Ngô Thị Huệ - D13CNPM5

Bài kiểm tra 1

Câu 1: Viết bài thu hoạch về các phương pháp mã hóa dữ liệu phạm trù category (dữ liệu giá trị cho bằng chữ)

Chuyển đổi dữ liệu phân loại sang dữ liệu số bao gồm hai bước:

* Mã hóa số nguyên
* Mã hóa một lần

1. Mã hóa số nguyên

Bước đầu tiên, mỗi giá trị danh mục duy nhất được gán một giá trị số nguyên.

Ví dụ: “đỏ” là 1, “xanh lá cây” là 2 và “xanh lam” là 3.

Đây được gọi là mã hóa nhãn hoặc mã hóa số nguyên và có thể dễ dàng đảo ngược. Đối với một số biến, điều này có thể là đủ. Các giá trị số nguyên có mối quan hệ có thứ tự tự nhiên giữa nhau và các thuật toán học máy có thể hiểu và khai thác mối quan hệ này.

Ví dụ: các biến thứ tự như ví dụ “địa điểm” ở trên sẽ là một ví dụ điển hình trong đó mã hóa nhãn là đủ.

2. Mã hóa một nóng

Đối với các biến phân loại không tồn tại mối quan hệ thứ tự như vậy, mã hóa số nguyên là không đủ.

Trên thực tế, việc sử dụng mã hóa này và cho phép mô hình giả định thứ tự tự nhiên giữa các danh mục có thể dẫn đến hiệu suất kém hoặc kết quả không mong muốn (dự đoán giữa các danh mục).

Trong trường hợp này, mã hóa một nóng có thể được áp dụng cho biểu diễn số nguyên. Đây là nơi biến được mã hóa số nguyên bị xóa và một biến nhị phân mới được thêm vào cho mỗi giá trị số nguyên duy nhất.

Trong ví dụ về biến “màu sắc”, có 3 danh mục và do đó cần có 3 biến nhị phân. Giá trị “1” được đặt trong biến nhị phân cho màu và giá trị “0” cho các màu khác.

Ví dụ:

|  |  |  |
| --- | --- | --- |
| red | green | blue |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

Các biến nhị phân thường được gọi là "biến giả" trong các trường khác, chẳng hạn như thống kê.

Câu 2: Trả lời câu hỏi: tại sao không thể dùng SVM cho bài toán phân lớp giá trị nhãn category được mã hóa bằng OneHotEncoding?

Bởi vì SVM không mang kiểu dữ liệu phạm trù là số

Tôi giả sử bạn đang hỏi về các tính năng phân loại, không phải biến mục tiêu, đã được giả định là phân loại (nhị phân) trong bộ phân loại SVM.

Đầu tiên, có hai loại phụ của đối tượng phân loại: Đối tượng địa lý thông thường và đối tượng địa lý danh nghĩa.

Ordinal có nghĩa là một “đơn đặt hàng” được ngụ ý. Ví dụ: số liệu về mức độ hài lòng của khách hàng {‘hài lòng’, ‘trung tính’, ‘không hài lòng’} là một biến số thứ tự vì chúng tôi có thể sắp xếp theo thứ tự: ‘hài lòng’> ‘trung tính’> ‘không hài lòng’. Ở đây, chúng ta có thể chỉ cần ánh xạ ký hiệu ‘string’ thành một ký hiệu số nguyên, ví dụ: ‘thoả mãn’ = 1, ‘trung tính’ = 0 và ‘không hài lòng’ = -1.

Nếu biến của chúng ta là danh nghĩa, thì một "order" không có ý nghĩa. Ví dụ, hãy nghĩ về ‘màu sắc’; Có một số trường hợp trong xử lý hình ảnh mà việc sắp xếp các giá trị màu có ý nghĩa, nhưng để đơn giản, chúng tôi không thể nói 'red> blue> yellow' hoặc tương tự. Để đối phó với các biến như vậy trong phân loại SVM, chúng tôi thường thực hiện mã hóa "một nóng". Ở đây, chúng tôi tạo cái gọi là biến giả có thể là giá trị nhị phân - chúng tôi tạo một biến giả cho mỗi giá trị có thể của biến đặc trưng danh nghĩa đó. Giả sử rằng biến màu của chúng tôi có thể có một trong ba giá trị: ‘red,’ ‘blue,’ ‘yellow.” Và Giả sử chúng ta có tập dữ liệu sau bao gồm 4 mẫu đào tạo:

mẫu 1: 'blue'

mẫu 2: 'màu vàng'

mẫu 3: 'red'

mẫu 4: 'màu vàng'

Sau đó, mã hóa một nóng của chúng tôi sẽ trông như thế này:

|  |  |  |  |
| --- | --- | --- | --- |
|  | blue | red | yellow |
| Mẫu 1 | 1 | 0 | 0 |
| Mẫu 2 | 0 | 0 | 1 |
| Mẫu 3 | 0 | 1 | 0 |
| Mẫu 4 | 0 | 0 | 1 |

Lưu ý rằng chỉ có một giá trị “true” (số nguyên 1) trong mỗi hàng, biểu thị cột cho mẫu đó trong tập huấn luyện. Mẫu 1 màu xanh lam; mẫu 2 là màu vàng, v.v.

Câu 3: Nêu ưu nhược điểm của phương pháp OHE.

* One-Hot-Encoding có lợi thế là kết quả là nhị phân chứ không phải thứ tự và mọi thứ nằm trong một không gian vectơ trực giao.
* Mã hóa one-hot là một cách biến đổi nhanh chóng từ dữ liệu dạng hạng mục sang dạng số.