Правдоподобные рассуждения

Современное состояние и перспективы

Дмитрий Виноградов

ФИЦ ИУ РАН

27 сентября 2023 г.

Правдоподобные рассуждения и ДСМ-метод

ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания (ред.: Аншаков О.М., Финн В.К.). М.: Эдиториал УРСС, 2009.

1. Правила правдоподобного вывода 1-го рода: индуктивное обобщение обучающих примеров в гипотезы о причинах целевого свойства.
2. Правила правдоподобного вывода 2-го рода: предсказание по аналогии.
3. Абдуктивное принятие гипотез на основании объяснения свойств обучающей выборки.
4. Пополнение обучающей выборки аналогами необъясненных обучающих примеров или добавление признаков.

Квазиаксиоматические системы - комбинация правил достоверного (дедуктивного) и правдоподобного вывода.

Описать окружность около выпуклого многоугольника

Многоугольники	**goal**	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Равносторонний треугольник	+	1	0	0	1	1	0	1	1
Прямоугольный треугольник	+	1	0	1	0	0	0	0	0
Квадрат	+	0	1	1	1	1	1	1	1
Прямоугольник	+	0	1	1	1	0	1	1	1
Ромб	-	0	1	0	1	1	1	1	0
Параллелограмм	-	0	1	0	1	0	1	1	0
Равнобедренная трапеция	+	0	1	0	1	0	1	1	0
Прямоугольная трапеция	-	0	1	1	0	0	1	1	0
Равнобедренный треугольник	?	1	0	0	1	0	0	1	0

f_1: треугольник; f_2: четырёхугольник; f_3: есть прямой угол; f_4: пара сторон равна; f_5: все стороны равны; f_6: есть пара параллельных сторон; f_7: пара углов равна; f_8: все углы равны.

- Сходство $\{f_1\}$ - вокруг любого треугольника можно
- Сходство $\{f_4, f_5, f_7, f_8\}$ - вокруг правильного многоугольника можно
- Сходство $\{f_2, f_3, f_4, f_6, f_7, f_8\}$ - вокруг прямоугольника можно
- Сходство $\{f_2, f_4, f_6, f_7\}$ имеет ромб как контр-пример!

Булева алгебра: экспоненциальный взрыв

Пусть $O = \{o_1, o_2, \ldots, o_n\}$ - множество объектов, а $F = \{f_1, f_2, \ldots, f_n\}$ - множество признаков, и обучающая выборка задаётся формулой $o_i I f_j \Leftrightarrow i \neq j$:

$O \mid F$	f_1	f_2	$\ldots$	f_n
o_1	0	1	$\ldots$	1
o_2	1	0	$\ldots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_n	1	1	$\ldots$	0

Ясно, что любое подмножество признаков $\{f_{j_1}, \ldots, f_{j_k}\}$ является сходством объектов $O \setminus \{o_{j_1}, \ldots, o_{j_k}\}$, поэтому мы имеем Булеву алгебру всех 2^n подмножеств (=битовых строк).
При $n = 32$ обучающая выборка занимает $n \cdot n = 2^{10}$ бит = 128 байт. Но чтобы записать результат требуется $n \cdot 2^n = 2^{37}$ бит, т.е. ровно 16 Гигабайт.
При $n = 64$ обучающая выборка занимает $n^2 = 2^{12}$ бит = 512 байт, но для сохранения всех кандидатов нужно $n \cdot 2^n = 2^{70}$ бит!

Случайные плотные обучающие выборки

Рассмотрим серию испытаний Бернулли $\delta_{1,1}, \ldots, \delta_{m,n}$ с вероятностью успеха $p = \left(1 - \frac{1}{m}\right)$, заполняющую обучающую выборку $I \subseteq O \times F$, где $O = \{o_1, \ldots, o_m\}$ и $F = \{f_1, \ldots, f_n\}$. Обозначим через c_i столбец высоты m, где $(m\text{-}1)$ позиция заполнены единицами, а единственный ноль встречается на i-ой позиции.
Значение для параметра p определяется как оценка максимального правдоподобия, чтобы обеспечить в среднем ровно один ноль в столбце высоты m. Тогда вероятность $\mathbb{P}[f_j = c_i]$ получить столбец c_i равна

$$s = p^{m-1} \cdot (1 - p) = \left(1 - \frac{1}{m}\right)^{m-1} \cdot \frac{1}{m}$$

для любых i и j, причём события $[f_j = c_i]$ и $[f_k = c_l]$ независимы для разных i, k, j и l.

Подрешетка, изоморфная Булевой алгебре

Для того, чтобы породилась m-мерная булева алгебра как решётка кандидатов из нашей обучающей выборки, необходимо, чтобы среди столбцов бинарной матрицы встречались c_i для всех $1 \leq i \leq m$. Тем самым наша ситуация становится аналогичной известной *задаче о собирателе купонов*.

Теорема

При $m \to \infty$ и $n \geq e \cdot m \cdot \ln m$ для вероятности порождения m-мерной булевой алгебры как решётки кандидатов имеем

$$\lim_{m \to \infty} \mathbb{P}\left[\forall j \exists i \left(f_j = c_i\right)\right] \geq 1 - m^{(1 - e \cdot m \cdot s)} \to 1.$$

Доказательство теоремы

Рассмотрим вероятность дополнительного события $P = \mathbb{P}\left[\exists j \forall i \left(f_j \neq c_i\right)\right]$. Из неравенства Буля получаем оценку

$$P \leq m \cdot \prod_{i=1}^{e \cdot m \cdot \ln m} \left(1 - \mathbb{P}\left[f_j = c_i\right]\right)$$

из-за независимости событий $[f_j = c_i]$ и $[f_k = c_l]$ для разных i, k, j и l. Используя неравенство $(1 - s) \leq e^{-s}$, имеем оценку

$$P \leq e^{\ln m - s \cdot e \cdot m \cdot \ln m} = m^{(1 - s \cdot e \cdot m)},$$

где s — общая вероятность событий $[f_j = c_i]$ для разных i и j.
Оценим теперь

$$e \cdot m \cdot s = e \cdot m \cdot \left(1 - \frac{1}{m}\right)^{m-1} \cdot \frac{1}{m} = e \cdot \left(1 - \frac{1}{m}\right)^{m-1} = e \cdot \left(1 + \frac{1}{m-1}\right)^{-(m-1)}.$$

Из известного неравенства $\left(1 + \frac{1}{m-1}\right)^{m-1} < e$ следует, что $e \cdot m \cdot s > 1$, откуда получаем $m^{(1 - e \cdot m \cdot s)} \to 0$ при $m \to \infty$, что завершает доказательство теоремы.

Кандидаты в гипотезы о причинах

(Обучающую) выборку можно понимать как бинарное отношение между элементами множества O, которые мы называем **именами объектов**, и элементами множества F, которые мы называем **признаками**.

$O \times F$	f_1	$\ldots$	f_{j_1}	f_{j_1+1}	$\ldots$	$f_{j_{m-1}}$	$\ldots$	f_{j_m}	$\ldots$	f_n
o_1	0	$\ldots$	0	0	$\ldots$	1	$\ldots$	0	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_1}	0	$\ldots$	**1**	1	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	1
o_{i_1+1}	0	$\ldots$	0	0	$\ldots$	1	$\ldots$	1	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_l-1}	1	$\ldots$	**1**	0	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_l}	1	$\ldots$	**1**	1	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_k	0	$\ldots$	1	0	$\ldots$	0	$\ldots$	0	$\ldots$	0

Формальное определение кандидатов

Для подмножества $A \subseteq O$ объектов его **общим фрагментом** называется подмножество $A' = \{f \in F : \forall o\,[o \in A \Rightarrow (oIf)]\} \subseteq F$. Тогда $\emptyset' = F$.
На самом деле, это определение совпадает с последовательным вычислением побитового умножения строк, соответствующих отобранным во множество A объектов.
Для подмножества $B \subseteq F$ признаков его **сходством** называется подмножество $B' = \{o \in O : \forall f\,[f \in B \Rightarrow (oIf)]\} \subseteq O$. Тогда $\emptyset' = O$.
Операции $' : 2^O \to 2^F$ и $' : 2^F \to 2^O$ называются **полярами** и задают **соответствие Галуа**.

Определение

*Пару $\langle A, B\rangle$ назовем **кандидатом**, если $A = B' \subseteq O$ и $B = A' \subseteq F$.*
*Подмножество $A \subseteq O$ называем **списком родителей** кандидата, а $B \subseteq F$ - **(общим) фрагментом** кандидата.*
*Равенство $A = B'$ соответствует принципу **исчерпываемости**.*

Обучающие выборки средней плотности

Рассмотрим серию испытаний Бернулли $\delta_{1,1}, \ldots, \delta_{n,n}$ с вероятностью успеха $p = \frac{1}{2}$, заполняющую обучающую выборку $I \subseteq O \times F$, где $O = \{o_1, \ldots, o_n\}$ и $F = \{f_1, \ldots, f_n\}$.
Каждый кандидат $\langle A, B \rangle, A \subseteq O, B \subseteq F$ с $|A| = m$ и $|B| = l$ может перестановками строк и столбцов быть преобразован к виду:

$O \times F$	f_1	$\ldots$	f_j	$\ldots$	f_l	f_{l+1}	$\ldots$	f_n
o_1	**1**	$\ldots$	**1**	$\ldots$	**1**	|	$\ldots$	|
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_i	**1**	$\ldots$	**1**	$\ldots$	**1**	|	$\ldots$	|
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_m	**1**	$\ldots$	**1**	$\ldots$	**1**	|	$\ldots$	|
o_{m+1}	$-$	$\ldots$	$-$	$\ldots$	$-$	$+$	$\ldots$	$+$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_n	$-$	$\ldots$	$-$	$\ldots$	$-$	$+$	$\ldots$	$+$

Здесь каждый столбец из $|$ и каждая строка из $-$ должны содержать хотя бы один 0, а прямоугольник из $+$ может содержать любые биты.

Среднее число кандидатов

Вероятность появления кандидата $\langle A, B\rangle, A \subseteq O, B \subseteq F$ с $|A| = m$, $|B| = l$, $|O| = |F| = n$ равна

$$\mathbb{P}[\langle A, B\rangle] = \binom{n}{m} \cdot \binom{n}{l} 2^{-m \cdot l} \cdot (1 - 2^{-m})^{n-l} \cdot (1 - 2^{-l})^{n-m}.$$

Сомножитель $\binom{n}{m} \cdot \binom{n}{l}$ соответствует условиям $|A| = m$, $|B| = l$, а $2^{-m \cdot l} \cdot (1 - 2^{-m})^{n-l} \cdot (1 - 2^{-l})^{n-m}$ возникает из-за требований: верхний левый прямоугольник содержит 1, каждый столбец из $|$ и каждая строка из $-$ должны содержать хотя бы один 0, а прямоугольник из $+$ может содержать любые биты.
Для кандидата $\langle A, B\rangle, A \subseteq O, B \subseteq F$ с $|A| = m$ и $|B| = l$ введём индикатор $[\langle A, B\rangle (I \subseteq O \times F)]$, принимающий значение 1, если для обучающей выборки $I \subseteq O \times F$ пара $\langle A, B\rangle$ является кандидатом, и 0 в противном случае. Ясно, что число кандидатов $|H(I \subseteq O \times F)|$ для $I \subseteq O \times F$ равно $\sum_{\langle A,B\rangle} [\langle A, B\rangle (I \subseteq O \times F)]$. Понятно, что $|H|$ является целочисленной случайной величиной.

Неполиномиальность числа кандидатов

Теорема

$$\mathbb{E}|H| \geq \frac{1}{2e^2\pi} \cdot \exp\{-\ln\log_2 n + 2\log_2 n - 2\log_2 n \cdot \ln\log_2 n + \log_2 n \cdot \ln n\} \cdot (1 + o(1)).$$

Эта теорема является вариантом результата, доказанного в работе Sakurai Taro. **On formal concepts of random formal contexts** // *Information Sciences*. — 2021. — Vol. 578. — p. 615–620.

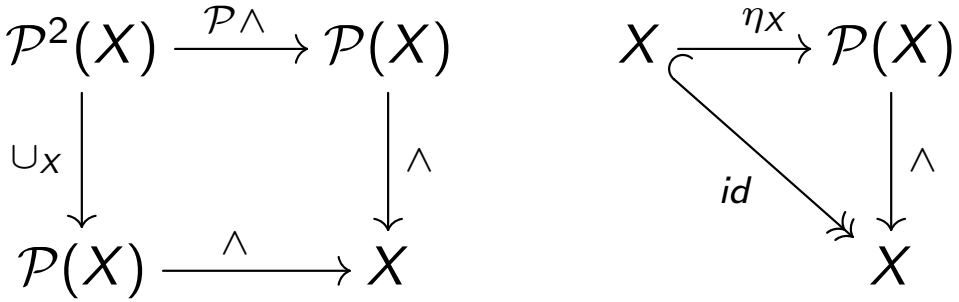
Выделяя главный член асимптотики $\exp\{\log_2 n \cdot \ln n\}$, получаем

$$\mathbb{E}|H| \geq \frac{1}{2e^2\pi} \cdot n^{\log_2 n} \cdot (1 + o(1)).$$

Алгебры над монадой $\langle \mathcal{P}, \eta, \cup \rangle$

Легко проверить, что тройка $\langle \mathcal{P}, \eta, \cup \rangle$, где естественное преобразование $\eta : I_{\mathbf{Set}} \to \mathcal{P}$ с компонентами $\eta_X : X \to \mathcal{P}X$ отображает каждый $x \in X$ в одноэлементное подмножество $\eta_X(x) = \{x\} \in \mathcal{P}X$, задает *монаду* в категории **Set**. Необходимые тождества: $\cup \cdot (\mathcal{P}\cup) = \cup \cdot (\cup\mathcal{P}) : \mathcal{P}^3 \to \mathcal{P}$ и $\cup \cdot (\eta\mathcal{P}) = id = \cup \cdot (\mathcal{P}\eta) : \mathcal{P} \to \mathcal{P}$ представляют собой равенство $\cup_{i \in \cup\{I_j : j \in J\}} S_i = \cup_{j \in J} \{\cup_{i \in I_j} S_i\}$ и равенства $\cup\{A : A \subseteq S\} = S = \cup\{\{x\} : x \in S\}$, соответственно.

Для монады $\langle \mathcal{P}, \eta, \cup \rangle$ в категории **Set** можно определить категорию *алгебр* $\mathbf{Set}^{\mathcal{P}}$ как множество пар $\langle X, \wedge \rangle$, где объект (множество) X называется *носителем алгебры*, а морфизм $\wedge : \mathcal{P}X \to X$ называется *структурным отображением*, причем должны выполняться тождества $\wedge \cdot \mathcal{P}\wedge = \wedge \cdot \cup_X : \mathcal{P}\mathcal{P}X \to X$ и $\wedge \cdot \eta_X = id_X : X \to X$



Полурешетки как алгебры над $\langle \mathcal{P}, \eta, \cup \rangle$

Лемма

*Класс алгебр $\langle X, \wedge \rangle$ над монадой $\langle \mathcal{P}, \eta, \cup \rangle$ в категории **Set** совпадает с полными полурешетками.*

Структурное отображение $\wedge : \mathcal{P}X \to X$ задает частичный порядок: $x \leq y \Leftrightarrow [\wedge\{x, y\} = x]$. Антисимметричность: $x \leq y$ и $y \leq x$ влекут $x = \wedge\{x, y\} = \wedge\{y, x\} = y$. Рефлексивность: $\wedge\{x\} = \wedge \cdot \eta_X(x) = id_X(x) = x$, т.е. $x \leq x$. Транзитивность: $x \leq y$ и $y \leq z$ влекут $\wedge\{x, z\} = \wedge\{\wedge\{x, y\}, z\} = \wedge(\{x, y\} \cup \{z\}) = \wedge\{x, y, z\} = \wedge(\{x\} \cup \{y, z\}) = \wedge\{x, \wedge\{y, z\}\} = \wedge\{x, y\} = x$.
Покажем, что $\wedge S$ - точная нижняя грань для $S \subseteq X$. Для всякого $x \in S$ верно $S \cup \{x\} = S$. Тождество $\wedge \cdot \mathcal{P}\wedge = \wedge \cdot \cup_X$ влечет $\wedge\{\wedge S, x\} = \wedge S$, что означает $\wedge S \leq x$. Пусть для всех $x \in S$ выполняется $\wedge\{x, y\} = y$ (т.е. $y \leq x$). Тогда $\wedge\{\wedge S, y\} = \wedge(S \cup \{y\}) = \wedge(\cup\{\{x, y\} : x \in S\}) = \wedge(\{\wedge\{x, y\} : x \in S\}) = \wedge\{y\} = y$.
То, что полная полурешетка является алгеброй над монадой $\langle \mathcal{P}, \eta, \cup \rangle$, легко проверяется. □

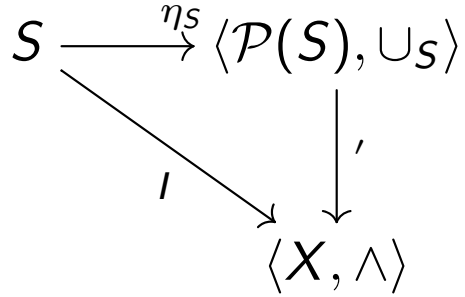
Свободные полные полурешетки

Ganter, Bernhard and Kuznetsov, S.O. **Pattern Structures and Their Projections** *Proc. 9th Conference ICSS 2001*, LNCS 2120. – 2001. – p. 129–142

Теорема

Для монады $\langle \mathcal{P}, \eta, \cup \rangle$ *в категории* ***Set*** *имеет место сопряжение* $\langle F, G; \eta, \varepsilon \rangle : \mathbf{Set} \rightharpoonup \mathbf{Lat}$, *где* $F\langle X, \wedge \rangle = X$ *- забывающий функтор* $F : \mathbf{Lat} \to \mathbf{Set}$, $GX = \langle \mathcal{P}X, \cup_X \rangle$ *- порождающий функтор* $G : \mathbf{Set} \to \mathbf{Lat}$, $\eta_X : X \to \mathcal{P}X$ *- естественное преобразование* $I_{\mathbf{Set}} \to FG$, *а* $\varepsilon_{\langle X, \wedge \rangle} = \wedge$ *- естественное преобразование* $GF \to I_{\mathbf{Lat}}$.

Сопряжение $\langle F, G; \eta, \varepsilon \rangle : \mathbf{Set} \rightharpoonup \mathbf{Lat}$ определяет изоморфизм $\phi : \mathbf{Lat}(\langle \mathcal{P}S, \cup_S \rangle, \langle X, \wedge \rangle) \simeq \mathbf{Set}(S, X)$, для которого $f : S \to X$ отображается в $' = \phi^{-1}(f) = \wedge \cdot \mathcal{P}f : \langle \mathcal{P}S, \cup_S \rangle \to \langle X, \wedge \rangle$.



Пример фантомного сходства

Пусть $O = \{o_1 = B737, o_2 = MC21, o_3 = SJ100, o_4 = A320\}$ будет множеством самолётов, находящихся на ремонте, каждый из которых описывается проблемами из списка
$F = \{f_1 = \text{оперение}, f_2 = \text{двигатель}, f_3 = \text{ругательство}\}$:

$O \mid F$	f_1	f_2	f_3
o_1	1	0	0
o_2	1	0	1
o_3	0	1	1
o_4	0	1	0

Если рассмотреть непустые сходства не менее двух объектов, то мы получим две «настоящие» причины: $\langle \{o_1, o_2\}, \{f_1\} \rangle$ «самолёт с повреждённым оперением не летает» и $\langle \{o_3, o_4\}, \{f_2\} \rangle$ «самолёт с повреждённым двигателем не летает», и одно «фантомное» сходство $\langle \{o_2, o_3\}, \{f_3\} \rangle$ «самолёт, на котором написано ругательство, не летает». Последний кандидат возник из-за случайного совпадения подмножества признаков $\{f_3\}$ у двух примеров o_2 и o_3, каждый из которых имеет свою отличную от других «настоящую» причину.

Контр-примеры и гипотезы

Контр-примером называется объект c, описываемый фрагментом $\{c\}' \subseteq F$ из заданного набора F признаков, но не имеющий целевого свойства.
Говорят, что кандидат $\langle A, B\rangle$, для которого выполняется условие $B \subseteq \{c\}'$, не проходит «запрет контр-примеров».
Любое такое вложение означает, что фрагмент B кандидата $\langle A, B\rangle$ вкладывается в описание $\{c\}'$ контр-примера c. Другими словами, гипотетический механизм есть, а эффект отсутствует. Поэтому сомнительно, что такой кандидат является причиной проявления целевого свойства.
Если кандидат преодолевает все контр-примеры, то он становится **гипотезой** (о причине наличия целевого свойства).
Дополнительно можно потребовать, чтобы число родителей превосходило заданный порог: $|A| \geq b$.
Однако такое ограничение может приводить к «недообучению», когда будут отброшены причины, для которых в обучающей выборке оказалось слишком мало примеров.

Фантомные сходства неустранимы

Теорема

Для $p \geq (-\ln(1-\varepsilon)/n)^{1/b}$ вероятность появления фантомного сходства b случайных p-примеров не меньше, чем $\varepsilon > 0$.

Пусть число n обозначает количество сопутствующих признаков, которыми мы ограничиваемся. Для каждого контр-примера или обучающего примера образуем последовательность n испытаний Бернулли с одинаковой вероятностью успеха p, причём последовательности для разных объектов независимы. Число m будет равно числу контр-примеров.

Теорема

При числе сопутствующих признаков $n \to \infty$ и вероятности появления этих признаков у контр-примеров и обучающих примеров, равной $p = \sqrt{\frac{a}{n}}$ $(a \leq 1)$, вероятность возникновения фантомного сходства двух обучающих примеров, не устранённого никаким из $m = c \cdot \sqrt{n}$ контр-примеров, будет стремиться к

$$1 - e^{-a} - a \cdot e^{-a} \cdot \left[1 - e^{-c \cdot \sqrt{a}}\right] > 0.$$

Переобучение неустранимо: эксперименты

Л.А. Якимова в рамках магистерской диссертации (ОИС РГГУ, 2020) исследовала вопрос о подозрительных на «фантомность» ДСМ-гипотез, порождённых из обучающей выборки Mushrooms.

- Массив взят из Репозитория данных для тестирования алгоритмов машинного обучения http://archive.ics.uci.edu/ml/datasets/Mushroom
- Оцифрованная книга: *Lincoff, G.H.* **The Audubon Society Field Guide to North American Mushrooms**. – NY: Knopf, 1981. – 926 pp.
- Содержит описания 8124 грибов двух видов (съедобные и ядовитые).
- Содержит описания 4208 съедобных и 3916 ядовитых грибов.
- Каждый пример описывался 22 признаками, описывающие различные характеристики грибов. Эти признаки - номинальные, принимающие одно из нескольких значений.

При случайном разделении массива на обучающую и тестовую выборки (с вероятностью $1/2$) количество поганок, предсказанных как съедобные грибы, составило от 7 до 22. В некоторых экспериментах доля ошибок превысила 1%.

Операции «Замыкай-по-одному»

Операция **замыкай-по-одному-вниз** на кандидате $\langle A, B\rangle$ и объекте $o \in O$ порождает кандидат

$$CbODown(\langle A, B\rangle, o) = \langle A, B\rangle \wedge \langle \{o\}'', \{o\}'\rangle = \langle (A \cup \{o\})'', B \cap \{o\}'\rangle.$$

Операция **замыкай-по-одному-вверх** на кандидате $\langle A, B\rangle$ и признаке $f \in F$ порождает кандидат

$$CbOUp(\langle A, B\rangle, f) = \langle A, B\rangle \vee \langle \{f\}', \{f\}''\rangle = \langle A \cap \{f\}', (B \cup \{f\})''\rangle.$$

Ускорение вычислений:
Если $o \in A$, то $CbODown(\langle A, B\rangle, o) = \langle A, B\rangle$. Аналогично, если $f \in B$, то $CbOUp(\langle A, B\rangle, f) = \langle A, B\rangle$.
В случае Булеана вычисления упрощаются
Если $o_j \notin A$, то $CbODown(\langle A, B\rangle, o_j) = \langle A \cup \{o_j\}, B \setminus \{f_j\}\rangle$, и если $f_j \notin B$, то $CbOUp(\langle A, B\rangle, f_j) = \langle A \setminus \{o_j\}, B \cup \{f_j\}\rangle$.

Алгоритм спаривающей цепи Маркова

Data: множество обучающих (+)-примеров; внешние функции
CbOUp(,) и *CbODown*(,) операций «замыкай-по-одному»
Result: кандидат $\langle A, B\rangle$
$O :=$ (+)-примеры, $F :=$ признаки; $I \subseteq O \times F$ - формальный контекст для (+)-примеров;
$R := O \cup F$; $Min := \langle O, O'\rangle$; $Max := \langle F', F\rangle$;
while $(Min \neq Max)$ **do**
  Выбираем случайный элемент $r \in R$;
  if $(r \in O)$ **then**
    $Min := CbODown(Min, r)$; $Max := CbODown(Max, r)$;
  end
  else
    $Min := CbOUp(Min, r)$; $Max := CbOUp(Max, r)$;
  end
end
$\langle A, B\rangle := Min$;

Algorithm 1: Спаривающая цепь Маркова

Спаривающая цепь Маркова

Состоянием изменяемых переменных в цикле (= состоянием цепи Маркова) является упорядоченная пара кандидатов $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$.

Определение

Порядок *на кандидатах:* $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$, *если* $B_1 \subseteq B_2$.

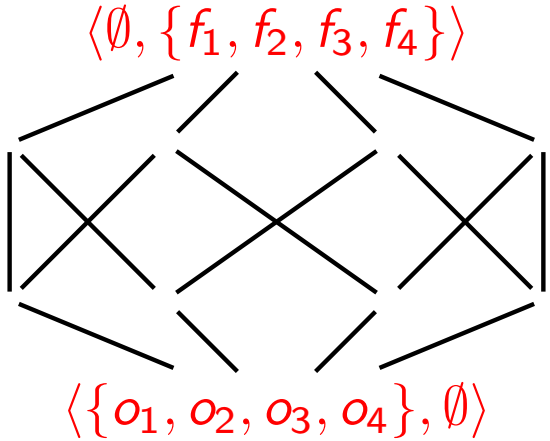
Первоначально меньший кандидат совпадает с наименьшим кандидатом $Min := \langle O, O'\rangle$, а больший - с наибольшим $Max := \langle F', F\rangle$.
В цикле к обоим кандидатам применяется одна и та же операция *CbODown* с выбранным объектом, или *CbOUp* с выбранным признаком.

Лемма

Для всякой упорядоченной пары кандидатов $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ *и любого* $o \in O$ *имеем* $CbODown(\langle A_1, B_1\rangle, o) \leq CbODown(\langle A_2, B_2\rangle, o)$.
Для всякой упорядоченной пары кандидатов $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ *и любого* $f \in F$ *имеем* $CbOUp(\langle A_1, B_1\rangle, f) \leq CbOUp(\langle A_2, B_2\rangle, f)$.

Процесс останавливается, когда меньший кандидат совпадет в большим. Тогда этот общий кандидат и выдается алгоритмом 1.

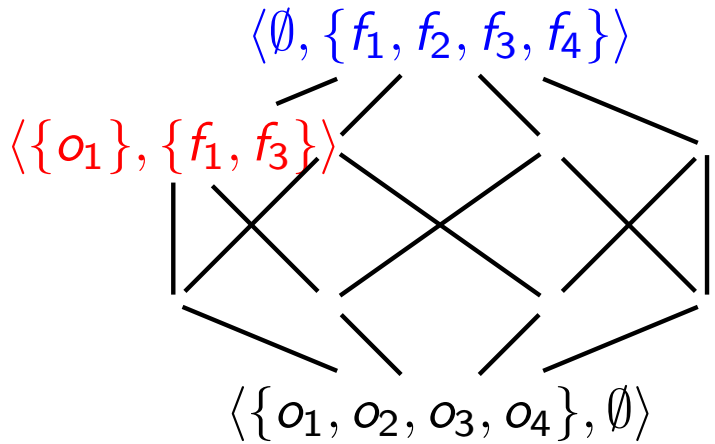
Как работает спаривающая цепь Маркова: шаг 0



Пример

верхний	f_1	f_2	f_3	f_4	нижний	f_1	f_2	f_3	f_4
o_1	1	0	1	0	o_1	***1***	***0***	***1***	***0***
o_2	1	0	0	1	o_2	***1***	***0***	***0***	***1***
o_3	0	1	1	0	o_3	***0***	***1***	***1***	***0***
o_4	0	1	0	1	o_4	***0***	***1***	***0***	***1***

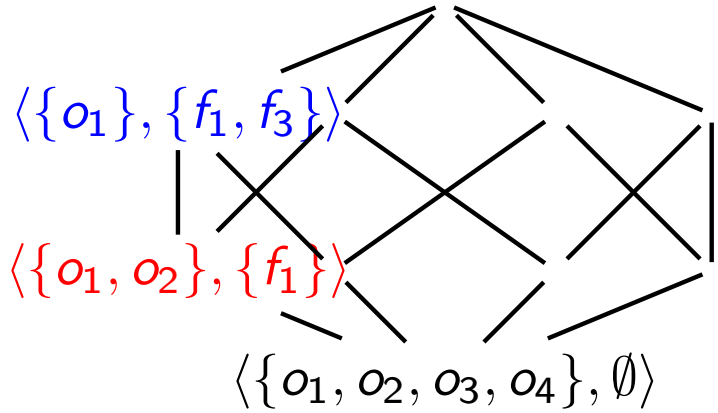
Как работает спаривающая цепь Маркова: выбор o_1



Пример

верхний	f_1	f_2	f_3	f_4	нижний	f_1	f_2	f_3	f_4
o_1	***1***	***0***	***1***	***0***	o_1	***1***	***0***	***1***	***0***
o_2	1	0	0	1	o_2	***1***	***0***	***0***	***1***
o_3	0	1	1	0	o_3	***0***	***1***	***1***	***0***
o_4	0	1	0	1	o_4	***0***	***1***	***0***	***1***

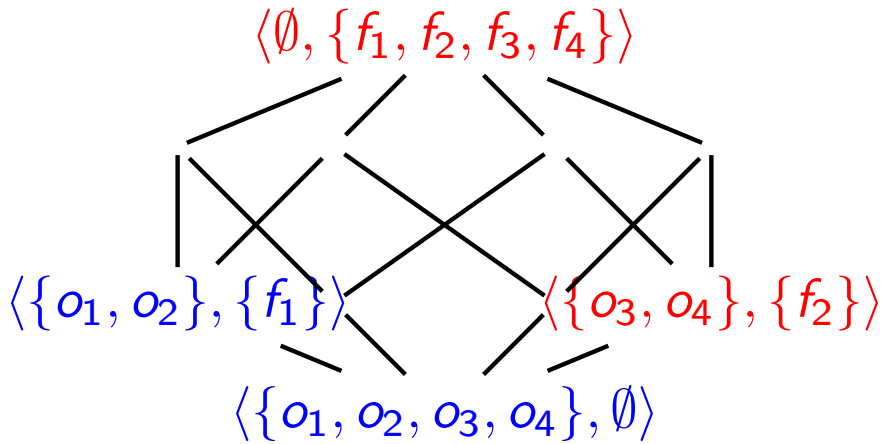
Как работает спаривающая цепь Маркова: выбор o_2



Пример

верхний	f_1	f_2	f_3	f_4	нижний	f_1	f_2	f_3	f_4
o_1	***1***	***0***	***1***	***0***	o_1	***1***	***0***	***1***	***0***
o_2	***1***	***0***	***0***	***1***	o_2	***1***	***0***	***0***	***1***
o_3	0	1	1	0	o_3	***0***	***1***	***1***	***0***
o_4	0	1	0	1	o_4	***0***	***1***	***0***	***1***

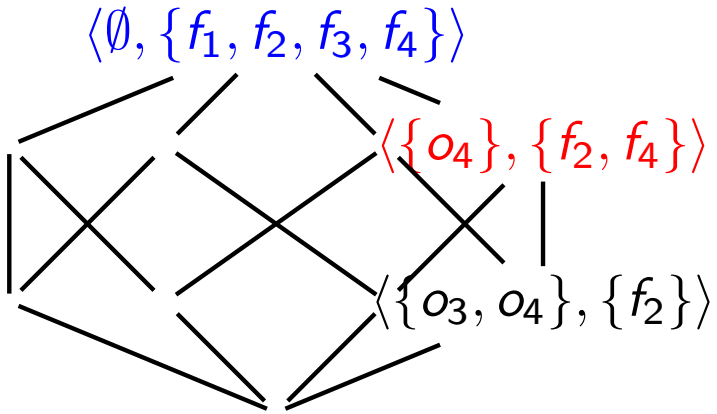
Как работает спаривающая цепь Маркова: выбор f_2



Пример

верхний	f_1	f_2	f_3	f_4	нижний	f_1	f_2	f_3	f_4
o_1	1	0	1	0	o_1	1	0	1	0
o_2	1	0	0	1	o_2	1	0	0	1
o_3	0	1	1	0	o_3	***0***	***1***	***1***	***0***
o_4	0	1	0	1	o_4	***0***	***1***	***0***	***1***

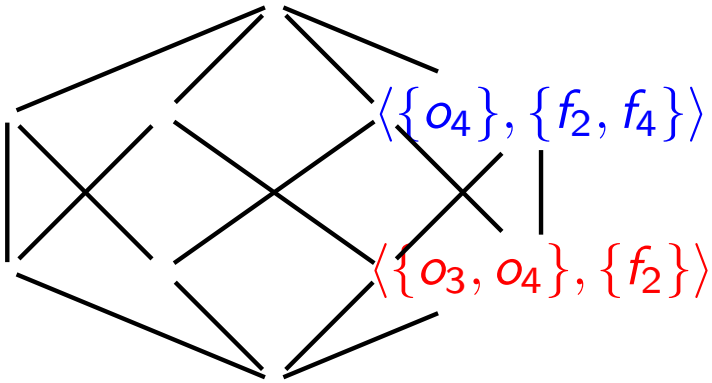
Как работает спаривающая цепь Маркова: выбор o_4



Пример

верхний	f_1	f_2	f_3	f_4	нижний	f_1	f_2	f_3	f_4
o_1	1	0	1	0	o_1	1	0	1	0
o_2	1	0	0	1	o_2	1	0	0	1
o_3	0	1	1	0	o_3	***0***	***1***	***1***	***0***
o_4	***0***	***1***	***0***	***1***	o_4	***0***	***1***	***0***	***1***

Как работает спаривающая цепь Маркова: выбор o_3



Пример

верхний	f_1	f_2	f_3	f_4	нижний	f_1	f_2	f_3	f_4
o_1	1	0	1	0	o_1	1	0	1	0
o_2	1	0	0	1	o_2	1	0	0	1
o_3	***0***	***1***	***1***	***0***	o_3	***0***	***1***	***1***	***0***
o_4	***0***	***1***	***0***	***1***	o_4	***0***	***1***	***0***	***1***

Состояния спаривающей цепи Маркова

Определение

Множество состояний вида $E = \{\langle A, B\rangle = \langle A, B\rangle\}$ *спаривающей цепи Маркова (совпадающих пар кандидатов) называется* ***эргодическим множеством****. Состояния* $s_i \in E$ *называются* ***эргодическими****. Состояние вида* $\{\langle A, B\rangle < \langle C, D\rangle\}$ *называется* ***невозвратным****.*

Заметим, что при первом моменте попадания в любое из эргодических состояний алгоритм спаривающейся цепи Маркова останавливается. Теорема о невозвратных состояниях цепи Маркова может быть сформулирована как утверждение:

$$\lim_{t\to\infty} \mathbb{P}\left[X_t \notin E \mid X_0 = s_i\right] \to 0 \tag{1}$$

для любого $s_i \notin E$.

Останавливаемость спаривающей цепи

Теорема

Алгоритм спаривающей цепи Маркова останавливается с вероятностью 1.

Мы докажем немного более общее утверждение о том, что моменты $T_i(E) = \min\{t : X_t \in E, X_0 = s_i\}$ первого попадания в E, стартуя с некоего невозвратного состояния $s_i = \{\langle A, B\rangle < \langle C, D\rangle\} \notin E$, являются *марковскими*, т.е. $\mathbb{P}[T_i(E) < \infty \mid X_0 = s_i] = 1$.
Имеем $\{X_t \in E, X_0 = s_i\} = \bigcup_{n \le t} U_n(s_i)$, где

$$U_n(s_i) = \{X_n \in E, X_{n-1} \notin E, \ldots, X_1 \notin E, X_0 = s_i\}.$$

Из-за дизъюнктности разных $U_n(s_i)$ и формулы (1) получаем

$$\mathbb{P}\{X_t \in E \mid X_0 = s_i\} = \sum_{n \le t} \mathbb{P}[U_n(s_i) \mid X_0 = s_i] \to 1$$

при $t \to \infty$. Так как $U_n(s_i) = \{T_i(E) = n\}$, то по σ-аддитивности получаем нужное утверждение. □

Длина траекторий для Булеана

Теорема

Средняя длина траекторий $\sum_{j=1}^{n} T_j$ *для* n*-мерного гиперкуба равна*

$$\mathbb{E}[\sum_{j=1}^{n} T_j] = \sum_{j=1}^{n} \frac{n}{j} \approx n \cdot \ln(n) + n \cdot \gamma + \frac{1}{2}.$$

При $n = 32$ средняя длина $\sum_{j=1}^{n} \frac{n}{j} \leq 130$.

Теорема

$\mathbb{P}\left[\sum_{j=1}^{n} T_j \geq (1+\varepsilon) \cdot n \cdot \ln(n)\right] \to 0$ *при* $n \to \infty$ *для любого* $\varepsilon > 0$.

При $n = 32$ Булев гиперкуб содержит $4,294,967,296$ вершин. 1000 траекторий спаривающей цепи Маркова породят около $260,000$ элементов гиперкуба.

Рекуррентные соотношения для средней длины

Получим рекуррентные соотношения для средней длины траектории спаривающей цепи Маркова с использованием формулы полной вероятности и марковости момента склеивания.
Из-за аддитивности среднего имеем $\mathbb{E}[T_i(E)] = \sum_{n=1}^{\infty} n \cdot \mathbb{P}[U_n(s_i)|X_0 = s_i]$, где $U_n(s_i) = \{X_n \in E, X_{n-1} \notin E, \ldots, X_1 \notin E, X_0 = s_i\}$. Тогда

$$\mathbb{E}[T_i(E)] = \sum_{n=1}^{\infty} n \cdot \mathbb{P}[U_n(s_i)|X_0 = s_i] =$$

$$= \sum_{n=1}^{\infty} \mathbb{P}[U_n(s_i)|X_0 = s_i] + \sum_{n=2}^{\infty} (n-1) \cdot \mathbb{P}[U_n(s_i)|X_0 = s_i] =$$

$$= 1 + \sum_{k=1}^{\infty} k \cdot \mathbb{P}[X_{k+1} \in E, X_k \notin E, \ldots, X_1 \notin E|X_0 = s_i] = 1+$$

$$+ \sum_{s_j \notin E} \sum_{k=1}^{\infty} k \cdot \mathbb{P}[X_{k+1} \in E, X_k \notin E, \ldots, X_2 \notin E|X_1 = s_j] \cdot \mathbb{P}[X_1 = s_j|X_0 = s_i] =$$

$$= 1 + \sum_{s_j \notin E} \mathbb{E}[T_j(E)] \cdot \mathbb{P}[X_1 = s_j|X_0 = s_i].$$

Линейный порядок из $n+1$ элемента

Применим теперь полученные рекуррентные соотношения к случаю линейного порядка из $n+1$ элемента. Обучающая выборка для этой решетки кандидатов имеет вид

$O \mid F$	f_1	f_2	f_3	$\cdots$	f_n
o_1	0	1	1	$\cdots$	1
o_2	0	0	1	$\cdots$	1
o_3	0	0	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_n	0	0	0	$\cdots$	0

Определение

Расстояние $\rho(\langle A, B\rangle, \langle C, D\rangle)$ *между кандидатами* $\langle A, B\rangle$ *и* $\langle C, D\rangle$ *определяется как число позиций, в которых отличаются битовые строки* B *и* D.

Соотношения для линейного порядка

Определение

Введем условный момент ***склеивания*** T_m, *когда впервые будет* $X_{T_m} \in E$ *при условии, что для* $X_0 = \langle A, B\rangle \leq \langle C, D\rangle$ *расстояние равно* $\rho(X_0) = \rho(\langle A, B\rangle, \langle C, D\rangle) = m$.

Момент T_m не зависит от выбора $X_0 = \langle A, B\rangle \leq \langle C, D\rangle$ с $\rho(X_0) = m$ из-за одинаковости числа вариантов для перехода.
Заметим, что для момента T остановки алгоритма спаривающей цепи Маркова выполняется $T = T_n$. Поэтому нужно вычислить $\mathbb{E}T = g_n(n)$, где введено обозначение $g_n(m) = \mathbb{E}T_m$.

Лемма

Для средних условных моментов $g_n(m) = \mathbb{E}T_m$ *имеет место рекуррентное соотношение*

$$g_n(m) = 1 + \frac{n-m}{2n} \cdot g_n(m) + \frac{1}{n} \cdot [g_n(m-1) + \cdots + g_n(1)]\,.$$

Средняя длина для линейного порядка

Теорема

Среднее время склеивания для $n+1$-элементного линейного порядка равно

$$\mathbb{E}[T] = \frac{2n(2n+1)}{(n+1)(n+2)} \leq 4.$$

Докажем сначала, что $g_n(1) = \frac{2n}{n+1}$. Имеем $g_n(1) = 1 + \frac{n-1}{2n} \cdot g_n(1)$, откуда получаем требуемую оценку для $g_n(1)$.
Умножив обе части рекуррентности из леммы 3 на $2n$, получим

$$(n+m) \cdot g_n(m) = 2n + 2[g_n(m-1) + g_n(m\text{-}2) + \ldots + g_n(1)].$$

Вычтем равенство $(n+m-1) \cdot g_n(m-1) = 2n + 2[g_n(m-2) + \ldots + g_n(1)]$.
Тогда $(n+m) \cdot g_n(m) - (n+m-1) \cdot g_n(m-1) = 2 \cdot g_n(m-1)$, т. е.
$g_n(m) = \frac{n+m+1}{n+m} \cdot g_n(m-1)$. Применяя телескопическое тождество, имеем

$$g_n(n) = \frac{2n+1}{2n} \cdot g_n(n-1) = \ldots = \frac{2n+1}{2n} \cdot \frac{2n}{2n-1} \cdot \ldots \cdot \frac{n+3}{n+2} \cdot h_n(1).$$

Но $g_n(1) = \frac{2n}{n+1}$, поэтому $\mathbb{E}\,T = g_n(n) = \frac{2n(2n+1)}{(n+1)(n+2)} < 4$. □

Обогащённые обучающие выборки

Обогатим обучающую выборку, добавляя к каждому бинарному признаку $f_j \in F$ его отрицание $\overline{f_j}$. Эта конструкция часто имеет полезный смысл: мы хотим, чтобы отсутствие признака могло бы быть частью причины проявления целевого свойства.
Обогащенное множество признаков будем обозначать F^+, и обозначим его мощность через $2n = |F^+|$. Обычно $2n \ll k$, что мы будем предполагать в дальнейшем. Обогатим выборку $I \subseteq O \times F^+$ по правилу:

$$oI\overline{f_j} \Leftrightarrow \neg(oIf_j).$$

Разделим все невозвратные состояния на 2 группы:

$$V = \{s = (\langle A, B\rangle < \langle C, D\rangle) : \exists f \in F^+ [f \in B]\}$$

и

$$W = \{s = (\langle A, B\rangle < \langle C, D\rangle) : \forall f \in F^+ [f \notin B]\}.$$

Ясно, что состояние $s_0 = (\bot < \top) \in W$.

Некоторые леммы

Определение

Введём частичный ***порядок между состояниями*** $s_i = (\langle A_i, B_i\rangle \leq \langle C_i, D_i\rangle)$ *и* $s_j = (\langle A_j, B_j\rangle \leq \langle C_j, D_j\rangle)$ *спаривающей цепи:*

$$s_j \leqslant s_i \Leftrightarrow \langle A_i, B_i\rangle \leq \langle A_j, B_j\rangle \leq \langle C_j, D_j\rangle \leq \langle C_i, D_i\rangle.$$

Лемма

Для любой упорядоченной пары состояний $s_j \leqslant s_i$ *и любых* $o \in O$ *выполняется* $CbODown(s_j, o) \leqslant CbODown(s_i, o)$, *а для любых* $f \in F$ *верно* $CbOUp(s_j, f) \leqslant CbOUp(s_i, f)$.

Лемма

Для любой упорядоченной пары невозвратных состояний $s_j \leqslant s_i$ *спаривающей цепи Маркова выполняется* $\mathbb{E}T_j(E) \leq \mathbb{E}T_i(E)$.

Доказательство второй леммы

Зададим спаренное блуждание упорядоченной пары состояний $X_t \leqslant Y_t$

$$\mathbb{P}\left[X_1 = s'_j, Y_1 = s'_i \mid X_0 = s_j, Y_0 = s_i\right] =$$

$$= \begin{cases} \frac{m}{n+k}, & m = |\{o \in O : s'_j = CbODown(s_j, o), s'_i = CbODown(s_i, o)\}| + \\ & +|\{f \in F : s'_j = CbOUp(s_j, f), s'_i = CbOUp(s_i, f)\}| \\ 0, & \neg\exists o \in O \left[s'_j = CbODown(s_j, o), s'_i = CbODown(s_i, o)\right] \& \\ & \&\neg\exists f \in F \left[s'_j = CbOUp(s_j, f), s'_i = CbOUp(s_i, f)\right] \end{cases}.$$

Так как из $\langle A_i, B_i\rangle = \langle C_i, D_i\rangle$ для $\langle A_i, B_i\rangle \leq \langle A_j, B_j\rangle \leq \langle C_j, D_j\rangle \leq \langle C_i, D_i\rangle$ следует $\langle A_i, B_i\rangle = \langle A_j, B_j\rangle = \langle C_j, D_j\rangle = \langle C_i, D_i\rangle$, то

$$\mathbb{P}\left[X_t = Y_t \in E \mid X_0 = s_j \leqslant Y_0 = s_i\right] \geq \mathbb{P}\left[Y_t \in E \mid X_0 = s_j \leqslant Y_0 = s_i\right].$$

Но $X_t \notin E \Leftrightarrow T_i(E) > t$ и $Y_t \notin E \Leftrightarrow T_j(E) > t$. Поэтому

$$\mathbb{E}T_j(E) = \sum_{t=0}^{\infty} \mathbb{P}\left[T_j(E) > t \mid X_0 = s_j, Y_0 = s_i\right] \leq$$

$$\leq \sum_{t=0}^{\infty} \mathbb{P}\left[T_i(E) > t \mid X_0 = s_j, Y_0 = s_i\right] = \mathbb{E}T_i(E). \quad \square$$

Спуск донизу

По предыдущей лемме для любого $s_j \in W$ выполнено $\mathbb{E}T_j(E) \leq \mathbb{E}T_0(E)$.
По определению множества V и предыдущей лемме для любого $s_j \in V$ выполнено $\mathbb{E}T_j(E) \leq \mathbb{E}T_i(E)$, где $s_i = (\langle \{f\}', \{f\}'' \rangle < \top) \in V$ для любого $f \in B$ при $s_j = \langle A, B \rangle$.
Введём целочисленную случайную величину Z, принимающую значение m на множестве $\{X_m = (\bot = \bot), X_{m-1} \notin V, \ldots, X_1 \notin V, X_0 = s_0\}$, которая определяет минимальное число шагов алгоритма спаривающей цепи Маркова по состояниям из $X_t \in W$ до тех пор, пока не получим $X_m = (\bot = \bot)$.

Лемма

Для выборки $I \subseteq O \times F^+$ с $2n = |F^+| \leq k = |O|$ имеем

$$\mathbb{E}Z = \sum_{l=1}^{\infty} \mathbb{P}[Z \geq l] \leq (k + 2n) \cdot \left(\ln(2n) + \frac{1}{1 - e^{-1}} \right).$$

Оценка средней длины спуска

Разобьём слагаемые в сумме на непересекающиеся подмножества $I_0 \sqcup \bigsqcup_{r=1}^{\infty} I_r$, где $I_0 = \{1 \le l < (k+2n)\cdot \ln(2n)\}$ и

$$I_r = \{(k+2n)\cdot(\ln(2n)+r-1) \le l < (k+2n)\cdot(\ln(2n)+r)\}.$$

Ясно, что $\sum_{l=1}^{(k+2n)\cdot\ln(2n)-1} \mathbb{P}[Z \ge l] \le (k+2n)\cdot\ln(2n)$.
Для того, чтобы произошло событие $Z \ge l$ необходимо, чтобы нашёлся хотя бы один признак (из $2n$), чтобы не был выбран ни один пример в серии длины l, в котором этого признака нет. Поэтому по неравенству Буля

$$\mathbb{P}[Z > l] \le 2n \cdot \left(1 - \frac{1}{k+2n}\right)^l.$$

$$\sum_{l=(k+2n)\cdot(\ln(n)+r-1)}^{(k+2n)\cdot(\ln(2n)+r)-1} \mathbb{P}[Z > l] \le \sum_{l=(k+2n)\cdot(\ln(2n)+r-1)}^{(k+2n)\cdot(\ln(2n)+r)-1} 2n \cdot \left(1 - \frac{1}{k+2n}\right)^l \le$$

$$\le (k+2n)\cdot 2n \cdot \left(1 - \frac{1}{k+2n}\right)^{(k+2n)\cdot(\ln(2n)+r-1)} \le$$

$$\le (k+2n)\cdot e^{\ln 2n}\cdot e^{-(\ln(2n)+r-1)} = (k+2n)\cdot e^{-r+1}.$$

Разбор случаев

Суммируя по r, получим

$$\sum_{l=(k+2n)\cdot\ln(2n)}^{\infty} \mathbb{P}[Z \geq l] \leq (k+2n) \cdot \sum_{r=1}^{\infty} e^{-r+1} = \frac{k+2n}{1-e^{-1}}.$$

□

Обозначим границу из предыдущей леммы через *tail*.
Рассмотрим непересекающиеся события

$$H_l(s_j) = \{X_l = s_j \in V, X_{l-1} \notin V, \ldots, X_1 \notin V, X_0 = s_0\}.$$

Обозначим событие $\{X_{t+l} \in E, X_{t+l-1} \notin E, \ldots, X_{l+1} \notin E\} \cap H_l(s_j)$ через $G_{t,l}(s_j)$, а $\bigsqcup_{s_j \in V} G_{t,l}(s_j)$ - через $U_{t,l}$.
Очевидно имеем разложение события на дизъюнктные части

$$\{X_{t+l} \in E, X_{t+l-1} \notin E, \ldots, X_0 = s_0\} =$$
$$= \bigsqcup_{s_j \in V} (\{X_{t+l} \in E, X_{t+l-1} \notin E, \ldots, X_{l+1} \notin E\} \cap H_l(s_j)) \sqcup$$
$$\sqcup \{X_{t+l} = (\bot = \bot), X_{t+l-1} \notin V, \ldots, X_1 \notin V, X_0 = s_0\}.$$

Средняя длина траектории

Нас интересует

$$\mathbb{E}T_0(E) = \sum_{m=1}^{\infty} m \cdot \mathbb{P}\left[X_m \in E, X_{m-1} \notin E, \ldots, X_1 \notin E \mid X_0 = s_0\right]. \qquad (2)$$

Ясно, что $\mathbb{E}T_0(E) = \mathbb{E}T_0'(E) + \mathbb{E}Z$, где $T_0'(E)$ - ограничение $T_0(E)$ на $\bigsqcup_{t=1}^{\infty}\bigsqcup_{l=1}^{\infty} G_{t,l}$.

Теорема

Для обогащённой выборки $I \subseteq O \times F^+$ имеем верхнюю границу на среднюю длину траектории алгоритма спаривающей цепи Маркова при $2n = |F^+|$ и $k = |O|$

$$\mathbb{E}T_0 \leq \frac{(k+2n)(k^2+k(2n+1)+4n^2+2n)}{2n(k^2+k+2n)} + \frac{(k+1)(k+2n)}{k^2+k+2n}\,tail.$$

Доказательство теоремы

Введём обозначения $R = \sum_{l=1}^{n} \frac{1}{k+2n} \left(T_{f_l} + T_{\bar{f}_l}\right)$, где $T_{f_l} = T_i(E)$ для $s_j = (\langle \{f_l\}', \{f_l\}'' \rangle < \top)$, аналогично для $T_{\bar{f}_l}$.
Тогда по марковскому свойству и однородности

$$\mathbb{E}T_0'(E) = \sum_{t=1}^{\infty}\sum_{l=1}^{\infty}(t+l)\cdot\mathbb{P}U_{t,l} =$$

$$= \sum_{t=1}^{\infty} t \cdot \sum_{s_j \in V} \mathbb{P}\left[X_t \in E, X_{t-1} \notin E, \ldots, X_1 \notin E \mid X_0 = s_j\right] \cdot \sum_{l=1}^{\infty} \mathbb{P}\left[H_l(s_j)\right] +$$

$$+ \sum_{l=1}^{\infty} l \cdot \sum_{s_j \in V} \mathbb{P}\left[H_l(s_j)\right] \cdot \sum_{t=1}^{\infty} \mathbb{P}\left[X_t \in E, X_{t-1} \notin E, \ldots, X_1 \notin E \mid X_0 = s_j\right] \leq$$

$$\leq \sum_{s_j \in V} \mathbb{E}T_j(E) \cdot \mathbb{P}\left[X_1 = s_j \mid X_0 = s_0\right] + \sum_{s_j \in V}\sum_{l=1}^{\infty} l \cdot \mathbb{P}\left[H_l(s_j)\right] \leq$$

$$\leq \mathbb{E}R + \frac{k+2n}{2n},$$

где последнее слагаемое - среднее для геометрически распределенной случайной величины ожидания выбора первого признака.

Окончание доказательства теоремы

По формуле полной вероятности имеем

$$\mathbb{E}T_{f_l} \leq 1 + \sum_{i=1}^{n} \frac{1}{k+2n}\left(\mathbb{E}T_{f_i} + \mathbb{E}T_{\bar{f}_i}\right) - \frac{1}{k+2n}\cdot \mathbb{E}T_{\bar{f}_l} + \frac{k}{k+2n}\mathbb{E}T_0(E).$$

Отсюда следует

$$\mathbb{E}R \leq \frac{2n}{k+2n}\left[1 + \mathbb{E}R + \frac{k}{k+2n}\mathbb{E}T_0(E)\right] - \frac{1}{k+2n}\mathbb{E}R.$$

Поэтому

$$\frac{k+1}{k+2n}\mathbb{E}R \leq \frac{2n}{k+2n} + \frac{2nk}{(k+2n)^2}\mathbb{E}T_0(E).$$

Подставляя $\mathbb{E}R \leq \frac{2n}{k+1} + \frac{2nk}{(k+1)(k+2n)}\mathbb{E}T_0(E)$ в неравенство

$$\mathbb{E}T_0(E) \leq \mathbb{E}R + \frac{k+2n}{2n} + tail,$$

получим

$$\frac{k^2+k+2n}{(k+1)(k+2n)}\mathbb{E}T_0(E) \leq \frac{2n}{k+1} + \frac{k+2n}{2n} + tail,$$

что и приводит к нужной оценке.

Ленивые вычисления кандидатов

Согласно определению

$$CbODown(\langle A, B\rangle, o) = \langle (A \cup \{o\})'', B \cap \{o\}'\rangle.$$

Вычисление сходства $B \cap \{o\}' = (A \cup \{o\})'$ соответствует побитовому умножению соответствующих строк, но операция $(A \cup \{o\})''$ формирования нового списка родителей может потребовать побитово перемножить с полученным ранее сходством почти все объекты, чтобы проверить, обладает ли еще какой-нибудь объект полученным сходством.
Для улучшения ситуации предлагается (лениво) откладывать вычисления второй производной, пока последовательный выбор нескольких объектов для *CbODown* не сменится выбором признака с переходом к операции *CbOUp*. Этом случае нужно использовать равенство $(A \cup \{o\})'' = (B \cap \{o\}')'$.
Используя равенство $(B \cup \{f\})'' = \left(A \cap \{f\}'\right)'$, можно аналогично откладывать вычисления второй производной до тех пор, пока выбор нескольких признаков для *CbOUp* не сменится выбором объекта с переходом к операции *CbODown*.

Алгоритм ленивой спаривающей цепи Маркова

Data: множество обучающих (+)-примеров
Result: кандидат $\langle A_1, B_1 \rangle$
$R := O \cup F$; $\langle A_1, B_1 \rangle := \langle O, O' \rangle$; $\langle A_2, B_2 \rangle := \langle F', F \rangle$; $moveUp := true$;
while $(\langle A_1, B_1 \rangle \neq \langle A_2, B_2 \rangle)$ **do**
  Выбираем случайный элемент $r \in R$;
  if $(r \in O \&\& moveUp)$ **then**
    $B_1 := A_1'$; $B_2 := A_2'$; $moveUp := false$;
  end
  if $(r \in O)$ **then**
    $B_1 := B_1 \cap (\{r\}')$; $B_2 := B_2 \cap (\{r\}')$;
  end
  if $(r \in F \&\& !moveUp)$ **then**
    $A_1 := B_1'$; $A_2 := B_2'$; $moveUp := true$;
  end
  if $(r \in F)$ **then**
    $A_1 := A_1 \cap (\{r\}')$; $A_2 := A_2 \cap (\{r\}')$;
  end
end

Algorithm 2: Ленивая спаривающая цепь Маркова

Выигрыш от ленивых вычислений: теория

Теорема

В ленивой схеме вычислений на каждую пару применений операции замыкания (одной в CbOUp и одной в CbODown) в среднем в классической схеме мы будем делать $\frac{(n+k)^2}{k \cdot n}$ операций замыкания, где k - число обучающих примеров, а n - число признаков, используемых для описания объектов.

Так как

$$\frac{(n+k)^2}{k \cdot n} - 4 = \frac{(n-k)^2}{k \cdot n} \geq 0,$$

то

$$\frac{(n+k)^2}{k \cdot n} \geq 4.$$

В худшем случае ($k = n$) это сокращение вызовов трудоемкой операции в среднем $\frac{4}{2} = 2$ раза.
Чем сильнее различаются k и n, тем больше средний выигрыш от применения ленивой схемы вычислений.

Выигрыш от ленивых вычислений: проверка

Л.А. Якимова в рамках выпускной квалификационной работы (ОИС РГГУ, 2018) исследовала вопрос фактическом преимуществе вычислений ВКФ-кандидатов с помощью ленивого варианта спаривающей цепи Маркова.

- Исходные данные включают описания 8124 грибов, разделенные на две категории (съедобные и ядовитые).
- Обучающая выборка содержит $k = 4208$ примеров (съедобные грибы).
- Контр-примеров (ядовитые грибы) содержится 3916 штук.
- Каждый пример описывался 22 признаками, описывающие различные характеристики грибов (цвет, форма шляпки, места произрастания,частота встречаемости и т.п.). ВКФ-система закодировала эти признаки битовыми строками длины $n = 124$ бит.
- Число $\frac{(n+k)^2}{k \cdot n} = 35,96$. Поэтому максимальное ускорение может составить примерно $\frac{36}{2} = 18$ раз.
- На практике ускорение вычислений по ленивой схеме превысило 17 раз!

Остановленная цепь Маркова

Для устранения слишком длинных траекторий спаривающей цепи Маркова:

Определение

Если $T_1, \ldots, T_r$ – независимые целочисленные случайные величины, имеющие распределение времени склеивания T, то ***верхняя граница склеивания*** *по r испытаниям определяется как $\hat{T} = T_1 + \cdots + T_r$.*

Остановленная цепь Маркова $\mu(\hat{T})$: если спаривающая цепь Маркова μ не склеивается до времени $\hat{T}$, то начинаем заново, иначе выдаём $\langle A_1(T), B_1(T)\rangle = \langle A_2(T), B_2(T)\rangle$ $(T \leq \hat{T})$.

Теорема

Для верхней границы $\hat{T}$ склеивания по $r > 1$ испытаниям любого $\| \mu - \mu(\hat{T}) \|_{TV} \leq \frac{1}{2^r - 1}$ в метрике тотальной вариации.

Алгоритм индуктивного обобщения

Data: множество обучающих (+)- и (-)-примеров; число N порождаемых ВКФ-гипотез
Result: выборка S ВКФ-гипотез
$O :=$ (+)-примеры, $F :=$ признаки; $I \subseteq O \times F$ (обучающая выборка из (+)-примеров); $C :=$ (-)-примеры; $S := \emptyset$; $i := 0$;
while $(i < N)$ **do**
  породить кандидата $\langle A, B \rangle$ с помощью цепи Маркова;
  $hasObstacle :=$ **false**;
  for $(c \in C)$ **do**
    if $(B \subseteq c')$ **then**
      $hasObstacle :=$ **true**;
    end
  end
  if $(hasObstacle =$ ***false***$)$ **then**
    $S := S \cup \{\langle A, B \rangle\}$;
    $i := i + 1$;
  end
end

Algorithm 3: Процедура индуктивного обобщения

Алгоритм предсказания по аналогии

Data: расширенная выборка S^+ ВКФ-гипотез, файл (τ)-примеров
Result: предсказанные свойства (τ)-примеров
$X := (\tau)$-примеры;
for $(o \in X)$ **do**
  $PredictPositively(o) :=$ **false**;
  for $(\langle A, B \rangle \in S^+)$ **do**
    if $(B \subseteq o')$ **then**
      $PredictPositively(o) :=$ **true**;
    end
  end
end

Algorithm 4: Процедура предсказания по аналогии

Надёжность гипотез

Зафиксируем $\varepsilon > 0$ - точность предсказания.

Определение

*Объект o назовем ε-**важным**, если суммарная вероятность появления таких гипотез $\langle A, B \rangle$, которые предсказывают его положительно, будет больше ε.*

Теорема

Для n признаков и любых $\varepsilon > 0$ и $1 > \delta > 0$ достаточно породить

$$N \geq \frac{n \cdot \ln 2 - \ln \delta}{\varepsilon}$$

гипотез, чтобы вероятностью $> 1 - \delta$ все ε-важные объекты могли быть предсказаны положительно.

Минимизация эмпирического риска

Так как априорная ВПК-оценка для N сильно завышена , предлагается попеременно запускать алгоритм индуктивного порождения гипотез (удваивая каждый раз их число) и следующего алгоритма

Data: расширенная выборка S^+ гипотез, файл (+)-примеров
Result: значение k эмпирического риска
$O :=$ (+)-примеры;
$k := 1$; $d = |O|$;
for $(o \in O)$ **do**
  for $(\langle A, B\rangle \in S^+)$ **do**
    if $(B \subseteq \{o\}')$ **then**
      $k = k - \frac{1}{d}$;
      break;
    end
  end
end

до тех пор, пока эмпирический риск не прекратит уменьшаться. Это - вариант метода минимизации эмпирического риска В.Н. Вапника–А.Я. Червоненкиса.

Абдуктивное объяснение по В.К. Финну

Для ДСМ-метода В.К. Финн предложил использовать следующую процедуру

Data: расширенная выборка S^+ гипотез, файл (+)-примеров
Result: набор необъясненных примеров
$O :=$ (+)-примеры;
for $(o \in O)$ **do**
  explained(o):=**false**;
  for $(\langle A, B\rangle \in S^+)$ **do**
    if $(B \subseteq \{o\}')$ **then**
      explained(o):=**true**;
      break;
    end
  end
end

Действия и увеличение числа признаков

Признаки, описывающие грибы

1. форма, поверхность и цвет шляпки
2. синяки
3. запах
4. присоединение, разреженность, размер и цвет пластинок
5. форма ножки, корень ножки
6. поверхность ножки над и под колечком, цвет ножки над и под колечком
7. тип и цвет пленки
8. число и тип колечек
9. цвет спор
10. частота встречаемости
11. места произрастания

Нужно ли сближаться??

Что видно издалека?

1. форма и цвет шляпки
2. тип и цвет пленки
3. частота встречаемости
4. места произрастания

Что видно вблизи?

1. поверхность шляпки
2. синяки
3. форма ножки
4. цвет ножки над и под колечком

Нужно ли срезать гриб?

Что видно после срезания гриба?

1. запах
2. присоединение, разреженность, размер и цвет пластинок
3. корень ножки
4. поверхность ножки над и под колечком
5. число и тип колечек

Как узнать?

1. цвет спор