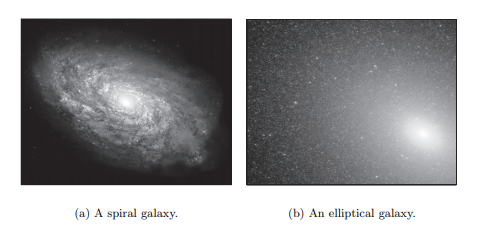
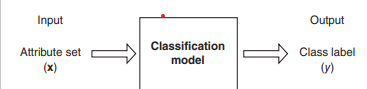
**Phân Loại : Các khái niệm và kỹ thuật cơ bản**

Con người vốn có khả năng bẩm sinh để phân loại mọi thứ thành các loại. Ví dụ như nhiệm vụ lọc thư rác từ email hay các nhiệm vụ chuyên ngành hơn như nhận diện các đối tượng thiên văn trong ảnh của kính viễn vọng (xem Hình 3.1). Việc phân loại này có thể làm thủ công nhưng chỉ có thể áp dụng cho các loại dữ liệu nhỏ và đơn giản vốn chỉ chứa một vài thuộc tính. Còn đối với các tập dữ liệu lớn và phức tạp cần một giải pháp tự động.



**Hình 3.1. Phân loại các thiên hà từ ảnh chụp bằng kính viễn vọng lấy từ trang web của NASA.**

**Chương 3: Phân Lớp**

****

**Hình 3.2. Sơ đồ phân loại nhiệm vụ của phân lớp**

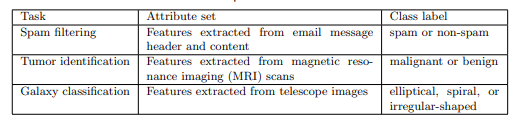
Chương này sẽ giới thiệu các khái niệm cơ bản về phân lớp và mô tả một số vấn đề chính của nó như model overfitting, model selection, và model evaluation. Mặc dù các chủ đề dưới đây được trình bày bằng cách sử dụng một kỹ thuật phân loại gọi là cây quyết định, nhưng hầu hết các phương thức trong chương này cũng đều có thể áp dụng cho các kỹ thuật phân loại khác, đa số được đề cập đến trong Chương 4.

**3.1 Khái niệm cơ bản.**

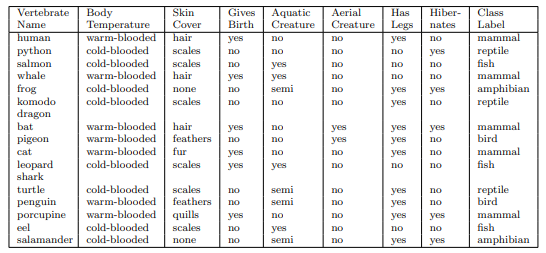
Hình 3.2 minh họa ý tưởng chung đằng sau việc phân lớp. Dữ liệu cho vệc phân loại bao gồm một tập hợp các instance. Mỗi đối tượng được biểu diễn bởi một tuple(x,y), trong đó x là một bộ giá trị của các thuộc tính mô tả instance đó, còn y thể hiện nhãn của đối tượng. Tập thuộc tính x có thể chứa nhiều kiểu giá trị, trong khi nhãn lớp là thuộc phân loại.

Mô hình phân lớp biểu diễn một cách trừu tượng về mối quan hệ giữa tập thuộc tính và nhãn lớp. Trong hai phần tiếp theo có thể thấy mô hình được biểu diễn theo nhiều cách, ví dụ dưới dạng cây, bảng xác suất hay đơn giản là một vectơ của tham số có giá trị thật. Chúng ta có thể biểu diễn nó bằng toán học như việc tạo ra một hàm mục tiêu f nhận đầu vào là các tập thuộc tính x, sau khi xử lý hàm f sẽ cho đầu ra tương ứng với nhãn lớp được dự đoán. Mô hình được cho là phân lớp chính xác một đối tượng (x,y) nếu f(x)=y.

**Bảng 3.1.1 Ví dụ cho việc Phân lớp**



**Bảng 3.2 Một mẫu dữ liệu sử dụng cho việc phân loại động vật có xương sống**

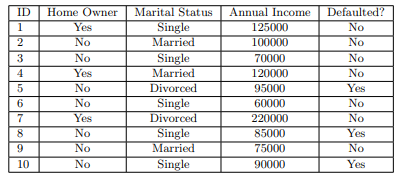
****

Bảng 3.1 cho thấy ví dụ về các tập thuộc tính và nhãn lớp của cho nhiệm vụ phân loại các loài. Lọc thư rác và xác định khối u là ví dụ về phân loại nhị phân, trong đó mỗi trường dữ liệu được phân vào một trong hai lớp. Còn trường hợp số lượng các lớp lớn hơn hai được gọi là phân loại đa lớp. Chúng ta sẽ minh họa các khái niệm cơ bản của phân lớp trong chương này thông qua hai ví dụ sau.

Ví dụ 3.1 [phân loại động vật có xương sống ] bảng 3.2 hiện thị tập dữ liệu mẫu được 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 này cũng có thể sử dụng cho phân loại nhi phân, chẳng hạn như phân loại động vật có vú và động vật không vú, thực hiện bằng cách gom các loài bò sát, chim, cá và lưỡng cư vào nhóm động vật không vú.

Ví dụ 3.2 [phân loại người vay tiền] để xem xét vấn đề dự đoán người vay tiền có khả năng chi trả hay không. Tập dữ liệu được sử dụng để xây dựng mô hình phân lớp được hiển thị trong bảng 3.3. Bộ thuộc tính sẽ bao gồm thông tin cá nhân người vay như tình trạng hôn nhân, thu nhập hàng năm, trong khi đó nhãn lớp cho biết liệu người vay có khả năng chi trả hay không.

**Bảng 3.3 một mẫu dữ liệu cho vấn đề phân loại người vay tiền**



Một mô hình phân lớp phục vụ cho hai chức năng quan trọng trong khai phá dữ liệu. Đầu tiên nó được sử dụng như một mô hình dự đoán để phân loại các trường hợp trước đó chưa được dán nhãn. Một mô hình phân loại tốt phải cung cấp dự đoán chính xác với thời gian phản hồi nhanh. Thứ hai nó được dùng như một mô hình mô tả để xác định các đặc điểm nhằm phân biệt các lớp khác nhau. Điều này rất hữu ích và được ứng dụng vào các lĩnh vực quan trọng, ví dụ như chẩn đoán y khoa, tuy nhiên không thể đưa ra một mô hình dự đoán mà không cần chứng minh làm thế nào để nó dự đoán.

Ví dụ, một mô hình dự đoán được tao ra từ tập dữ liệu động vật có xương sống trong bảng 3.2. Mô hình này có thể sử dụng để dự đoán nhãn lớp của động vật có xương sống sau.



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

Có một số điểm đáng chú ý liên quan tới 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 về loại thuộc tính có thể sử dung làm biến dự báo. Mặt khác nhãn lớp phải được định nghĩa. Điều này phân biệt Phân lớp khác với mô hình dự đoán khác như hồi quy. Tìm hiểu them trong phụ lục D

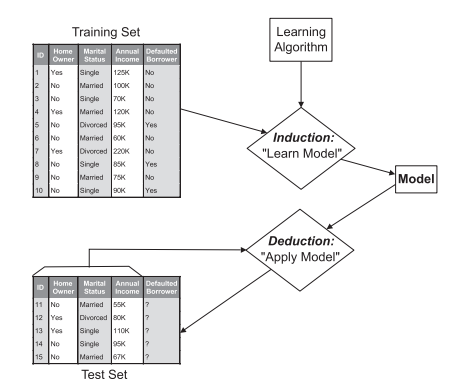
Một điểm chú ý khác là không phải tất cả các thuộc tính có thể liên quan đến nhiệm vụ phân loại, ví dụ như chiều dài trung bình hoặc trọng lượng của một động vật có xương sống có thể không hữu ích để phân loại động vật có vú và động vật không vú. Một thuộc tính như vậy thường được loại bỏ trong quá trình tiền xử lí. Các thuộc tính còn lại có thể không tự phân biệt được các lớp, do đó cần được kết hợp với các thuộc tính khác. Chẳng hạn như thuộc tính nhiệt độ cơ thể không đủ để phân biệt động vật có vú với các động vật có xương sống khác, nhưng khi được sử dụng với thuộc tính sinh con, việc phân loại động vật có vú được cải thiện đáng kể. Tuy nhiên khi bao gồm các thuộc tính bổ sung như độ che phủ của da, mô hình trở nên quá cụ thể và không còn chỉ bao gồm các loại động vật có vú. Tìm ra sự kết hợp tối ưu giữa các thuộc tính phân biệt của các lớp khác nhau là vấn đề chính trong việc xây dựng mô hình phân lớp

**3.2 framework** **cho phân lớp**

Phân loại là nhiệm vụ gán nhãn cho các trường hợp dữ liệu chưa có nhãn và trình phân loại được sử dụng để thực hiện một tác vụ như vậy. Mộ trình phân loại thường được mô tả theo các mô hình đã được minh họa trong phần trước. Mô hình được tạo ra bằng cách sử dụng một tập hợp các tập thuộc tính được gọi là tập huấn luyện, tập huấn luyện chứa các giá trị thuộc tính và nhãn lớp của mỗi đối tượng. Phương pháp để một hệ thống có thể học một mô hình phân loại từ tập huấn luyện được gọi là thuật toán học tập. Qúa trình sử dụng thuật toán học tập để xây dựng một mô hình phân loại từ dữ liệu huấn luyện được gọi là quy nạp. Qúa trình này cũng thường được mô tả như “học tập mô hình” hoặc “xây dựng mô hình”. Qúa trình áp dụng một mô hình phân loại trên các trường hợp thử nghiệm không có nhãn để dự đoán nhãn lớp của họ được gọi là khấu trừ. Do đó, quá trình phân loại bao gồm hai bước: áp dụng thuật toán học tập để học mô hình từ dữ liệu huấn luyện và sau đó sữ dung mô hình để gán nhãn cho những trường hợp không có nhãn. Hình 3.3 minh họa framework để phân loại

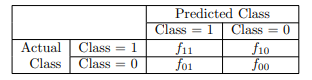
Mỗi kỹ thuật phân loại đều có một cách tiếp cận chung. Ví dụ kỹ thuật cây quyết định mà chúng ta sẽ nghiên cứu trong chương này. Đây là một kỹ thật phân loại như hầu hết các kỹ thuật khác, bao gồm tổng hợp của các mô hình liên quan và số lượng thuật toán được dùng để học một mô hình. Trong chương 4, chúng ta sẽ nghiên cứu các kỹ thuật phân loại khác bao gồm mạng thần kinh và các máy vectơ hỗ trợ

Một vài lưu ý về thuật ngữ, đầu tiên, thuật ngữ “trình phân loại” và “mô hình” thường được coi là đồng nghĩa. Nếu một kỹ thuật phân loại xây dựng được một mô hình tổng quan duy nhất thì điều đó là tốt. Tuy nhiên mọi mô hình định nghĩa một trình phân loại chứ không phải mọi trình phân loại được định nghĩa bằng một mô hình. Một số việc phân loại chẳng hạn như phân loại lân cận K-nearest, không xây dựng được một mô hình rõ ràng, trong khi các phân loại khác, chẳng hạn như phân loại đồng bộ sẽ kết hợp đầu ra của tập hợp một số các mô hình (mục 4.10). Thứ hai thuật ngữ “trình phân loại” thường được sử dụng theo nghĩa chung hơn để chỉ một kỹ thuật phân loại. Trong đó ví dụ , “trình phân loại cây quyết định” có thể dùng ám chỉ cho kỹ thuật phân loại cây quyết định hoặc một bộ phận cụ thể được xây dựng cụ thể bằng kỹ thuật ấy, may thay ý nghĩa của “trình phân loại” thường rất rõ ràng



**Hình 3.3. Khung chung cho việc xây dựng mô hình phân loại**

Trong Khung chung được biểu diễn ở hình 3.3, việc khấu trừ và quy nạp nên được thực hiện riêng. Trong thực tế, như sẽ được thảo luận sau trong Mục 3.6, các bộ huấn luyện và bộ kiểm tra phải độc lập với nhau để đảm bảo rằng mô hình quy nạp có thể dự đoán chính xác các nhãn lớp của các trường hợp mà nó chưa bao giờ gặp phải trước đây. Các mô hình cung cấp dự đoán như vậy được cho là có hiệu suất khái quát tốt. Hiệu suất của một mô hình (trình phân loại) có thể được đánh giá bằng cách so sánh kết quả dự đoán nhãn với nhãn thực tế đã biết từ đầu của các trường hợp. Thông tin này có thể được tóm tắt trong một bảng gọi là ma trận hỗ hợp. Bảng 3.4 mô tả ma trận hỗn hợp cho việc phân loại nhị phân. Mỗi đầu vào fij biểu thị số lượng trường hợp lớp i được dự đoán từ lớp j. Ví dụ, f01 là số lượng trường lớp 0 được dự đoán sai thành lớp 1. Số lượng trường hợp dự đoán đúng được thực hiện bởi mô hình là (f11 + f00) và số lượng dự đoán không chính xác là (f10 + f01).



**Bảng 3.4. Ma trẫn hỗn hợp của phân lọai nhị phân**

Mặc dù ma trận hỗn cung cấp thông tin cần thiết để xác định mô hình phân loại có tốt hay không, đồng thời tổng các trường hợp thành một con số duy nhất giúp thuận tiện hơn trong việc 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 =

Đối với phân loại nhị phân. độ chính xác của một mô hình được đưa ra bởi

Độ chính xác =

Tỷ lệ lỗi là một số liệu liên quan khác, được tính theo cách sau và áp dụng cho phân loại nhi phân.

Tỷ lệ lỗi = =

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 và có tỷ lệ lỗi thấp nhất khi áp dụng cho bộ thử nghiệm. Chúng ta sẽ xem xét lại chủ đề đánh giá mô hình trong Mục 3.6.