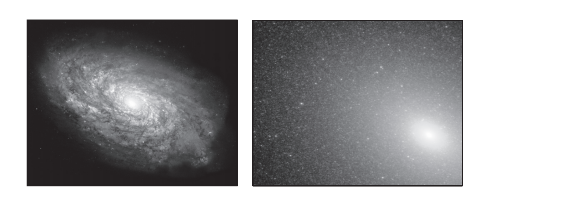
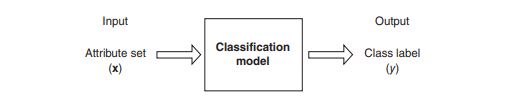
**Phân Lớp: Các Khái Niệm Và Kỹ Thuật Cơ Bản**

Con người có khả năng bẩm sinh để phân mọi thứ thành các loại, ví như việc lọc thư rác, hoặc đặc biệt hơn một chút là nhận ra các vật thể trên trời từ ảnh của kính viễn vọng(Hình 3.1). Việc phân loại thủ công thường chỉ dành cho những tập dữ liệu nhỏ và đơn giản, nhưng khi dữ liệu lớn và phức tạp hơn thì cần một giải pháp tự động.



1. Một ngân hà hình xoắn ốc. (b) Một ngân hà hình elip.

**Hình 3.1**. Phân loại ngân hà từ ảnh của kính viễn vọng được lấy từ website NASA.



**Hình 3.2**. Biểu đồ minh họa cho việc phân lớp.

Chương này giới thiệu những khái niệm cơ bản về phân loại và mô tả vài vấn đề chính như model overfitting, model selection, và model evaluation. Phương pháp phân loại - Cây quyết định sẽ được dùng để minh họa cho các chủ đề ở trên, phần lớn nội dung trong chương này cũng có thể áp dụng cho các phương pháp phân loại khác, nhiều thứ sẽ được thảo luận ở chương 4.

**3.1 Các khái niệm cơ bản**

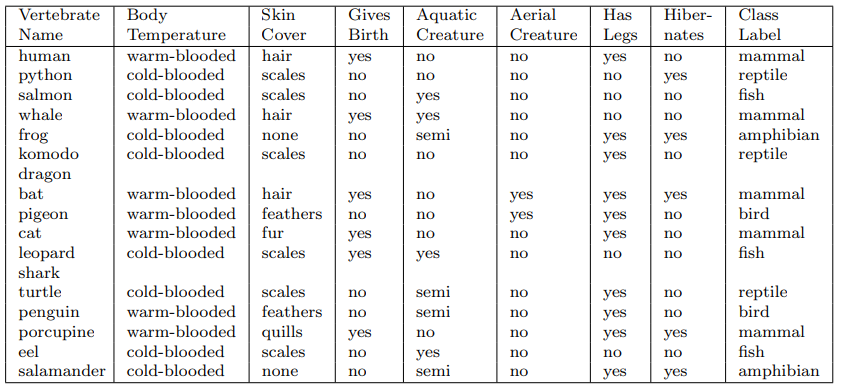
Hình 3.2 minh họa ý tưởng chung đằng sau việc phân loại. Dữ liệu cho một nhiệm vụ phân loại bao gồm một tập các **instance**(bản ghi). Mỗi instance được đặc tả bởi một tuple(x, y), trong đó x là tập giá trị của các thuộc tính mô tả instance đó, và y là **class label**(nhãn lớp) của instance. Tập thuộc tính x có thể chứa bất kỳ loại thuộc tính nào, trong khi nhãn lớp y phải là thuộc tính phân loại.

Một **classification model**(Mô hình phân loại) là một biểu diễn trừu tượng cho mỗi quan hệ giữa tập thuộc tính và nhãn lớp. Chúng ta sẽ thấy trong 2 chương tới, một mô hình có thể biểu diễn nhiều cách, như cây, bảng xác xuất, hoặc thường là một vector các giá trị thực. Cụ thể hơn, ta có thể biểu diễn nó dưới dạng toán học là một hàm mục tiêu f, đầu vào là tập thuộc tính x, cho ra kết quả tương ứng là dự đoán nhãn lớp. Mô hình được cho là phân loại đúng nếu f(x) = y.

**Bảng 3.1**. Các ví dụ về phân lớp.

|  |  |  |
| --- | --- | --- |
| Nhiệm vụ | Tập thuộc tính | Nhãn lớp |
| Lọc thư spam | Những đặc tính trích xuất từ tiêu đề và nội dung của các email. | Spam hoặc không spam |
| Nhận định khối u | Những đặc tính trích xuất từ quét ảnh cảm ứng từ (MRI) | Ác tính hay lành tính |
| Phân loại ngân hà | Những đặc tính lấy từ ảnh của kính viễn vọng | Hình elip, xoắn óc, hoặc bất thường |

**Bảng 3.2** Một dữ liệu mẫu cho việc phân loại các loài có xương sống



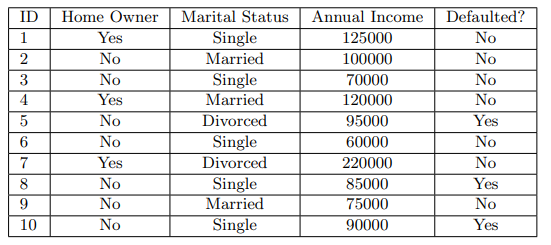
Bảng 3.1 gồm những ví dụ về tập thuộc tính và nhãn lớp cho các nhiệm vụ phân lớp khác nhau. Lọc thư spam và nhận định khối u là những ví dụ về bài toán phân loại nhị phân, mỗi mẫu của dữ liệu có thể phân vào 1 trong 2 lớp. Nếu số lượng lớp lớn hơn 2, như trong ví dụ phân loại ngân hà, thì nó được gọi là bài toán phân loại đa lớp.

Chúng ta minh họa những khái nhiệm cơ bản của phân lớp trong chương này với 2 ví dụ sau.

**Ví dụ 3.1 [Phân loại các loài có xương sống]** Bảng 3.2 gồm một tập có dữ liệu mẫu cho việc phân lớp các loài thành động vật có vú, bò sát, chim, cá và loài lưỡng cư. Tập thuộc tính gồm những đặc điểm của động vật có xương sống như nhiệt độ thân thể, phủ trên da và khả năng bay. Tập dữ liệu có thể được xử dụng cho bài toán phân lớp nhị phân như phân loại động vật có vú, bằng cách nhóm các loài bò sát, chim, cá và lưỡng cư thành một loại duy nhất tên là động vật không vú.

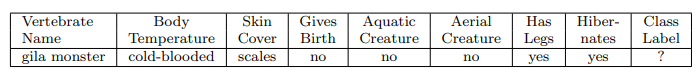
**Ví dụ 3.2 [Phân loại người vay]** Xem xét bài toán dự đoán người vay nợ có thể trả được khoản vay hay không. Tập dữ liệu sử dụng để xây dựng mô hình phân lớp được hiển thị trong Bảng 3.3. Tập thuộc tính bao gồm những thông tin cá nhân của người vay như tình trạng hôn nhân và thu nhập hằng năm, trong khi nhãn lớp chỉ ra người vay nợ có thể trả nợ hay không.

**Bảng 3.3**. Dữ liệu mẫu cho bài toán phân loại người vay nợ.



Một mô hình phân loại đóng 2 vai trò qua trọng trong khai phá dữ liệu. Đầu tiên, nó được sử dụng như một mô hình dự đoán để phân loại các instance trước đó chưa được đánh nhãn. Một mô hình phân loại tốt phải cho ra những dự đoán chính xác với tốc độ phản hồi nhanh chóng. Thứ hai, nó đóng vai trò như một mô hình mô tả để nhận ra những đặc tính dùng để phân biệt các instance từ các lớp khác nhau. Điều này rất hữu dụng cho những chương trình đánh giá, như chuẩn đoán y khoa, nó không đủ để có một mô hình mà đưa ra những dự đoán khi không chứng minh được làm thế nào để đưa ra quết định.

Ví dụ, một mô hình phân loại từ tập dữ liệu động vật có xương sống được đưa ra ở Bảng 3.2 có thể sử dụng để dự đoán nhãn lớp của động vật có xương sống sau:



Ngoài ra, nó có thể sử dụng như một mô hình miêu tả giúp xác định những đặc điểm để định nghĩa một loài có xương sống là loài động vật có vú, bò sát, chim, cá hay động vật lưỡng cư. Ví dụ, mô hình có thể chỉ ra động vật có vú là loài máu lạnh, sinh con.

Có vài điểm đáng lưu ý về ví dụ trước. Đầu tiên, mặc dù tất cả những thuộc tính trong Bảng 3.2 là định tính, nhưng không có hạn chế về kiểu của thuộc tính có thể sử dụng như các biến dự đoán. Ngược lại, nhãn lớp phải là kiểu định danh. Điều này phân biệt phân loại này với các kiểu mô hình dự đoán khác, ví như hồi quy, giá trị được dự đoán thường là định lượng. Thông tin thêm về hồi quy có thể thấy ở Appendix D.

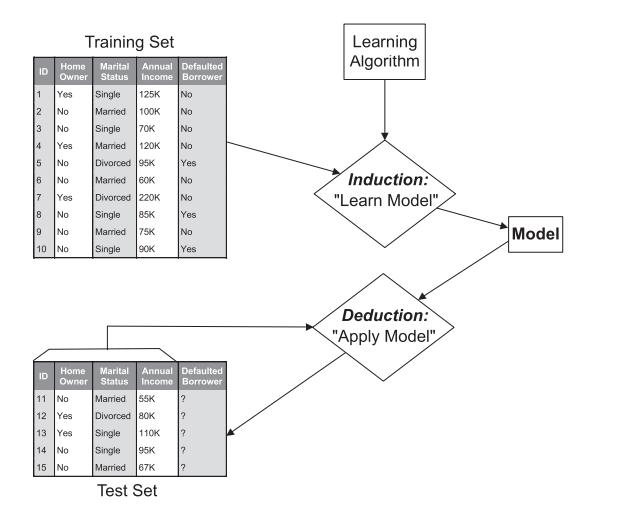
Một điểm đáng lưu ý khác là không phải tất cả các thuộc tính phải liên quan đến bài toán phân lớp. Ví dụ, chiều dài hoặc chiều cao trung bình của một loài có xương sống có thể không hữu dụng để phân loại động vật có vú, vì có thể mang cùng giá trị cho cả động vật có vú và không vú. Nó sẽ bị loại bỏ trong giai đoạn tiền xử lý. Các thuộc tính còn loại có thể không tự mình phân biệt được các lớp, nên phải phối hợp với các thuộc tính khác. Ví dụ, thuộc tính nhiệt độ thân thể không đủ để phân biệt loài có vú với các loài còn lại. Khi nó sử dụng chung với sinh con, việc phân loại loài có ý cải thiện một cách đáng kể. Tuy nhiên, khi thêm vào các thuộc tính khác như phủ da, mô hình sẽ chi tiết quá mức và không còn phủ hết tất cả những động vật có vú. Tìm ra một tổ hợp các thuộc tính tối ưu để phân biệt tốt nhất các instance từ các lớp khác nhau là một thách thức lớn trong việc xây dựng các mô hình phân loại.

**3.2 Khung chung cho phân lớp**

Phân lớp là một nhiệm vụ chỉ ra được nhãn của các instance chưa rõ nhãn và **classifier** được sử dụng để thi hành nhiệm vụ. Một classifier được mô tả trong những giới hạn của một mô hình đã được minh họa trong phần trước. Mô hình được tạo ra sử dụng một tập các instance cho trước, được gọi là **training set**, chứa các giá trị thuộc tính cũng như nhãn lớp của mỗi instance. Một cách tiếp cận có hệ thống để học một mô hình phân lớp được cho bởi training set gọi là **learning algorithm.** Quá trình sử dụng một learning algorithm để xây dụng một mô hình phân loại gọi là **induction**. Quá trình này áp dụng cho một mô hình phân lớp trên những mẫu thử để đoán nhãn lớp gọi là **deduction**. Do đó, quá trình phân loại bao gồm 2 bước: Áp dụng learning algorithm vào training data để học model, và sau đó áp dụng model để phân lớp các instance chưa gắn nhãn. Hình 3.3 minh họa khung chung cho phân lớp.

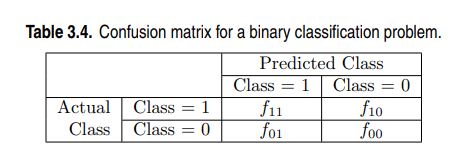
**Classification technique**(Phương pháp phân lớp) nói đến một phương pháp chung để phân lớp. Ví dụ, phương pháp cây quyết định chúng ta sẽ học trong chương này. Phương pháp này khá giống những cái khác, bao gồm một họ của những mô hình liên quan và một số thuật toán để học các mô hình đó. Trong chương 4, chúng ta sẽ học thêm những phương pháp phân lớp, bao gồm mạng neural và hỗ trợ vector machine.

Những cặp thuật ngữ cần lưu ý. Đầu tiên, “classifier” và “model” thường được cho là đồng nghĩa. Nếu một phương pháp phân lớp xây dựng độc lập, mô hình toàn cục, thì điều trên ổn. Tuy nhiên, mặc dù mọi mô hình đều định nghĩa một classifier, nhưng không phải mọi classifier được định nghĩa bởi một mô hình đơn. Một vài classifier, như k-nearest neighbor classifier, không được xây dựng mô hình rõ ràng, trong khi những mô hình khác, như ensemble classifier, kết hợp tập kết quả của một bộ các mô hình. Thứ hai, từ “classifier” thường được sử dụng trong nhiều nghĩa chung để nói về phương pháp phân lớp decision tree hoặc một classifier cụ thể xử dụng phương pháp đó. May mắn thay, nghĩa của “classifier” thường rõ ràng từ ngữ cảnh.

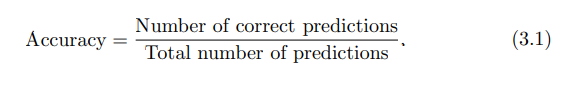


**Hình 3.3** Khung chung để xây dựng một mô hình phân lớp.

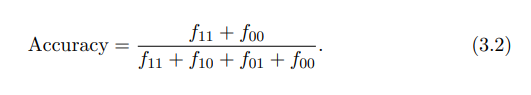
Trong khung chung ở Hình 3.3, Bước induction và deduction nên được tiến hành độc lập. Thực tế, training và test set nên độc lập với nhau để đảm bảo mô hình kết quả có thể dự đoán chính xác nhãn của các instance chưa bao giờ gặp. Những mô hình đưa ra các dự đoán rõ ràng gọi là có **generalization performance** tốt. Hiệu năng của một mô hình có thể được đánh giá bằng cách so sánh nhãn giữa true label và predicted label của instance. Thông tin này có thể được tóm gọn trong một bảng gọi là **confusion matrix**. Bảng 3.4 vẽ ma trận hỗn hợp cho bài toán phân lớp nhị phân. Mỗi mục fij biểu diễn số lượng instance từ lớp được dự đoán là j của lớp i. Ví dụ f01 là số lượng instance từ lớp 0 dự đoán sai vào lớp 1. Số lượng dự đoán đúng từ mô hình là (f11+f00) và số lượng đoán sai là (f10+f01).



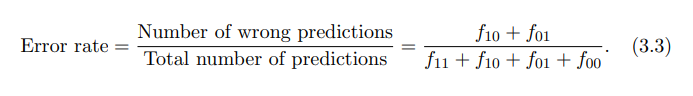
Mặc dù confusion matrix cung cấp thông tin cần thiết để xác định hiệu năng của mô hình tốt đến đâu, tổng hợp thông tin này thành một số duy nhất để thuận tiện so sánh với hiệu năng của các mô hình khác. Ta sử dụng một đại lượng để đánh giá (**evaluation metric)** như là độ chuẩn xác(**accuracy),** được tính bằng cách:



Về bài toán phân lớp nhị phân, độ chính xác của mô hình được cho bởi



Tỉ lệ sai(**Error rate**) là một đại lượng liên quan khác, được định nghĩa như sau



Các learning algorithm của các phương pháp phân loại hầu hết được thiết kế để học những mô hình thu được độ chính xác cao nhất, tương đương error rate thấp nhất khi áp dụng cho bộ test.