**Chương 3: Phân Loại**

Bộ thuộc tính đầu vào(x) Mô hình phân loại Nhãn lớp đầu ra (y)

***Hình 3.2*** *Một minh họa sơ đồ của một nhiệm vụ phân loại*

Chương này giới thiệu các khái niệm cơ bản về phân loại và mô tả một số vấn đề chính của nó như quá mức mô hình, lựa chọn mô hình và mô hình đánh giá. Trong khi các chủ đề này được minh họa bằng cách sử dụng một kỹ thuật phân loại được gọi là cảm ứng cây quyết định, hầu hết các cuộc thảo luận trong chương này là cũng áp dụng cho các kỹ thuật phân loại khác, nhiều trong số đó được bảo hiểm trong Chương 4.

**3.1 Khái niệm cơ bản**

Hình 3.2 minh họa ý tưởng chung đằng sau phân loại. Dữ liệu cho một nhiệm vụ phân loại bao gồm một tập hợp các trường hợp (hồ sơ). Mỗi cái như vậy cá thể được đặc trưng bởi tuple (x, y), trong đó x là tập hợp thuộc tính các giá trị mô tả thể hiện và y là nhãn lớp của thể hiện. Các tập thuộc tính x có thể chứa các thuộc tính thuộc bất kỳ loại nào, trong khi nhãn lớp y phải được phân loại. Một mô hình phân loại là một đại diện trừu tượng của mối quan hệ giữa tập thuộc tính và nhãn lớp. Như sẽ thấy trong phần tiếp theo hai chương, mô hình có thể được trình bày theo nhiều cách, ví dụ, như một cái cây, một bảng xác suất, hoặc đơn giản, một vectơ của các tham số có giá trị thực. Chính thức hơn, chúng ta có thể biểu diễn nó một cách toán học như là một hàm mục tiêu f dùng làm đầu vào tập thuộc tính x và tạo ra một đầu ra tương ứng với lớp dự đoán nhãn. Mô hình được cho là phân loại chính xác một thể hiện (x, y) nếu f (x) = y.

***Bảng 3.1.*** *Ví dụ về các nhiệm vụ phân loại*

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

***Bảng 3.2****. Một dữ liệu mẫu cho vấn đề phân loại động vật có xương sống.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tên động vật có xương sống | Thân nhiệt | Bao da | Sinh con | Sinh vật dưới nước | Sinh vật trên không | Có chân | Ngủ đông | Nhãn lớp |
| Nhân loại  Trăn  cá hồi  cá voi  ếch  komodo  rồng  con dơi  chim bồ câu  con mèo  báo  cá mập  rùa  chim cánh cụt  nhím  lươn  kỳ nhông | Máu nóng  Máu lạnh  máu lạnh  máu nóng  máu lạnh  máu lạnh  máu nóng  máu nóng  máu nóng  máu lạnh  máu lạnh  máu nóng  máu nóng  máu lạnh  máu lạnh | Tóc  quy mô  quy mô  tóc  không có  quy mô  tóc  quy mô  quy mô  quy mô  không có | Đúng  Không  Không  Đúng  Không  Không  Đúng  Không  Đúng  Đúng  Không  Không  Đúng  Không  không | Không  Không  Đúng  Đúng  Cả hai  Không  Không  Không  Không  Đúng  Cả hai  Cả hai  Không  Đúng  Cả hai | Không  Không  Không  Không  Không  Không  Đúng  Đúng  Không  Không  Không  Không  Không  Không  Không | Đúng  Không  Không  Không  Đúng  Đúng  Đúng  Đúng  Đúng  Không  Đúng  Đúng  Đúng  Không  Đúng | Không  Đúng  Không  Không  Đúng  Không  Đúng  Không  Không  Không  Không  Không  Đúng  Không  Đúng | động vật có vú  bò sát  Cá  động vật có vú  động vật lưỡng cư  bò sát  động vật có vú  chim  động vật có vú  cá  bò sát  chim  động vật có vú  cá  động vật lưỡng cư |

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

**Ví dụ 3.1. [Phân loại động vật có xương sống]** Bảng 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 của 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 không trả được nợ

***Bảng 3.3.*** *Một dữ liệu mẫu cho vấn đề phân loại người vay*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Chủ nhà | *Tình trạng hôn nhân* | *Thu nhập hàng năm* | *Mặc định?* |
| *1* | *Đúng* | *Độc thân* | *125000* | *không* |
| *2* | *Không* | *Đã kết hôn* | *100000* | *Không* |
| *3* | *Không* | *Độc thân* | *70000* | *Không* |
| *4* | *Có* | *Đã kết hôn* | *120000* | *Không* |
| *5* | *Không* | *Đã ly hôn* | *95000* | *Có* |
| *6* | *Không* | *Độc thân* | *60000* | *Không* |
| *7* | *Có* | *Đã ly hôn* | *220000* | *Không* |
| *8* | *Không* | *Độc thân* | *85000* | *Có* |
| *9* | *Không* | *Đã kết hôn* | *75000* | *Không* |
| *10* | *Không* | *Độc thân* | *90000* | *có* |

Một mô hình phân loại phục vụ hai vai trò quan trọng trong khai thác dữ liệu. Đầu tiên, nó là được sử dụng như một mô hình dự đoán để phân loại các trường hợp chưa được gắn nhãn trước đó. Tốt mô hình phân loại phải cung cấp dự đoán chính xác với phản ứng nhanh thời gian. Thứ hai, nó phục vụ như một mô hình mô tả để xác định các đặc điểm phân biệt các trường hợp từ 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 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 có xương sống | Thân nhiệt | Bao da | Sinh con | Sinh vật dưới nước | Sinh vật trên không | Có chân | Ngủ đông | Nhãn lớp |
| quái vật gila | máu lạnh | quy mô | Không | Không | Không | Đúng | Đúng | ? |

Ngoài ra, nó có thể được sử dụng như một mô hình mô tả để giúp xác định các đặc điểm xác định động vật có xương sống là động vật có vú, bò sát, chim, cá hoặc cá động vật lưỡng cư. Ví dụ, mô hình có thể xác định động vật có vú là máu nóng động vật có xương sống sinh con

Có một số điểm đáng chú ý liên quan đến ví dụ trước. Đầu tiên, mặc dù tất cả các thuộc tính được hiển thị trong Bảng 3.2 là định tính, nhưng có không hạn chế về loại thuộc tính có thể được sử dụng làm biến dự đoán. Mặt khác, nhãn lớp phải là loại danh nghĩa. Điều này phân biệt phân loại từ các nhiệm vụ mô hình dự đoán khác như hồi quy, trong đó giá trị dự đoán thường là định lượng. Thêm thông tin về hồi quy có thể được tìm thấy trong Phụ lục D.

Một điểm đáng chú ý là không phải tất cả các thuộc tính đều có liên quan đến nhiệm vụ phân loại. Ví dụ: chiều dài hoặc trọng lượng trung bình của một động vật có xương sống có thể không hữu ích để phân loại động vật có vú, vì những thuộc tính này có thể hiển thị cùng một giá trị cho cả động vật có vú và không có vú. Thật là một thuộc tính thường bị loại bỏ trong quá trình tiền xử lý. Các thuộc tính còn lại có thể không thể tự phân biệt các lớp và do đó, phải được sử dụng trong hòa nhạc với các thuộc tính khác. Chẳng hạn, thuộc tính Nhiệt độ cơ thể không đủ để phân biệt động vật có vú với các động vật có xương sống khác. Khi nó được sử dụng cùng với sinh ra, việc phân loại động vật có vú giúp cải thiện đáng kể không ngừng Tuy nhiên, khi bao gồm các thuộc tính bổ sung, chẳng hạn như Skin Cover, mô hình trở nên quá cụ thể và không còn bao gồm tất cả các động vật có vú. Phát hiện sự kết hợp tối ưu của các thuộc tính phân biệt tốt nhất các trường hợp từ các lớp khác nhau là thách thức chính trong việc xây dựng các mô hình phân loại.

**3.2 Khung chung cho phân loại**

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 thuộc tính các giá trị cũng như nhãn lớp cho từng 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 này kỹ thuật phân loại như hầu hết những người khác, bao gồm một gia đình của các mô hình liên quan và một số thuật toán để học các mô hình này. Trong Chương 4, chúng tôi sẽ nghiên cứu các kỹ thuật phân loại bổ sung, bao gồm các mạng lưới thần kinh và hỗ trợ máy vector. Một vài lưu ý về thuật ngữ. Đầu tiên, các thuật ngữ phân loại của Wikipedia thường 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 phân loại, chẳng hạn như k-nearest phân loại lân cận, không xây dựng một mô hình rõ ràng (Phần 4.3), trong khi

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Chủ nhà | Tình trạng hôn nhân | Thu nhập hàng năm | người vay mặc định |
| 1  2  3  4  5  6  7  8  9  10 | Đúng  Không  Không  Đúng  Không  Không  Đúng  Không  Không  Không | Độc thân  Đã kết hôn  Độc thân  Đã kết hôn  Đã ly hôn  Đã kết hôn  Đã ly hôn  Độc thân  Đã kết hôn  Độc thân | 125k  100k  70k  120k  95k  60k  220k  85k  75k  90k | Không  Không  Không  Không  Có  Không  Không  Có  Không  Có |

Thuật toán học

**Giới thiệu**:

“Mô hình học”

Mô hình

**Sự phá hủy**

“áp dụng mô hình”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Chủ nhà | Tình trạng hôn nhân | Thu nhập hàng năm | người vay mặc định |
| 11  12  13  14  15 | Không  Đúng  Đúng  Không  Không | Đã kết hôn  Đã ly hôn  Độc thân  Độc thân  Đã kết hôn | 55k  80k  110k  95k  67k | ?  ?  ?  ?  ? |

các phân loại khác, chẳng hạn như phân loại đồng bộ, kết hợp đầu ra của một bộ sưu tập của các mô hình (Mục 4.10). Thứ hai, thuật ngữ phân loại phân loại trực tuyến, thường được sử dụng nhiều hơn ý nghĩa chung để đề cập đến một kỹ thuật phân loại. Vì vậy, ví dụ, quyết định của người Viking phân loại cây có thể tham khảo kỹ thuật phân loại cây quyết định hoặc cụ thể phân loại được xây dựng bằng cách sử dụng kỹ thuật đó. May mắn thay, ý nghĩa của phân loại trên mạng thường rõ ràng từ bối cảnh. Trong khung chung được chỉ ra trong Hình 3.3, quy nạp và khấu trừ các bước nên được thực hiện riêng biệt. Trong thực tế, như sẽ được thảo luận sau trong Mục 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 lớp trường hợp nó chưa bao giờ gặp phải trước đây. Các mô hình cung cấp dự đoán như vậy hiểu biết được cho là có hiệu suất khái quát tốt. Buổi biểu diễn 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 trường hợp từ lớp tôi dự đoán là của lớp j. Ví dụ: f01 là

***Bảng 3.4****. Ma trận nhầm lẫn cho một vấn đề phân loại nhị phân.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Lớp dự đoán | |
| *Lớp = 1* | *Lớp = 0* |
| *Lớp học thực tế* | *Lớp = 1* | *f11* | *f10* |
| *Lớp = 0* | *f01* | *f00* |

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:

Độ chính xác = (3.1)

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

Độ chính xác = (3.2)

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 == (3.3)