

• Введение в машинное обучение •
**Логические методы
машинного обучения**

Воронцов Константин Вячеславович

`k.v.vorontsov@phystech.edu`

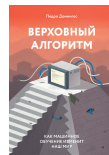
`http://www.MachineLearning.ru/wiki?title=User:Vokov`

Этот курс доступен на странице вики-ресурса

`http://www.MachineLearning.ru/wiki`

«Введение в машинное обучение (курс лекций, К.В.Воронцов)»

- 1 **символизм** – поиск логических закономерностей
 - Decision Tree, Rule Induction
- 2 **коннекционизм** – обучаемые нейронные сети
 - BackPropagation, Deep Belief Nets, Deep Learning
CNN, ResNet, LSTM, GRU, Attention, Transformer
- 3 **эволюционизм** – саморазвитие сложных моделей
 - Genetic Algorithms, Genetic Programming, Symbolic Regression
- 4 **байесионизм и вероятностно-статистические методы**
 - MLE, EM, GLM, LR, OBC, Naive Bayes, QD, LDF
Bayesian Networks, Bayesian Learning, Graphical Models
- 5 **аналогизм** – «близким объектам близкие ответы»
 - kNN, RBF, SVM, KDE, Kernel Smoothing
- ⊕ **композиционизм** – кооперация моделей
 - Weighted Voting, Boosting, Bagging, Stacking,
Random Forest, Яндекс.CatBoost



Объяснимый ИИ (XAI, eXplainable Artificial Intelligence)

- **Interpretability**

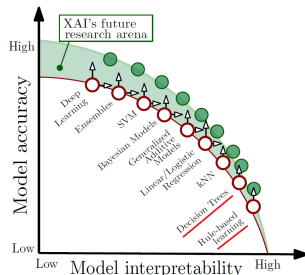
— пассивная интерпретируемость
внутреннего строения модели
или предсказания на объекте

- **Understandability, Transparency**

— понятность, самоочевидность,
прозрачность строения модели

- **Explainability** — активная генерация объяснений
на объекте как дополнительных выходных данных модели

- **Comprehensibility** — возможность представить выученные
закономерности в виде понятного людям знания



“Do you want an interpretable model, or the one that works?”

[Yann LeCun, NIPS'17]

V.Belle, I.Papantonis. Principles and practice of explainable machine learning. 2020

1 Индукция правил

- Научная школа М. М. Бонгарда
- Понятие закономерности
- Алгоритмы поиска закономерностей

2 Критерии информативности

- Двухкритериальный отбор правил в плоскости (p, n)
- Логические и статистические закономерности
- Сравнение критериев информативности

3 Решающие деревья

- Обучение решающих деревьев
- CART: деревья регрессии и классификации
- Преобразование дерева в набор конъюнкций

Первые программы для поиска логических правил

- 1958: Программа «Открой закон» восстанавливала зависимость полным комбинаторным перебором формул
- 1959: Программа «Арифметика» для сокращения перебора использовала оценки *информативности*
- 1961: Программа «КоРа» перебирала *информативные тройки* признаков



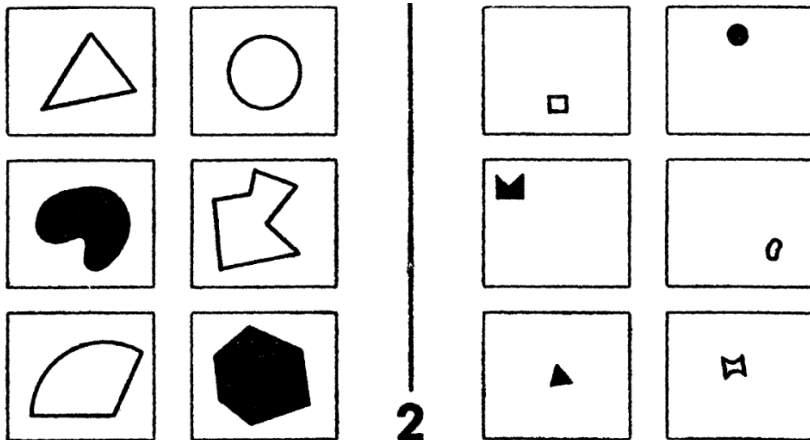
Михаил Моисеевич
Бонгард
(1924–1971)

«КоРа-3»: первое применение распознавания незрительных образов для распознавания границы нефть-вода в скважине. Введены принципы *голосования, скользящего контроля*, понятия *информативности* и *предрассудка* (переобучения).

Бонгард М. М., Вайнцвайг М. Н., Губерман Ш. А. Извекова М. Л., Смирнов М. С. Использование обучающейся программы для выявления нефтеносных пластов. 1966.

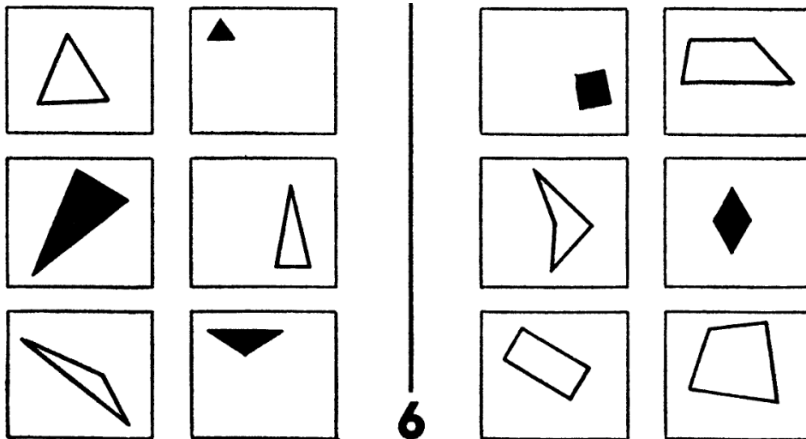
Тесты Бонгарда [Проблема узнавания, 1967]

Обучающая выборка: по 6 объектов каждого из двух классов.
Требуется найти правило классификации.



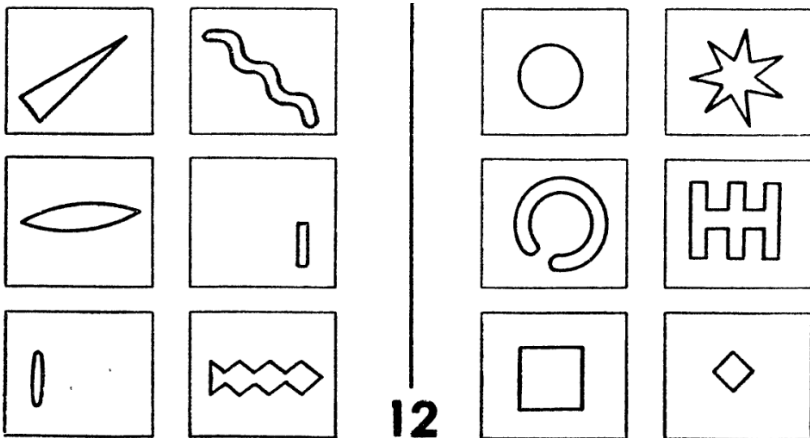
Тесты Бонгарда [Проблема узнавания, 1967]

Что даёт нам уверенность, что мы нашли верное правило?
Безошибочная классификация примеров обучающей выборки.



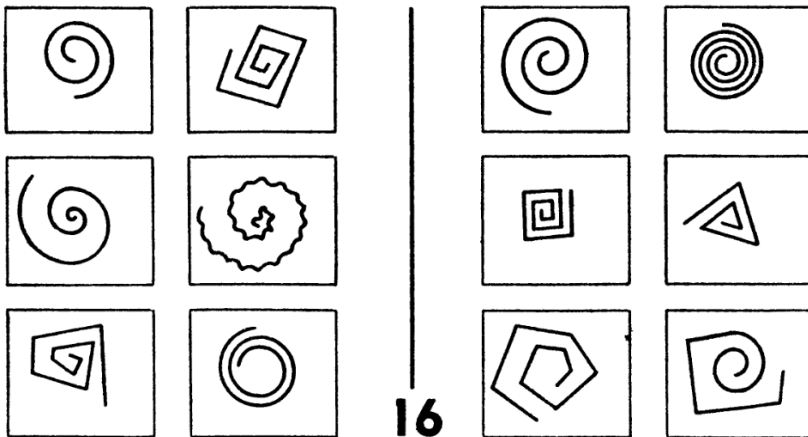
Тесты Бонгарда [Проблема узнавания, 1967]

Что ещё даёт нам уверенность, что мы нашли верное правило?
Простота, общность, изящество найденного правила.



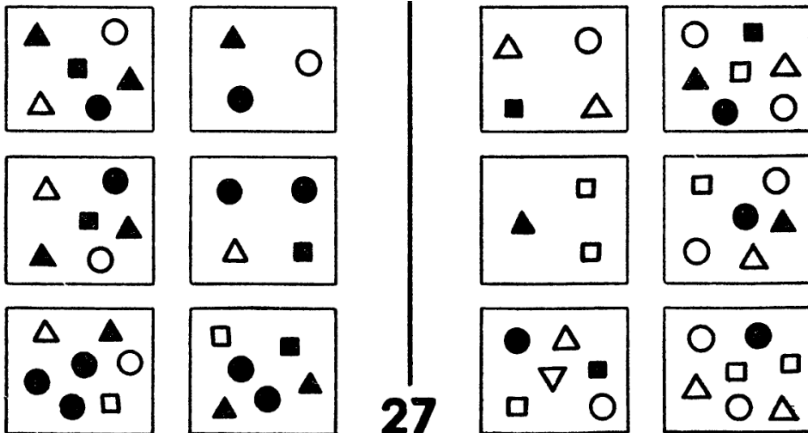
Тесты Бонгарда [Проблема узнавания, 1967]

Мы решаем эти задачи почти мгновенно. Чем мы пользуемся?
Однако для компьютера они сложны. Чего ему не хватает?



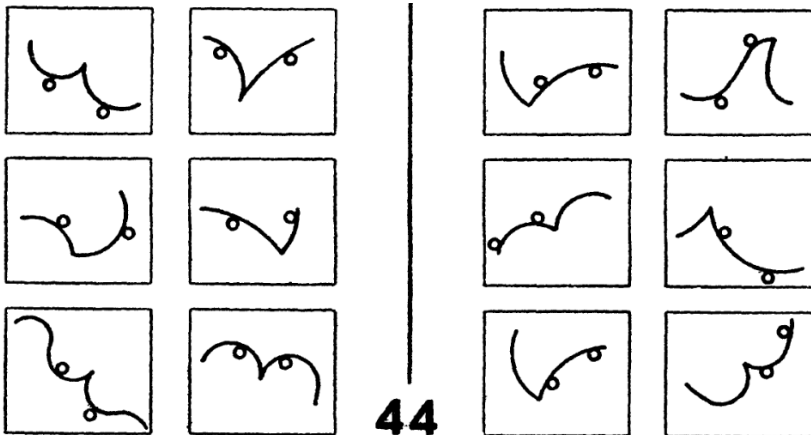
Тесты Бонгарда [Проблема узнавания, 1967]

Нужно ли закладывать знания геометрии в явном виде?
Или возможно выработать необходимые понятия на примерах?



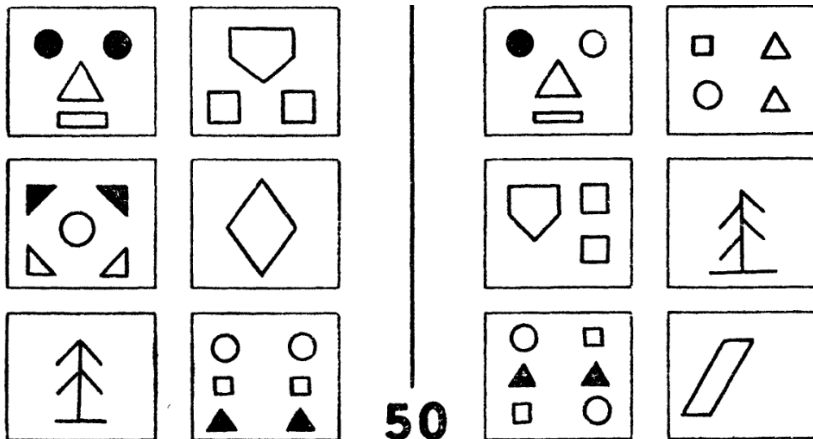
Тесты Бонгарда [Проблема узнавания, 1967]

Как вычислять полезные признаки по «сырым» данным?
Возможно ли поручить перебор признаков и моделей машине?



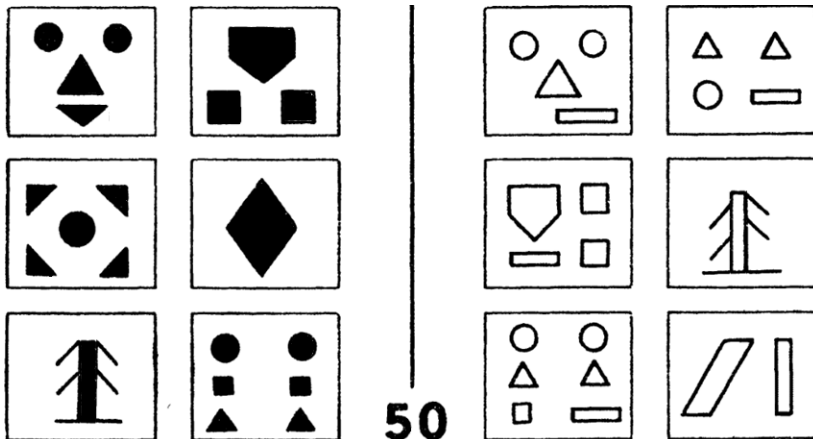
Тесты Бонгарда [Проблема узнавания, 1967]

Каков риск вывести из данных ложное правило, *предрассудок*?
Как этот риск зависит от числа примеров и сложности правил?



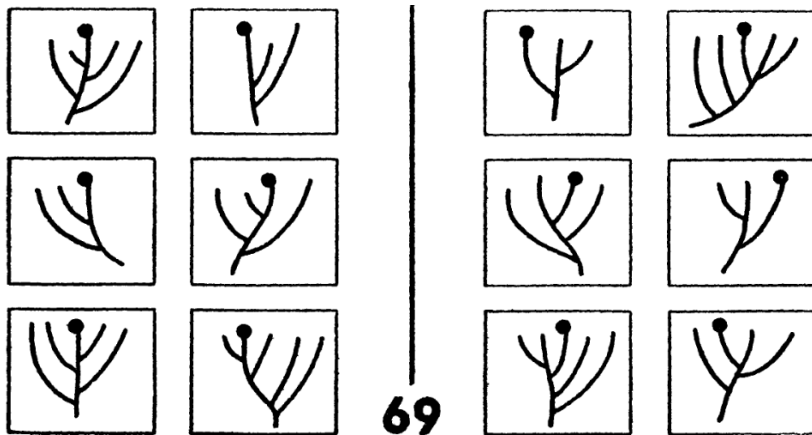
Тесты Бонгарда [Проблема узнавания, 1967]

Какого числа примеров достаточно для выработки правила?
Что делать, если к выборке подходит много разных правил?



Тесты Бонгарда [Проблема узнавания, 1967]

Эти вопросы составляют основу машинного обучения сегодня.
М.М.Бонгард поставил все эти проблемы в середине 60-х!



Логические закономерности в задачах классификации

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

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

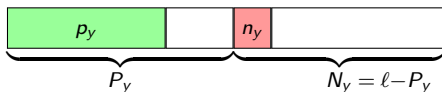
- 1) *интерпретируемость* (понятность + простота):
 - 1) R записывается на естественном языке
 - 2) R зависит от небольшого числа признаков (не более 7)

- 2) *информативность* относительно одного из классов $y \in Y$:

$$p_y(R, X^\ell) = \#\{x_i \in X^\ell : R(x_i)=1 \text{ и } y_i=y\} \rightarrow \max;$$

$$n_y(R, X^\ell) = \#\{x_i \in X^\ell : R(x_i)=1 \text{ и } y_i \neq y\} \rightarrow \min;$$

$$\frac{p_y}{P_y} \gg \frac{n_y}{N_y}$$



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

Требование интерпретируемости

- 1) $R_y(x)$ записывается на естественном языке
- 2) $R_y(x)$ зависит от небольшого числа признаков (не более 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%

Замечание. *Риск* — частотная оценка вероятности класса, вычисляемая, как правило, по отложенной контрольной выборке

Обучение логических классификаторов

Алгоритмов *индукции правил* (rule induction) очень много!

Основные шаги их построения — надо выбрать:

- ① семейство правил, в котором будем искать закономерности
- ② способ порождения правил (rule generation)
- ③ критерий отбора информативных правил (rule selection)
- ④ модель классификации с правилами в роли признаков, например, линейный классификатор (weighted voting):

$$a(x) = \arg \max_{y \in Y} \sum_{j=1}^{n_y} w_{yj} R_{yj}(x)$$

Две трактовки понятия «логическая закономерность» $R_y(x)$:

- обучаемый информативный интерпретируемый признак
- классификатор одного класса y с отказами

Часто используемые семейства правил

- Пороговое условие (решающий пень, decision stump):

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

- Конъюнкция пороговых условий:

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

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

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

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

Часто используемые семейства правил

- *Полуплоскость* — линейная пороговая функция:

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

- *Шар* — пороговая функция близости:

$$R(x) = \left[\rho(x, x_0) \leq w_0 \right]$$

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

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

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

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

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

Обоснование перебора троек признаков («КоРа-3» и др.)

Оценим вероятности выбрать конъюнкцию-предрассудок.

$x_i \in \{0, 1\}^n$ — объекты описываются n бинарными признаками;

$P(x) = 2^{-n}$ — равномерное распределение на $\{0, 1\}^n$;

$K_r(x)$ — конъюнкция ранга r , из r признаков или их отрицаний;

$|\mathcal{K}_n^r| = C_n^r 2^r$ — число различных таких конъюнкций.

$P\{K_r(x)=0\} = 1 - 2^{-r}$ — вероятность, что K_r случайно ложна;

$P\{K_r(X^\ell)=0\} = \prod_i P\{K_r(x_i)=0\} = (1 - 2^{-r})^\ell$ — вероятность, что K случайно ложна на всех объектах из $X^\ell = (x_1, \dots, x_\ell)$.

Верхняя оценка вероятности, что в результате поиска $K_r \in \mathcal{K}_n^r$ найденная K_r окажется случайно ложной на выборке X^ℓ :

$$\begin{aligned} P\{\exists K_r: K_r(X^\ell)=0\} &= P(\cup_{K_r} \{K_r(X^\ell)=0\}) \leq \text{неравенство Буля} \\ &\quad \text{(union bound)} \\ &\leq \sum_{K_r} P\{K_r(X^\ell)=0\} = C_n^r 2^r (1 - 2^{-r})^\ell. \end{aligned}$$

Закревский Аркадий Дмитриевич. Логика распознавания. Минск, 1988.

Оценка вероятности выбрать конъюнкцию-предрассудок

Верхняя оценка вероятности $P(r)$ найти случайно ложную конъюнкцию заданного ранга r , при $\ell = 200$, $n = 100$:

r	P	r	P
1	$1.24 \cdot 10^{-58}$	4	$1.56 \cdot 10^2$
2	$2.04 \cdot 10^{-21}$	5	$4.21 \cdot 10^6$
3	$3.26 \cdot 10^{-6}$	6	$3.27 \cdot 10^9$

Зависимость граничного значения r , после которого резко увеличивается верхняя оценка вероятности $P(r)$:

n	$\ell = 20$	50	100	200	500	1000
10	1	2	3	4	5	6
30	1	2	2	3	4	5
100	1	1	2	3	4	5

Закревский Аркадий Дмитриевич. Логика распознавания. Минск, 1988.

Мета-эвристики для поиска информативных правил

Вход: обучающая выборка X^ℓ ;

Выход: множество закономерностей Z ;

инициализировать начальное множество правил Z ;

повторять

$Z' :=$ множество *локальных модификаций* правил из Z ;

$Z :=$ наиболее *информативные* правила из $Z \cup Z'$;

пока правила продолжают улучшаться;

вернуть Z ;

Частные случаи (см. лекцию про методы отбора признаков):

- стохастический локальный поиск (stochastic local search)
- генетические (эволюционные) алгоритмы
- усечённый поиск в ширину (beam search)
- поиск в глубину (метод ветвей и границ)

Локальные модификации правил

Пример. Семейство конъюнкций пороговых условий:

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

Локальные модификации конъюнктивного правила:

- добавление признака f_j в J с варьированием порогов a_j , b_j
- удаление признака f_j из J
- варьирование одного из порогов a_j и b_j
- варьирование обоих порогов a_j , b_j одновременно

При удалении признака (pruning) информативность обычно оценивается по контрольной выборке (hold-out)

Вообще, для оптимизации множества J подходят те же методы, что и для отбора признаков (feature selection)

Обучение классификатора на закономерностях

Взвешенное голосование (Weighted Voting):

много-классовый линейный классификатор с весами w_{yt}
(возможна L_1 -регуляризация для отбора закономерностей)

$$a(x) = \arg \max_{y \in Y} \sum_{t=1}^{T_y} w_{yt} R_{yt}(x)$$

Простое голосование (Simple Voting, комитет большинства):

$$a(x) = \arg \max_{y \in Y} \frac{1}{T_y} \sum_{t=1}^{T_y} R_{yt}(x)$$

Решающий список (Majority Voting, комитет старшинства):

обучаемая система продукций — правил «если $R_{yt}(x)$ то y »

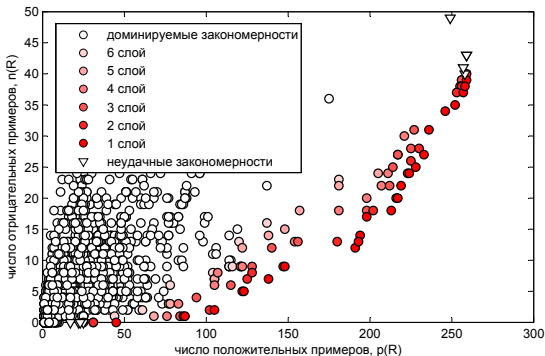
для всех $t = 1, \dots, T$: если $R_{yt}(x) = 1$ то вернуть y ;

Двухкритериальный отбор закономерностей в плоскости (p, n)

позитивные: $p_y(R) = \#\{x_i: R(x_i)=1 \text{ и } y_i=y\} \rightarrow \max;$

негативные: $n_y(R) = \#\{x_i: R(x_i)=1 \text{ и } y_i \neq y\} \rightarrow \min;$

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



UCL:german

Логические и статистические закономерности

Предикат $R(x)$ — логическая закономерность класса $y \in Y$:

$$\text{Precision} = \frac{p_y(R)}{p_y(R) + n_y(R)} \geq \pi_0 \quad \text{Recall} = \frac{p_y(R)}{P_y} \geq \rho_0$$

Если $n_y(R) = 0$, то R — непротиворечивая закономерность

Предикат $R(x)$ — статистическая закономерность класса $y \in Y$:

$$|\text{Stat}(p_y(R), n_y(R)) \geq \sigma_0$$

$|\text{Stat}$ — минус-лог вероятности реализации (p, n) при условии нулевой гипотезы, что $y(x)$ и $R(x)$ — независимые случайные величины (точный тест Фишера, Fisher's Exact Test):

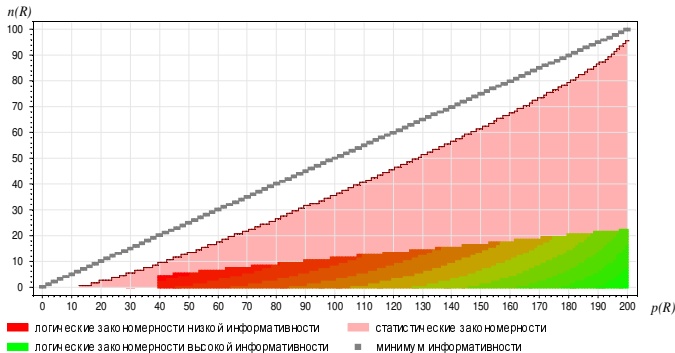
$$|\text{Stat}(p, n) = -\frac{1}{\ell} \log_2 \frac{C_P^p C_N^n}{C_{P+N}^{p+n}} \rightarrow \max,$$

где $P = \#\{x_i: y_i=y\}$, $N = \#\{x_i: y_i \neq y\}$, $C_N^n = \frac{N!}{n!(N-n)!}$

Критерии поиска закономерностей в плоскости (p, n)

Логические закономерности: Precision ≥ 0.9 , Recall ≥ 0.2

Статистические закономерности: IStat ≥ 3



$P = 200$

$N = 100$

Логический критерий удобнее для финального отбора правил;
статистический критерий — в процессе модификации правил

Зоопарк критериев информативности

Очевидные, но не вполне адекватные критерии:

- $I(p, n) = \frac{p}{p+n} \rightarrow \max$ (precision)
- $I(p, n) = p/P \rightarrow \max$ (recall)
- $I(p, n) = p/P - n/N \rightarrow \max$ (relative accuracy)

Адекватные, но не очевидные критерии:

- энтропийный критерий прироста информации:

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

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

- критерий Джини (Gini impurity):

$$IGain(p, n) \text{ при } h(q) = 4q(1 - q)$$

- критерий бустинга и его нормированный вариант:

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

J.Fürnkranz, P.Flach. ROC'n'rule learning – towards a better understanding of covering algorithms // Machine Learning, 2005.

Нетривиальность проблемы свёртки двух критериев

Пример: в каждой паре правил первое гораздо лучше второго, однако простые эвристики не различают их по качеству (при $P = 200$, $N = 100$).

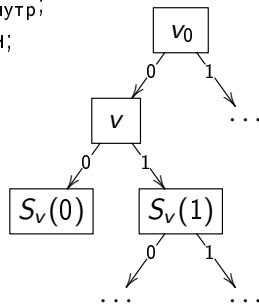
p	n	$p-n$	$p-5n$	$\frac{p}{P} - \frac{n}{N}$	$\frac{p}{n+1}$	$IStat \cdot \ell$	$IGain \cdot \ell$	$\sqrt{p} - \sqrt{n}$
50	0	50	50	0.25	50	22.65	23.70	7.07
100	50	50	-150	0	1.96	2.33	1.98	2.93
50	9	41	5	0.16	5	7.87	7.94	4.07
5	0	5	5	0.03	5	2.04	3.04	2.24
100	0	100	100	0.5	100	52.18	53.32	10.0
140	20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

Замечание. Критерии $IStat$ и $IGain$ асимптотически эквивалентны: $IStat(p, n) \rightarrow IGain(p, n)$ при $\ell \rightarrow \infty$

Определение решающего дерева (Decision Tree)

Решающее дерево — модель классификации/регрессии $a(x)$, задающаяся *деревом* (связным ациклическим графом) с корнем $v_0 \in V$ и множеством вершин $V = V_{\text{внутр}} \sqcup V_{\text{лист}}$;
 $f_v: X \rightarrow D_v$ — дискретный признак, $v \in V_{\text{внутр}}$;
 $S_v: D_v \rightarrow V$ — множество дочерних вершин;
 $y_v \in Y$ — ответ в вершине $v \in V_{\text{лист}}$;

$v := v_0$;
пока ($v \in V_{\text{внутр}}$): $v := S_v(f_v(x))$;
вернуть $a(x) := y_v$;

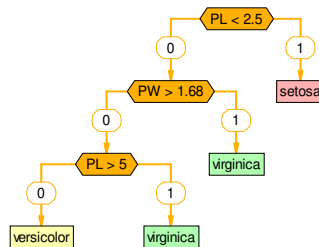
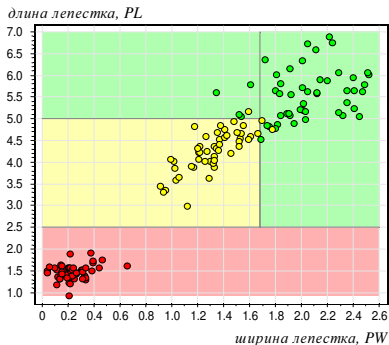


На практике часто используются бинарные признаки вида $f_v(x) = [f_j(x) > \theta]$, $j = 1, \dots, n$

Если $D_v \equiv \{0, 1\}$, то решающее дерево называется *бинарным*

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.

Обучение решающего дерева: ID3 (Iterative Dichotomiser)

Обучающая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$, признаки $F = \{f_1, \dots, f_n\}$
 $v_0 := \text{TreeGrowing}(X^\ell)$ — функция рекурсивно вызывает себя

$\text{TreeGrowing}(\text{Вход: } U \subseteq X^\ell) \mapsto \text{Выход:}$ корень дерева v ;

$f_v := \arg \max_{f \in F} \text{Gain}(f, U)$ — критерий ветвления дерева;

если $\text{Gain}(f_v, U) < G_0$ **то**

┌ создать новый лист v ; $y_v := \text{Major}(U)$; **вернуть** v ;
 создать новую внутреннюю вершину v с функцией f_v ;

для всех $k \in D_v$

┌ $U_k := \{x \in U : f_v(x) = k\}$;
 ┌ $S_v(k) := \text{TreeGrowing}(U_k)$;

вернуть v ;

Мажоритарное правило: $\text{Major}(U) = \arg \max_{y \in Y} \#\{x_i \in U : y_i = y\}$.

John Ross Quinlan. Induction of Decision Trees // Machine Learning, 1986.

Критерий ветвления $\text{Gain}(f, U)$

$\mathcal{L}(a, y)$ — потеря от ответа модели a при верном ответе y .
Суммарная потеря модели $a(x)$ на выборке $U \subseteq X^\ell$:

$$Q(a, U) = \sum_{x_i \in U} \mathcal{L}(a(x_i), y_i)$$

Пусть U — множество объектов x_i , дошедших до вершины v .
Суммарная потеря, если вершину v сделать листом:

$$Q_0(U) = \min_{a \in Y} Q(a, U)$$

Суммарная потеря при ветвлении $U_k = \{x \in U : f(x) = k\}$,
если все дочерние вершины $S_v(k)$ сделать листьями:

$$Q(f, U) = \sum_{k \in D_v} \min_{a \in Y} Q(a, U_k)$$

$\text{Gain}(f, U) = Q_0(U) - Q(f, U)$ — выигрыш от ветвления по f

Жадная нисходящая стратегия: достоинства и недостатки

Достоинства:

- Интерпретируемость и простота классификации
- Правила $[f_j(x) > \theta]$ не требуют масштабирования признаков
- Допустимы разнотипные данные (через бинаризацию)
- Объём вычислений линейный по длине выборки $O(|F|\ell h)$

Недостатки:

- Жадная стратегия переусложняет структуру дерева, и, как следствие, ведёт к переобучению
- Фрагментация выборки: чем дальше v от корня, тем меньше статистическая надёжность выбора f_v, y_v
- Высокая чувствительность результата обучения дерева к шуму в данных, составу выборки, критерию ветвления

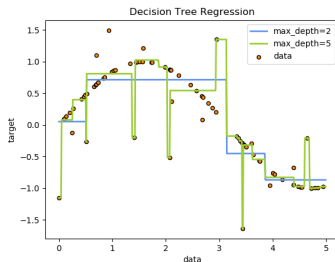
CART: деревья регрессии и классификации

При квадратичной функции потерь $\mathcal{L}(a, y) = (a - y)^2$, $Y = \mathbb{R}$, задача наименьших квадратов решается аналитически; оптимальное значение y_v в листовой вершине $v \in V_{\text{лист}}$:

$$Q(a, U) = \sum_{x_i \in U} (a - y_i)^2; \quad y_v = \arg \min_{a \in Y} Q(a, U) = \frac{1}{|U|} \sum_{x_i \in U} y_i$$

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

Пример. Чем сложнее дерево (чем больше его глубина), тем выше влияние шумов в данных и выше риск переобучения



Leo Breiman et al. Classification and regression trees. 1984.

scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

CART: критерий Minimal Cost-Complexity Pruning

Среднеквадратичная ошибка со штрафом за сложность дерева:

$$Q_\alpha(a) = \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \alpha |V_{\text{лист}}| \rightarrow \min_a$$

При увеличении α дерево последовательно упрощается, причём последовательность вложенных деревьев единственна. Из неё выбирается дерево с \min ошибкой на тесте (Hold-Out).

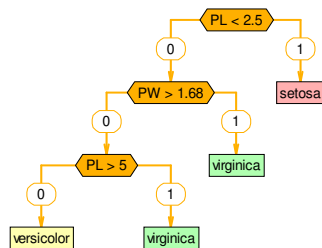
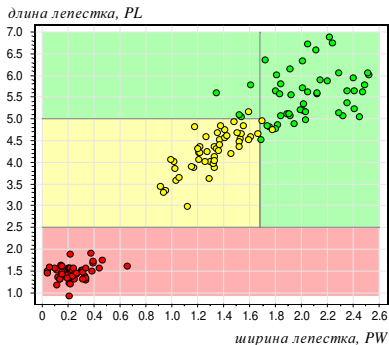
Дерево — линейный классификатор над бинарными признаками $B_v(x) = [\text{объект } x \text{ прошёл путь от корня } v_0 \text{ до листа } v]$:

$$a(x) = \sum_{v \in V_{\text{лист}}} y_v B_v(x),$$

y_v — вес признака B_v , для регрессии среднее $\{y_i : B_v(x_i) = 1\}$

Leo Breiman et al. Classification and regression trees. 1984.

Решающее дерево → покрывающий набор конъюнкций



setosa
virginica
virginica
versicolor

$$r_1(x) = [PL \leq 2.5]$$

$$r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$$

$$r_3(x) = \cancel{[PL > 2.5]} \wedge [PW \leq 1.68] \wedge [PL > 5]$$

$$r_4(x) = [PL > 2.5] \wedge [PW < 1.68] \wedge [PL \leq 5]$$

- Эмпирическая индукция — вывод знаний из данных:
 - индукция правил (Rule Induction)
 - решающие деревья, списки, таблицы
- Преимущества логических методов:
 - интерпретируемость
 - возможность обработки разнотипных данных
 - возможность обработки данных с пропусками
- Недостатки логических методов:
 - ограниченное качество классификации
 - перебор правил либо долгий, либо неполный
 - решающие деревья неустойчивы, склонны к переобучению
- Способы устранения этих недостатков:
 - упрощение (pruning), как правило, по тестовым данным
 - композиции правил, леса деревьев (в следующих лекциях)