Что нужно сделать

Согласно книге "Interpretable Machine Learning. A Guide for Making Black Box Models Explainable" by Christoph Molna,

- Интерпретируемость - это степень, до которой человек может понять причину решения (Interpretability is the degree to which a human can understand the cause of a decision)
- Интерпретируемость - это степень, до которой человек может стабильно предсказывать результат работы модели (Interpretability is the degree to which a human can consistently predict the model’s result)

Наша задача - объяснить работу Чёрного ящика через методы FCA. В частности - гипотезы и импликации.

Главным образом нам интересны следующие вопросы:

1. Можем ли мы предсказать результат модели зная лишь значения отдельного подмножества признаков? Если да, то насколько уверенными мы можем быть в этом предсказании (сильные и слабые гипотезы)?
2. Исходя из п.1, какие признаки и какие значения признаков наиболее "важны" для Чёрного ящика (насколько они влияют на результат предсказания)?

Алгоритм работы

1. Бинаризуем исходные данные (каким образом?)
2. Находим структуры, шаблоны в бинаризованных данных - т.е. определяем понятия
3. Определяем результат предсказания для каждого понятия (среднее предсказание для объектов из объёма понятия)
4. Убрать "лишние" ассоциативные связи (асс. связь "лишняя" - если после её применения предсказание понятия меняется меньше, чем на $\delta \to 0$)
5. Построить решётку понятий

Объёдинение понятий

Идея

Пусть есть следующие данные:

$$\text{Признаки} = \{\text{Форма}, \text{Цвет}\}$$
$$\text{Форма} = \{\text{Круглая}, \text{Квадратная}\}$$
$$\text{Цвет} = \{\text{Зелёный}, \text{Салатовый}, \text{Красный}\}$$

Задача бинарной классификации:

$$\text{Класс} = \{0, 1\}$$

Гипотезы:

$$\text{Форма\_\_Круглая} \cap \text{Цвет\_\_Зелёный} \Rightarrow 1$$
$$\text{Форма\_\_Круглая} \cap \text{Цвет\_\_Салатовый} \Rightarrow 1$$

Хотелось бы объединить две гипотезы в одну:

$$\text{Форма\_\_Круглая} \cap \{\text{Цвет\_\_Зелёный} \cup \text{Цвет\_\_Салатовый}\} \Rightarrow 1$$

Или же даже создать новый признак:

$$\text{Оттенок} = \begin{cases} \text{Зелёный, если Цвет} \in \{\text{Зелёный}, \text{Салатовый}\} \\ \text{Красный, если Цвет} \in \{\text{Красный}\} \end{cases}$$
$$\text{Форма\_\_Круглая} \cap \text{Оттенок\_\_Зелёный} \Rightarrow 1$$

При этом, изначально в датасете не было признака Оттенок, т.е. мы создали его самостоятельно, т.к. он делает модель более интерпретируемой.

Реализация

Вариант 1

Сравнивать все понятия на предмет похожести их содержаний. Если два понятия дают похожий результат предсказания и похожи по содержнию - объединять их.

Вариант 2

- Записать все текущие понятия и исходные атрибуты в новый Формальный Контекст.
- Построить на этом контексте монотонные понятия, которые по своей природе объединяют аттрибуты.
- Прочистить получившуюся решётку. Оставить только наиболее общие понятия.
- Визуализировать получившуюся решётку и понятия.

Пример работы

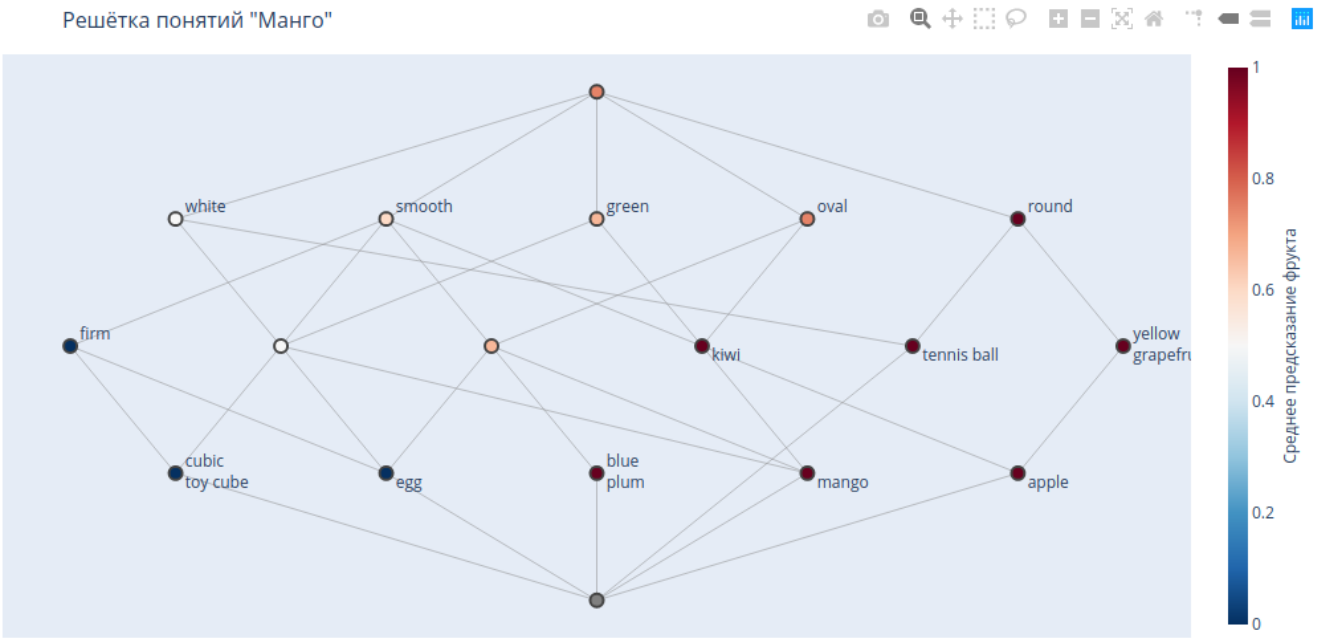
Контекст "Манго"

Контекст (вместе с предсказанием модели)

title	color	firm	smooth	form	fruit	prediction
apple	yellow	False	True	round	True	1
grapefruit	yellow	False	False	round	True	1
kiwi	green	False	False	oval	True	1
plum	blue	False	True	oval	True	1
toy cube	green	True	True	cubic	False	0
egg	white	True	True	oval	False	0
tennis ball	white	False	False	round	False	1
mango	green	False	True	oval	True	1

Анализ гипотез

Решётка понятий



Список всех гипотез (17\ штук):

Базовая гипотеза

- 0: _ -> 0.75

Позитивные гипотезы

- ***3: round -> 1.00***
- ***7: green,oval -> 1.00***
- ***10: white,round -> 1.00***
- ***11: yellow,round -> 1.00***
- ***12: smooth,blue,oval -> 1.00***
- ***13: smooth,green,oval -> 1.00***
- ***14: smooth,yellow,round -> 1.00***

Отрицательные гипотезы

- 1: green -> 0.67
- 4: smooth -> 0.60
- 5: white -> 0.50
- ***6: firm,smooth -> 0.00***
- 8: smooth,green -> 0.50
- 9: smooth,oval -> 0.67
- ***15: firm,smooth,green,cubic -> 0.00***
- ***16: firm,smooth,white,oval -> 0.00***

Формат записи:
$<$Номер понятия/гипотезы$>$: $<$содержание гипотезы$>$ $\rightarrow$ $<$Среднее предсказание целевого признака$>$
В нашем случае:
$<$Номер понятия/гипотезы$>$: $<$характеристики объекта$>$ $\rightarrow$ $<$вер-ть, что Чёрный ящик назовёт объект фру

Сильные гипотезы (выделены жирным курсивом) в бинарной классификации - те гипотезы, которые определяют целевой признак однозначно в 1 или однозначно в 0. Например:

- 3; round -> 1
- 7; firm $\cap$ smooth $\cap$ green $\cap$ cubic -> 0

Слабые гипотезы в бинарной классификации - те гипотезы, которые не определяют целевой признако однозначно в 1 или однозначно в 0. Например:

- 1: green -> 0.67 (вер-ть что модель определит объект как фрукт - 67%)
- 5: white -> 0.5

Что хотелось бы улучшить?

Родственные сильные гипотезы (вторая - частный случай первой). Например:

- 3: round -> 1
- 10: white ∩ round -> 1
- 11: yellow ∩ round -> 1

Гипотезы 10 и 11 являются частными случаями гипотезы №3, приэтом все они - сильные. Поэтому гипотезы 10 и 11 можно исключить из анализа. Аналогично можно поступить с сильными негативными гипотезами 6, 15, 16.

Оставшиеся гипотезы (11 штук):

Базовая гипотеза

- 0: _ -> 0.75

Позитивные гипотезы

- ***3: round -> 1.00***
- ***7: green,oval -> 1.00***
- ***12: smooth,blue,oval -> 1.00***
- ***14: smooth,yellow,round -> 1.00***

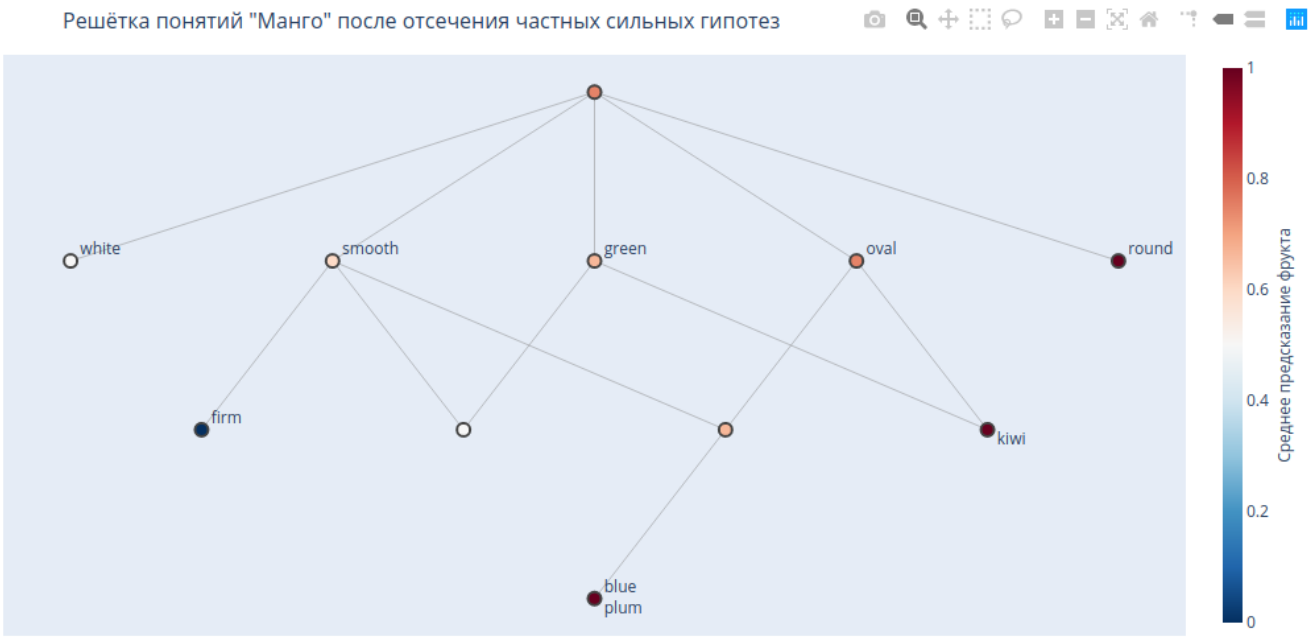
Отрицательные гипотезы

- 1: green -> 0.67
- 4: smooth -> 0.60
- 5: white -> 0.50
- ***6: firm,smooth -> 0.00***
- 8: smooth,green -> 0.50
- 9: smooth,oval -> 0.67

Также, при желании, можно объединить гипотезы 12 и 14:

$$\text{smooth} \cap \left(\text{blue} \cap \text{oval} \bigcup \text{yellow} \cap \text{round}\right) \rightarrow 1$$

Прореживаем гипотезы автоматически



На решётке видно, что после отсечения частных сильных гипотез, некоторые пути в графе остаются "незавершёнными". Например,

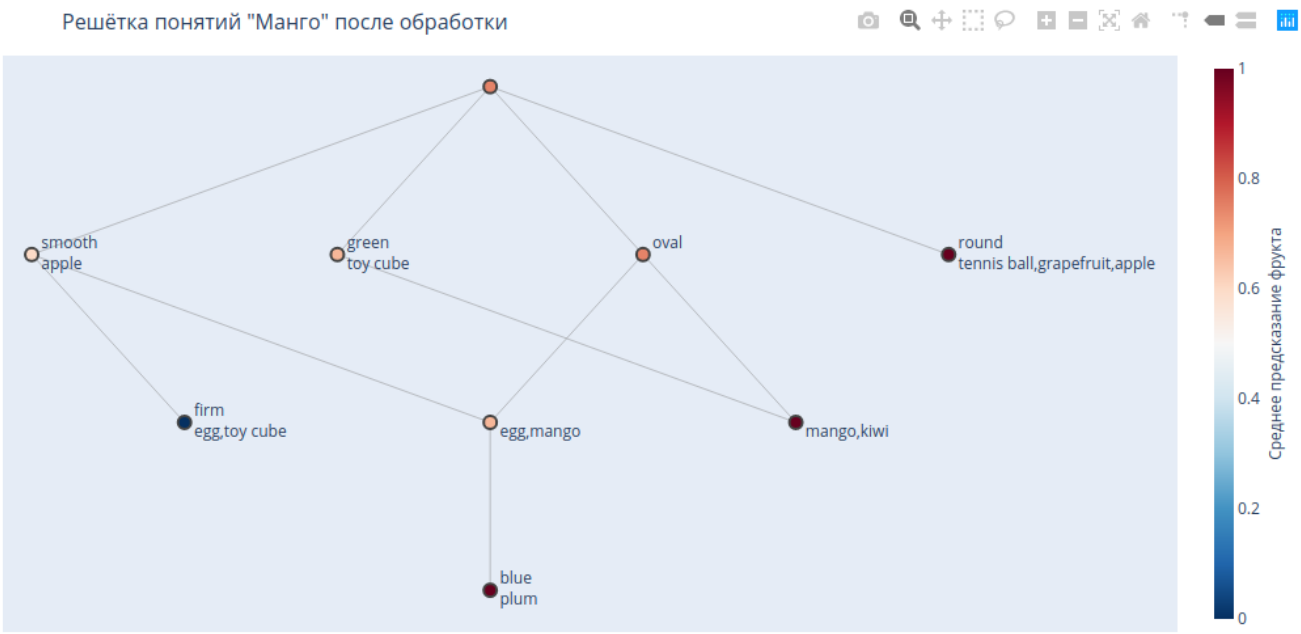
- Понятие 5: (egg, tennis ball; white) -> 0.5

При этом, объект "egg" также встречается в "сильном" понятии 6:

- Понятие 6: (toy cube, egg; firm, smooth) -> 0

Поэтому удалим все пути, которые не приводят к сильным гипотезам. По крайней мере, именно сильные гипотезы нам и интересны.

Финальный вид решётки понятий Манго



Оставшиеся гипотезы (8 штук):

Базовая гипотеза

- 0: _ -> 0.75

Позитивные гипотезы

- ***3: round -> 1.00***
- ***6: green,oval -> 1.00***
- ***8: smooth,blue,oval -> 1.00***

Отрицательные гипотезы

- 1: green -> 0.67
- 4: smooth -> 0.60
- ***5: firm,smooth -> 0.00***
- 7: smooth,oval -> 0.67

Итого: мы сократили 17 изначальных гипотез к 8 основным, которые хорошо описывают работу модели Чёрный ящик на имеющихся данных

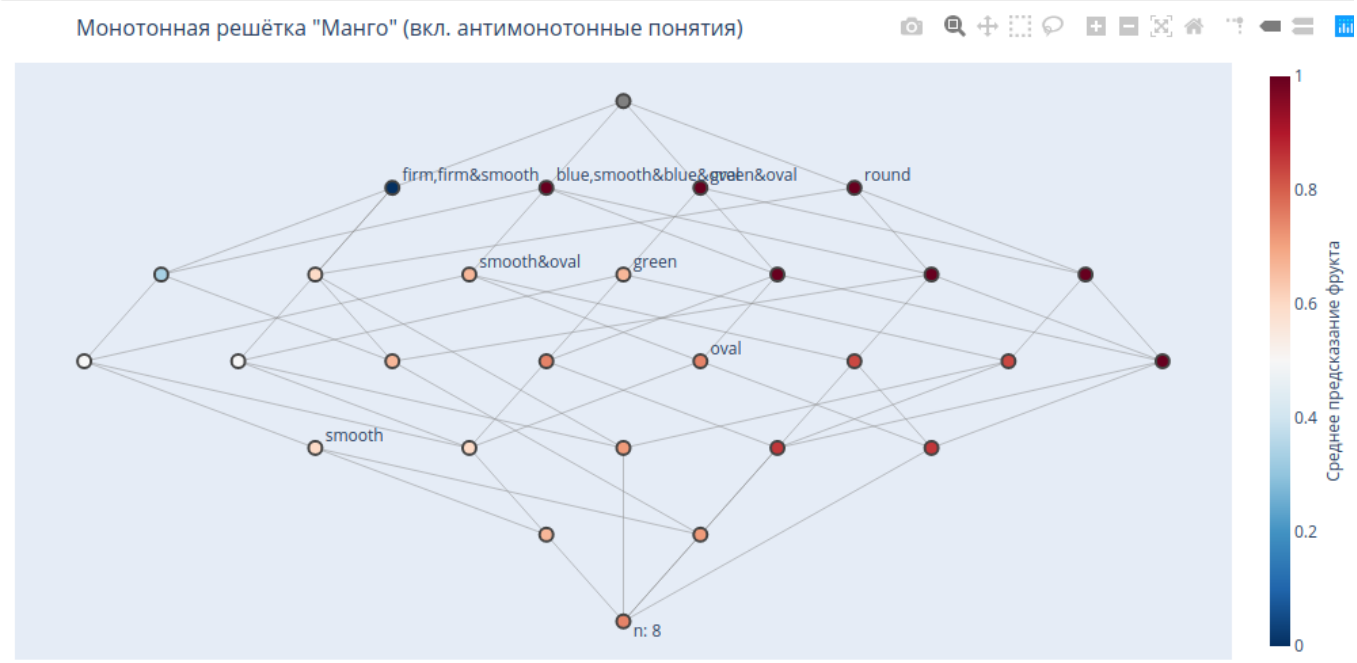
Монотонная решётка

Что если текущие гипотезы можно каким-либо образом объединить и т.о. ещё сильнее сократить их количество?

Составим новый Формальный контекст на основе полученных ранее понятий и входящих в них отдельных атрибутов.

title	oval	round	green	firm	blue	smooth	smooth&oval	green&oval	firm&smooth	smooth&blue&oval	prediction
apple	False	True	False	False	False	True	False	False	False	False	1
grapefruit	False	True	False	False	False	False	False	False	False	False	1
kiwi	True	False	True	False	False	False	False	True	False	False	1
plum	True	False	False	False	True	True	True	False	False	True	1
toy cube	False	False	True	True	False	True	False	False	True	False	0
egg	True	False	False	True	False	True	True	False	True	False	0
tennis ball	False	True	False	False	False	False	False	False	False	False	1
mango	True	False	True	False	False	True	True	True	False	False	1

Монотонная решётка:



Получилось 28 понятий, поэтому все гипотезы лучше не выводить. Однако, интересно посмотреть на следующие гипотезы:

- $5 : \text{blue} \bigcup \text{smooth} \cap \text{blue} \cap \text{oval} \bigcup \text{green} \cap \text{oval} \to 1$
- $9 : \text{round} \bigcup \text{blue} \bigcup \text{smooth} \cap \text{blue} \cap \text{oval} \to 1$
- $11 : \text{round} \bigcup \text{green} \cap \text{oval} \to 1$

- $15 : \text{round} \bigcup \text{blue} \bigcup \text{smooth} \cap \text{blue} \cap \text{oval} \bigcup \text{green} \cap \text{oval} \to 1$

При этом, гипотеза 15 - ближайщий нижний сосед гипотез 5, 9 и 11.

Перепишем гипотезы с учётом свойств бинарных операций

- $5 : \text{blue} \bigcup \text{green} \cap \text{oval} \to 1$
- $9 : \text{round} \bigcup \text{blue} \to 1$
- $11 : \text{round} \bigcup \text{green} \cap \text{oval} \to 1$

- $15 : \text{round} \bigcup \text{blue} \bigcup \text{green} \cap \text{oval} \to 1$

При этом, гипотеза 15 - ближайщий нижний сосед гипотез 5, 9 и 11.

Видно, что одной сложносочинённой гипотезой 15 можно выразить условия гипотез 5, 9, 11. Автоматизируем данный процесс

Алгоритм:

1. Пусть есть два понятия C_1, C_2 где C_2 - ближайщий нижний сосед C_1. Если оба понятия соответствуют сильным гипотезам одинакового знака (полож. или отриц.), то понятие C_1 - частное понятие от C_2 и его можно удалить.
2. Повторить п.1 пока существуют понятия, удовлетворящие условиям из п.1.
3. Удалить все понятия, соответствующие слабым гипотезам.

Скорее всего существует вариант не удалять все слабые гипотезы, но он пока не придуман.

Оставшиеся гипотезы (3 штуки):

Базовая гипотеза

- $0 : \emptyset \to 0.75$

Позитивные гипотезы

- $2 : \text{round} \bigcup \text{blue} \bigcup \text{smooth} \cap \text{blue} \cap \text{oval} \bigcup \text{green} \cap \text{oval} \to 1$

Отрицательные гипотезы

- $1 : \text{firm} \bigcup \text{firm} \cap \text{smooth} -> 0$

Оставшиеся гипотезы (после упрощения) (3 штуки):

Базовая гипотеза

- $0 : \emptyset \to 0.75$

Позитивные гипотезы

- $2 : \text{round} \bigcup \text{blue} \bigcup \text{green} \cap \text{oval} \to 1$

Отрицательные гипотезы

- $1 : \text{firm} -> 0$

Итого

Получилось два набора гипотез:

- полученные через антимонотонные понятия (8 штук)
- полученные через комбинацию антимонотонных и монотонных понятий (3 штуки)

Базовая гипотеза

- $0 : \emptyset \to 0.75$

Позитивные гипотезы

Антимонотонные

- $3 : \mathbf{round} \to \mathbf{1}$
- $6 : \mathbf{green} \cap \mathbf{oval} \to \mathbf{1}$
- $8 : \mathbf{smooth} \cap \mathbf{blue} \cap \mathbf{oval} \to \mathbf{1}$

Монотонные (только сильные)

- $2 : \mathbf{round} \bigcup \mathbf{blue} \bigcup \mathbf{green} \cap \mathbf{oval} \to \mathbf{1}$

Отрицательные гипотезы

Антимонотонные

- $1 : \mathrm{green} \to 0.67$
- $4 : \mathrm{smooth} \to 0.60$
- $5 : \mathbf{firm} \cap \mathbf{smooth} \to \mathbf{0}$
- $7 : \mathrm{smooth} \cap \mathrm{oval} \to 0.67$

Монотонные (только сильные)

- $1 : \mathbf{firm} -> \mathbf{0}$

И положительный и отрицательный класс описываются одной составной сильной гипотезой, при этом, все исходные объекты попадают либо под первую, либо под вторую гипотезу. Иначе говоря, предсказания для всех имеющихся объектов можно получить с помощью двух найденных монотонных гипотез.

Почему получилось именно так? Возможно контекст "Манго" слишком прост, т.к. содержит всего 8 объектов. Надо провести похожий анализ на каком-нибудь более реальном, практичном датасете.