Bẫy biến giả(Dummy Variable Trap) và Các Bước Tiền Xử Lý Dữ Liệu Trong Machine Learning

MỤC TIÊU BÀI HỌC

- Hiểu rõ hiện tượng **Dummy Variable Trap** là gì.
- Biết cách xử lý biến phân loại khi dùng trong Machine Learning.
- Nắm được các bước khởi đầu để chuẩn bị dữ liệu huấn luyện cho mô hình (Step-by-step).

I. Bẫy biến giả(DUMMY VARIABLE TRAP) LÀ GÌ?

Khái niệm:

Biến giả(**Dummy Variable):** là các biến nhị phân (chỉ nhận giá trị 0 hoặc 1) được tạo ra từ biến phân loại (categorical variable).

Bẫy biến giả xảy ra khi **có quá nhiều biến giả liên quan đến cùng một đặc tính**, gây ra hiện tượng **đa cộng tuyến hoàn hảo** (perfect multicollinearity), dẫn đến lỗi hoặc mô hình hoạt động không chính xác.

Ví dụ cụ thể:

Giả sử bạn có biến **“Thương hiệu nước ngọt ưa thích”** với 3 giá trị:

- Coca-Cola
- Pepsi
- 7Up

Bạn tạo các biến giả:

CocaCola	Pepsi	7Up
1	0	0
0	1	0
0	0	1
1	0	0

Bây giờ nếu bạn **dùng cả 3 biến giả này cùng lúc**, thì một biến có thể **dự đoán được** từ hai biến còn lại:

```
7Up = 1 - CocaCola - Pepsi
```

Điều này gây ra **đa cộng tuyến hoàn hảo** → mô hình học máy như hồi quy tuyến tính hoặc một số thuật toán sẽ **bị lỗi hoặc giảm hiệu suất**.

Vậy nên làm gì?

Giải pháp: **Loại bỏ 1 biến giả** → gọi là "biến chuẩn" hay **baseline**.

Ví dụ:

- Giữ lại: `CocaCola`, `Pepsi`
- Loại bỏ: `7Up`

Thông tin của `7Up` vẫn còn:
Khi cả CocaCola và Pepsi đều bằng 0 → người đó thích 7Up.

TẠI SAO KHÔNG MẤT THÔNG TIN?

Không mất thông tin vì:

- Biến bị loại bỏ (baseline) vẫn có thể suy ra từ các biến còn lại.
- Trong hồi quy: thông tin này được mã hóa trong **hệ số chặn (intercept)**.
- Trong XGBoost: baseline là điểm so sánh gốc để các cây quyết định phân tách.

TÓM TẮT PHẦN 1

Vấn đề	Giải pháp
Dummy variable trap	Loại bỏ 1 biến giả để tránh đa cộng tuyến
Mất thông tin?	□ Không → thông tin được mã hóa ở baseline

II. BẮT ĐẦU HƯỚNG DẪN TỪNG BƯỚC TIỀN XỬ LÝ

Bước 1: Xác định bài toán & Tạo bộ dữ liệu

Phải có giả thuyết/hướng nghiên cứu rõ ràng (hypothesis-driven)

- Mỗi bài toán nên tự xây dựng dataset phù hợp thay vì chỉ lấy sẵn từ nguồn khác.

> Ví dụ: Nếu bạn muốn dự đoán liệu khách hàng có mua hàng không, bạn cần xác định những yếu tố nào có thể ảnh hưởng đến hành vi này (tuổi, thu nhập, kênh tiếp cận,...)

Bước 2: Chuyển đổi biến phân loại thành biến giả (dummy variables)

- **Sử dụng thư viện như `pandas.get_dummies()` trong Python.**
- **Nhớ tránh Dummy Variable Trap bằng cách loại bỏ 1 biến**

Ví dụ

```
import pandas as pd

df = pd.DataFrame({
    'drink': ['CocaCola', 'Pepsi', '7Up', 'CocaCola']
})

dummies = pd.get_dummies(df['drink'], drop_first=True)
print(dummies)
```

Kết quả:

Pepsi	7Up
0	0
1	0
0	1
0	0

CocaCola là baseline.

Bước 3: Chia dữ liệu thành tập huấn luyện và kiểm tra

- Tập huấn luyện (train): ~70–80%
- Tập kiểm tra (test): ~20–30%

```
from sklearn.model_selection import train_test_split

X = df_features  # tập biến đầu vào
y = df_target    # biến mục tiêu

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

KẾT LUẬN

Bước	Mô tả
1	Xác định giả thuyết và tạo dữ liệu phù hợp
2	Chuyển đổi các biến phân loại thành dummy, tránh dummy trap
3	Chia dữ liệu thành tập huấn luyện và kiểm tra

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #1
Được tạo 15 tháng 5 2025 10:01:47 bởi Đỗ Ngọc Tú
Được cập nhật 15 tháng 5 2025 10:10:33 bởi Đỗ Ngọc Tú