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

Con người bẩm sinh đã có khả năng phân chia mọi thứ thành nhiều loại, ví dụ như những việc bình thường như phân loại thư email rác cho đến các công việc chuyên môn hơn như nhận diện các thiên thể trong các hình chụp từ kính viễn vọng. Trong khi phân loại thủ công thường chỉ làm được tốt với các tập dữ liệu nhỏ và đơn giản chỉ chứa vài thuộc thính, các tập dữ liệu lớn và phức tạp hơn lại đòi hỏi một giải pháp tự động.

Chương này giới thiệu các khái niệm cơ bản của phân lớp và mô tả một số vấn đề chính như model overfitting, model selection và model evaluation. Dù các chủ đề này được minh hoạ bằng cách sử dụng một kỹ thuật phân lớp được gọi là quy nạp cây quyết định (decision tree induction), hầu hết các lý luận trong chương này cũng áp dụng cho các kỹ thuật phân loại khác, nhiều trong số đó sẽ được đề cập ở Chương 4.

**3.1 Khái niệm cơ bản**

Hình 3.2 minh hoạ ý tưởng đại khái bên dưới phân lớp. Dữ liệu cho việc phân lớp bao gồm một bộ các instances(bản ghi). Mỗi instance được đặc tả bởi một tập các biến (x, y), trong đó x là tập giá trị của các thuộc tính còn y là nhãn lớp của instance. Tập thuộc tính x có thể chứa các thuộc tính thuộc bất kỳ loại nào, trong khi nhãn lớp y phải là thể loại.

Một mô hình phân lớp là một đại diện trừu tượng của mối quan hệ giữa tập các thuộc tính và nhãn lớp. Như sẽ thấy ở 2 chương tiếp theo, mô hình có thể được diễn tả theo nhiều cách, ví dụ như một cây, một bảng xác suất, hoặc chỉ đơn giản là một véc tơ các tham số giá trị thực. Một cách chính quy hơn, ta có thể biểu diễn nó theo toán học, trong đó hàm mục tiêu f nhận các giá trị đầu vào là tập các thuộc tính x và sinh ra giá trị đầu ra, tương ứng ở đây là nhãn lớp dự đoán. Mô hình được phát biểu là phân loại một instance (x,y) đúng nếu f(x) = y.

Bảng 3.1 chỉ ra nhiều ví dụ về các tập thuộc tính và các nhãn lớp cho nhiều công việc phân lớp khác nhau. Lọc rác và xác định khối u là những ví dụ về bài toán phân lớp nhị phân, trong đó mỗi dữ liệu instance có thể được chia thành một trong 2 lớp. Nếu số lượng lớp lớn hơn 2, như ví dụ về phân lớp thiên hà, thì ta gọi nó là bài toán phân lớp đa lớp.

Ta minh hoạ các khái niệm cơ bản của chương này với 2 ví dụ sau đây.

**Ví dụ 3.1 [Phân loại động vật có xương sống]** Bảng 3.1 chỉ ra các tập dữ liệu mẫu cho việc phân lớp các động vật có xương sống thành động vật có vú, bò sát, chim, cá và động vật lưỡng cư. Tập thược tính bào gồm các nét đặc trưng của động vật có xương sống như thân nhiệt, bao phủ da và khả năng bay. Tập dữ liệu cũng có thể đước sử dụng cho việc phân loại nhị phân như phân lớp động vật có vú, bằng cách nhóm bò sát, chim, các và lưỡng cư thành 1 loại riêng gọi là nhóm không có vú.

**Ví dụ 3.2 [Phân loại người vay]** Xem xét bài toán dự đoán xem liệu người vay sẽ trả lại được nợ hay sẽ vỡ nợ. Tập dữ liệu sử dụng để xây dựng mô hình phân lớp được biểu thị ở bảng 3.3. Tập thuộc tính bao gồm thông tin cá nhân của người vay như trạng thái hơn nhân và thu nhập hàng năm, trong khi nhãn lớp chỉ ra rằng người nợ đã từng phá sản vì các khoản vay trước đó hay chưa.

Mô hình phân lớp phục vụ 2 vai trò vai trò quan trọng trong khai phá dữ liệu. Đầu tiên, nó được sử dụng như mô hình dự đoán nhàm phân lớp các instance chưa được gán nhãn trước đó. Một mô hình phân lớp tốt phải cũng cấp khả năng dự đoán chính xác đi cùng với thời gian phản hồi nhanh. Thứ hai, nó phục vụ như là mô hình mô tả xác định các nét đặc trưng mà phân biệt instance khỏi các lớp khác. Điều này đặc biệt hữu dụng cho các ứng dụng quan trọng, như chẩn đoán y khoa, ở đó một mô hình có thể đưa ra dự đoán mà không chứng minh làm thế nào nó đạt được quyết định đó là không đủ.

Ví dụ, một mô hình phân lớp được lấy ra từ tập dữ liệu động vật có xương sống ở bảng 3.2 có thể được dùng để dự đoán nhãn lớp của các động vật có xương sống bên dưới đây:

Ngoài ra, nó có thể được sử dụng như là mô hình mô tả giúp xác định các nét đặc trưng định nghĩa một động vật có xương sống là động vật có vú, bò sát, chim, chuột hay lưỡng cư. Ví dụ, mô hình có thể nhận dạng động vật có vú là động vật xương sống máu nóng sinh con.

Có nhiều điều đáng chú ý liên quan đến các ví dụ trước. Đầu tiên, mặc dù tất cả các thuộc tính được biểu diễn ở bảng 3.2 là định tính, không có các hạn chế trên kiểu của thuộc tính mà có thể được dùng như biến dự đoán. Nhãn lớp, nói cách khác, phải là loại danh nghĩa. Điều này phân biệt phân lớp với các việc mô hình dự đoán khác như hồi quy, ở đó giá trị dự đoán thường là định lượng.

Một điều đáng chú ý khác là không phải tất cả các thuộc tính đều có thể thích hợp với việc phân lớp. Ví dụ, độ dài hoặc cân nặng trung bình của một động vật có xương sống có thể không hữu dựng cho việc phân lớp động vật có vú, do những thuộc tính này biểu thị giá trị giống hệt nhau cho cả động vật có vú và không có vú. Một thuộc tính như vậy thường bị loại đi trong suốt quá trình xử lý. Các thuộc tính còn lại có thể không thể tự phân biệt các lớp, do đó, phải được sử dụng hoà với các thuộc tính khác. Ví dụ, thân nhiệt là không đủ để phân biệt động vật có vú với các động vật có xương sống khác. Khi được sử dụng cùng cách thức sinh con, việc phân lớp động vật có vú được cải thiện đáng kể. Tuy nhiên, khi các thuộc tính bổ sung, như bao phủ da được thêm vào, mô hình trở nên quá cụ thể và không còn bao phủ toàn bộ động vật có vú. Tìm tổ hợp thuộc tính tối ưu phân việt tốt nhất các instance khỏi các lớp khác là thử thách mang tính bước ngoặt để xây dựng mô hình phân lớp.

**3.2 Framework Chung cho phân lớp**

Phân lớp là việc chỉ định nhãn cho một instance dữ liệu vô nhãn và **bộ phân lớp** được sử dụng để thực hiện công việc đó. Một bộ phân lớp thường được mô tả trong các điều khoản của một mô hình (được minh hoạ ở phần trước). Mô hình có được tạo bằng một tập các instance cho trước, được gọi là **tập dữ liệu huấn luyện**, bao gồm giá trị của các thuộc tính cũng như nhân của mỗi instance. Cách tiếp cận có hệ thống cho việc học một mô hình phân lớp nhờ tập dữ liệu huấn luyện được gọi là **thuật toán học**. Qúa trình sử dụng thuật toán học để xây dụng mô hình phân lớp từ dữ liệu huấn luyện được gọi là **quy nạp** (induction). Quá trình này cũng thường được miêu tả là “học mô hình” hoặc “xây dựng mô hình”. Quá trình áp dụng mô hình phân lớp vào các instance kiểm thử chưa biết đến để dự đoán nhãn lớp của chúng được gọi là **suy đoán**(deduction). Do đó, quá trình phân lớp liên quan đển 2 bước: áp dụng thuật toán học với dữ liệu huấn luyện để học mô hình, và áp dụng mô hình để chỉ định nhãn cho các instance vô nhãn.

**Kỹ thuật phân lớp** ám chỉ đến cách tiếp khái quát của phân lớp, ví dụ kỹ thuật cây quyết định. Kỹ thuật này giống với phần lớn các kỹ thuật khác, bao gồm một gia đình của các mô hình liên quan với nhau và một số thuật toán để học các mô hình đó. Trong chương 4, chúng ta sẽ học thêm các kỹ thuật khác, bao gồm mạng nơ ron và máy hỗ trợ véc tơ.

A vài lưu ý về thuật ngữ. Thứ nhất, thuật ngữ bộ phân lớp và mô hình thường được đánh đồng. Nếu kỹ thuật phân lớp xây dựng được một mô hình đơn giản và tổng quát, thì thế là ổn. Tuy nhiên, dù mọi mô hình đều xác định một bộ phân lớp, không phải mọi bộ phân lớp được xác định bởi chỉ một mô hình đơn lẻ. Một vài bộ phân lớp, như bộ phân lớp K-nearest neighbor, không xây dựng một mô hình tường minh mà kết hợp đầu ra của một bộ các mô hình, mặc dù các bộ phân lớp khác như ensemble không như vậy. Thứ hai, thuật ngữ “bộ phân lớp” thường được sử dụng mang nhiều ý nghĩa hơn để ám chỉ đến kỹ thuật phân loại. Do đó, lấy ví dụ, “bộ phân lớp cây quyết định” có thể ám chỉ đến kỹ thuật phân lớp cây quyết định hoặc cũng có thể ám chỉ bộ phân lớp riêng biệt được xây dựng bằng cách sử dụng kỹ thuật đó. May thay, nghĩa của “bộ phân lớp” thường rõ ràng từ ngữ cảnh.

Trong framwork chung biểu thị ở hình minh hoạ 3.3, các bước quy nạp và suy luận nên được thực hiện riêng biệt. Thực tế, điều này sẽ được bàn luận sau đó ở phần 3.6, tập huấn luyện và kiểm thử nên là độc lập với nhau để đảm bảo rằng mô hình xuất ra có thể dự đoán chính xác nhãn của các instance mà nó chưa bao giờ gặp trước kia. Các mô hình mà đưa ra được dự đoán minh mẫn như vậy được phát biểu là có hiệu suất khái quát tốt. Hiệu suất của một mô hình (bộ phân lớp) có thể được đánh giá bằng cách so sánh nhãn dự đoán với nhãn thực của instance. Thông tin này có thể được tóm tắt lại thành một bảng gọi là **ma trận sai (matrix confusion).** Bảng 3.4 miêu tả ma trận sai cho bài toán phân lớp nhị phân. Mỗi bản ghi fij biểu lộ số instance từ lớp i được dự đoán thành lớp j. Ví dụ, f01 là số instance từ lớp 0 dự đoán sai thành lớp 1. Số lượng dự đoán đúng nhờ mô hình là (f11 + f00) và số lượng dự đoán sai là (f10 + f01).

Mặc dù ma trận sai cung cấp thông tin cần để xác định một mô hình phân lớp thực hiện tốt như thế nào, nhưng việc tóm tắt những thông tin này thành một số đơn lẻ làm nó thuận tiện hơn để so sánh hiệu suất tương đối của các mô hình khác. Điều này có thực được thực hiện bằng cách sử dụng một **hệ** **đánh giá (evaluation metric)** như **accuracy**, được tính toán như sau:

Đối với bài toán phân lớp nhị nhân, độ chính xác của mô hình được cho bằng:

**Error rate** là hệ tương đối khác, được xác định theo bài toán phân lớp nhị phân như sau:

Thuật toán học của phần lớn kỹ thuật phân lớp được thiết kế để học mô hình sao cho đạt được độ chính xác cao nhất, hoặc tương đương là tỉ lệ sai thấp nhất khi áp dụng vào tập kiểm thử. Ta sẽ gặp lại chủ đề đánh giá mô hình vào phần 3.6