

Наумов П. 3413

Тема. Применение условных вероятностей в анализе данных

На некотором наборе данных (выбрать самостоятельно, можно:

<https://www.kaggle.com/datasets> (<https://www.kaggle.com/datasets>)) продемонстрировать применение условных вероятностей (см. пример с рейтингом автомобилей по типу кузова, Notebook с примером "Титаника"). Рассматриваемый набор данных необходимо описать (назначение, описание признаков).

15. Network Intrusion Detection

Обнаружение сетевых атак

Назначение: это данные сетевого трафика, используемые для обнаружения вторжений в компьютерных сетях. Набор содержит характеристики сетевых соединений для классификации нормального и вредоносного трафика.

Описание признаков:

- duration - длительность соединения
- protocol_type - тип протокола (tcp, udp, icmp)
- service - сетевая служба (http, ftp, telnet и т.д.)
- flag - статус соединения
- src_bytes, dst_bytes - объем переданных данных
- logged_in - статус авторизации (1 = вошел, 0 = не вошел)
- и многие другие технические параметры соединения

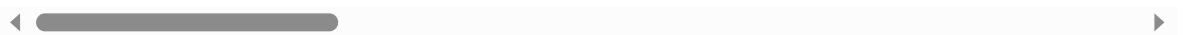
```
Ввод [1]: import numpy as np
import pandas as pd
```

```
Ввод [2]: df = pd.read_csv('Test_data.csv')
df
```

Out[2]:

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment |
|-------|----------|---------------|----------|------|-----------|-----------|------|----------------|
| 0 | 0 | tcp | private | REJ | 0 | 0 | 0 | 0 |
| 1 | 0 | tcp | private | REJ | 0 | 0 | 0 | 0 |
| 2 | 2 | tcp | ftp_data | SF | 12983 | 0 | 0 | 0 |
| 3 | 0 | icmp | eco_i | SF | 20 | 0 | 0 | 0 |
| 4 | 1 | tcp | telnet | RSTO | 0 | 15 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 22539 | 0 | tcp | smtp | SF | 794 | 333 | 0 | 0 |
| 22540 | 0 | tcp | http | SF | 317 | 938 | 0 | 0 |
| 22541 | 0 | tcp | http | SF | 54540 | 8314 | 0 | 0 |
| 22542 | 0 | udp | domain_u | SF | 42 | 42 | 0 | 0 |
| 22543 | 0 | tcp | sunrpc | REJ | 0 | 0 | 0 | 0 |

22544 rows × 41 columns



Условная вероятность

```
Ввод [3]: # Функция для вычисления условных вероятностей (как в примере с Титаником)
def conditional_probability(df, condition_col, condition_val, target_col, target_val):
    subset = df[df[condition_col] == condition_val]
    if len(subset) == 0:
        return 0
    return (subset[target_col] == target_val).mean()

# Для демонстрации выберем несколько ключевых признаков
print("АНАЛИЗ УСЛОВНЫХ ВЕРОЯТНОСТЕЙ ДЛЯ СЕТЕВОГО ТРАФИКА")
print("=" * 60)

# По типу протокола
print("\nВероятность по типу протокола:")
for protocol in df['protocol_type'].unique():
    prob = conditional_probability(df, 'protocol_type', protocol, 'logged_in', 1)
    print(f" P(logged_in|protocol_type={protocol}) = {prob:.3f}")

# По сервису (топ-5 самых частых)
print("\nВероятность по сервису (топ-5):")
top_services = df['service'].value_counts().head(5).index
for service in top_services:
    prob = conditional_probability(df, 'service', service, 'logged_in', 1)
    print(f" P(logged_in|service={service}) = {prob:.3f}")

# По флагу соединения
print("\nВероятность по флагу соединения:")
top_flags = df['flag'].value_counts().head(5).index
for flag in top_flags:
    prob = conditional_probability(df, 'flag', flag, 'logged_in', 1)
    print(f" P(logged_in|flag={flag}) = {prob:.3f}")
```

АНАЛИЗ УСЛОВНЫХ ВЕРОЯТНОСТЕЙ ДЛЯ СЕТЕВОГО ТРАФИКА

=====

Вероятность по типу протокола:

P(logged_in|protocol_type=tcp) = 0.528

P(logged_in|protocol_type=icmp) = 0.000

P(logged_in|protocol_type=udp) = 0.000

Вероятность по сервису (топ-5):

P(logged_in|service=http) = 0.955

P(logged_in|service=private) = 0.000

P(logged_in|service=telnet) = 0.065

P(logged_in|service=pop_3) = 0.872

P(logged_in|service=smtp) = 0.970

Вероятность по флагу соединения:

P(logged_in|flag=SF) = 0.624

P(logged_in|flag=REJ) = 0.000

P(logged_in|flag=S0) = 0.000

P(logged_in|flag=RST0) = 0.172

P(logged_in|flag=RSTR) = 0.798