Tên: Trần Sỹ Lộc : 17CNTT2

Chương 3 Classification

Chương này giới thiệu các khái niệm cơ bản về phân loại và mô tả các vấn đề chính của nó như quá mức mô hình, lựa chọn mô hình và đánh giá mô hình. Mặc dù các chủ đề này được minh họa bằng cách sử dụng một kỹ thuật phân loại được gọi là cảm ứng cây quyết định, hầu hết các cuộc thảo luận trong chương này cũng có thể áp dụng cho các kỹ thuật phân loại khác, nhiều trong số đó được đề cập trong Chương 4.

3.1 Basic Concepts

Hình 3.2 minh họa ý tưởng chung đằng sau phân loại. Dữ liệu cho nhiệm vụ phân loại bao gồm một tập hợp các thể hiện (bản ghi). Mỗi trường hợp này được đặc trưng bởi bộ tuple (x, y), trong đó tập hợp các thuộc tính mô tả thể hiện và là nhãn lớp của thể hiện. Dramtribution setxcan chứa các thuộc tính của bất kỳ loại nào, trong khi lớp labelymustbe phân loại.

Mô hình phân loại là một biểu diễn trừu tượng của mối quan hệ giữa tập thuộc tính và nhãn lớp. Như sẽ thấy trong các chương tiếp theo, mô hình có thể được biểu diễn theo nhiều cách, ví dụ, như một cái cây, bảng aprobability hoặc đơn giản là 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, nó sẽ lấy đầu vào là tập hợp theattribution setxand tạo ra một đầu ra tương ứng với nhãn lớp được dự đoán. Mô hình được cho là phân loại một thể hiện (x, y) chính xác iff (x) = y.

Bảng 3.1

Bảng 3.1 cho thấy các ví dụ về các tập thuộc tính và nhãn lớp cho các nhiệm vụ phân loại khác nhau. Lọc thư rác và xác định khối u là các ví dụ về các vấn đề phân loại nhị phân, trong đó mỗi trường hợp dữ liệu có thể được phân loại theo một trong hai lớp. Nếu số lượng các lớp lớn hơn 2, như trong ví dụ về galaxyclassization, thì nó được gọi là một vấn đề 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 am-phibian. 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 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 vào một loại duy nhất gọi là không phải động vật có vú.

Ví dụ 3.2. [Phân loại người vay tiền] Xem xét vấn đề dự đoán liệu người vay có trả được khoản vay hay không trả được nợ. Tập dữ liệu được sử dụng để xây dựng mô hình phân loại được hiển thị trong Bảng3.3. Bộ thuộc tính bao gồm thông tin cá nhân của người đi vay 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 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 tiên 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 với thời gian phản hồi nhanh. Thứ hai, nó phục vụ như mô hình mô tả để xác định các trường hợp đặc trưng phân biệt với 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 đủ để đưa ra một mô hình đưa ra dự đoán mà không chứng minh làm thế nào nó đạt được 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 các loài động vật có xương sống sau:

Bảng

Ngoài ra, nó có thể được sử dụng như một mô hình mô tả để giúp xác định các đặc tính địa hình xác định động vật có xương sống là động vật có vú, bò sát, chim, cá hoặc anamphibian. Ví dụ, mô hình có thể xác định động vật có vú là động vật có xương sống ấm áp sinh ra con non.

Mặt khác, có một số điểm đáng chú ý về ví dụ trước. Mặc dù tất cả các thuộc tính được hiển thị trong Bảng 3.2 là định tính, nhưng có các hạn chế areno đối với loại thuộc tính có thể được sử dụng làm biến dự báo. Mặt khác, nhãn lớp phải thuộc loại danh nghĩa. Điều này phân biệt phân loại với các nhiệm vụ 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 nhiệm vụ phân loại. 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 các thuộc tính này cho thấy giá trị tương tự cho cả động vật có vú và không 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 và do đó, phải được sử dụng inconcert với các thuộc tính khác. Chẳng hạn, Temperatureattributionis không đủ để phân biệt động vật có vú với các động vật có xương sống khác. Khi nó được sử dụng cùng với Sinh ra, việc phân loại động vật có vú sẽ cải thiện đáng kể. Tuy nhiên, khi bao gồm các thuộc tính bổ sung, chẳng hạn như Cover Cover, mô hình trở nên quá cụ thể và không còn bao gồm tất cả các động vật có vú. Tìm kiếm sự kết hợp tối ưu của các thuộc tính phân biệt tốt nhất các thể hiện từ các lớp khác nhau là thách thức chính trong việc xây dựng các mô hình phân loại.

3.2 General Framework for Classification

Phân loại là nhiệm vụ gán nhãn cho các trường hợp dữ liệu không được gắn nhãn và aclassifieris được sử dụng để thực hiện một tác vụ như vậy. Một bộ phân loại thường được mô tả xen kẽ của một 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 xác định, chứa các giá trị at-Tribute cũng như nhãn lớp cho mỗi thể hiện. Phương pháp tiếp cận có hệ thống để học một mô hình phân loại được cung cấp một tập huấn luyện được gọi là thuật toán học. Quá trình sử dụng một thuật toán học tập để xây dựng một mô hình phân loại từ dữ liệu đào tạo được gọi là thuyết minh. Quá trình này cũng được mô tả như là học hỏi một mô hình xây dựng mô hình hoặc hoặc xây dựng mô hình. Quá trình này áp dụng một mô hình phân loại trên các trường hợp thử nghiệm chưa thấy để dự đoán các lớp học của chúng được biết đến như là sự kết hợp. Do đó, quá trình phân loại bao gồm hai phân đoạn: áp dụng thuật toán học tập để đào tạo dữ liệu để tìm hiểu một mô hình và từ đó áp dụng mô hình để gán nhãn cho các trường hợp không được gắn nhãn. Hình 3.3 duy trì khung chung để phân loại.

Kỹ thuật phân loại theo cách tiếp cận chung để phân loại, ví dụ, kỹ thuật cây quyết định 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, wewill nghiên cứu các kỹ thuật phân loại bổ sung, bao gồm các mạng thần kinh và máy vectơ hỗ trợ.

Figure 3.3.General framework for building a classification model.

Một vài lưu ý về thuật ngữ. Đầu tiên, các thuật ngữ phân loại và các mô hình khác, các mô hình phân loại trực tuyến, thường được coi là đồng nghĩa. Nếu một kỹ thuật phân loại 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 loại, không phải mọi phân loại đều được xác định bởi một mô hình duy nhất. Một số bộ phân loại, như bộ phân loại lân cận hỏi gần nhất, không xây dựng mô hình rõ ràng (Mục 4.3), trong khi các bộ phân loại khác, chẳng hạn như bộ phân loại tập hợp, kết hợp đầu ra của mô hình bộ sưu tập (Mục 4.10). Thứ hai, thuật ngữ phân loại trực tuyến, thường được sử dụng theo nghĩa đa nghĩa để chỉ một kỹ thuật phân loại. Do đó, ví dụ, bộ phân loại decisiontree, có thể tham khảo kỹ thuật phân loại cây quyết định hoặc một bộ phân loại cụ thể được xây dựng bằng kỹ thuật đó. May mắn thay, ý nghĩa của bộ phân loại 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 nên 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 của lớp mà nó chưa từng gặp trước đây. Các mô hình cung cấp khả năng dự báo như vậy được cho là có hiệu suất tăng trưởng tốt. Có thể đánh giá mô hình của mô hình (phân loại) bằng cách so sánh các dự đoán với các nhãn thực tế. Thông tin này có thể được tóm tắt trong một bảng gọi là ma trận aconfusion. Bảng 3.4 mô tả ma trận nhầm lẫn cho một vấn đề phân loại nhị phân. Mỗi mục nhập số lượng các yếu tố từ phân loại được phân loại là của classj. Ví dụ: f01 là số lượng phiên bản từ lớp 0 được dự đoán không chính xác là lớp 1. Số dự đoán chính xác của mô hình là (f11 + f00) và số dự đoán không chính xác là (f10 + f01).

Predicted: dự đoán

Actual : thực tế ;

Mặc dù ma trận nhầm lẫn cung cấp thông tin cần thiết để xác định mô hình phân loại thực hiện tốt như thế nào, việc tóm tắt thông tin này thành số đơn lẻ 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 tính không chính xác của số liệu đánh giá, được tính theo cách sau:

Accuracy = (Number of correct predictions) / (Total number of predictions.)

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

Accuracy =(f11+f00) / (f11+f10+f01+f00);

Lỗi tỷ lệ là một số liệu liên quan khác, được xác định như sau cho các vấn đề phân loại nhị phân:

(Number of correct predictions) / (Total number of predictions.) / (f11+f00) / (f11+f10+f01+f00); (3.3)

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 được áp dụng cho bộ kiểm tra. Chúng tôi sẽ xem xét lại chủ đề đánh giá mô hình trong Phần 3.6.