Логический анализ данных в распознавании: вводная лекция

д.ф.-м.н. Елена Всеволодовна Дюкова
edjukova@mail.ru
к.ф.-м.н. Пётр Александрович Прокофьев
p_prok@mail.ru, https://t.me/pprok

МГУ, Москва

2 октября 2023 г.

О чём этот спецкурс?

Спецкурс посвящён вопросам применения аппарата дискретной математики в задачах интеллектуального анализа данных. Излагаются:

- общие принципы, лежащие в основе логического подхода к задачам машинного обучения
- методы конструирования процедур классификации по прецедентам с использованием понятий теории булевых функций и теории покрытий булевых матриц
- основные модели логических процедур классификации, вопросы сложности их реализации и качества решения прикладных задач

Содержание вводной лекции

Логический подход в распознавании

Задача распознавания по прецедентам

- M — множество объектов, $M = K_1 \sqcup \ldots \sqcup K_l$
- $\{x_1, \ldots, x_n\}$ — система признаков
- $(x_1(S), \ldots, x_n(S))$ — признаковое описание объекта $S \in M$
- $y : M \to \{1, \ldots, l\}$ — целевая функция, истинные номера классов объектов
- $T = \{S_1, \ldots, S_m\}$ — обучающая выборка (прецеденты)
- $y_i = y(S_i)$ — номер класса, которому принадлежит прецедент S_i
- $A_T : M \to \{0, 1, \ldots, l\}$ — алгоритм распознавания
- если $A_T(S_i) = y_i$, $\forall S_i \in T$, то A_T — *корректный алгоритм*
- Обобщающая способность A_T характеризуется качеством распознавания объектов из $M \setminus T$

Признаковые описания объектов

$x_j : M \to D_j$, D_j — множество значений (домен) признака x_j.
Домены признаков, использующиеся в логическом подходе:

- $D_j = \{0, 1\}$ — бинарный признак
- $D_j = \{1, ..., k\}$ — целочисленный (номинальный) признак, k-значный
- $D_j = \mathcal{P}$ — частично упорядоченное множество, то есть $\forall a, b, c \in \mathcal{P}$:
  - $a \preccurlyeq a$ *(рефлексивность)*
  - если $a \preccurlyeq b$ и $b \preccurlyeq a$, то $a = b$ *(антисимметричность)*
  - если $a \preccurlyeq b$ и $b \preccurlyeq c$, то $a \preccurlyeq c$ *(транзитивность)*
- $D_j = \mathbb{R}$ действительнозначный (количественный) признак
  - используют корректные перекодировки, переходя, например, к бинарным признакам $[l_{jr} \leqslant x_j(S) \leqslant u_{jr}]$
  - ограничиваются упорядоченным множеством значений на выборке $\{x_j(S_1), \dots x_j(S_m)\}$

Векторы и матрицы признаковых описаний

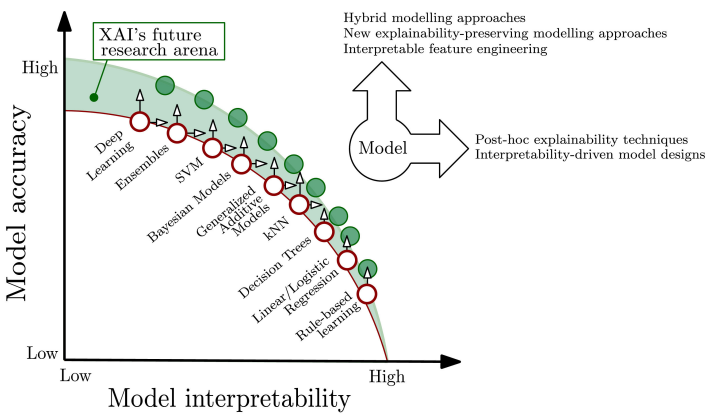
- Вектор $(a_{i1}, \ldots, a_{in})$, $a_{ij} = x_j(S_i)$ — признаковое описание прецедента S_i
- Матрица «объекты–признаки»

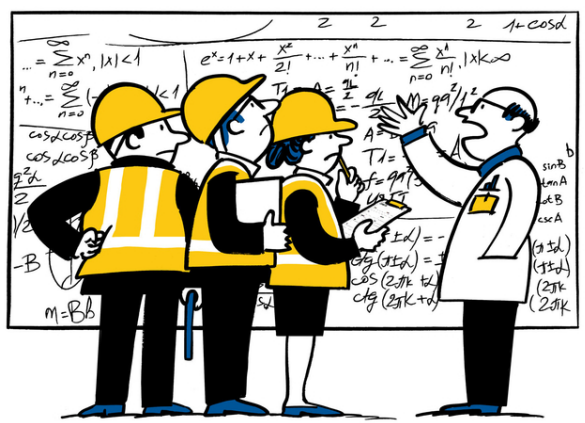
$$L = \|a_{ij}\|_{m \times n} = \begin{pmatrix} x_1(S_1) & \ldots & x_n(S_1) \\ \ldots & \ldots & \ldots \\ x_1(S_m) & \ldots & x_n(S_m) \end{pmatrix}$$

- Объект S фактически отождествляется с его вектором значений признаков $(x_1(S), \ldots, x_n(S))$

Когда целесообразно применять логический подход?

- **Интерпретируемость распознающих процедур**, когда эксперту предметной области задачи требуется на естественном языке объяснить причину отнесения распознаваемого объекта к определенному классу
- **Небольшие выборки**, когда нет возможность применить статистические методы или глубокое обучение
- **Естественные логические закономерности в данных**, обусловленные предметной областью задачи, должны быть учтены при построении процедур распознавания

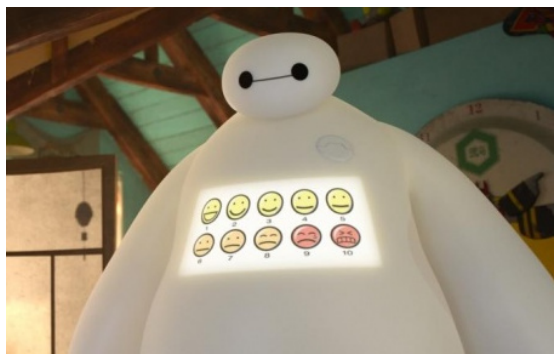




Прикладные области применения

Задача медицинской диагностики

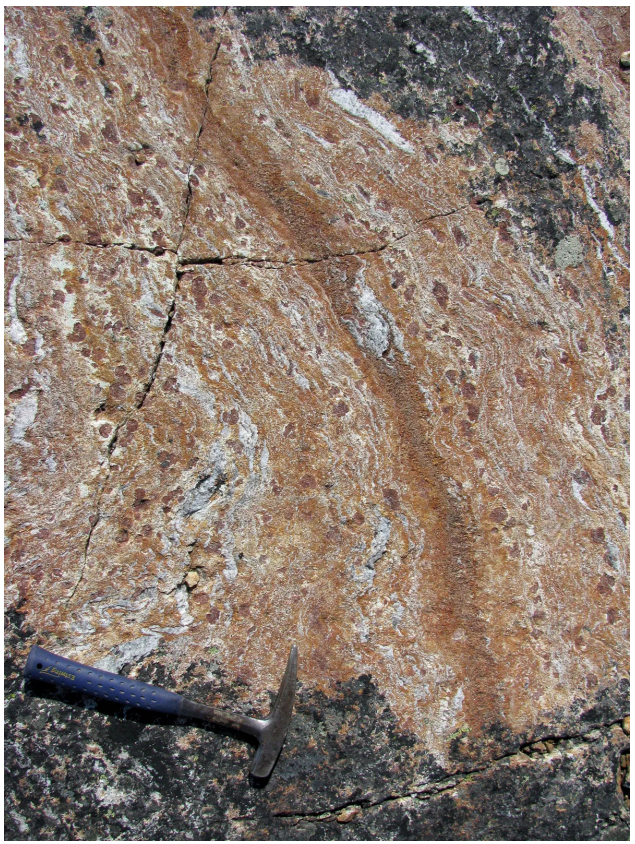
- **объекты:**
  пациенты в определенный момент времени
- **классы:**
  диагнозы, исходы болезни, способы лечения
- **признаки:**
  возраст, вес, рост, слабость, тошнота и т.д.
- **особенности:** важна интерпретируемость, значения многих признаков субъективны, много пропусков



Прикладные области применения

Задача распознавания месторождений

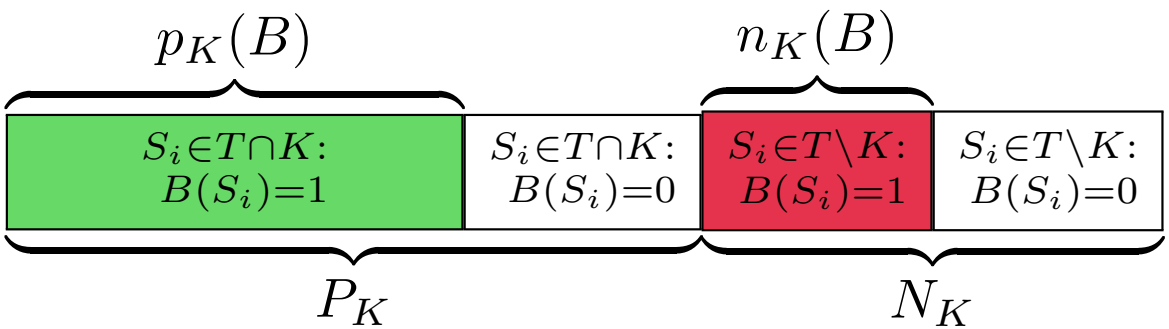
- **объект** — геологический район
- **классы:**
  факт присутствия определенного ископаемого в достаточном количестве
- **признаки:** присутствие
  крупных зон смятия и рассланцевания, минеральное разнообразие и т.д.
- **особенности:** малые выборки



Предикаты как базовые элементы логических процедур

- $B : M \to \{0, 1\}$ — предикат на множестве объектов (множестве векторов значений признаков)
- Если $B(S) = 1$, то говорят, что предикат B выделяет объект S
- Для класса K обозначим:

$$P_K = |T \cap K|, \qquad N_K = |T \setminus K|,$$
$$p_K(B) = \sum_{S_i \in T \cap K} B(S_i), \qquad n_K(B) = \sum_{S_i \in T \setminus K} B(S_i)$$



Логические закономерности

- Предикат называется *логической закономерностью*, если он:
  - записывается на естественном языке и обычно зависит от небольшого числа признаков (*интерпретируемость*)
  - выделяет преимущественно объекты одного класса (*информативность*):

$$p_K(B) \to \max, \quad n_K(B) \to \min, \quad \frac{p_K(B)}{P_K} \gg \frac{n_K(B)}{N_K}$$

- В логическом подходе часто от закономерностей требуется *корректность* (непротиворечивость): $n_K(B) = 0$
- Также применяются «антипредставительные» предикаты со свойством $p_K(B) = 0$, $n_K(B) > 0$, которые также называются корректными

Часто использующиеся семейства предикатов

- Фиксированное значение номинального признака

$$B(S) = [x_j(S) = \sigma_j]$$

- Пороговое условие

$$B(S) = [a_j \preccurlyeq x_j(S) \preccurlyeq b_j], \qquad a_j, b_j \in D_j$$

- Конъюнкция пороговых условий

$$B(S) = \bigwedge_{j \in H} [a_j \preccurlyeq x_j(S) \preccurlyeq b_j], \qquad H \subset \{1, \ldots, n\}$$

- Синдром — выполняется не менее d условий

$$B(S) = \left[\sum_{j \in H} [a_j \preccurlyeq x_j(S) \preccurlyeq b_j] \geqslant d \right], \quad d \in \{1, \ldots, |H|\}$$

Часто использующиеся семейства предикатов (2)

- Совпадение признаковых подописаний объекта и прецедента S_i

$$B(S) = \bigwedge_{j \in H} [x_j(S_i) = x_j(S)]$$

- Сравнение признаковых подописаний объекта и прецедента S_i

$$B(S) = \bigwedge_{j \in H} [x_j(S_i) \preccurlyeq x_j(S)]$$

Элементарный классификатор

Определение

Элементарным классификатором (эл.кл.) ранга r, $1 \leqslant r \leqslant n$, *называется пара* (H, σ), *где* $H = (x_{j_1}, \ldots, x_{j_r})$ — *набор различных признаков и* $\sigma = (\sigma_1, \ldots, \sigma_r)$ — *вектор целых чисел, в котором* σ_t — *допустимое значение признака* x_{j_t}, $t \in \{1, \ldots, r\}$.

- $H(S) = (x_{j_1}(S), \ldots, x_{j_r}(S))$ — признаковое подописание объекта S
- эл.кл. (H, σ), выделяет объект, если $H(S) = \sigma$, т.е. фактически задаёт конъюнкцию $[x_{j_1}(S) = \sigma_1 \wedge \ldots \wedge x_{j_r}(S) = \sigma_r]$

Определение

Эл.кл. (H, σ) *называется корректным для класса* K, *если не существует двух прецедентов* S_i *и* S_t, *выделяемых эл.кл.* (H, σ), *таких, что* $S_i \in K$ *и* $S_t \notin K$.

Алгоритм голосования по представительным наборам

Определение

Представительным набором класса K называется корректный для K эл.кл., выделяющий хотя бы один прецедент из K.

Голосование по представительным наборам

- На этапе обучения для каждого класса $K \in \{K_1, \ldots, K_l\}$ строится семейство C_K представительных наборов
- При распознавании объекта S для каждого класса $K \in \{K_1, \ldots, K_l\}$ вычисляется оценка принадлежности S к K

$$\Gamma(S, K) = \sum_{(H,\sigma) \in C_K} w_{(H,\sigma)}[H(S) = \sigma],$$

Другие виды голосования по корректным элементарным классификаторам

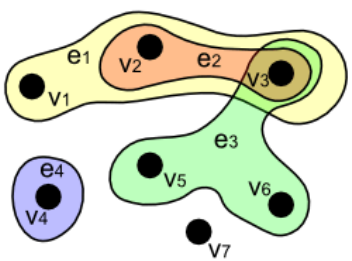
1. Голосование по тестам (*Дмитриев, Журавлёв, Кренделев, 1968*).
2. Голосование по представительным наборам (*Вайнцвайг, 1973*) (*Баскакова, Журавлёв, 1981*).

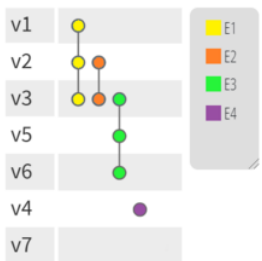
   Зарубежный аналог представительного набора — *emerging pattern* (*Dong, Zhang, Wong, Li, 1999*).
3. Голосование по корректным элементарным классификаторам (*Дюкова, Песков, 2002*).

Дискретные задачи при обучении

Трансверсали гиперграфов

- *Гиперграф* — пара множеств (V, E), в которой V — множество вершин и E — множество ребер, являющихся подмножествами V. То есть гиперграф — это обобщение графа: каждое его ребро может содержать произвольное число вершин.
- Множество вершин $H, H \subset V$, называется *трансверсалью* (вершинным покрытием), если оно пересекается со всеми ребрами: $H \cap J \neq \varnothing$, $\forall J \in E$.
- Трансверсаль называется *минимальной*, если любое ее подмножество трансверсалью не является.
- Множество всех минимальных трансверсалей является множеством ребер так называемого *двойственного* гиперграфа





Представительный набор = трансверсаль гиперграфа

Задача. Найти представительные наборы, задаваемые признаковыми подописаниями фиксированного прецедента S_i из класса K.

Решение:

1. Рассмотрим гиперграф с вершинами $V = \{1, \ldots, n\}$ и ребрами вида

$$\{j : x_j(S_k) \neq x_j(S_i)\}, \quad S_k \in T \setminus K.$$

Каждое ребро содержит номера признаков, по которым S_i можно отличить от некоторого объекта другого класса.

2. Рассмотрим трансверсаль построенного гиперграфа: трансверсаль $H = \{j_1, \ldots, j_r\}$ определяет представительный набор (H, σ) с признаковым подописанием прецедента S_i из условия задачи:

$$\sigma = (\sigma_1, \ldots, \sigma_r), \quad \sigma_s = x_{j_s}(S_i), \quad s \in \{1, \ldots, r\}.$$

Представительные наборы для минимальных трансверсалей

3. Представительный набор, соответствующий минимальной трансверсали называется неприводимым. Неприводимый набор всегда обладает информативностью не ниже любого его расширения, поэтому предпочтительнее использовать неприводимые наборы.
4. Двойственный гиперграф определяет все неприводимые представительные наборы, являющиеся признаковыми подописаниями прецедента S_i

Задача построения двойственного гиперграфа

Задача

- **Вход:** Гиперграф (V, E)
- **Выход:** Двойственный к (V, E) гиперграф (V, E')

Недостатки постановки:

- для размера выхода $N = |E'|$ невозможно указать зависимость от размера входа $n = |V|$, $m = |E|$ или хотя бы разумно оценить $N = \mathcal{O}(f(n, m))$ (в общем случаем размер выхода растет экспоненциально от размера входа $N = \mathcal{O}(2^n)$);
- также невозможно оценить ёмкостную и временную сложности алгоритма от размера входа.

Задача дуализации гиперграфа

Задача

- **Вход:** Гиперграф (V, E) и E' — набор ранее построенных минимальных трансверсалей (V, E)
- **Выход:** Минимальная трансверсаль $H \notin E'$ гиперграфа (V, E), если она существует, либо $\varnothing$

- **Достоинства:**
  - \+ можно реализовать алгоритм *дуализации*, в котором трансверсали *перечисляются* одна за другой;
  - \+ сложность алгоритма дуализации фактически оценивается сложностью одного шага перечисления от размера входа.
- **Недостатки:**
  - \- входом кроме гиперграфа являются ранее найденные решения E' и сложность шага может существенно зависеть от величины $N = |E'|$, которая по-прежнему оценивается как $N = \mathcal{O}(2^n)$

Теоретические оценки сложности дуализации «в худшем случае»

- Эффективность алгоритма дуализации оценивается временем выполнения одного шага. На каждом шаге находится в точности одно решение. Временные оценки даются в «худшем случае», то есть для самой сложной задачи (*Johnson, Yannakakis, Papadimitriou, 1988*).
- Вопрос о полиномиальной разрешимости дуализации открыт.
- Для немногих частных случаев построены алгоритмы с полиномиальными задержками, имеющие временные оценки вида $\mathcal{O}(f)$, где $f(m, n)$ — полином от m и n.
- Наилучший теоретический результат получен в (*Fredman, Khachiyan, 1996*). Построен *инкрементальный* алгоритм с квазиполиномиальной временной оценкой $\mathcal{O}(g^{\log g})$, где $g(m, n, N)$ — полином от размера входа m, n и числа выполненных шагов N.

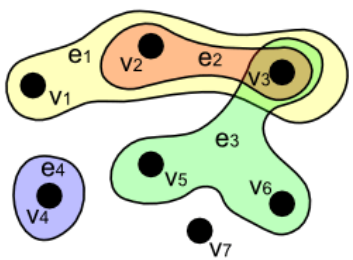
Теоретические оценки сложности дуализации «в среднем»

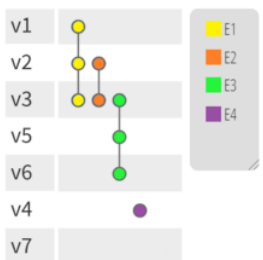
- В (*Дюкова, 1977*) предложен подход к построению *асимптотически оптимальных алгоритмов дуализации* (алгоритмов, эффективных в среднем).
- Асимптотически оптимальные алгоритмы отличаются от алгоритмов с полиномиальными задержками тем, что имеют «лишние» полиномиальные шаги.
- При определенных ограничениях на размер входа число шагов такого алгоритма (включая «лишние») асимптотически эквивалентно числу решений для *почти всех входов* такого же размера

От гиперграфа к булевой матрице

Гиперграф задаётся булевой матрицей. Перечисление минимальных трансверсалей сводится к перечислению неприводимых покрытий булевой матрицы.

- Рассмотрим гиперграф с вершинами $V = \{1, \ldots, n\}$ и рёбрами $E = \{J_1, \ldots, J_m\}$
- Закодируем каждое ребро J_i бинарным вектором $a_i = (a_{i1}, \ldots, a_{in})$, где $a_{ij} = [j \in J_i]$
- Булева матрица, составленная из этих вектором, однозначно задаёт исходный гиперграф





$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Покрытия булевой матрицы

- Трансверсаль H задает *покрытие* этой матрицы:

$$\forall i \in \{1, \ldots, m\}, \exists j \in H : a_{ij} = 1 \tag{1}$$

- Минимальной трансверсали соответствует *неприводимое покрытие* булевой матрицы
- На множестве бинарных вектором естественным образом определяется частичный порядок:

$$(\alpha_1, \ldots, \alpha_n) \preccurlyeq (\beta_1, \ldots, \beta_n) \Leftrightarrow \forall j \in V, \alpha_j \leqslant \beta_j$$

- Если трансверсаль закодировать бинарным вектором $h = (h_1, \ldots, h_n)$, где $h_j = [j \notin H]$, то (1) эквивалентно (2):

$$\forall i \in \{1, \ldots, m\}, a_i \not\preccurlyeq h \tag{2}$$

От бинарного случая к частичным порядкам

Перечисление неприводимых покрытий булевой матрицы — это частный случай перечисления максимальных независимых векторов, на множестве которых задан покоординатный частичный порядок.

- Пусть заданы частичные порядки $\mathcal{P}_1, \ldots, \mathcal{P}_n$
- На множестве векторов $\mathcal{P} = \mathcal{P}_1 \times \ldots \times \mathcal{P}_n$ естественным образом определяется частичный порядок:

$$(\alpha_1, \ldots, \alpha_n) \preccurlyeq (\beta_1, \ldots, \beta_n) \Leftrightarrow \forall j \in \{1, \ldots, n\}, \alpha_j \preccurlyeq \beta_j$$

- Рассматривается множество векторов $A = \{a_1, \ldots, a_m\}, a_i \in \mathcal{P}$

Дуализация над произведением частичных порядков

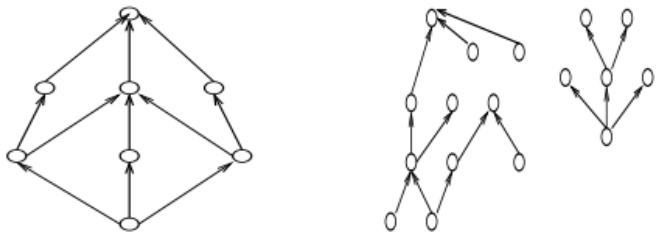
- Вектор $h \in \mathcal{P}$ называется *независимым* от множества векторов $\{a_1, \ldots, a_m\}, a_i \in \mathcal{P}$, если

$$\forall i \in \{1, \ldots, m\}, a_i \not\preccurlyeq h.$$

- Если для любого $h' : h \preccurlyeq h'$ вектор h' не обладает свойством независимости, то h называется *максимально независимым*
- Перечисление максимально независимых элементов называется *дуализацией над произведением частичных порядков*

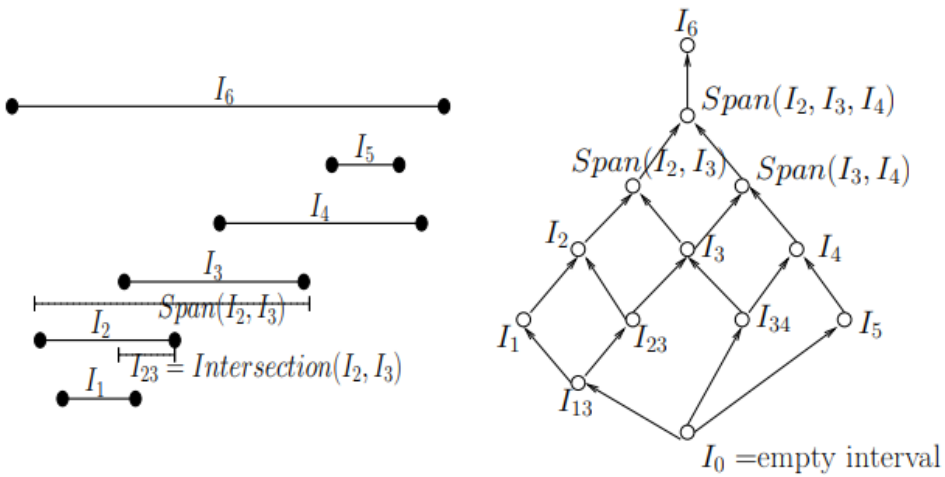
Полезные частичные порядки

Частичные порядки из прикладных задач (*Khaled Elbassioni. Algorithms for Dualization over Products of Partially Ordered Sets. 2009, Siam Journal on Discrete Mathematics*)



a: A lattice with $W = 3$. b: A forest with $d = 3$.

Figure 1: Lattices and forests.



a: A set of intervals $\mathbb{I}_1$. b: The corresponding lattice of intervals $\mathcal{L}_1$.

Figure 2: The lattice of intervals.

Дополнительные вопросы курса

Корректные vs некорректные эл.кл.

- Достоинства и недостатки процедуры голосования по представительным наборам:
  - \+ Процедура не ошибается на прецедентах, т.к. корректны все голосующие эл.кл.
  - \- В прикладных задачах представительные наборы зачастую могут иметь очень большой ранг, что усложняет алгоритм и приводит к *переобучению* (плохо распознает те объекты, которые не встречаются в выборке).
- Попробуем отказаться от непротиворечивости эл.кл. и скорректировать выделяемые ими объекты в совокупности.

Алгебраический подход

- В *алгебраическом* подходе строится набор операций $\{B_1, \ldots, B_p\}$, для которого можно найти «хорошую» *корректирующую функцию* $F : R^p \to R$ и простое *решающее правило* $C : R \to Y$ (*Журавлев, 19??*):

$$\begin{array}{ccc} M & \xrightarrow{A_T} & Y \\ {\scriptstyle B_1,\ldots,B_p}\downarrow & & \uparrow{\scriptstyle C} \\ R^p & \xrightarrow[F(B_1,\ldots,B_p)]{} & R \end{array}$$

- Алгоритм распознавания — это композиция

$$A_T(S) = C(F(B_1(S), \ldots, B_p(S)))$$

Алгебро-логический подход

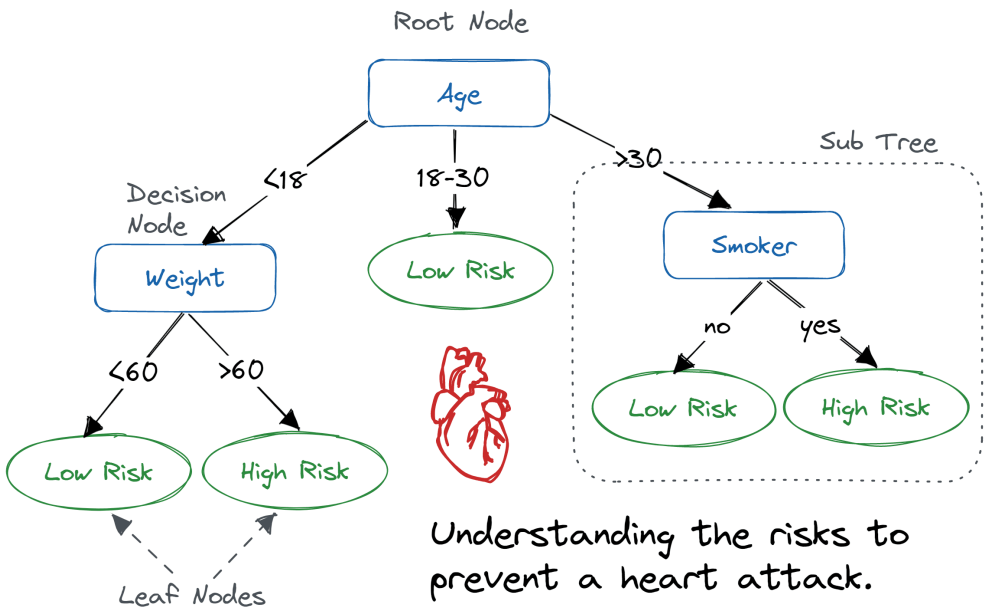
- Рассмотрим в качестве операций $\{B_1, \ldots, B_p\}$ логические закономерности $B_k : M \to \{0, 1\}$
- Корректирующие функции выбираются среди булевых функций

$$F : \{0, 1\}^p \to \{0, 1\}.$$

- Распознающие процедуры, построенные путем коррекции логических закономерностей, называются *логическими корректорами*
- Корректное распознавание на базе произвольных, не обязательно корректных эл.кл., основано на идее *алгебро-логического подхода*, которая предложена в (*Дюкова, Журавлёв, Рудаков, 1996*).
  Зарубежный аналог — Logical Analysis of Data (LAD) (*Boros, Hammer et al., 2000*).

Решающее дерево (Decision Tree)

Решающее дерево — это алгоритм распознавания, задающийся деревом, в каждой внутренней вершине которого принимается решение о переходе к очередной вершине на основе значения одного из признаков, а каждая листовая вершина помечена номером класса, к которому следует отнести объект



Покрывающие наборы конъюнкций:

- High risk
  $[\textbf{Age} < 18 \wedge \textbf{Weight} > 60]$
  $[\textbf{Age} > 30 \wedge \textbf{Smoker} = \textbf{Yes}]$
- Low risk
  $[\textbf{Age} < 18 \wedge \textbf{Weight} < 60]$
  $[18 \leqslant \textbf{Age} \leqslant 30]$
  $[\textbf{Age} > 30 \wedge \textbf{Smoker} = \textbf{No}]$

Полное решающее дерево

- Дерево решений строится рекурсивно, путем выбора признака (логической закономерности), наилучшим образом разделяющего прецеденты (обычно используется информационный критерий деления *IGain, Gini*).
- В случае, когда несколько признаков имеют сравнимые с наилучшим показатели качества разделения, можно поступить разными способами.
  1. Случайным образом выбирают один из наилучших признаков (*ID3, C4.5*)
  2. Построить *полное решающее дерево*, в котором деление во внутренней вершине может быть связана с двумя и более признаками. При этом распознаваемый объект может дойти до нескольких листьев дерева и решение принимается голосованием.
- Полные решающие деревья предложены и разработаны в (*Дюкова Е. В., Песков Н. В., 2007*) и (*Генрихов И. Е., Дюкова Е. В., 2012*)

Частые наборы признаков

- Пусть признаки объектов бинарны $x_j : M \to \{0, 1\}$
- *Частота совместной встречаемости* признаков $H = \{x_{j_1}, \ldots, x_{j_r}\}$ оценивается по набору прецедентов величиной, называемой *поддержкой* (*support*):

$$\nu(H) = \frac{1}{m} \sum_{i=1}^{m} x_{j_1}(S_i) \wedge \ldots \wedge x_{j_r}(S_i)$$

- Если $\nu(H) > \delta$, то набор H называется *частым* (*frequent itemset*), где δ — минимальный уровень поддержки

Определение ассоциативного правила

Ассоциативным правилом (association rule) $H \to G$ называется пара непересекающихся наборов признаков H, G таких, что

1. совместно H и G встречаются часто:
$$\nu(H \cup G) \geqslant \delta;$$
2. если встречается H, то часто встречается и G:
$$\nu(G|H) = \frac{\nu(H \cup G)}{\nu(H)} \geqslant \varkappa,$$

$\nu(G|H)$ — *значимость правила* (*confidence*),
$\varkappa$ — минимальный уровень значимости

Пример использования ассоциативных правил

Анализ рыночных корзин (*1993*)

- объекты — чеки покупателей (*transactions*)
- признак — наличие определенного товара из магазина в чеке (*items*)

Пример. Если куплен хлеб, то молоко будет куплено с вероятностью 60%, причем вместе хлеб с молоком покупаются с вероятностью 2%.

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B\&C \Rightarrow D$	1/5	1/3	5/9

Поиск ассоциативных правил

Этапы поиска ассоциативных правил

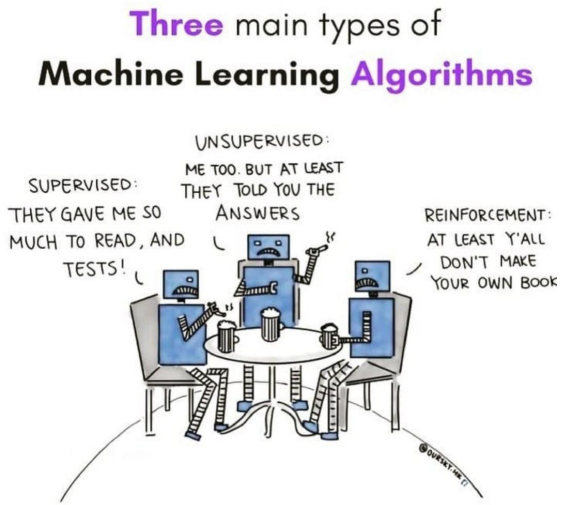
1. Поиск частых наборов по набору объектов (базе транзакций).
2. Выделение ассоциативных правил путем комбинирования частых наборов

- Наиболее известный алгоритм поиска ассоциативных правил **APriory** достаточно эффективно реализует эти два этапа.
- Учениками Дюковой Е.В. также разработаны эффективные алгоритмы для поиска обобщенных ассоциативных правил, когда признаки объектов не бинарны, а принимают значения из частичных порядков.

Практическая часть курса

Курс имеет в том числе практическую направленность.
Мы с вами будем:

- программировать алгоритмы дуализации и дискретной оптимизации (C/C++, python);
- обучать алгоритмы распознавания и проверять их работу на реальных и модельных данных;
- экспериментировать с модификациями распознающих процедур;
- извлекать знания из данных построением ассоциативных правил;
- добиваться ускорения за счет распараллеливания и GPU.



Выводы

- Логический анализ данных в распознавании решает задачи извлечения знаний из данных с учителем и без учителя методами дискретной математики
- Базовыми элементами логических распознающих процедур являются логические закономерности, обладающие свойствами интерпретируемости и информативности
- Существуют различные конструкции логических процедур распознавания: голосование по эл. кл., логические корректоры, деревья решений
- Обучение распознающих процедур и извлечение ассоциативных правил зачастую сводится к применению алгоритмов дуализации над различными дискретными структурами