Bayesian Networks

ДАТАСЕТ: MUSHROOM CLASSIFICATION

Ткаченко Елизавета

М8О-309Б-23

Что такое Bayesian Networks?

Байесовская сеть — это ориентированный ациклический граф, который моделирует вероятностные зависимости между случайными переменными.

Основная формула (теорема Байеса):

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

Датасет Mushroom Classification

Цель: Классифицировать грибы как съедобные (e) или ядовитые (p)

Размер: 8124 записи × 23 признака

```
data.head(3)
```

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring
0	p	x	s	n	t	p	f	c	n	k	...	s	w	w
1	e	x	s	y	t	a	f	c	b	k	...	s	w	w
2	e	b	s	w	t	l	f	c	b	n	...	s	w	w

3 rows × 23 columns

Основные признаки: odor (запах) — один из самых информативных признаков, cap-color (цвет шляпки), gill-color (цвет жабр), spore-print-color (цвет отпечатка спор), gill-size (размер жабр), bruises (наличие синяков) и ещё 17 других характеристик

Label Encoding (кодирование категорий)

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
for col in data.columns:
    data[col] = le.fit_transform(data[col])

data.head(3)
```

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type
0	1	5	2	4	1	6	1	0	1	4	...	2	7	7	0
1	0	5	2	9	1	0	1	0	0	4	...	2	7	7	0
2	0	0	2	8	1	3	1	0	0	5	...	2	7	7	0

3 rows × 23 columns

Почему это важно:

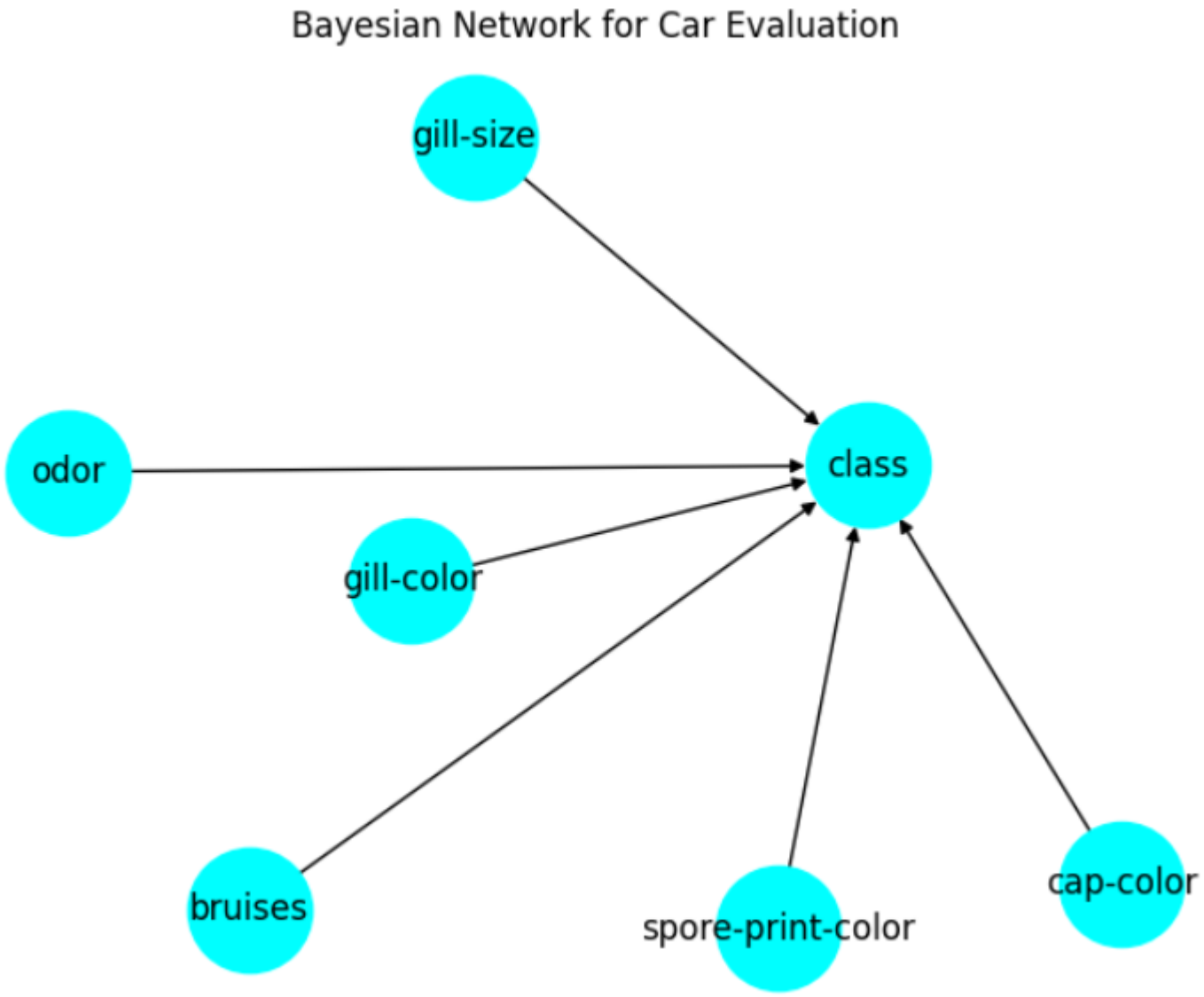
- pgmpy работает только с дискретными числовыми значениями
- Label Encoding преобразует категории в коды 0, 1, 2, ... N

Построение Bayesian Network

Данный метод представляет собой конструктивное построение структуры дискретной байесовской сети путем явного задания направленных ребер между узлами. В этом подходе структура сети создается декларативно - мы напрямую указываем список кортежей (родительский узел, дочерний узел), что определяет направленные зависимости между переменными.

```
# Задаём структуру сети вручную
network = [
    ('odor', 'class'),
    ('gill-color', 'class'),
    ('cap-color', 'class'),
    ('spore-print-color', 'class'),
    ('gill-size', 'class'),
    ('bruises', 'class')
]
```

Визуализация сети



Оценка параметров и CPT

CPT for odor:

odor(0)	0.0492368
odor(1)	0.0236337
odor(2)	0.265879
odor(3)	0.0492368
odor(4)	0.00443131
odor(5)	0.434269
odor(6)	0.0315116
odor(7)	0.070901
odor(8)	0.070901

bruises	...	bruises(1)
cap-color	...	cap-color(9)
gill-color	...	gill-color(11)
gill-size	...	gill-size(1)
odor	...	odor(8)
spore-print-color	...	spore-print-color(8)
class(0)	...	0.5
class(1)	...	0.5

CPT for cap-color:

cap-color(0)	0.0206795
cap-color(1)	0.00541605
cap-color(2)	0.184638
cap-color(3)	0.226489
cap-color(4)	0.281142
cap-color(5)	0.0177253
cap-color(6)	0.00196947
cap-color(7)	0.00196947
cap-color(8)	0.128016
cap-color(9)	0.131955

Вероятностный вывод (Inference)

```
from pgmpy.inference import VariableElimination

infer = VariableElimination(model)
# Запрос для вероятности класса при конкретном запахе (например, odor == 5)
query = infer.query(variables=['class'], evidence={'odor': 5})
print(query)
```

```
+----------+--------------+
| class    |   phi(class) |
+==========+==============+
| class(0) |       0.5789 |
+----------+--------------+
| class(1) |       0.4211 |
+----------+--------------+
```

Вывод — распределение вероятности по классам:

- class(0) (вероятнее всего съедобный): 0.5789
- class(1) (ядовитый): 0.4211.

Сравнение результатов с baseline-моделью по метрике accuracy

```
Accuracy наивного байесовского классификатора: 0.9459
Accuracy байесовской сети: 0.9975
```

Вывод:
- Обе модели показывают высокую точность
- Признак "odor" почти полностью определяет класс гриба
- Bayesian Network: лучше для интерпретации, видны явные связи между признаками
- Naive Bayes: проще и быстрее, хорошо работает как baseline