**Trang 114**

Đầu vào (input) đầu ra (output)

Nhãn lớp

Phân loại thuộc tính

Thuộc tính (x)

Hình 3.2

* SƠ ĐỒ minh họa sợ đồ của một máy phân loại thuộc tính
* Chương này giới thiệu các khái niệm về phân loại , mô tả và một số vấn đề chính như giới hạn mô hình , chọn lựa mô hình và mô hình đánh giá .Các chủ đề này được minh họa bằng một số kĩ thuật phân loại gọi là cảm ứng quyết định .

**3.1 .một số khái niệm cơ bản**

- hình 3.2 mô tả ý tưởng của việc phân loại . Dữ liệu cho nhiệm vụ phân loại gồm tập bản ghi . mỗi trường hợp thì đặc trưng bởi tuple(x,y), trong đó x là tập các thuộc tính của bản ghi , y là nhãn lớp của bản ghi . tập thuộc tính x có thể chứa các thuộc tính thuộc bất kì loại nào , trong khi nhãn lớp y phải được phân loại.

- một mô hình phân loại được biểu diễn mô phỏng về quan hệ giữa thuộc tính và nhãn lớp . trong 2 chương tiếp theo , mô hình có thể biểu diễn bằng nhiều cách , ví dụ như dạng cây , bảng xác suất hoặc đơn giản là vecto các tham số có giá trị thực. Cụ thể hơn , chúng ta có thể biểu diễn nó một cách toán học như là một hàm mục tiêu f có đầu vào là thuôc tính x và tạo ra đầu ra tương ứng là các nhãn lớp dự đoán .mô hình được phân loại cụ thể một bản ghi (x,y) nếu f(x) = y

3.1

|  |  |  |
| --- | --- | --- |
| Bài tập | Thuộc tính | Nhãn lớp |
| Lọc thư rác | Các chức năng trích xuất từ thông điệp email tiêu đề và nội dung | thư rác hoặc không phải thư rác |
| Nhận dạng | Các tính năng được trích xuất từ quét cộng hưởng từ (MRI) | ác tính hoặc lành tính |
| Phân loại | Các tính năng được trích xuất từ hình ảnh kính viễn vọng | hình elip, xoắn ốc, hoặc  hình dạng không đều |

**Trang 115**

Bảng 3.1 cho thấy các ví dụ về các tập thuộc tính và nhãn lớp cho các nhiệm vụ phân loại khác nhau. Lọc thư rác và xác định khối u là ví dụ về các vấn đề phân loại nhị phân, trong đó mỗi trường hợp dữ liệu có thể được phân loại thành một trong hai lớp. Nếu số lượng lớp lớn hơn 2, như trong ví dụ phân loại thiên hà, thì nó được gọi là vấn đề phân loại đa lớp. Chúng tôi minh họa các khái niệm cơ bản của phân loại trong chương này với hai ví dụ sau.

Ví dụ 3.1. Bảng phân loại động vật có xương sống 3.2 cho thấy dữ liệu mẫu thiết lập để phân loại động vật có xương sống thành động vật có vú, bò sát, chim, cá và lưỡng cư. Tập thuộc tính bao gồm các đặc điểm của động vật có xương sống như nhiệt độ cơ thể, độ che phủ của da và khả năng bay. Tập dữ liệu cũng có thể là được sử dụng cho một nhiệm vụ phân loại nhị phân như phân loại động vật có vú, bằng cách nhóm các loài bò sát, chim, cá và động vật lưỡng cư thành một loại duy nhất được gọi là nonmammals.

Ví dụ 3.2. Phân loại khách hàng vay Xem xét vấn đề dự đoán liệu một người vay tiền sẽ trả nợ hoặc mặc định cho khoản vay thanh toán. Tập dữ liệu được sử dụng để xây dựng mô hình phân loại được hiển thị trong Bảng 3.3. Bộ thuộc tính bao gồm thông tin cá nhân của người đi vay, chẳng hạn như tình trạng hôn nhân và thu nhập hàng năm, trong khi nhãn lớp cho biết liệu người vay đã mặc định thanh toán khoản vay.

**Trang 116**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | Chủ nhà | Tình trạng hôn nhân | Thu nhập thường xuyên | Mặc định |
| 1 | Có | 1 mình | 125000 | Không |
| 2 | Không | Đã cưới | 100000 | Không |
| 3 | Không | 1 mình | 70000 | Không |
| 4 | Có | Đã cưới | 120000 | Không |
| 5 | Không | Ly hôn | 95000 | Có |
| 6 | Không | 1 mình | 60000 | Không |
| 7 | Có | Ly hôn | 220000 | Không |
| 8 | Không | 1 mình | 85000 | Có |
| 9 | Không | Đã cưới | 75000 | Không |
| 10 | không | 1 mình | 90000 | có |

Một mô hình phân loại có 2 nhiệm vụ quan trọng trong khai thác dữ liệu .đầu tiên , nó được sử dụng như 1 mô hình dự đoán để phân loại các trường hợp chưa được gắn nhãn trước đó . 1 mô hình tốt phải cung cấp dự đoán chính xác với thời gian phản hồi nhanh. Thứ 2 , nó phục vụ như một mô hình mô tả để xác định các đặc điểm phân biệt các thể hiện với các lớp khác nhau. Điều này đặc biệt hữu ích cho các ứng dụng quan trọng, chẳng hạn như chẩn đoán y tế, nơi không đủ để có một mô hình đưa ra dự đoán mà không cần chứng minh làm thế nào nó đạt được như vậy một quyết định. Ví dụ: mô hình phân loại được tạo ra từ tập dữ liệu động vật có xương sống thể hiện trong Bảng 3.2 có thể được sử dụng để dự đoán nhãn lớp sau đây động vật có xương sống:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tên động vật | Thân nhiệt | da | Sinh con | Sinh vật dưới nước | Sinh vật trên không | Có chân | Ngủ đông | Nhãn lớp |
| Quái vật gila | Máu lạnh |  | không | Không | không | Có | Có | ? |

**Trang 117**

Phân loại là nhiệm vụ gán nhãn cho các trường hợp dữ liệu chưa được gắn nhãn và phân loại được sử dụng để thực hiện một nhiệm vụ như vậy. Một phân loại thường được mô tả trong các điều khoản của một mô hình như được minh họa trong phần trước. Mô hình được tạo sử dụng một tập hợp các thể hiện đã cho, được gọi là tập huấn luyện, chứa các giá trị thuộc tính cũng như nhãn lớp cho mỗi thể hiện. Cách tiếp cận có hệ thống để học một mô hình phân loại được cung cấp một tập huấn luyện được gọi là học tập thuật toán. Quá trình sử dụng thuật toán học tập để xây dựng một phân loại mô hình từ dữ liệu đào tạo được gọi là cảm ứng. Quá trình này cũng thường được mô tả như là học tập mô hình của người Viking hoặc xây dựng mô hình. Quá trình này của áp dụng một mô hình phân loại trên các trường hợp kiểm tra không nhìn thấy để dự đoán lớp của họ nhãn được gọi là khấu trừ. Vì vậy, quá trình phân loại liên quan đến hai các bước: áp dụng thuật toán học tập để đào tạo dữ liệu để học mô hình và sau đó áp dụng mô hình để gán nhãn cho các trường hợp không được gắn nhãn. Hình 3.3 minh họa khung chung để phân loại.

Một kỹ thuật phân loại đề cập đến một cách tiếp cận chung để phân loại,ví dụ: kỹ thuật cây quyết định mà chúng ta sẽ nghiên cứu trong chương này. Điều nàykỹ thuật phân loại như hầu hết những người khác, bao gồm một gia đình của các mô hình liên quanvà một số thuật toán để học các mô hình này. Trong Chương 4, chúng tôisẽ nghiên cứu các kỹ thuật phân loại bổ sung, bao gồm các mạng lưới thần kinh vàhỗ trợ máy vector.Một vài lưu ý về thuật ngữ. Đầu tiên, các thuật ngữ phân loại của Wikipediathường được coi là đồng nghĩa. Nếu một kỹ thuật phân loại xây dựng một,mô hình toàn cầu, sau đó điều này là tốt. Tuy nhiên, trong khi mọi mô hình định nghĩa một trình phân loại,không phải mọi phân loại được định nghĩa bởi một mô hình duy nhất. Một số trình phân loại, chẳng hạn như trình phân loại lân cận knearest, không xây dựng mô hình rõ ràng (Mục 4.3), trong khi các phân loại khác, chẳng hạn như phân loại đồng bộ, kết hợp đầu ra của các mô hình bộ sưu tập (Mục 4.10). Thứ hai, thuật ngữ phân loại trực tuyến, thường được sử dụng theo nghĩa chung hơn để chỉ một kỹ thuật phân loại. Do đó, ví dụ, bộ phân loại cây quyết định, có thể tham khảo kỹ thuật phân loại cây quyết định hoặc một bộ phân loại cụ thể được xây dựng bằng kỹ thuật đó. May mắn thay, ý nghĩa của phân loại trên mạngthường rõ ràng từ bối cảnh.

Trang 118

Hướng dẫn mô hình học tập

Áp dụng mô hình

Mô hình

Học tập

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | Chủ nhà | Tình trạng hôn nhân | Thu nhập thường xuyên | Mặc định |
| 11 | Có | 1 mình | 125000 | ? |
| 12 | Không | Đã cưới | 100000 | ? |
| 13 | Không | 1 mình | 70000 | ? |
| 14 | Có | Đã cưới | 120000 | ? |
| 15 | Không | Ly hôn | 95000 | ? |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | Chủ nhà | Tình trạng hôn nhân | Thu nhập thường xuyên | Mặc định |  |  |  |  |  |
| 1 | Có | 1 mình | 125000 | Không |  |  |  |  |  |
| 2 | Không | Đã cưới | 100000 | Không |  |  |  |  |  |
| 3 | Không | 1 mình | 70000 | Không |  |  |  |  |  |
| 4 | Có | Đã cưới | 120000 | Không |  |  |  |  |  |
| 5 | Không | Ly hôn | 95000 | Có |  |  |  |  |  |
| 6 | Không | 1 mình | 60000 | Không |  |  |  |  |  |
| 7 | Có | Ly hôn | 220000 | Không |  |  |  |  |  |
| 8 | Không | 1 mình | 85000 | Có |  |  |  |  |  |
| 9 | Không | Đã cưới | 75000 | Không |  |  |  |  |  |
| 10 | không | 1 mình | 90000 | có |  |  |  |  |  |

Trang 119

Trong khung chung được chỉ ra trong Hình 3.3, các bước cảm ứng và khấu trừ nên được thực hiện riêng. Trong thực tế, như sẽ được thảo luận sau trong Phần 3.6, các bộ huấn luyện và kiểm tra phải độc lập với nhau để đảm bảo rằng mô hình cảm ứng có thể dự đoán chính xác các nhãn của lớp mà nó chưa từng gặp trước đây. Các mô hình cung cấp khả năng dự đoán như vậy được cho là có hiệu suất khái quát tốt. Hiệu suất của một mô hình (phân loại) có thể được đánh giá bằng cách so sánh dự đoán

nhãn chống lại nhãn thực sự của các trường hợp. Thông tin này có thể được tóm tắt trong một bảng gọi là ma trận nhầm lẫn. Bảng 3.4 mô tả ma trận nhầm lẫn cho một vấn đề phân loại nhị phân. Mỗi mục nhập fij biểu thị số lượng phiên bản từ lớp i được dự đoán là của lớp j. Ví dụ: f01 là số trường hợp từ lớp 0 dự đoán không chính xác là lớp 1. Số dự đoán đúng được thực hiện bởi mô hình là (f11 + f00) và số lượng dự đoán không chính xác là (f10 + f01).

Mặc dù ma trận nhầm lẫn cung cấp thông tin cần thiết để xác định mô hình phân loại thực hiện tốt như thế nào, tóm tắt thông tin này thành một số duy nhất giúp thuận tiện hơn để so sánh hiệu suất tương đối của các mô hình khác nhau. Điều này có thể được thực hiện bằng cách sử dụng một số liệu đánh giá như độ chính xác, được tính theo cách sau:

**Sự chính xác = Số lượng dự đoán đúng/Tổng số dự đoán.**

Đối với các vấn đề phân loại nhị phân, độ chính xác của một mô hình được đưa ra bởi:

**Sự chính xác= f11+f00 /f11+f10+f01+f00**

Tỷ lệ lỗi là một số liệu liên quan khác, được xác định như sau cho nhị phân vấn đề phân loại:

**Tỷ lệ lỗi = Số dự đoán sai/ Tổng số dự đoán = f10+f01/f11+f10+f01+f00**

Các thuật toán học tập của hầu hết các kỹ thuật phân loại được thiết kế để tìm hiểu các mô hình đạt độ chính xác cao nhất hoặc tương đương, tỷ lệ lỗi thấp nhất khi áp dụng cho bộ kiểm tra