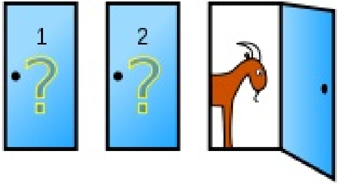
Байсовские подходы и объяснимый ИИ

Александр Сироткин

Сириус, 12 июля 2022

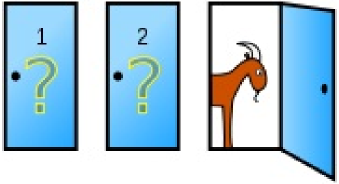
Парадокс Монти Холла (Monty Hall)



- Три двери. За одной автомобиль
- Игрок выбирает дверь, $P(\text{Авто}) = 1/3$
- Ведущий знает, где автомобиль.
  Он открывает 1 из двух дверей и предлагает поменять свой изначальный выбор.

Стоит ли менять дверь?

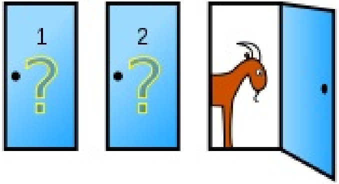
Парадокс Монти Холла (Monty Hall)



- Три двери. За одной автомобиль
- Игрок выбирает дверь, $P(\text{Авто}) = 1/3$
- Ведущий знает, где автомобиль.
  Он открывает 1 из двух дверей и предлагает поменять свой изначальный выбор.

Стоит ли менять дверь? Меняя дверь, выигрываем в 2 раза чаще!

Парадокс Монти Холла (Monty Hall)



- Три двери. За одной автомобиль
- Игрок выбирает дверь, $P(\text{Авто}) = 1/3$
- Ведущий знает, где автомобиль.
  Он открывает 1 из двух дверей и предлагает поменять свой изначальный выбор.

Стоит ли менять дверь? Меняя дверь, выигрываем в 2 раза чаще!
Симуляции этого пари: http://stayorswitch.com.

Парадокс Монти Холла. Решение.

Не умаляя общности, выберем дверь № 1

- $A =$ {выиграем авто, если выберем № 1 и поменяем выбор}
- {№1} = {авто за № 1}
  - Ведущий откроет №2 или №3, меняя выбор, проиграем с вероятностью 1
  - $P(A|\text{№}1) = 0$
- {№2} = {авто за № 2}
  - Ведущий откроет №3, меняя выбор, выиграем с вероятностью 1
  - $P(A|\text{№}2) = 1$
- {№3} = {авто за № 3}
  - Ведущий откроет №2, меняя выбор, выиграем с вероятностью 1
  - $P(A|\text{№}3) = 1$

$$P(\mathrm{A}) = P(A|\text{№}1)P(\text{№}1) + P(A|\text{№}2)P(\text{№}2) + P(A|\text{№}3)P(\text{№}3) = 2/3$$

Пример 2

Медицинский тест на выявление гепатита эффективен в 98% случаев

- “ложно положителен” в 1% случаев
- $0,5\%$ популяции имеют гепатит
- Пусть A ваш тест положителен
- Пусть B вы действительно больны
- $P(B|A) = ?$

Пример 2

Медицинский тест на выявление гепатита эффективен в 98% случаев

- "ложно положителен" в 1% случаев
- $0,5\%$ популяции имеют гепатит
- Пусть A ваш тест положителен
- Пусть B вы действительно больны
- $P(B|A) = ?$

Решение:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

$$P(B|A) = \frac{0.98 \cdot 0,005}{0.98 \cdot 0,005 + 0.01(1 - 0,005)} \approx 0,33$$

- "Перестраховка"

Пример 2

	Hepatitis +	Hepatitis -
Test +	$0.98 = P(A\|B)$	$0,01 = P(A\|B^c)$
Test -	$0.02 = P(A^c\|B)$	$0.99 = P(A^c\|B^c)$

- Пусть A^c ваш тест отрицательный
- Пусть B вы действительно больны
- $P(B|A^c) =?$

$$P(B|A^c) = \frac{P(A^c|B)P(B)}{P(A^c|B)P(B) + P(A^c|B^c)P(B^c)}$$

$$P(B|A^c) = \frac{0.02 \cdot 0,005}{0.02 \cdot 0,005 + 0.99(1 - 0,005)} \approx 0,0001$$

- “Пропуск сигнала”

Что общего у примеров?

- Есть предшествующий опыт и оценка ситуации
- Поступает дополнительная информация
- Наша оценка меняется

Рассмотрим события A и B на Ω

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A|B)}{P(A)} \cdot P(B).$$

$P(B)$ априорная вероятность
$P(B|A)$ апостериорная вероятность

Формула Байеса.

Томас Байес (1702 — 1761) математик, философ.

Байесовский подход использует субъективные вероятности.

Полная группа событий

Конечный набор попарно несовместных событий $H_1, H_2, ...H_n$ таких, что $P(H_i) > 0$, называется полной группой событий или разбиением пространства Ω, если

$$\sum_{i=1}^{n} P(H_i) = 1.$$

События, образующие полную группу событий, часто называют гипотезами.

Формула полной вероятности. Общий вид.

Пусть $H_1, H_2, ...H_n$ — полная группа событий. Тогда вероятность любого события может быть вычислена по формуле:

$$P(A) = \sum_{i=1}^{n} P(A|H_i)P(H_i).$$

Объединяя две формулы:

$$P(H_i|A) = \frac{P(A|H_i) \cdot P(H_i)}{\sum_{i=1}^{n} P(A|H_i)P(H_i)}$$

Наивный байесовский классификатор

- У нас есть гипотезы H_i
- И набор признаков A_j
- Предполагаем, что признаки условно независимы, при условии гипотезы:

$$P(A_1 A_2 \dots A_n | H_i) = \prod_j P(A_j | H_i)$$

- Надо найти $\operatorname{argmax} P(H_i | A_1 A_2 \dots A_n)$

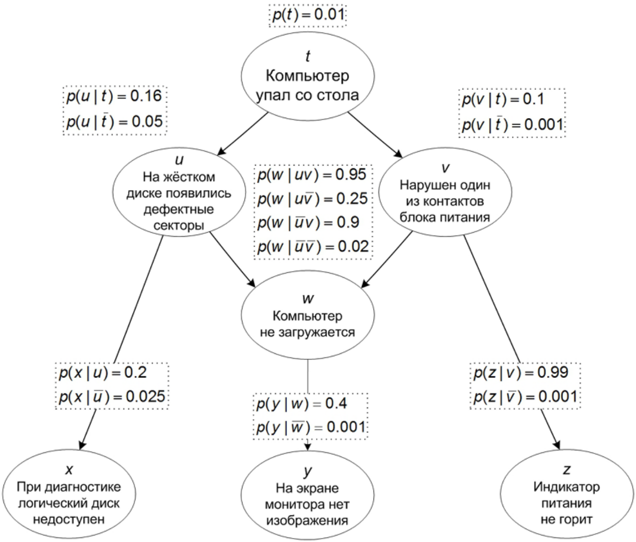
Наивный байесовский классификатор: что не так

- Наивный байесовский классификатор часто хорошо работает.
- Но он основывается на очень серьёзном предположении, а именно на условной независимости атрибутов при условии данного целевого значения.
- Зачастую такое предположение делает аппарат неприменимым.
- Что делать?

От классификатора к сети

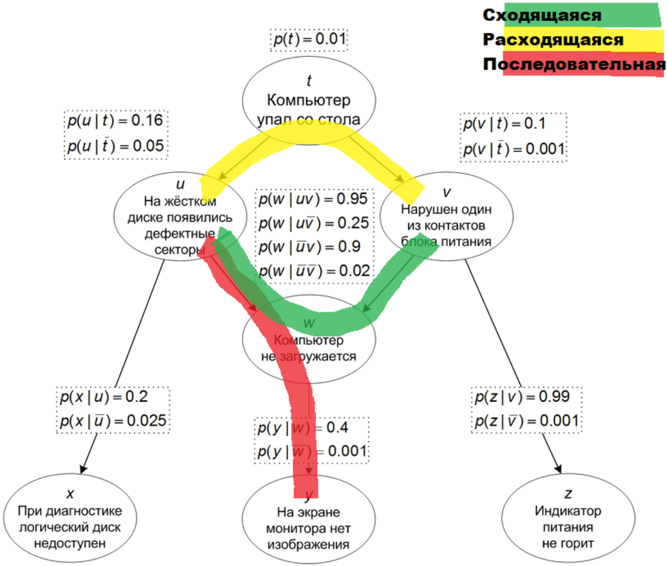
- Нужно научиться представлять множество (не)зависимостей между имеющимися переменными.
- Достаточно естественная идея: направленный граф, в котором стрелки показывают причинно-следственную связь.

Пример



$p(t) = 0.01$
t
Компьютер упал со стола
$p(u \mid t) = 0.16$
$p(u \mid \bar{t}) = 0.05$
$p(v \mid t) = 0.1$
$p(v \mid \bar{t}) = 0.001$
u
На жёстком диске появились дефектные секторы
$p(w \mid uv) = 0.95$
$p(w \mid u\bar{v}) = 0.25$
$p(w \mid \bar{u}v) = 0.9$
$p(w \mid \bar{u}\bar{v}) = 0.02$
v
Нарушен один из контактов блока питания
w
Компьютер не загружается
$p(x \mid u) = 0.2$
$p(x \mid \bar{u}) = 0.025$
$p(y \mid w) = 0.4$
$p(y \mid \bar{w}) = 0.001$
$p(z \mid v) = 0.99$
$p(z \mid \bar{v}) = 0.001$
x
При диагностике логический диск недоступен
y
На экране монитора нет изображения
z
Индикатор питания не горит

Виды связи



Сходящаяся
Расходящаяся
Последовательная
p(t) = 0.01
t
Компьютер упал со стола
p(u | t) = 0.16
p(u | t̄) = 0.05
p(v | t) = 0.1
p(v | t̄) = 0.001
u
На жёстком диске появились дефектные секторы
p(w | uv) = 0.95
p(w | uv̄) = 0.25
p(w | ūv) = 0.9
p(w | ūv̄) = 0.02
v
Нарушен один из контактов блока питания
w
Компьютер не загружается
p(x | u) = 0.2
p(x | ū) = 0.025
p(y | w) = 0.4
p(y | w̄) = 0.001
p(z | v) = 0.99
p(z | v̄) = 0.001
x
При диагностике логический диск недоступен
y
На экране монитора нет изображения
z
Индикатор питания не горит

d-разделимость

Два узла направленного графа x и y называются d-разделенными, если для всякого пути из x в y (здесь не учитывается направление ребер) существует такой промежуточный узел z (не совпадающий ни с x, ни с y), что либо связь в пути в этом узле последовательная или расходящаяся, и узел z зафиксирован (есть информация о его состоянии), либо связь сходящаяся, и ни узел z, ни какой-либо из его потомков не зафиксирован. В противном случае узлы называются d-связанными.

Цепное правило

Самое главное предположение байесовских сетей – цепное правило, вытекает из предположение, что вероятности любых d-разделенных узлов, условно независимы, при заданном наборе свидетельств (зафиксированных узлов) и задается следующим образом:

$$P(x_1 x_2 \dots x_n) = \prod P(x_i | pa(x_i))$$

А как считать?

- Цепное правило позволяет существенно упростить вычисления
- Например, для нашего примера общее распределение определяется как:

$$p(tuvwxyz) = p(t)p(u|t)p(v|t)p(w|uv)p(x|u)p(y|w)p(z|v)$$

- А с другой стороны, мы знаем, что, например, $p(z) = \sum_{tuvwxy} p(tuvwxyz)$
- Если же нет никакой информации об y, то по формуле полной вероятности $\sum_y p(y|w) = 1$

Если нет циклов

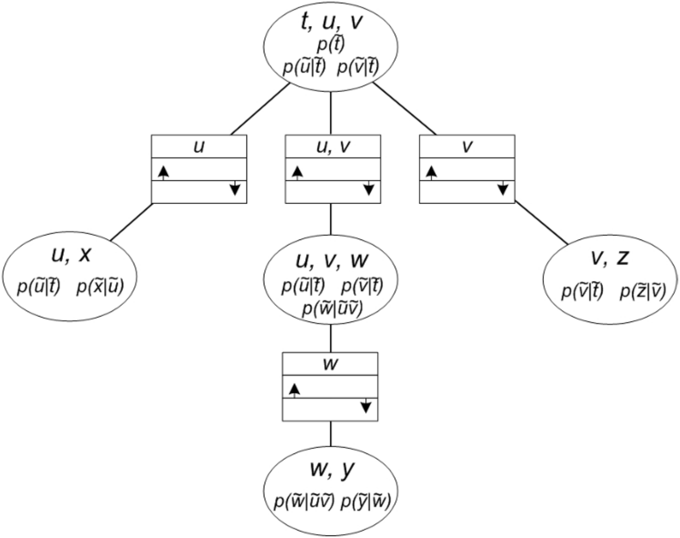
- Используют алгоритм передачи сообщений
- Для любого узла можно разбить свидетельства на те что «выше», и те что «ниже»
- Они не влияют друг на друга и суммирование в каждой части можно проводить независимо

Если есть ненаправленные циклы

- Необходимо привести структуре к подобию дерева
- Для этого будем объединять некоторые переменные в один «блок»
- После таких преобразований, можно использовать алгоритм для дерева

В общем случае поиск оптимального объединения в блоки (триангуляция) NP-сложная задача.

Виды связи



t, u, v
p(t̃)
p(ũ|t̃) p(ṽ|t̃)
u
u, v
v
u, x
p(ũ|t̃) p(x̃|ũ)
u, v, w
p(ũ|t̃) p(ṽ|t̃)
p(w̃|ũṽ)
v, z
p(ṽ|t̃) p(z̃|ṽ)
w
w, y
p(w̃|ũṽ) p(ỹ|w̃)

А как обучать байесовскую сейть?

- Обучение вероятностей
- Обучение структуры

Обучение вероятностей

- На основе структуры выделяем каздый из элементов цепного правила
- Каждое из распределений в цепном правиле можно обучить просто на статистических данных
- Можно предположить «априорное распределение на распределениях», например, гамма распределения, если мы рассматриваем каждое возможное сочетание значений переменной в блоке, как значение мультиномиальной переменной.

Обучение вероятностей: дефицит информации

- Не всегда известны состояния всех переменных одновременно
- Можно попробовать построить EM-алгоритм
- Мы находим оптимальное для текущей модели состояние скрытых переменных, после чего для найденных значений, пересчитываем параметры распределений

- По сути это задача аналогична casual inference
- Можно искать пары независимых и условно независимых переменных.
- Очень сложно найти «сложные» связи, объединяющие большое число переменных.
- Можно перебирать все тройки X, Y, Z и проверять, что X и Y независимы при условии Z.

Объяснимый ИИ

- Байесовские сети идеальный пример объяснимого ИИ
- Сама по сети структура сети – описывает причинно-следственные связи
- Построение такой структуры, позволяет нам объяснять почему результат работы алгоритма такой, а не иной

Байесовский подход, больше чем байесовская сеть

- Байесовская сеть – это пример структуры, базирующейся на байесовском выводе
- Ключевая особеннойсть это уточнение или обновление оценок (наших предположений) в соответствии с новыми данными
- В классических байесовских сетях используются бинарные или мультиномиальные переменные, но можно использовать и непрерывные
- Один из вариантов сетей с непрерывными переменными строится на основе связей вида

$$P(X|u_1 \dots u_n) \sim N(a_0 + \sum_i a_i \cdot u_i, \sigma^2)$$