Chương 3: Classification (Phân lớp)

Input output

Thực hiện phân lớp

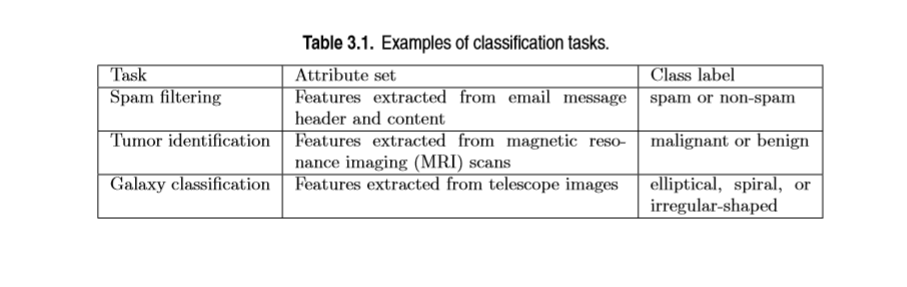
Tập thuộc thính Lớp được gán nhãn

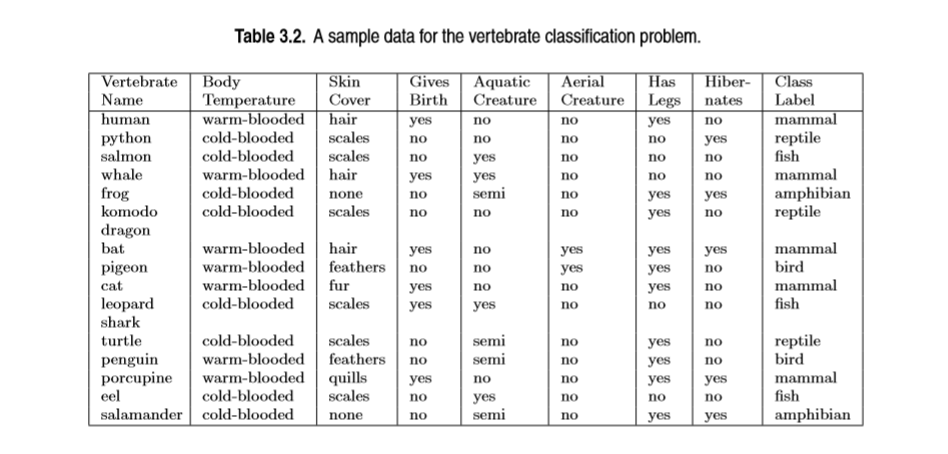
Bài học này giới thiệu về những khái niệm cơ bản của Classification và những vấn đề quan trong của nó như là model overfitting, model selection, và model evaluation. Ngoài những đề tài được minh họa sử dụng kỹ thuật phân lớp là Decision tree thì trong bài học này, chúng ta cũng sẽ thảo luận về những kỹ thuật phân lớp khác, một vài trong số đó sẽ được đề cập tại chương 4.

3.1 Những khái niệm cơ bản

Hình 3.2 minh họa ý tưởng chung của việc phân loại. Dữ liệu cho một kỹ thuật phân loại bao gồm một tập hợp các instances (bản ghi). Mỗi instance được thể hiện bởi tuple(x, y), trong đó x là tập hợp các thuộc tính mô tả instance và y là lớp được gán nhãn của instance. Tập thuộc tính x có thể chứa thuộc tính thuộc bất kỳ loại nào, trong khi lớp y phải được gán nhãn để phân loại.

Một mô hình phân loại là một biểu diễn trừu tượng về mối quan hệ giữa tập thuộc tính và nhãn lớp. Như sẽ thấy trong hai chương tiếp theo, mô hình có thể được biểu diễn theo nhiều cách, ví dụ như dưới dạng cây, bảng xác suất hoặc đơn giản hơn là một vectơ của các tham số có giá trị thực. Một cách thông dụng hơn, chúng ta có thể biểu diễn nó theo cách toán học như là một hàm mục tiêu f lấy đầu vào là tập thuộc tính x và tạo ra một đầu ra tương ứng với nhãn lớp dự đoán. Mô hình được cho là phân loại chính xác một thể instance (x, y) nếu f (x) = y.

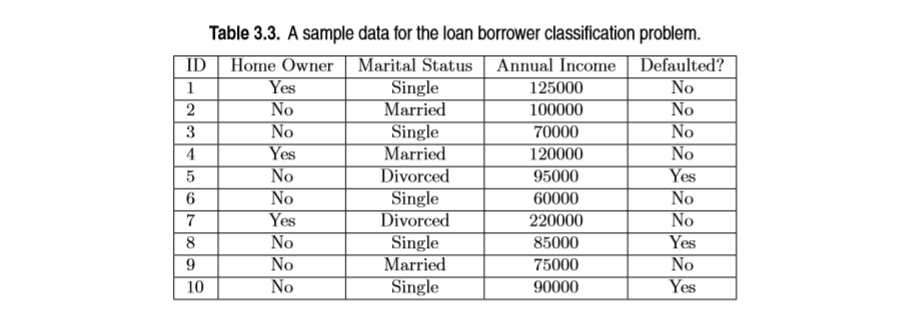




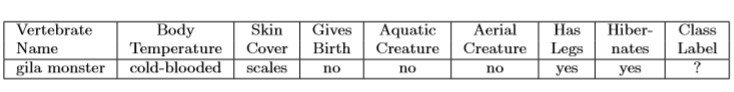
Bảng 3.1 cho thấy các ví dụ về các tập thuộc tính và nhãn lớp cho các công việc phân loại khác nhau. Lọc thư rác và nhận dạng khối u là các ví dụ về các vấn đề phân loại nhị phân, mỗi dữ liệu instance có thể được phân loại thành một trong hai lớp. Nếu số lượng các lớp lớn hơn 2, như trong ví dụ phân loại thiên hà, thì nó được gọi là 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. [Phân loại động vật có xương sống] Bảng 3.2 cho thấy một bộ dữ liệu mẫu để phân loại động vật có xương sống thành động vật có vú, bò sát, chim, cá và lưỡng cư. Bộ thuộc tính bao gồm các đặc điểm của động vật có xương sống như nhiệt độ cơ thể, lớp da và khả năng bay. Tập dữ liệu cũng có thể được sử dụng công việc phân loại nhị phân, chẳng hạn như phân loại động vật có vú, bằng cách nhóm các loài bò sát, chim, cá và động vật lưỡng cư vào một loại duy nhất được gọi là nonmammals.

Ví dụ 3.2. [Phân loại đói tượng vay tiền] Xem xét vấn đề dự đoán liệu người vay có trả được khoản vay hay không. 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 vay như tình trạng hôn nhân và thu nhập hàng năm, trong khi nhãn lớp cho biết liệu người vay có thể thanh toán các khoản vay hay không.



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

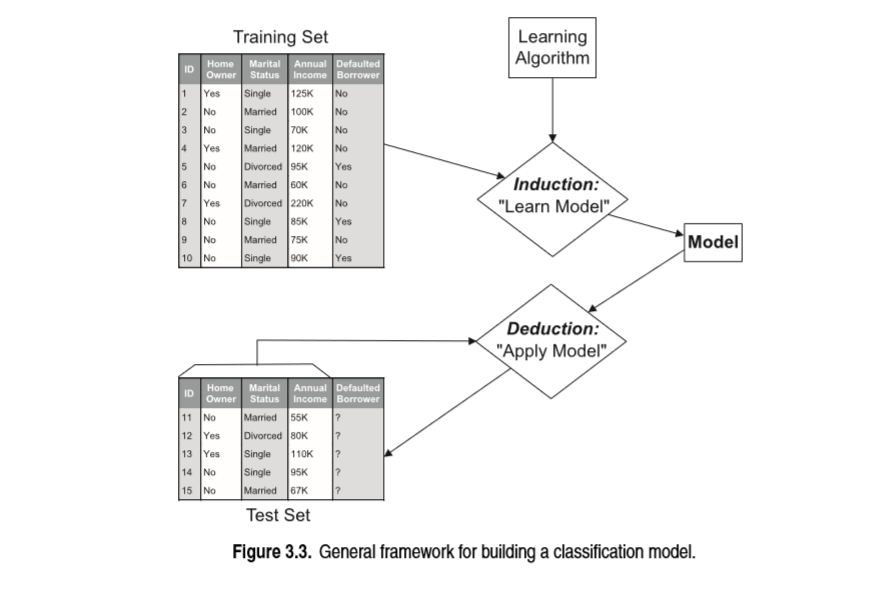


Thêm vào đó, nó có thể được sử dụng như một mô hình mô tả để giúp xác định các đặc điểm xác định động vật có xương sống như động vật có vú, bò sát, chim, cá hoặc lưỡng cư. Ví dụ, mô hình có thể xác định động vật có vú là động vật có xương sống, máu ấm, sinh ra con non. Có một số điểm đáng chú ý liên quan đến ví dụ trước. Đầu tiên, mặc dù tất cả các thuộc tính được hiển thị trong Bảng 3.2 là định tính, không có hạn chế nào về loại thuộc tính có thể được sử dụng làm biến dự báo. Nhãn được gán lớp mặt khác phải là loại danh nghĩa. Điều này phân biệt Classification với các mô hình dự đoán khác như hồi quy, trong đó giá trị dự đoán thường là định lượng. Thông tin thêm về hồi quy có thể được tìm thấy trong Phụ lục D.

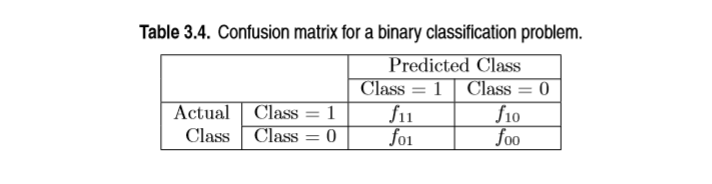
Một điểm đáng chú ý khác là không phải tất cả các thuộc tính có thể liên quan đến công việc phân loại. Ví dụ, chiều dài hoặc trọng lượng trung bình của động vật có thể không hữu ích để phân loại động vật có vú, vì các thuộc tính này có thể hiển thị cùng một giá trị cho cả động vật có vú và không phải động vật có vú. Một thuộc tính như vậy thường bị loại bỏ trong quá trình tiền xử lý. Các thuộc tính còn lại có thể không thể tự phân biệt các lớp dựa vào bản thân chúng và do đó, phải được sử dụng cùng với các thuộc tính khác. Chẳng hạn, thuộc tính Body Temperature không cần thiết để phân biệt động vật có vú với các động vật có xương sống khác. Nhưng khi nó được sử dụng cùng với Birth, việc phân loại động vật có vú giúp cải thiện đáng kể. Tuy nhiên, khi bao gồm các thuộc tính bổ sung, chẳng hạn như Skin Cover, mô hình sẽ trở nên quá đặc biệt và không còn bao gồm tất cả các động vật có vú. Tìm ra sự kết hợp tối ưu của các thuộc tính phân biệt tốt nhất của các instance từ các lớp khác nhau là thách thức lớn trong việc xây dựng các mô hình Classification.

3.2 Framework chung cho Classification.

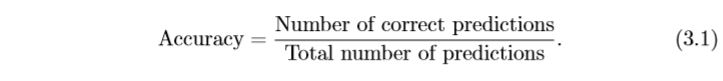
Classification là công việc gán nhãn cho các trường hợp dữ liệu chưa được gắn nhãn. Một Classiﬁer thường được mô tả như các mô hình được minh họa trong phần trước. Mô hình được tạo bằng cách sử dụng một tập hợp các instance đã cho, được gọi là tập huấn luyện, chứa các giá trị thuộc tính cũng như nhãn lớp cho mỗi instance. Cách tiếp cận có hệ thống để học một mô hình Classification được cung cấp một tập huấn luyện được gọi là thuật toán học tập. Quá trình sử dụng thuật toán học tập để xây dựng mô hình phân loại từ dữ liệu đào tạo được gọi là quy nạp. Quá trình này cũng thường được mô tả như là học tập mô hình của người Viking hoặc xây dựng mô hình. Quá trình áp dụng mô hình phân loại này vào các trường hợp thử nghiệm chưa thấy để dự đoán nhãn lớp của chúng được gọi là khấu trừ. Do đó, quá trình phân loại bao gồm hai bước: áp dụng thuật toán học tập để Training set có thể 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 Framework chung cho phân Classification. Một kỹ thuật Classification đề cập đến một cách tiếp cận chung cho phân loại, ví dụ, kỹ thuật Decision tree mà chúng ta sẽ nghiên cứu trong chương này. Kỹ thuật phân loại này giống như hầu hết các kỹ thuật khác, bao gồm một nhóm các mô hình liên quan và một số thuật toán để học các mô hình này. Trong Chương 4, chúng ta sẽ nghiên cứu các kỹ thuật phân loại bổ sung, bao gồm neural networks và support vector machines. Một vài lưu ý về thuật ngữ. Đầu tiên, các thuật ngữ của “Classiﬁer” và “Model” thường được coi là đồng nghĩa.



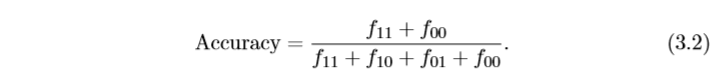
Trong khung chung được chỉ ra trong Hình 3.3, các bước cảm ứng và khấu trừ phải được thực hiện riêng. Trong thực tế, như sẽ được thảo luận sau trong Phần 3.6, các bộ huấn luyện và kiểm tra phải độc lập với nhau để đảm bảo rằng mô hình cảm ứng có thể dự đoán chính xác các nhãn lớp của các thể hiện mà nó chưa từng gặp trước đây. Các mô hình cung cấp những hiểu biết dự đoán như vậy được cho là có hiệu suất khái quát tốt. Hiệu năng của một mô hình (phân loại) có thể được đánh giá bằng cách so sánh các nhãn dự đoán với các nhãn thực thể của các thể hiện. 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 được dự đoán không chính xác là lớp 1. Số lượng dự đoán đúng được thực hiện bởi mô hình là (f11 + f00) và số lượng dự đoán không chính xác là (f10 + f01).



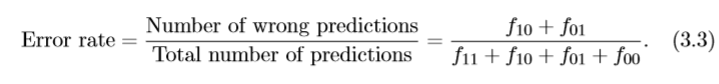
Mặc dù ma trận nhầm lẫn cung cấp thông tin cần thiết để xác định mô hình phân loại hoạt động tốt như thế nào, việc tóm tắt thông tin này thành một số giúp thuận tiện hơn khi so sánh hiệu suất tương đối của các mô hình khác nhau. Điều này có thể được thực hiện bằng cách sử dụng một số liệu đánh giá, chẳng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



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



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



Các thuật toán học tập của hầu hết các kỹ thuật phân loại được thiết kế để học các mô hình đạt độ chính xác cao nhất hoặc tương đương, tỷ lệ lỗi thấp nhất khi áp dụng cho bộ kiểm tra. Chúng tôi sẽ xem xét lại chủ đề đánh giá mô hình trong Phần 3.6.