

► Байесовские сети

Солодова София Михайловна М8О-309Б-23
Mushroom Classification Dataset

► СОДЕРЖАНИЕ РАБОТЫ

01

Подготовка

- Загрузка датасета
- Обработка данных
- Label Encoding

02

Модель

- Построение сети
- Оценка параметров
- Анализ СРТ

► Что такое Bayesian Networks?

ОПРЕДЕЛЕНИЕ

Ориентированный ациклический граф (DAG), в котором каждый узел представляет случайную величину, а рёбра кодируют условные зависимости между ними. Сеть совмещает теорию графов и теорию вероятностей для компактного представления совместного распределения вероятностей множества переменных.

КОМПОНЕНТЫ

- **Узлы (Nodes)** → случайные переменные (признаки грибов)
- **Рёбра (Edges)** → причинные/вероятностные влияния между переменными
- **CPT (Conditional Probability Tables)** → таблицы условных вероятностей для каждого узла

ПРЕИМУЩЕСТВА

- Компактное представление больших распределений вероятностей
- Логический вывод (Inference) — вычисление вероятностей при частично известных данных
- Высокая интерпретируемость — видны причинные связи между признаками
- Обучение структуры — автоматическое открытие зависимостей из данных

► Датасет – Основная информация

Размер: 8,124 образцов

Признаков: 23 (22 после удаления veil-type)

Целевой класс: class (e=съедобный, p=ядовитый)

Распределение: 51.8% / 48.2% (сбалансировано)

Признак	Описание	Значения
class	Класс гриба	e=съедобный, p=ядовитый
cap-shape	Форма шляпки	b=колокольчатая, c=коническая, x=выпуклая, f=плоская, k=с бугорком, s=вогнутая
cap-surface	Поверхность шляпки	f=волокнистая, g=бородавчатая, u=чешуйчатая, s=гладкая
cap-color	Цвет шляпки	n=коричневый, b=бежевый, c=коричневый, g=серый, r=зеленый, p=розовый, u=фиолетовый, e=красный, w=белый, y=желтый
bruises	Синяки	t=есть, f=нет
odor	Запах	a=миндальный, l=анисовый, c=креозотовый, y=рыбный, f=неприятный, m=затхлый, n=нет запаха, p=едкий, s=пряный
gill-attachment	Крепление пластинок	a=прикрепленные, d=нисходящие, f=свободные, p=выемчатые
gill-spacing	Расстояние пластинок	c=блико, w=тесно, d=далеко
gill-size	Размер пластинок	b=широкие, n=узкие
gill-color	Цвет пластинок	k=черный, n=коричневый, b=бежевый, h=шоколадный, g=серый, r=зеленый, o=оранжевый, p=розовый, u=фиолетовый, e=красный, w=белый, y=желтый
stalk-shape	Форма ножки	e=расширяющаяся, t=сужающаяся
stalk-root	Корень ножки	b=луковичный, c=булавовидный, u=чашевидный, e=равномерный, z=ризоморфный, r=корневидный, ?=отсутствует
stalk-surface-above-ring	Поверхность ножки над кольцом	f=волокнистая, y=чешуйчатая, k=шелковистая, s=гладкая
stalk-surface-below-ring	Поверхность ножки под кольцом	f=волокнистая, y=чешуйчатая, k=шелковистая, s=гладкая
stalk-color-above-ring	Цвет ножки над кольцом	n=коричневый, b=бежевый, c=коричневый, g=серый, o=оранжевый, p=розовый, e=красный, w=белый, y=желтый
stalk-color-below-ring	Цвет ножки под кольцом	n=коричневый, b=бежевый, c=коричневый, g=серый, o=оранжевый, p=розовый, e=красный, w=белый, y=желтый
veil-type	Тип покрывала	p=частичное, u=универсальное
veil-color	Цвет покрывала	n=коричневый, o=оранжевый, w=белый, y=желтый
ring-number	Количество колец	n=нет, o=одно, t=два
ring-type	Тип кольца	c=паутинистое, e=исчезающее, f=расширяющееся, l=большое, n=нет, p=висячее, s=влагалищное, z=зона
spore-print-color	Цвет спорового отпечатка	k=черный, n=коричневый, b=бежевый, h=шоколадный, r=зеленый, o=оранжевый, u=фиолетовый, w=белый, y=желтый
population	Популяция	a=обильная, c=скученная, n=многочисленная, s=рассеянная, v=несколько, u=одиночные
habitat	Среда обитания	g=трава, l=листва, m=луга, p=тропы, u=городская, w=отходы, d=леса

► ЗАГРУЗКА И ОБРАБОТКА ДАННЫХ



ЗАГРУЗКА ДАТАСЕТА:

```
import pandas as pd  
data = pd.read_csv('mushrooms.csv')  
print(data.shape) #(8124, 23)
```



Label Encoding

Что было: строки (a, b, c, ...)

Что стало: целые числа (0, 1, 2, ...)

```
from sklearn.preprocessing import LabelEncoder
```

```
for col in data.columns:
```

```
    le = LabelEncoder()
```

```
    data[col] = le.fit_transform(data[col])
```



КАЧЕСТВО ДАТАСЕТА

- ✓ Нет пропусков
- ✓ Нет дубликатов
- ✓ Все признаки категориальны

Удалена колонка '**'veil-type'** (содержала только одно значение)
Финальный размер: (8124, 22)

► Ручное построение Байесовской сети



Принцип работы

Граф строится вручную на основе экспертных знаний и понимания предметной области
Рёбра определяются логически на основе предположений о зависимостях между признаками



Преимущества

- Полный контроль над структурой
- Можно использовать знания экспертов
- Интерпретируемость гарантирована



Построение сети

```
from pgmpy.models import DiscreteBayesianNetwork
network = [
    ('odor', 'class'),      # Запах влияет на съедобность
    ('bruises', 'class'),   # Наличие синяков
]
model = DiscreteBayesianNetwork(network)
print("Ручная структура сети:")
print(model.edges())
```



Недостатки

- Требует глубокого понимания данных
- Может упустить скрытые зависимости
- Субъективный процесс

► Автоматическое построение Байесовской сети



Принцип работы

Алгоритм автоматически находит оптимальную структуру из данных, используя ВІС как метрику качества



Построение сети

```
from pgmpy.estimators import HillClimbSearch, BIC

hc = HillClimbSearch(data)
best_model = hc.estimate(scoring_method=BIC(data))
model = DiscreteBayesianNetwork(best_model.edges())
print("Автоматическая структура:")
print(model.edges())
```



Преимущества

- Открывает скрытые зависимости из данных (53 ребра найдено)
- Объективный процесс
- Не требует предварительных знаний
- Обнаруживает неожиданные корреляции



Недостатки

- Может быть вычислительно дорогим
- Требует достаточного количества данных
- Результат зависит от качества данных

► Обучение модели: выбор оценщика параметров



MLE

(Maximum Likelihood Estimation)

MLE — это метод, который вычисляет вероятности прямо из данных:

$$P(A)=NnA$$



BDeu

(Bayesian Dirichlet Equivalent Uniform)

BDeu — это метод Байесовского оценивания, который добавляет "виртуальные образцы" к реальным данным:

$$P(A)=N+\sum\alpha nA+\alpha$$



Байессовский оценщик

```
from pgmpy.estimators import BayesianEstimator  
model.fit(data, estimator=BayesianEstimator,  
prior_type='BDeu', equivalent_sample_size=0.1)
```

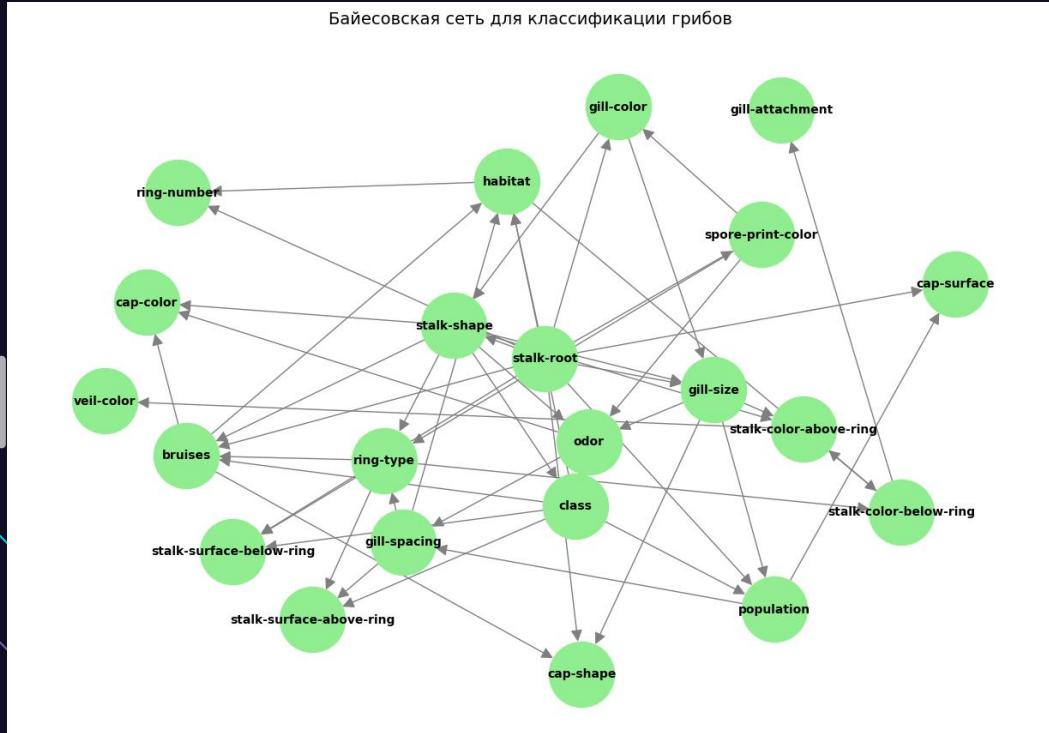


Параметр `equivalent_sample_size=0.1`

Контролирует $\sum \alpha$ (виртуальные образцы):

- $0.1 \rightarrow$ слабый prior (почти MLE)
- $5 \rightarrow$ средний prior
- $10+ \rightarrow$ сильный prior

► Визуализация структуры сети



На графике каждый узел представляет признак гриба, а направленные стрелки показывают вероятностные зависимости между ними.

Граф содержит 22 узла и 53 ребра,
что отражает сложность
взаимосвязей в данных.

► Анализ уверенности предсказаний

Для проверки качества предсказаний модели выбрали тестовые образцы с разными признаками. Модель анализирует две группы: грибы с нейтральным запахом ($\text{odor}=5$) и с приятным запахом ($\text{odor}=6$).

Строка 1: $\text{odor}=5$, $\text{bruises}=0$

$P(\text{съедобный}) = 0.058$

Реальный класс: ЯДОВИТЫЙ

МОДЕЛЬ УВЕРЕНА

Строка 2: $\text{odor}=5$, $\text{bruises}=0$

$P(\text{съедобный}) = 0.058$

Реальный класс: ЯДОВИТЫЙ

МОДЕЛЬ УВЕРЕНА

Все предсказания совпали с реальным классом. Модель показала высокую уверенность (вероятности близки к 0 или 1), что означает: выученные зависимости корректны и предсказания надежны.

▶ Сравнение с Naive Bayes

Байесовскую сеть сравнили с классическим Naive Bayes на тестовом наборе (2438 образцов).

```
== СРАВНЕНИЕ С BASELINE-МОДЕЛЬЮ ==
Размеры данных: train=(5686, 21), test=(2438, 21)

=====
РЕЗУЛЬТАТЫ СРАВНЕНИЯ:
=====

📊 ТОЧНОСТЬ (Accuracy):
Наивный Байес: 0.946
Наша Байесовская сеть: 0.984

📈 LOG-LOSS (чем меньше, тем лучше):
Наивный Байес: 0.161
Наша Байесовская сеть: 0.053

🔍 АНАЛИЗ:
 Наша модель точнее на 0.038
 Наша модель дает более увереные предсказания
```

Вывод:

Байесовская сеть даёт более точные и уверенные предсказания, потому что моделирует настоящие зависимости между признаками грибов.

В отличие от наивного байесовского метода, который считает признаки независимыми, сеть учитывает их взаимосвязи и строит классификацию на реальных связях данных.

Это делает её решения более обоснованными и интерпретируемыми.

► Итоги и достижения

01

Успешное построение Байесовской сети — использован метод Hill Climb Search с BIC-критерием для автоматического обнаружения структуры сети.

02

Правильная оценка параметров — применен метод BDeu с слабым prior для получения устойчивых и интерпретируемых таблиц условных вероятностей.

03

Превосходная точность классификации — 98.4% accuracy на тестовом наборе, что на 3.8% выше baseline-модели (Naive Bayes).

04

Высокая интерпретируемость — визуализирована структура сети, наглядно показаны причинные зависимости между признаками грибов (запах, синяки, цвет пластинок и т.д.).

05

Практическое применение — модель может использоваться для надежной оценки съедобности грибов, даже при частичной информации о их характеристиках.