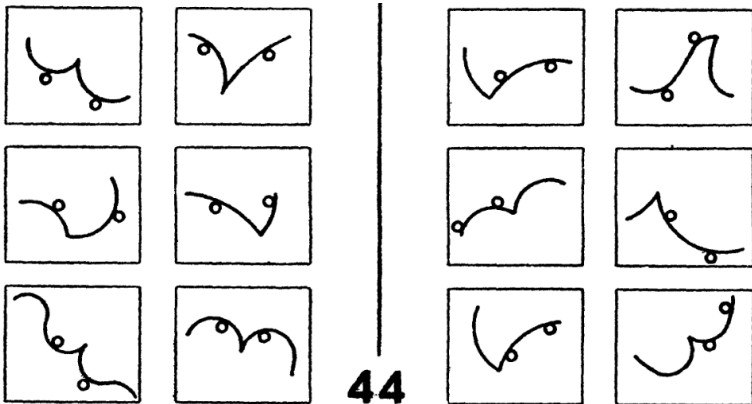
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



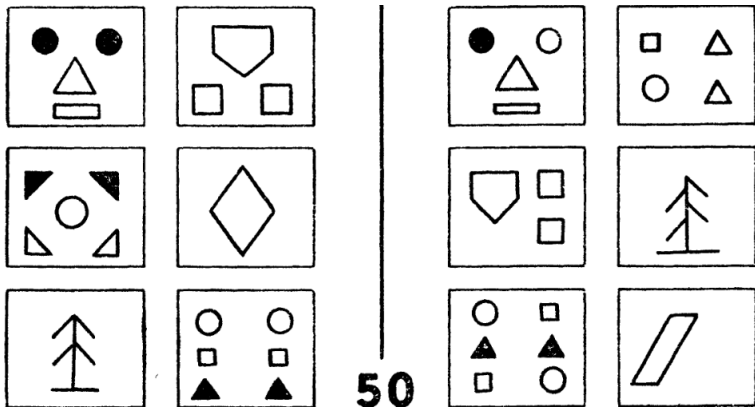
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



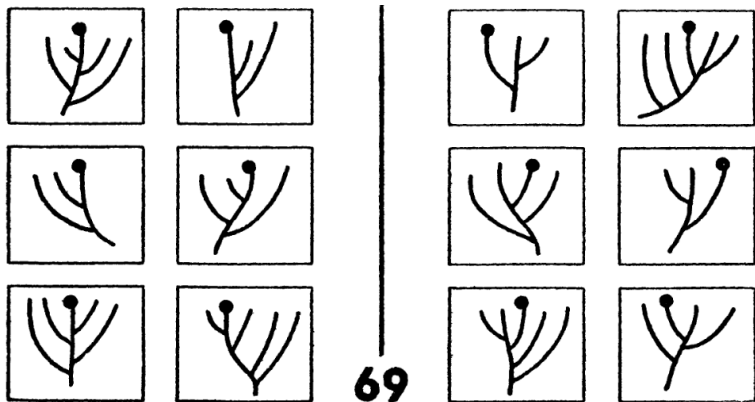
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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



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



Основные вопросы построения логических алгоритмов

- 1 Как изобретать признаки $f_1(x), \dots, f_n(x)$?
— 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 $p_c \rightarrow \max$, $n_c \rightarrow \min$
- 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. *Конъюнкция* пороговых условий (термов):

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

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

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

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:

$$R(x) = \left[\sum_{j \in J} w_j f_j(x) \geq w_0 \right].$$

4. *Шар* — пороговая функция близости:

$$R(x) = [r(x, x_0) \leq w_0],$$

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

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

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

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

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

Часто используемые критерии информативности

Проблема: надо сравнивать закономерности R .

Как свернуть два критерия в один критерий информативности?

$$\begin{cases} p(R) \rightarrow \max \\ n(R) \rightarrow \min \end{cases} \xRightarrow{?} I(p, n) \rightarrow \max$$

Очевидные, но не всегда адекватные свёртки:

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

$P = \#\{x_i: y_i=c\}$ — число «своих» во всей выборке;

$N = \#\{x_i: y_i \neq c\}$ — 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	$I\text{Gain}_c$	$\sqrt{p} - \sqrt{n}$
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:

$$h\left(\frac{P}{\ell}\right) - \frac{p+n}{\ell} h\left(\frac{p}{p+n}\right) - \frac{\ell-p-n}{\ell} h\left(\frac{P-p}{\ell-p-n}\right) \rightarrow \max,$$

где $h(q) = -q \log_2 q - (1-q) \log_2 (1-q)$;

- критерий неслучайности FET (Fisher's Exact Test):

$$C_P^p C_N^n / C_{P+N}^{p+n} \rightarrow \min;$$

- критерий бустинга [Cohen, Singer, 1999]:

$$\sqrt{p} - \sqrt{n} \rightarrow \max$$

- нормированный критерий бустинга:

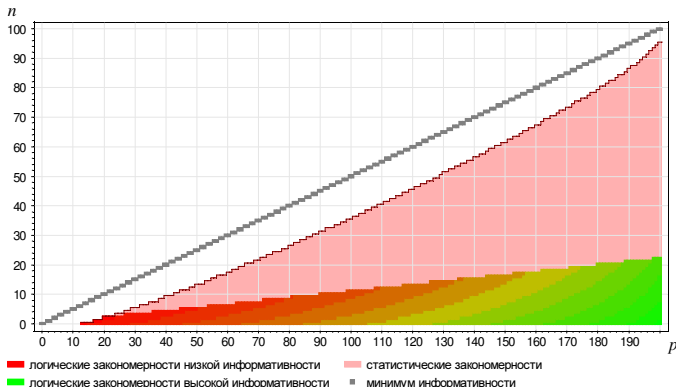
$$\sqrt{p/P} - \sqrt{n/N} \rightarrow \max;$$

- перестановочные тесты.

Где находятся закономерности в (p, n) -плоскости

Логические закономерности: $\frac{n}{p+n} \leq 0.1$, $\frac{p}{P+N} \geq 0.05$.

Статистические закономерности: $C_P^p C_N^n / C_{P+N}^{p+n} \leq 0.05$.



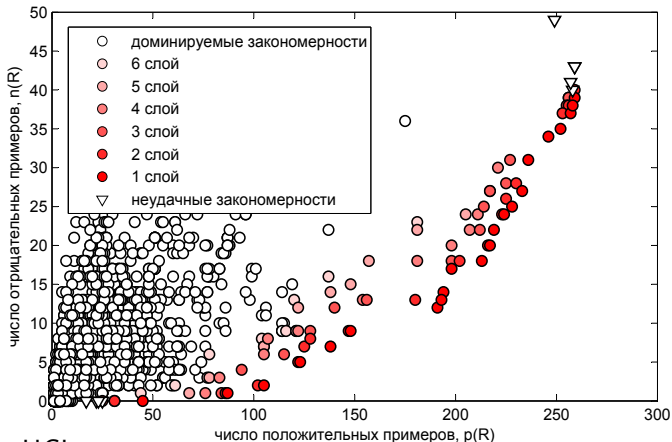
$P = 200$

$N = 100$

Вывод: неслучайность — ещё не значит закономерность.

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

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



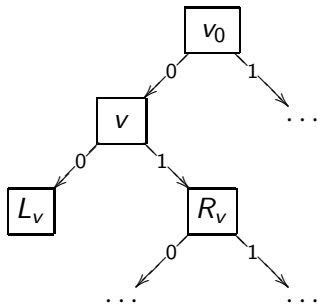
задача UCI:german

Определение бинарного решающего дерева

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

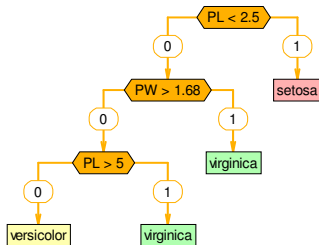
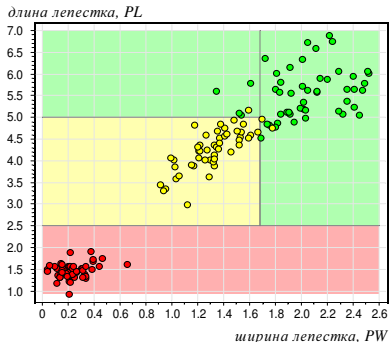
- 1) $\forall v \in V_{\text{внутр}} \rightarrow$ предикат $\beta_v : X \rightarrow \{0, 1\}$, $\beta \in \mathcal{B}$
- 2) $\forall v \in V_{\text{лист}} \rightarrow$ имя класса $c_v \in Y$.

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



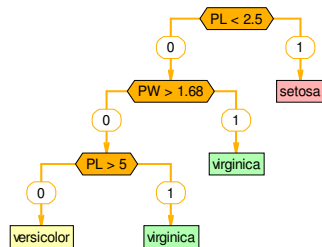
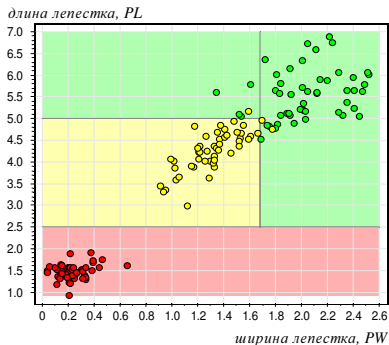
Пример решающего дерева

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



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

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:

$$\beta := \arg \max_{\beta \in \mathcal{B}} I(\beta, U);$$
- 5: разбить выборку на две части $U = U_0 \cup U_1$ по предикату β :

$$U_0 := \{x \in U : \beta(x) = 0\};$$

$$U_1 := \{x \in U : \beta(x) = 1\};$$
- 6: **если** $U_0 = \emptyset$ или $U_1 = \emptyset$ **то**
- 7: **вернуть** новый лист v , $c_v := \text{Мажоритарный класс}(U)$;
- 8: создать новую внутреннюю вершину v : $\beta_v := \beta$;
 построить левое поддерево: $L_v := \text{LearnID3}(U_0)$;
 построить правое поддерево: $R_v := \text{LearnID3}(U_1)$;
- 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 $P_c = \#\{x_i: y_i = c\}$, $p = \#\{x_i: \beta(x_i) = 1\}$, $h(z) \equiv -z \log_2 z$.

3. Критерий Джини:

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

4. D -критерий В.И.Донского:

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

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:

- $\hat{P}_v(y|x) = \beta_v(x)\hat{P}_{L_v}(y|x) + (1-\beta_v(x))\hat{P}_{R_v}(y|x)$;
- $\beta_v(x)$ 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

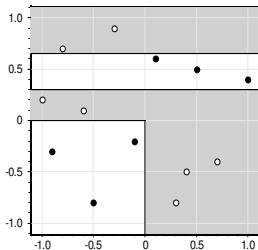
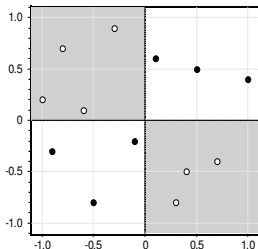
Достоинства:

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

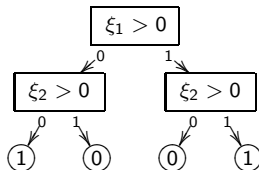
Недостатки:

- 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.

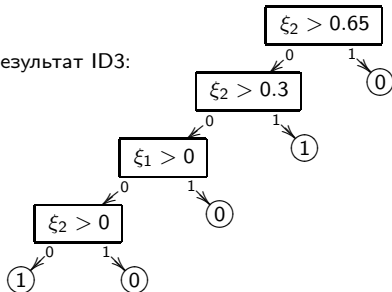
Жадный ID3 переусложняет структуру дерева



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** $S_v = \emptyset$ **то**
- 4: **вернуть** новый лист v , $c_v := \text{Мажоритарный класс}(U)$;
- 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, $c_v := \arg \min_{c \in Y} r_c(v)$.

Небрежные решающие деревья — ODT (Oblivious Decision Tree) [1993]

Решение проблемы фрагментации:

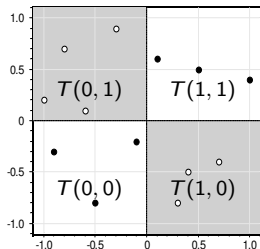
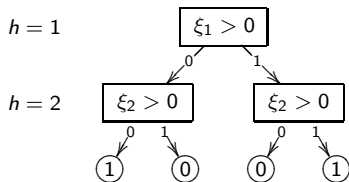
строится сбалансированное дерево высоты H ;

для всех узлов уровня h условие ветвления $\beta_h(x)$ одинаково;
на уровне h ровно 2^{h-1} вершин; X делится на 2^H ячеек.

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

$$a(x) = T(\beta_1(x), \dots, \beta_H(x)).$$

Пример: задача XOR, $H = 2$.



Алгоритм обучения ODT

Вход: выборка X^ℓ ; семейство правил \mathcal{B} ; глубина дерева H ;

Выход: условия β_h , $h = 1, \dots, H$; таблица $T: \{0, 1\}^H \rightarrow Y$;

1: **для всех** $h = 1, \dots, H$

2: найти предикат с максимальной информативностью:

$$\beta_h := \arg \max_{\beta \in \mathcal{B}} I(\beta_1, \dots, \beta_{h-1}, \beta; X^\ell);$$

3: **для всех** $b \equiv (b_1, \dots, b_H) \in \{0, 1\}^H$

4: классификация по мажоритарному правилу:

$$T(b_1, \dots, b_H) := \arg \max_{c \in Y} \sum_{i=1}^{\ell} [y_i = c] \prod_{h=1}^H [\beta_h(x_i) = b_h];$$

$$I(\beta_1, \dots, \beta_h) = \sum_{c \in Y} h \left(\frac{P_c}{\ell} \right) - \sum_{b \in \{0,1\}^h} \frac{|X_b|}{\ell} \sum_{c \in Y} h \left(\frac{|X_b \cap X_c|}{|X_b|} \right);$$

$$X_b = \{x_i: \beta_s(x_i) = b_s, s = 1, \dots, h\}, \quad X^\ell = \bigsqcup_{b \in \{0,1\}^h} X_b.$$

Вспомогательная задача бинаризации вещественного признака

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

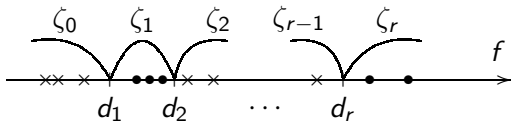
Дано: выборка значений вещественного признака $f(x_i)$, $x_i \in X^\ell$.

Найти: наилучшее (в каком-то смысле) разбиение области значений признака на относительно небольшое число зон:

$$\zeta_0(x) = [f(x) < d_1];$$

$$\zeta_s(x) = [d_s \leq f(x) < d_{s+1}], \quad s = 1, \dots, r-1;$$

$$\zeta_r(x) = [d_r \leq f(x)].$$



Способы разбиения области значений признака на зоны

- 1 Жадная максимизация информативности путём слияний
- 2 Разбиение на равномошные 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: $D = \{d_1 < \dots < d_r\}$ — последовательность порогов;

-
- 1: $D := \emptyset$; упорядочить выборку X^ℓ по возрастанию $f(x_i)$;
 - 2: **для всех** $i = 2, \dots, \ell$
 - 3: **если** $f(x_{i-1}) \neq f(x_i)$ и $[y_{i-1} = c] \neq [y_i = c]$ **то**
 - 4: добавить порог $\frac{1}{2}(f(x_{i-1}) + f(x_i))$ в конец D ;
 - 5: **повторять**
 - 6: **для всех** $d_j \in D, j = 1, \dots, |D| - 1$
 - 7: $\delta l_j := I_c(\zeta_{i-1} \vee \zeta_i \vee \zeta_{i+1}) - \max\{I_c(\zeta_{i-1}), I_c(\zeta_i), I_c(\zeta_{i+1})\}$;
 - 8: $i := \arg \max_s \delta l_s$;
 - 9: **если** $\delta l_i > \delta_0$ **то**
 - 10: слить зоны $\zeta_{i-1}, \zeta_i, \zeta_{i+1}$, удалив d_j и d_{j+1} из D ;
 - 11: **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;

Yandex MatrixNet = градиентный бустинг над ODT.