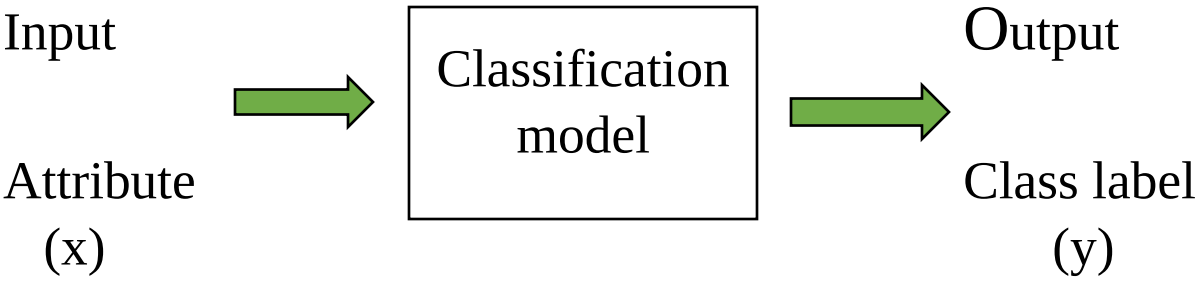
Phân lớp:Các khái niệm cơ bản và kĩ thuật

Chương 3:Phân lớp(Classification)



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

Chương này giới thiệu các khái niệm cơ bản về phân lớp 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 lớp được gọi là cảm ứng cây quyết định, hầu hết các nội dung trong chương này cũng áp dụng cho các kỹ thuật phân lớp khác, nhiều trong số đó được bao gồm trong Chương 4.

3.1 Khái niệm cơ bản(Basic Concepts)

Hình 3.2 minh họa ý tưởng chung đằng sau phân lớp. Dữ liệu cho một nhiệm vụ phân lớp bao gồm một tập hợp các trường hợp (records). Mỗi trường hợp như vậy cá thể được đặc trưng bởi tuple (x, y), trong đó x là tập hợp thuộc tính các giá trị mô tả và y là nhãn lớp của trường hợp. Các tập thuộc tính x có thể chứa các thuộc tính thuộc bất kỳ loại nào, trong khi nhãn lớp y phải được phân lớp.

Một mô hình phân lớp là một đại diện trừu tượng của mối quan hệ giữa tập thuộc tính và nhãn lớp. Như sẽ thấy trong phần tiếp theo của 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 trường hợp (x, y) nếu f (x) = y

Bảng 3.1 . Những ví dụ về nhiệm vụ Phân lớp

Nhiệm vụ (Task)	Tập thuộc tính(Attribute set)	Nhãn lớp(Class label)
Lọc thư rác(Spam filtering)	Các đặc tính được trích xuất từ thông điệp tiêu đề và nội dung của email	Spam or non-spam
Nhận dạng khối u(Tumor identification)	Các đặc tính được trích xuất từ quét cộng hưởng từ (MRI)	Ác tính hoặc lành tính(malignant or benign)
Phân loại thiên hà(Galaxy classification)	Các đặc tính được trích xuất từ hình ảnh kính viễn vọng	Hình elip, xoắn ốc, hoặc hình dạng không đều (elliptical, spiral, or irregular-shaped)

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

Vertebrate Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
Human	warm-blooded	hair	yes	no	no	yes	no	mammal
Python	cold-blooded	scales	no	no	no	no	yes	reptile
Salmon	cold-blooded	scales	no	yes	no	no	no	fish
Whale	warm-blooded	hair	yes	yes	no	no	no	mammal
Frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
Komodo Dragon	cold-blooded	scales	no	no	no	yes	no	reptile
Bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
Pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
Cat	warm-blooded	fur	yes	no	no	yes	no	mammal
Leopard Shark	cold-blooded	scales	yes	yes	no	no	no	fish
Turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
Penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
Porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
Eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

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

Ví dụ 3.1. [Phân lớp 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 lớp 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 mặc định thanh toán khoản vay.

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

ID	Home Owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125000	No
2	No	Married	100000	No
3	No	Single	70000	No
4	Yes	Married	120000	No
5	No	Divorced	95000	Yes
6	No	Single	60000	No
7	Yes	Divorced	220000	No
8	No	Single	85000	Yes
9	No	Married	75000	No
10	No	Single	90000	Yes

Một mô hình phân loại phục vụ hai vai trò quan trọng trong khai thác dữ liệu. Đầu tiên, nó được sử dụng như một **mô hình dự đoán(predictive model)** 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 với phản ứng thời gian nhanh. Thứ hai, nó phục vụ như **một mô hình mô tả(descriptive model)** 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 một quyết định như vậy.

Ví dụ: mô hình phân lớp 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 động vật có xương sống sau đây :

Vertebrate Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
gila monster	cold-blooded	scales	no	no	no	yes	yes	?

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 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 lớp 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.

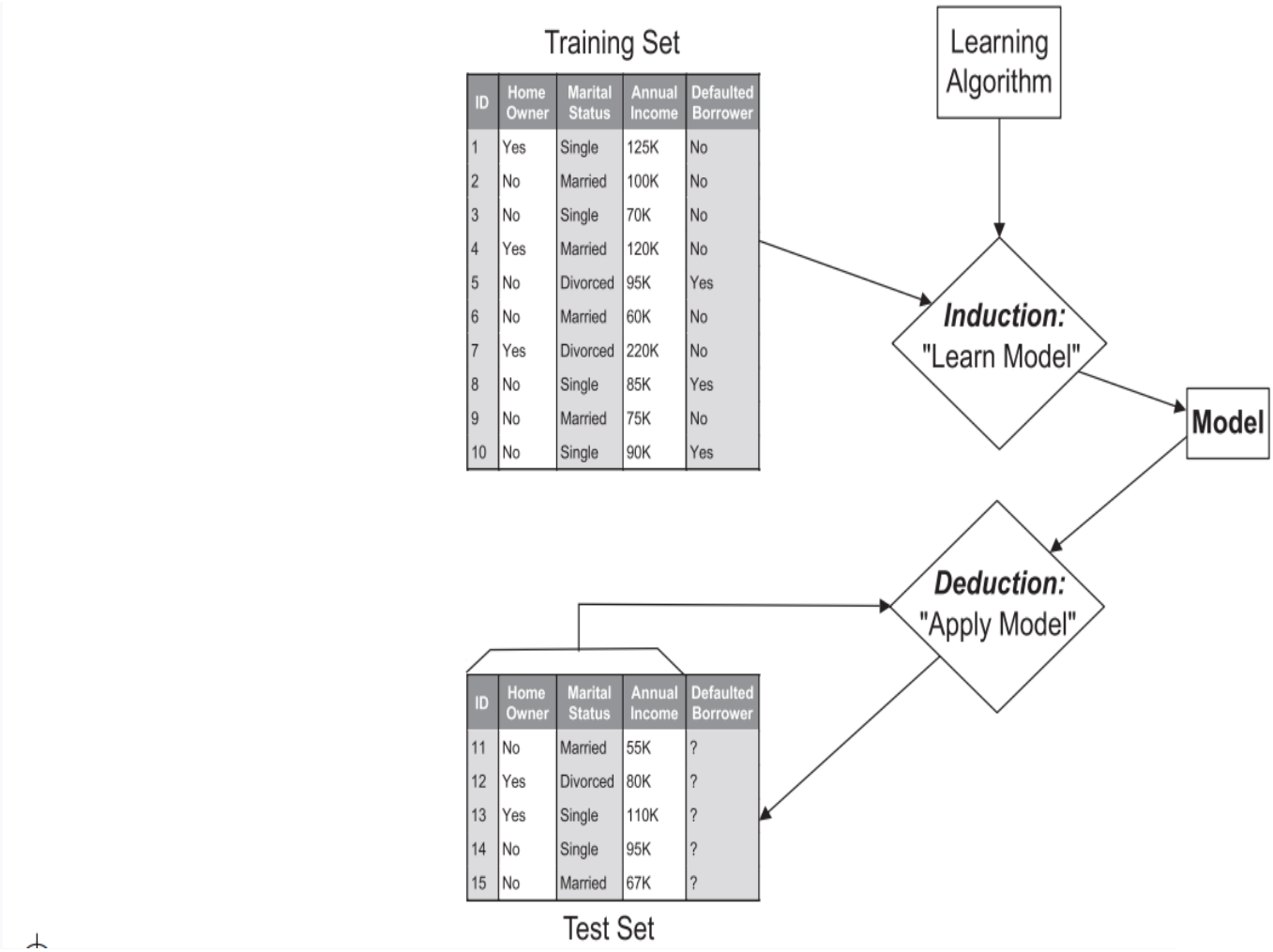
Một điểm đáng chú ý là không phải tất cả các thuộc tính đều có liên quan đến nhiệm vụ phân lớp. Ví dụ, chiều dài hoặc trọng lượng trung bình của động vật có xương sống có thể không hữu ích để phân loại động vật có vú, vì những thuộc tính này có thể hiển thị cùng một giá trị cho cả động vật có vú và không 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 **concert** với các thuộc tính khác. Chẳng hạn, thuộc tính **Body Temperature** không đủ để phân biệt động vật có vú với các động vật có xương sống khác. Khi nó được sử dụng cùng với **Gives Birth**, việc phân loại động vật có vú được cải thiện đáng kể. 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 lớp.

3.2 Khung phân lớp chung(General Framework for Classification)

Phân lớp là nhiệm vụ gán nhãn cho các trường hợp dữ liệu chưa được gắn nhãn và **trình phân lớp(classifier)** được sử dụng để thực hiện một tác vụ như vậy. Một bộ phân loại thường được mô tả theo các mô hình như được minh họa trong phần trước. Mô hình được tạo bằng cách sử dụng một tập hợp các thể hiện đã cho, được gọi là **tập huấn luyện(training set)**, chứa các giá trị thuộc tính cũng như nhãn lớp cho mỗi thể hiện. Phương pháp hệ thống để học một mô hình phân loại được cung cấp một tập huấn luyện được gọi là **thuật toán học tập(learning algorithm)**. 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(induction)**. 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 một mô hình phân loại trên các trường hợp thử nghiệm chưa thấy để dự đoán nhãn lớp của chúng được gọi là **khấu trừ(deduction)**. Do đó, quá trình phân loại bao gồm hai 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 lớp đề cập đến một cách tiếp cận chung để phân lớp, ví dụ, kỹ thuật cây quyết định mà chúng ta sẽ nghiên cứu trong chương này. Kỹ thuật phân lớp 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 tôi sẽ nghiên cứu các kỹ thuật phân loại bổ sung, bao gồm các mạng thần kinh và các máy vectơ hỗ trợ.

Một vài lưu ý về thuật ngữ. Đầu tiên, các thuật ngữ **"phân lớp(classifier)"** và **"các mô hình(model)"**, các mô hình phân tích thường được sử dụng để đồng nghĩa. Nếu một kỹ thuật phân lớp xây dựng một mô hình toàn cầu duy nhất, thì điều này là tốt. Tuy nhiên, trong khi mọi mô hình định nghĩa một trình phân lớp, không phải mọi phân lớp đều được xác định bởi một mô hình duy nhất. Một số trình phân lớp, chẳng hạn như **trình phân lớp K-nearest neighbor**, không xây dựng mô hình rõ ràng (Mục 4.3), trong khi



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	55K	?
12	Yes	Divorced	80K	?
13	Yes	Single	110K	?
14	No	Single	95K	?
15	No	Married	67K	?

Figure 3.3. Khung chung để xây dựng mô hình phân lớp.

các phân lớp khác, chẳng hạn như phân lớp đồng bộ, kết hợp đầu ra của một tập hợp các mô hình (Mục 4.10). Thứ hai, thuật ngữ **phân lớp(classifier)** thường được sử dụng theo nghĩa chung hơn để chỉ một kỹ thuật phân lớp. Vì vậy, ví dụ, bộ phân lớp cây quyết định, có thể tham khảo kỹ thuật phân lớp cây quyết định hoặc cụ thể

phân lớp được xây dựng bằng cách sử dụng kỹ thuật đó. May mắn thay, ý nghĩa của **"bộ phân lớp(classifier)"** trên YouTube thường rõ ràng từ ngữ cảnh.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 trường hợp mà nó chưa từng gặp trước đây. Các mô hình cung cấp dự đoán như vậy hiểu biết được cho là có **hiệu suất khái quát(generalization performance)** tốt. Hiệu suất của một mô hình **(phân lớp)** có thể được đánh giá bằng cách so sánh các nhãn dự đoán với các nhãn thực tế của các thể hiện. Thông tin này có thể được tóm tắt trong một bảng gọi là **ma trận lỗi(confusion matrix). Bảng 3.4** mô tả **Confusion matrix** cho một vấn đề phân lớp 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

		Predicted Class	
		Class = 1	Class = 0
Actual Class	Class = 1	*f*11	*f*10
	Class = 0	*f*01	*f*00

Bảng 3.4 . Confusion Matrix cho một vấn đề phân lớp nhị phân.

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 **Confusion Matrix** cung cấp thông tin cần thiết để xác định mô hình phân lớp 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(evaluation metric)**, chẳng hạn như **độ chính xác(accuracy)**, được tính theo cách sau:

$$Accuracy = \frac{\textbf{Number of correct predictions}(\textbf{Số lượng dự đoán đúng})}{\textbf{Total number of predictions}(\textbf{Tổng số lượng dự đoán})}$$

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

$$Accuracy = \frac{f11+f00}{f11+f10+f01+f00}$$

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

$$Error\ rate = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions} = \frac{f10+f01}{f11+f10+f01+f00}$$

Các thuật toán học tập của hầu hết các kỹ thuật phân lớp 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.