Байесовские сети

Mushroom Classification

Штыхно Илья

М8О-309Б

Что такое Байесовские сети?

Определение

Вероятностная графическая модель, представляющая зависимости между переменными через направленный ациклический граф (DAG)

Применение

Медицина, классификация, прогнозирование, принятие решений

Преимущества

Интерпретируемость, вероятностный вывод, работа с неопределенностью

Теорема Байеса

$$P(A|B) = P(B|A) \times P(A) / P(B)$$

Основа для вероятностного вывода в сети

Датасет: Mushroom Classification

Всего грибов
8124

Съедобные (e)
4208

Ядовитые (p)
3916

Изначально: 23 признака

cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, habitat...

Выбрано 6 ключевых признаков:

class (целевая), odor (запах), gill-color (цвет жабр), spore-print-color, population, habitat (среда)

Обработка данных

Label Encoding

Преобразование категориальных значений в числа для работы с pgmpy

До:

'p', 'e', 'x', 'n'...

После:

1, 0, 5, 4...

Дубликаты

До: 8124

После: 8124

Не найдено

Результат

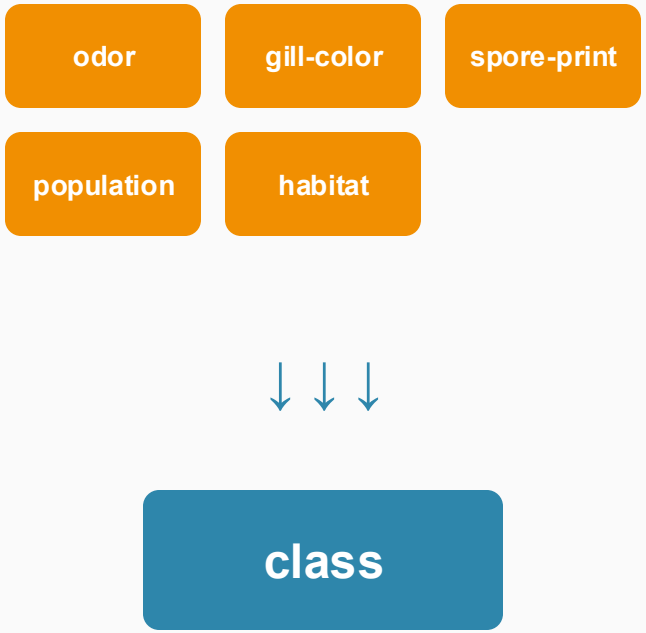
8124 строки

6 признаков

Готово для pgmpy

Структура сети

5 родителей → 1 ребенок



Оценка параметров (CPT)

Maximum Likelihood Estimator

Оценка условных вероятностей на основе частот в данных

CPT таблицы

Всего параметров: **6**

odor, gill-color, spore-print-color, population, habitat, class

Что показывают CPT?

P(class | odor, gill-color, spore-print-color, population, habitat)

Вероятность класса при различных комбинациях признаков

Визуализация сети

Граф сети (NetworkX)

Направленный ациклический граф

Структура графа

5 родительских узлов → 1 целевой узел

Узлы: признаки

Ребра: зависимости

Inference: Вероятностный вывод

Метод: **Variable Elimination**

Пример 1: odor = almond

Съедобный	Ядовитый
53.05%	**46.95%**

Пример 2: odor=5, gill-color=4

Съедобный	Ядовитый
52.01%	**47.99%**

Пример 3: odor=6, gill=5, spore=2

Съедобный	Ядовитый
39.92%	**60.08%**

Сравнение с Baseline

Naive Bayes (sklearn)

Baseline модель

97.83%

Accuracy на тесте (2438 примеров)

Bayesian Network

Наша модель (pgmpy)

100%

Accuracy на 2000 примерах

✨ Преимущества Байесовской сети

✓ Явная структура зависимостей

Видим как признаки влияют на класс

✓ Гибкий вероятностный вывод

Любые комбинации признаков

Выводы

Достигнуто

✓ Построена байесовская сеть

✓ Оценены CPT таблицы

✓ Выполнен inference

✓ Сравнение с baseline (100%)

Ключевые признаки

Запах (odor), цвет жабр и спор - наиболее важные для классификации

Интерпретируемость

Явная структура зависимостей

Применение

Предсказание съедобности