Модель машинного обучения,
основанного на операции сходства

Дмитрий Виноградов

ФИЦ ИУ РАН

13 декабря 2017 г.

Машинное обучение и проблема индукции

Машинное обучение - область прикладной математики, целью которой является разработка методов восстановления функциональных зависимостей из эмпирических данных.
Математическая статистика решает эту задачу, выбирая вероятностную модель из семейства распределений, заданных с точностью до параметров.
Нейронные сети строят аппроксимацию функциональной зависимости путем выбора топологии сети, функций активации и весов.
Алгебраические методы пытаются построить приближенную зависимость как элемент некоторой конечно-порожденной алгебры.
Метрические методы пытаются сгруппировать данные и найти «идеального» представителя в каждом классе.
Логические методы классификации пытаются найти общие закономерности, записанные в логическом языке, чтобы из них выводились эмпирические данные. «Проблема индукции» состоит в невозможности обосновать схемы вывода средствами семантики классической логики.

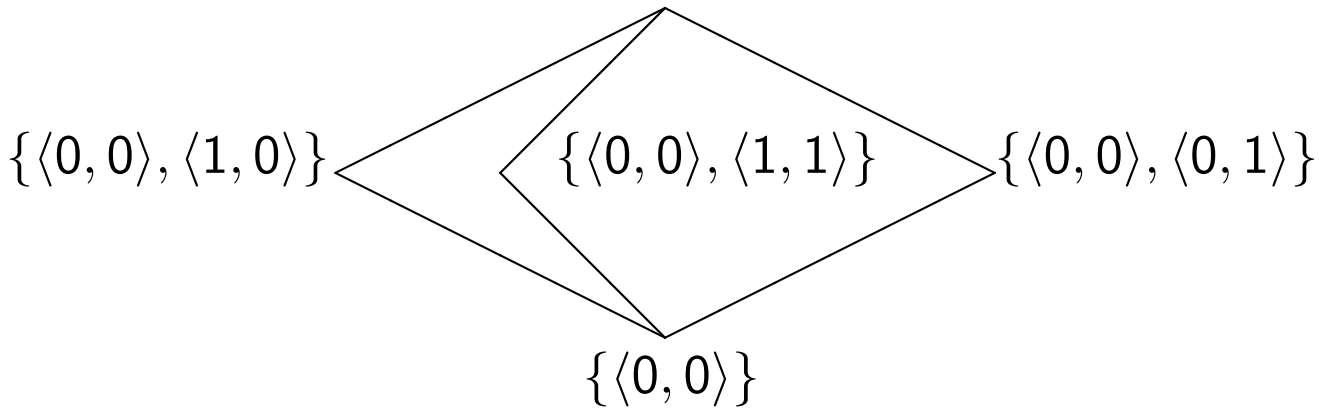
Что значит быть похожими?

Методы кластер-анализа, многомерного шкалирования используют бинарное отношение между объектами (близость, различие, расстояние).
Логические методы используют бинарную операцию сходства, выделяющую общую часть в логическом описании (антиунификация).
Более общая постановка идет из прикладной теории решеток:
Сходство - это бинарная операция $\cap : X \times X \to X$, задающая структуру нижней полурешетки с наименьшим элементом $\emptyset$. Элементы X называются *фрагментами*. Мотивировка этого названия идет из фармакологии, где ищутся «фармакофоры» - специальные структурные фрагменты химических молекул, вызывающие проявления лекарственных или контр-продуктивных свойств. Мы применяем операцию сходства к нескольким объектам последовательно, чтобы найти их общий фрагмент.

Сходство

Свойства нижней полурешетки (идемпотентность, коммутативность и ассоциативность) нужны для независимости результата от порядка применения операции сходства $\cap : X \times X \to X$.
Конечную нижнюю полурешетку легко превратить в решетку (с операцией $\cup : X \times X \to X$), добавив наибольший элемент, если его нет.
Без ограничения общности (по теореме Р. Вилле) фрагменты можно представлять битовыми строками с операцией побитового умножения как сходством. Но операция $\cup$ не будет побитовой дизъюнкцией!

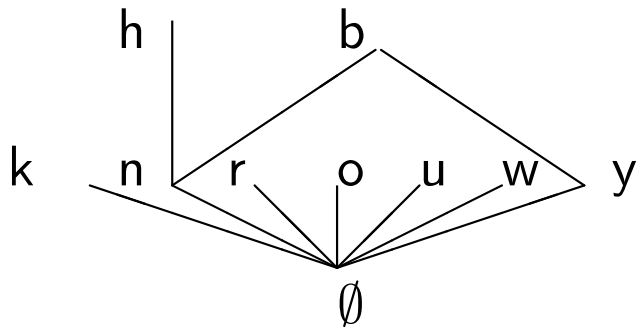
$$Z_2^2 = \{\langle 0,0\rangle, \langle 1,0\rangle, \langle 0,1\rangle, \langle 1,1\rangle\}$$



Пример кодирования: грибы

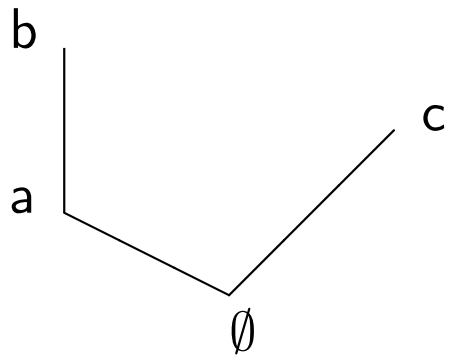
1. форма, поверхность и цвет шляпки
2. синяки?
3. запах
4. присоединение, разреженность, размер и цвет пластинок
5. форма, корень ножки,
6. поверхность ножки над и под колечком, цвет ножки над и под колечком
7. тип и цвет пленки
8. число и тип колечек
9. цвет спор
10. частота встречаемости
11. места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

Пример кодирования: цвет спор



$black = k$	10000000
$brown = n$	01000000
$buff = b$	01000001
$chocolate = h$	01100000
$green = r$	00010000
$orange = o$	00001000
$purple = u$	00000100
$white = w$	00000010
$yellow = y$	00000001

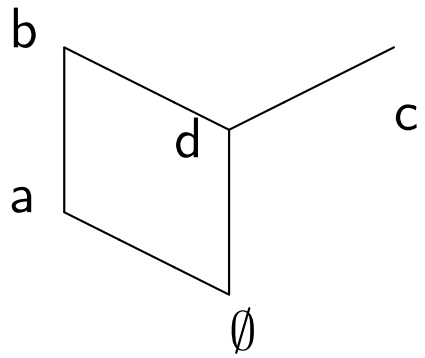
Как кодировать?



	a	b	c
c	0	0	1
b	1	1	0
a	1	0	0

В матрицу пишется 1, если метка строки расположена выше (более специфична), чем метка столбца

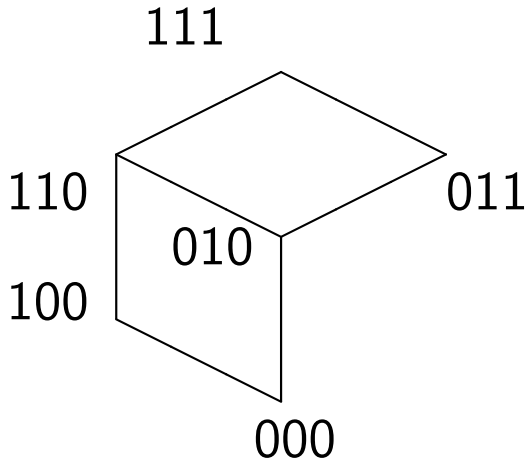
Как кодировать?



	a	**b**	c	d		a	c	d
d	0	**0**	0	1		0	0	1
c	0	**0**	1	1	$\Rightarrow$	0	1	1
b	1	**1**	0	1		1	0	1
a	1	**0**	0	0		1	0	0

Из матрицы удаляются те столбцы, которые являются побитовым умножением нового и одного из старых столбцов.

Как кодировать?



	a	c	d	**e**		a	c	d
e	1	1	1	**1**		1	1	1
d	0	0	1	**0**		0	0	1
c	0	1	1	**0**	$\Rightarrow$	0	1	1
b	1	0	1	**0**		1	0	1
a	1	0	0	**0**		1	0	0

Из матрицы удаляется новый столбец, если он является побитовым умножением какой-то пары старых столбцов.

ВКФ-метод и его истоки

Автор предлагает **вероятностно-комбинаторный формальный метод** (ВКФ-метод), развивая когнитивные процедуры логико-комбинаторного ДСМ-метода, модифицируя их:

1. индуктивное обобщение обучающих примеров в вероятностно порождаемых ВКФ-гипотезах;
2. абдуктивное уточнение и принятие ВКФ-гипотез (порождая дополнительные гипотезы для объяснения обучающих примеров);
3. предсказание целевого свойства по аналогии с обучающими примерами.

ДСМ-метод был предложен более 35 лет назад в работах В.К.Финна (ВИНИТИ РАН), который уточнил и логически формализовал (в многозначных логиках) идеи логиков и философов XIX-XX веков Д.С.Милля (индуктивная логика), Ч.С.Пирса (абдукция) и К.Поппера (фальсификация).

Понятие ВКФ-кандидата

Контекст можно понимать как бинарное отношение между элементами множества O, которые мы называем *именами объектов*, и элементами множества F, которые мы называем *признаками*. Если в строчке, соответствующей объекту $o \in O$, и столбце, соответствующим фрагменту $f \in F$, стоит единица, то мы говорим, что *объект o обладает признаком f*, и обозначаем это через oIf. В противном случае, говорим, что *объект o не имеет признака f*.

Для подмножества $A \subseteq O$ объектов его *сходством* называется подмножество $A' = \{f \in F : \forall o \in A\,[oIf]\} \subseteq F$. Полагаем $\emptyset' = F$.

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 $B' = \{o \in O : \forall f \in B\,[oIf]\} \subseteq O$. Полагаем $\emptyset' = O$.

Определение

*Пару $\langle A, B\rangle$ назовем **ВКФ-кандидатом**, если $A = B' \subseteq O$ и $B = A' \subseteq F$.*

Контекст для Булеана

Пусть $O = \{o_1, o_2, \ldots, o_n\}$ - множество объектов, а $F = \{f_1, f_2, \ldots, f_n\}$ - множество признаков, и контекст равен:

$O \mid F$	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 любая пара $\langle \{o_{j_1}, \ldots, o_{j_k}\}, F \setminus \{f_{j_1}, \ldots, f_{j_k}\} \rangle$ будет ВКФ-кандидатом, поэтому мы имеем Булеву алгебру всех 2^n подмножеств (=битовых строк).

В чем проблемы?

1. Потенциальный комбинаторный взрыв: экспоненциальное число ВКФ-гипотез для Булеана.
2. NP-полнота и $\#P$-полнота различных задач теории сходства (С.О. Кузнецов, М.И. Забежайло и др.).
3. Переобучение: возникновение случайных сходств.

	f_1	$\ldots$	f_{j_1}	$\ldots$	f_{j_2}	$\ldots$	f_{j_m}	$\ldots$	f_n
o_1	0	$\ldots$	0	$\ldots$	1	$\ldots$	0	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_1}	0	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_2}	1	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_l}	1	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	**1**	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_k	0	$\ldots$	1	$\ldots$	0	$\ldots$	0	$\ldots$	0

Пример случайного сходства

Пусть $O = \{o_1 = B737, o_2 = SSJ100, o_3 = IL76, o_4 = A320\}$ будет множеством самолетов, 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: $\langle \{o_1, o_2\}, \{f_1\} \rangle$ «самолет с поврежденным оперением не летает» и $\langle \{o_3, o_4\}, \{f_2\} \rangle$ «самолет с поврежденным двигателем не летает», и одно «случайное» сходство $\langle \{o_2, o_3\}, \{f_3\} \rangle$ «самолет, на котором написано ругательство, не летает». Последний ВКФ-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.

Случайные сходства неустранимы

Теорема

Для $p \geq (-\ln(1-\varepsilon)/n)^{1/b}$ вероятность появления случайного 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, причем последовательности для разных объектов независимы. Число m будет равно числу контр-примеров. Для двух родителей ($b = 2$) случайного сходства верна

Теорема

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

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

Фиксированное число контр-примеров

Определение

Назовем ***выжившими*** *на шаге t контр-примеры $\left\langle y_1^k, \ldots, y_t^k \ldots, y_n^k \right\rangle$, для которых* $\forall j \leq t \left[a_j = 1 \Rightarrow y_j^k = 1\right]$.

Будем следить за числом $X_t^{(m)}$ контр-примеров, выживших после одновременного нахождения t-ых признаков m контр-примеров и случайного сходства. Ясно, что это число должно быть элементом множества $S = \{0, 1, \ldots, m\}$. Нас интересует вероятность $\mathbf{P}\left[X_n^{(m)} = 0\right]$

Производящие функции (многочлены) для распределений $\mathbf{P}\left[X_t^{(m)} = s\right]$ будем обозначать через $\varphi_t^{(m)}(z) = \sum_{j=0}^{m} \mathbf{P}\left[X_t^{(m)} = j\right] \cdot z^j$.

Теорема

$$\varphi_n^{(m)}(z) = \sum_{j=0}^{m} \binom{m}{j} \cdot \prod_{t=1}^{n} \left[p_t^{b+j} + \left(1 - p_t^b\right)\right] \cdot (z-1)^j$$

Произвольное число контр-примеров

Определение

Двойной производящей функцией *для* $\boldsymbol{P}\left[X_n^{(m)} = s\right]$ *назовем формальный ряд*

$$\varphi_n(z,u) = \sum_{m=0}^{\infty}\sum_{s=0}^{m} \boldsymbol{P}\left[X_n^{(m)} = s\right] \cdot z^s \cdot u^m = \sum_{m=0}^{\infty} \varphi_n^{(m)}(z) \cdot u^m.$$

Теорема

$\varphi_n(0,u) = \sum_{j=0}^{\infty} \prod_{t=1}^{n} \left[p_t^{b+j} + \left(1 - p_t^b\right)\right] \cdot \frac{(-u)^j}{(1-u)^{j+\mathbf{1}}}.$

Операции «Замыкай-по-одному»

Лемма

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

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

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

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

В случае Булеана: если $o_j \notin A$, то $CbODown(\langle A, B\rangle, o_j) = \langle A \cup \{o_j\}, B \setminus \{f_j\}\rangle$, и $CbODown(\langle A, B\rangle, o_j) = \langle A, B\rangle$ в противном случае.
Аналогично, если $f_j \notin B$, то $CbOUp(\langle A, B\rangle, f_j) = \langle A \setminus \{o_j\}, B \cup \{f_j\}\rangle$, и $CbOUp(\langle A, B\rangle, f_j) = \langle A, B\rangle$ иначе

Алгоритм спаривающей цепи Маркова

Data: множество обучающих (+)-примеров; внешние функции
CbOUp(,) и *CbODown*(,) операций «замыкай-по-одному»
Result: ВКФ-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 ВКФ-кандидатов $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$.

Определение

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

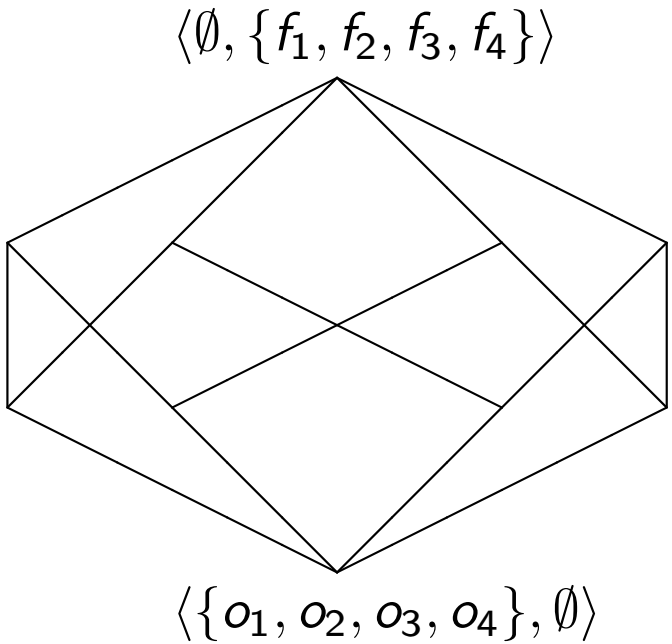
Первоначально меньший ВКФ-кандидат совпадает с наименьшим ВКФ-кандидатом $Min := \langle O, O'\rangle$, а больший - с наибольшим $Max := \langle F', F\rangle$. В цикле к обоим ВКФ-кандидатам применяется одна и та же операция *CbODown* с выбранным объектом, или *CbOUp* с выбранным признаком.

Лемма

Для всякой упорядоченной пары ВКФ-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 ВКФ-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 ВКФ-кандидат совпадет в большим. Тогда этот общий ВКФ-кандидат и выдается алгоритмом 1.

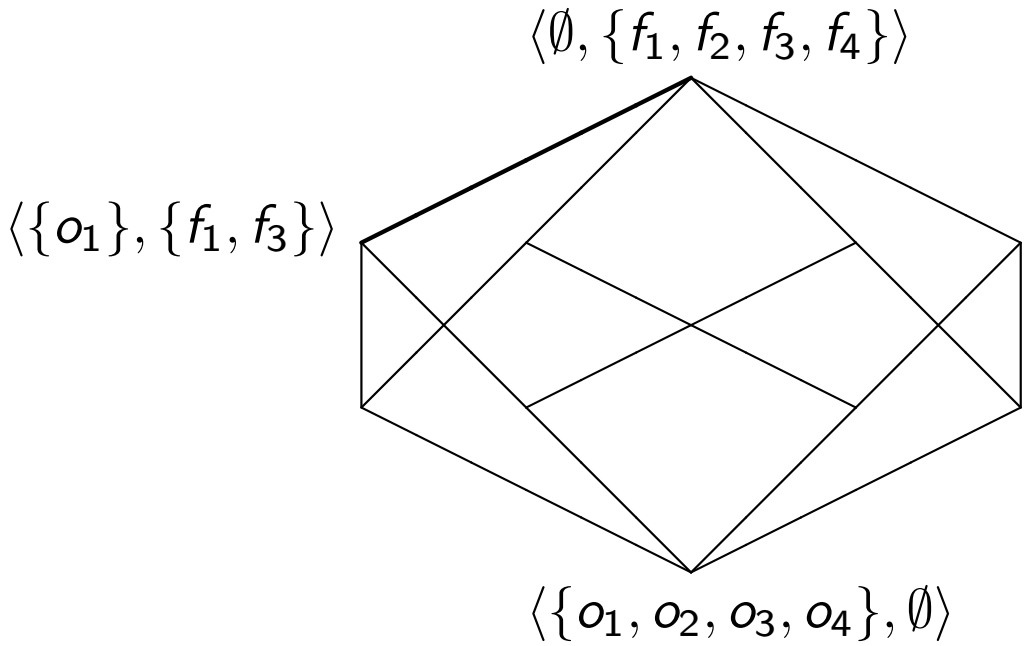
Как работает спаривающая цепь Маркова: шаг 0



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***

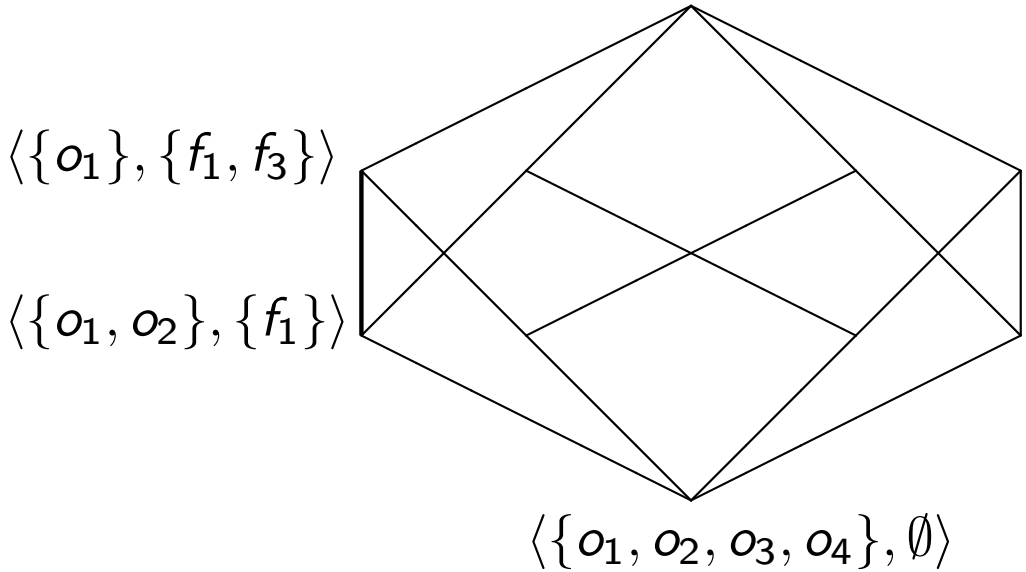
Как работает спаривающая цепь Маркова: выбор o_1



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***

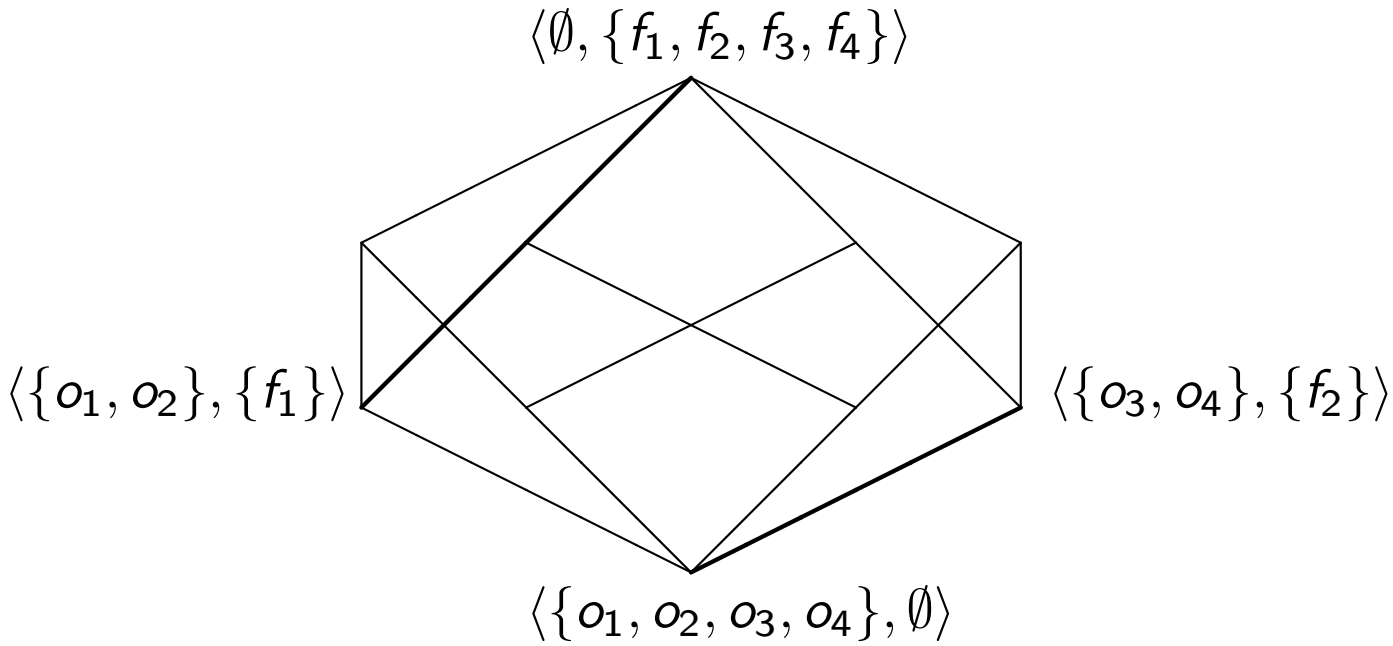
Как работает спаривающая цепь Маркова: выбор o_2



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***

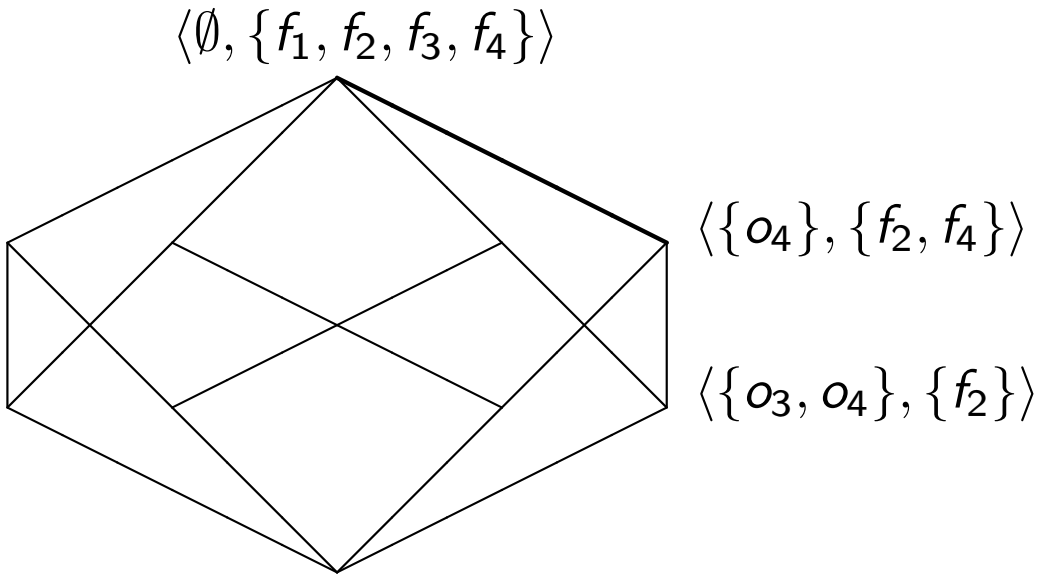
Как работает спаривающая цепь Маркова: выбор f_2



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***

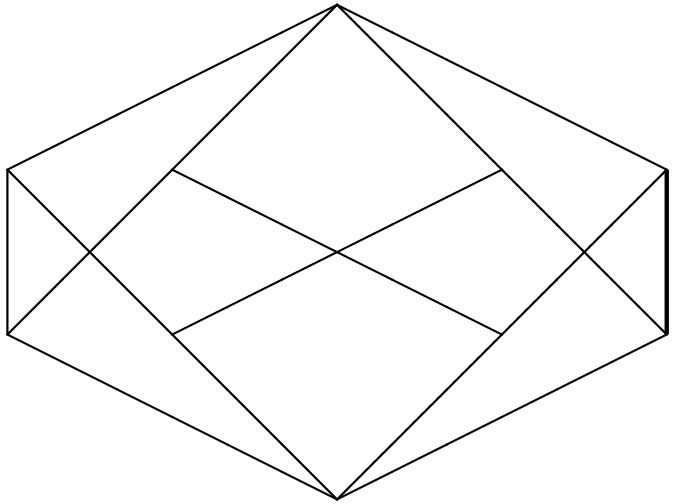
Как работает спаривающая цепь Маркова: выбор o_4



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



$\langle\{o_4\}, \{f_2, f_4\}\rangle$

$\langle\{o_3, o_4\}, \{f_2\}\rangle$

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***

Свойства спаривающей цепи Маркова

Теорема

Алгоритм 1 соответствует цепи Маркова.

Определение

Состояние вида $\langle A, B\rangle = \langle A, B\rangle$ спаривающей цепи Маркова для совпадающей пары ВКФ-кандидатов называется ***эргодическим****. Состояние вида $\langle A_1, B_1\rangle < \langle A_2, B_2\rangle$ называется* ***невозвратным****.*

Теорема

Вероятность того, что состояние

$$\langle A_1(t), B_1(t)\rangle \leq \langle A_2(t), B_2(t)\rangle$$

спаривающей цепи Маркова окажется невозвратным, стремится к нулю, когда $t \to \infty$.

Быстрая остановка

Во время проведения экспериментов с ВКФ-системой был обнаружен феномен очень быстрого нахождения очередного ВКФ-кандидата. Хотя мы не смогли получить оценку в общем виде, для случая Булеана имеются результаты о среднем времени склеивания и сильной концентрации времени склеивания около своего среднего.

Теорема

Среднее время склеивания для n-мерного гиперкуба равно

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

Теорема

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

Остановленная цепь Маркова

Для устранения слишком длинных траекторий спаривающей цепи Маркова (алгоритма 1) полезно применение следующей техники:

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 $\hat{T} = T_1 + \cdots + T_r$.*

На практике предлагается сделать r прогонов спаривающей цепи Маркова и взять оценку $t_1 + \ldots + t_r$ верхней границы склеивания.
Если спаривающая цепь Маркова не склеивается до времени $\hat{T}$, то начинаем заново, иначе выдаем $\langle A_1(T), B_1(T)\rangle = \langle A_2(T), B_2(T)\rangle$.

Теорема

Для любого $R \subseteq U$ с $\mu(R) = \rho$ и $r > log_2(\rho + 1) - log_2(\rho)$ имеем $\mu(\hat{T})(R) \geq \rho - \frac{1}{2^r - 1}$ для верхней границы $\hat{T}$ склеивания по $r > 1$ испытаниям.

Алгоритм индуктивного обобщения

Data: множество обучающих (+)- и (-)-примеров; число N порождаемых ВКФ-гипотез
Result: выборка S ВКФ-гипотез
$O :=$ (+)-примеры, $F :=$ признаки; $I \subseteq O \times F$ формальный контекст для (+)-примеров; $C :=$ (-)-примеры; $S := \emptyset$; $i := 0$;
while $(i < N)$ **do**
  породить ВКФ-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 2: Процедура индуктивного обобщения

Индуктивное обобщение данных

Проверка условия ($B \subseteq c'$) в алгоритме 2 означает, что фрагмент B ВКФ-кандидата $\langle A, B \rangle$ вкладывается в фрагмент (множество признаков) контр-примера c. Любое такое вложение означает, что ВКФ-кандидат нарушает условие «запрета на контр-пример». Если ВКФ-кандидат преодолевает все такие проверки, то он становится ВКФ-гипотезой (о причине наличия целевого свойства).
Для выбора числа N запусков спаривающей цепи Маркова (алгоритма 1) полезно применение следующей теоремы (мы используем объекты, представленные для предсказания):

Надежность ВКФ-гипотез

Зафиксируем $\varepsilon > 0$ - точность предсказания.

Определение

*Объект o назовем ε-**важным**, если суммарная вероятность появления таких ВКФ-гипотез $\langle A, B \rangle$, которые предсказывают его положительно, будет больше ε.*
*Семейство ВКФ-гипотез назовем ε-**сетью**, если для каждого ε-важного объекта найдется хотя бы одна ВКФ-гипотеза из этого семейства, которая предскажет этот объект положительно.*

Теорема

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

$$N \geq \frac{2 \cdot (n+1) - 2 \cdot \log_2 \delta}{\varepsilon}$$

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

Алгоритм абдуктивного уточнения

Data: выборка S ВКФ-гипотез, внешняя функция $CbODown(\ ,\ )$ операции «замыкай-по-одному-вниз»
Result: расширенная выборка S^+ ВКФ-гипотез
$S^+ := \emptyset$; $O :=$ (+)-примеры; $C :=$ (-)-примеры;
for $(o \in O$ *and* $\langle A, B\rangle \in S)$ **do**
  вычислить $\langle X, Y\rangle := CbODown(\langle A, B\rangle, o)$;
  $Explained(o) :=$ **false**; $hasObstacle :=$ **false**;
  for $(c \in C)$ **do**
    if $(Y \subseteq c')$ **then**
      $hasObstacle :=$ **true**;
    end
  end
  if $(hasObstacle =$ ***false***$)$ **then**
    $S^+ := S^+ \cup \{\langle X, Y\rangle\}$;
    $Explained(o) :=$ **true**;
  end
end

Algorithm 3: Процедура абдуктивного уточнения

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

Программная реализация

Автором была создана программная система, получившей название ВКФ-система:

- Программа реализована как библиотека разделяемого доступа. Она была создана в среде Netbeans C++(version 8.1) с использованием библиотеки boost (version 1_65_1). Компилятор - GNU C++ toolset (version 7.2.0).
- Примеры (обучающие, контр- и представленные для предсказания целевого свойства) представляются объектами класса *boost* :: *dynamic_bitset*. Они сохраняются в контейнерах типа *std* :: *vector* и *std* :: *list* стандартной библиотеки C++. Реализована сериализация результатов.
- Программа использует классы *std* :: *random* для датчиков случайных чисел. Это нужно для спаривающей цепи Маркова (алгоритм 1).
- Для реализации многопоточности используются классы *std* :: *thread*.
- Библиотека платформенно независима: она собирается и линкуется под Windows и под Linux (с использованием классов *boost* :: *dll*).

Достоинства ВКФ-системы

По сравнению с классическим ДСМ-подходом:

- Так как каждая ВКФ-гипотеза порождается независимым запуском цепи Маркова, то ВКФ-программа использует несколько потоков для вычисления индуктивного обобщения. Для ДСМ-системы подобное распараллеливание индукции невозможно.
- ВКФ-система вычисляет процедуру абдуктивного уточнения и принятия ВКФ-гипотез тоже в несколько потоков. В ДСМ-системе распараллеливание шага абдукции возможно, но пока не реализовано.
- Предсказание свойств по аналогии осуществляется в один поток, так как вычислительная сложность этого шага мала в сравнении с шагом индукции.
- На ЦПУ с 4 потоками (i5-4220Y) максимальная нагрузка процессора при вычислении в 4 потока достигает 90%. Для существующей параллельной версии ДСМ-системы она не превосходит 50%.

Ленивые вычисления

В настоящее время ВКФ-кандидаты находятся согласно алгоритму 1 с использованием операций *CbODown* и *CbOUp*.
Согласно определению

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

Если вычисление пересечения $(A \cup \{o\})' = B \cap o'$ фрагмента текущего ВКФ-кандидата с фрагментом выбранного объекта o соответствует побитовому умножению соответствующих строк, то операция $(A \cup \{o\})'' = (B \cap o')'$ формирования нового списка родителей может потребовать побитово перемножить с полученным ранее пересечением почти все объекты, чтобы проверить, обладает ли еще какой-нибудь объект полученным пересечением.
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*.

Ленивые вычисления-2

Аналогично, операция *CbOUp* имеет в своем составе потребляющую много времени компоненту $(B \cup \{f\})'' = (A \cap f')'$. Здесь тоже можно лениво откладывать вычисления этой части до тех пор, пока выбор нескольких признаков для *CbOUp* не сменится выбором объекта с переходом к операции *CbODown*.

Теорема

Выигрыш от введения ленивых вычислений для k обучающих примеров, описываемых n признаками, составляет

$$\frac{n}{k} + \frac{k}{n} \geq 2. \tag{1}$$

Массив SPECT Hearts

- Обучающая выборка содержит 40 (+)- и 40 (-)-примеров.
- Тестовая выборка содержит 172 (+)- и 15 (-)-примеров.
- Каждый пример описывался 22 бинарными атрибутами.
- ВКФ-система добавила отрицания исходных признаков, чтобы отсутствие атрибута могло быть частью причины проявления свойства. Поэтому обучающая выборка - это матрица 40×44.
- Точность предсказания простейшей ВКФ-системы достигла 86.1% (151 из 172 (+)-примеров и 10 из 15 (-)-примеров).
- Авторы массива SPECT достигли 84.0% точности своей программой CLIP3, которая реализует обучение покрытию средствами целочисленного программирования.

Массив Mushrooms

- Исходные данные включают описания 8124 грибов, разделенные на две категории (съедобные и ядовитые). Мы случайным образом разделили их на обучающую и тестовую выборки.
- Обучающая выборка содержит 4032 объекта.
- Тестовая выборка содержит 2120 (+)- (съедобные грибы) и 1972 (-)-примеров (ядовитые грибы).
- Каждый пример описывался 22 признаками, описывающие различные характеристики грибов (цвет, форма шляпки, места произрастания,частота встречаемости и т.п.). Эти признаки - номинальные, принимающие одно из нескольких значений.
- ВКФ-система закодировала эти признаки битовыми строками длины 124 бит.
- Точность предсказания простейшей ВКФ-системы достигла 100% для 100 ВКФ-гипотез и их абдуктивного уточнения.

Ваши теоремы ничего не доказывают! (с)

1. Теорема об оценке снизу вероятности возникновения случайного сходства без учета контр-примеров.
2. Оценка асимптотической вероятности появления случайного сходства при наличии контр-примеров.
3. Явный вид производящих функций для вероятности возникновения случайного сходства при наличии контр-примеров.
4. Доказательство того, что алгоритмы вероятностного нахождения сходств задают цепи Маркова.
5. Теорема об остановке с вероятностью единица алгоритма спаривающей цепи Маркова.
6. Оценка среднего времени склеивания и теорема о сильной концентрации времени склеивания около его среднего для случая Булеана.
7. Теорема об оценке изменения вероятностей результатов спаривающей цепи Маркова, остановленной по границе, вычисляемой по r прогонам.
8. Теорема о числе ВКФ-кандидатов, чтобы с заданной надежностью можно было предсказать положительно все ε-важные объекты.
9. Оценка эффективности ленивых вычислений на шаге индукции.

Направления будущих исследований

Теперь мы сформулируем открытые проблемы:

- Исследовать вопрос о времени перемешивания для монотонной цепи Маркова. Следует отметить, что в частном случае Булеана подобный результат мной получен.
- Получить оценку среднего времени склеивания в общем случае. Полезно указать, что метрика Хэмминга между верхним и нижним ВКФ-кандидатом не является функцией Ляпунова (может возрастать) в спаривающей цепи Маркова.
- Исследовать асимптотическую вероятность возникновения случайного сходства, когда число контр-примеров растет, а число признаков сохраняется. Автор надеется, что производящие функции, которые он получил окажутся при этом полезными.

Надежность предсказания: комментарии

1) Цепь Маркова порождает гипотезы, при этом независимые траектории порождают независимые элементы решетки ВКФ-кандидатов. Тестовые примеры задают подмножества точек (отсекаемые гиперплоскостями) на гиперкубе, где должны оказаться ВКФ-гипотезы. Это дуально парадигме Вапника-Червоненкиса (там гипотезы определяют подмножества, куда должны попасть обучающие точки).
2) Специфика заключается в рассмотрении точек (обучающих и тестовых примеров) в вершинах единичного гиперкуба. Нижние линейные полупространства на точках гиперкуба имеет очень малую емкость. Поэтому метод повторной выборки не дает дополнительного завышающего множителя. Хотя оценка все равно завышена по другим причинам.
3) Мы рассматриваем только ошибки первого рода, когда положительный тестовый пример не предсказывается положительным. Про неправильное предсказание отрицательных примеров речь не идет.

Приложение 1: Надежность предсказания-1

Определение

Объект o, описываемый фрагментом $o' \subseteq F$ (множеством признаков), ***предсказывается положительным*** *с помощью ВКФ-гипотезы $\langle A, B \rangle$, если $B \subseteq o'$.*

Если число признаков равно $n = |F|$, то можно рассматривать вершины n-мерного гиперкуба $\{0,1\}^n \subseteq \boldsymbol{R}^n$.
Каждый объект o, предъявляемый для предсказания, задает семейство нижних полупространств в $\boldsymbol{R}^n$:

Определение

Нижнее полупространство *$H^{\downarrow}(o)$, определяемое объектом o с фрагментом $o' \subseteq F$, задается линейным неравенством $x_{j_1} + \ldots + x_{j_k} < \frac{1}{2}$, где $F \setminus o' = \{f_{j_1}, \ldots, f_{j_k}\}$. Допускается также вырожденное нижнее полупространство $0 < \frac{1}{2}$, соответствующее $o' = F$, и совпадающее со всем $\boldsymbol{R}^n$.*

Приложение 1: Надежность предсказания-2

Лемма

Объект o предсказывается положительным тогда и только тогда, когда в любом нижнем полупространстве $H^{\downarrow}(o)$ содержится фрагмент B хотя одной ВКФ-гипотезы $\langle A, B\rangle$.

Определение

*Объект o назовем ε-**важным**, если суммарная вероятность появления таких ВКФ-гипотез $\langle A, B\rangle$, что $B \in H^{\downarrow}(o)$ будет больше ε.*
*Семейство ВКФ-гипотез назовем ε-**сетью**, если для каждого ε-важного объекта найдется хотя бы одна ВКФ-гипотеза из этого семейства, которая предскажет этот объект положительно.*

Теперь нас будет интересовать только вероятность ошибки «первого рода» (отказ от положительного предсказания): требуется найти такое число N, зависящее от ε и δ, чтобы с вероятностью, большей $1 - \delta$, случайная выборка объема N будет образовывать ε-сеть.

Приложение 1: Надежность предсказания-3

Лемма

$\boldsymbol{P}\{B_{p,N} \geq \boldsymbol{E}[B_{p,N}] - 1\} \geq \frac{1}{2}$ *для биномиальной случайной величины* $B_{p,N}$.

Для полупространства $\boldsymbol{P}H > \varepsilon$ выполняется

$$\boldsymbol{P}^N\{S_2 : N \cdot \boldsymbol{P}H - |S_2 \cap H| \leq \frac{\varepsilon \cdot N}{2}\} \leq \boldsymbol{P}^N\{S_2 : |S_2 \cap H| > \frac{\varepsilon \cdot N}{2}\}.$$

Лемма

Для любого ε *при* $N > \frac{2}{\varepsilon}$ *для независимых случайных выборок* S_1 *и* S_2 *ВКФ-кандидатов объемов* N *имеем оценку:*

$$\boldsymbol{P}^N\{S_1 : \exists H \in (Sub\downarrow)\,[S_1 \cap H = \emptyset, \boldsymbol{P}H > \varepsilon]\} \leq$$
$$\leq 2 \cdot \boldsymbol{P}^{2N}\{S_1 S_2 : \exists H \in (Sub\downarrow)\,[S_1 \cap H = \emptyset, |S_2 \cap H| > \varepsilon \cdot N/2]\}.$$

Приложение 1: Надежность предсказания-4

Лемма

Для любого ε для двух независимых случайных выборок S_1 и S_2 ВКФ-кандидатов объемов N имеем оценку:

$$\boldsymbol{P}^{2N}\{S_1 S_2 : \exists H \in (Sub\downarrow)\left[S_1 \cap H = \emptyset, |S_2 \cap H| > \frac{\varepsilon \cdot N}{2}\right]\} \leq 2^n \cdot 2^{-\frac{\varepsilon N}{2}}.$$

Теорема

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

$$N \geq \frac{2 \cdot (n+1) - 2 \cdot \log_2 \delta}{\varepsilon}$$

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

Решаем неравенство $2 \cdot 2^n \cdot 2^{-\varepsilon N/2} \leq \delta$ относительно N, чтобы получить утверждение теоремы.

Приложение 2: Остановленная цепь-1

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 $\hat{T} = T_1 + \cdots + T_r$.*

Если спаривающая цепь Маркова не склеивается до времени $\hat{T}$, то начинаем заново, иначе выдаем $\langle A_1(T), B_1(T)\rangle = \langle A_2(T), B_2(T)\rangle$.

Определение

Для целочисленной случайной величины $\hat{T}$, независимой от целочисленной случайной величины T, ***условное распределение*** *состояний относительно события $B = \{T \le \hat{T}\}$ есть распределение*

$$\mu(\hat{T})_i = \frac{\boldsymbol{P}[X_T = i, T \le \hat{T}]}{\boldsymbol{P}[T \le \hat{T}]}$$

для любого эргодического состояния i.

Приложение 2: Остановленная цепь-2

Определение

Расстояние ***тотального изменения*** *между распределениями вероятностей* $\mu = (\mu_i)_{i \in U}$ *и* $\nu = (\nu_i)_{i \in U}$ *на конечном пространстве* U *определяется правилом:* $\|\mu - \nu\|_{TV} = \frac{1}{2} \cdot \sum_{i \in U} |\mu_i - \nu_i|$.

Это расстояние является половиной метрики l_1, следовательно, само является метрикой (в частности, симметрично).

Лемма

$\|\mu - \nu\|_{TV} = \boldsymbol{max}_{R \subseteq U} |\mu(R) - \nu(R)|$.

В этой лемме подмножество R, на котором достигается максимум, определяется так: $R = \{i \in U \mid \mu_i > \nu_i\}$.

Приложение 2: Остановленная цепь-3

Лемма

$\boldsymbol{P}[T > \sum_{j=1}^{k} T_j] \leq \boldsymbol{P}[T > \sum_{j=1}^{k-1} T_j] \cdot \boldsymbol{P}[T > T_k]$ *для всех* $1 < k \leq r$.

Это следует из формулы условной вероятности, так как если $T > \sum_{j=1}^{k-1} T_j$, то, применяя ко всем четырем ВКФ-кандидатам $Min \leq \langle A_1(\sum_{j=1}^{k-1} T_j), B_1(\sum_{j=1}^{k-1} T_j)\rangle < \langle A_2(\sum_{j=1}^{k-1} T_j), B_2(\sum_{j=1}^{k-1} T_j)\rangle \leq Max$ одинаковые операции *CbODown* и *CbOUp*, имеем, что если $\langle A_1(t + \sum_{j=1}^{k-1} T_j), B_1(t + \sum_{j=1}^{k-1} T_j)\rangle < \langle A_2(t + \sum_{j=1}^{k-1} T_j), B_2(t + \sum_{j=1}^{k-1} T_j)\rangle$ спаривается позднее момента $T_k + \sum_{j=1}^{k-1} T_j = \sum_{j=1}^{k} T_j$, то и спаривание $Min < Max$ совершается позднее момента T_k.

Лемма

$\|\mu - \mu(\hat{T})\|_{TV} \leq \frac{\boldsymbol{P}[T > \hat{T}]}{1 - \boldsymbol{P}[T > \hat{T}]}$, *где* $\mu(\hat{T})$ - *распределение остановленной на верхней границе* $\hat{T}$ *склеивания по* $r > 1$ *испытаниям, а* μ - *распределение выдачи неостановленной цепи.*

Приложение 2: Остановленная цепь-4

Из определения $T, T_1, \ldots, T_r$ как независимых одинаково распределенных случайных величин, следует, что $\boldsymbol{P}[T > T_j] \leq \frac{1}{2}$ для всех $1 \leq j \leq r$.

Лемма

$\|\mu - \mu(\hat{T})\|_{TV} \leq \frac{2^{-r}}{1-2^{-r}} = \frac{1}{2^r-1}$, *где* $\mu(\hat{T})$ - *распределение остановленной на верхней границе склеивания по* $r > 1$ *испытаниям, а* μ - *распределение выдачи неостановленной цепи.*

Теорема

Для любого $R \subseteq U$ *с* $\mu(R) = \rho$ *и* $r > log_2(\rho + 1) - log_2(\rho)$ *имеем* $\mu(\hat{T})(R) \geq \rho - \frac{1}{2^r-1}$ *для верхней границы* $\hat{T}$ *склеивания по* $r > 1$ *испытаниям.*