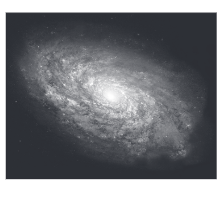
Name:Phạm Phương Nam,class:17CNTT3.

Bài 1:Dịch sách tiếng anh:

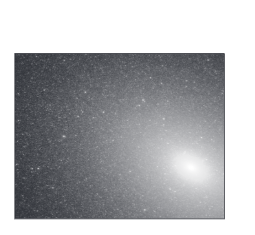
Chương 3:Khái niệm và các kĩ thuật phân loại.

Con người có một khả năng bẩm sinh để phân loại nhiều thứ thành các loại .ví dụ:Các tác vụ thông thường giống như lọc các thông báo thư rác hoặc tác vụ chuyên sâu hơn như xác nhận các đối tượng thiên thể trong bức ảnh của kính thiên văn .(xem bức ảnh 3.1).Trong khi việc phân loại thủ công thường đủ cho các tập dữ liệu nhỏ và đơn giản với chỉ có 1 vài thuộc tính ,các tập dữ liệu lớn hơn và phức tạp nhiều hơn yêu cầu một giải pháp tự động.

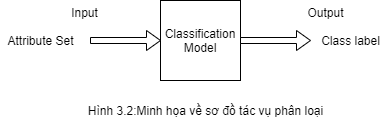
Hình 3.1 .Phân loại về các ngân hà từ các bức ảnh kính thiên văn lấy từ các website NASA.



a.Ngân hà xoắn ốc .



b.Ngân hà eclip.



Chương này ,giới thiệu các khái niệm cơ bản về phân loại và mô tả một số vấn đề chính của nó .model overfitting(mô hình quá mức),model selection(lựa chọn mô hình),model evalution(đánh giá mô hình).Trong khi những chủ đề này là minh họa sử dụng một kĩ thuật phân loại biết đến như hướng dân cây quyết định,hầu hết các cuộc thảo luận trong chương này là củng đả áp dụng đến các kĩ thuật phân loại khác ,nhiều trong số đó là được đề cập trong chương 4.

3.1 Khái niệm cơ bản:

Hình 3.2 mình họa về ý tưởng chung phía sau phân loại.Các dữ liệu cho một tác vụ phân loại bao gồm một sưu tập về các trường hợp(bản ghi).Mỗi trường hợp như vậy là đặc trưng bởi các tuple(x,y),với x là tập các giá trị thuộc tính mà ở đó mô tả các trường hợp ,và y nhãn class cho các trường hợp.Các tập thuộc tính x có thể chứa bất kì loại thuộc tính ,trong khi các nhãn class y phải được phân loại.

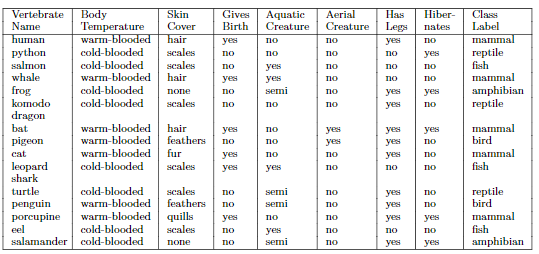
Một mô hình phân loại là 1 đại diện trừu tượng về các mối quan hệ giữa các tập thuộc tính và các nhãn class .Như sẽ nhìn thấy trong chương tiếp theo chương 2 ,các mô hình có thể đại diện trong nhiều cách như :1 cái cây, 1 cái bảng xác xuất ,đơn giản ,hoặc các tham số thực của 1 vector.Chính thức hơn ,Chúng ta có thể biểu thị nó về mặt toán học như một hàm mục tiêu f mà ở đó nhận các dữ liệu đầu vào là các tập thuộc tính x và trình ra dữ liệu ra tương ứng để phỏng đoán nhãn class .

Các mô hình nói để phân loai một instances(x,y) chính xác if(x)=y

Bảng 3.1.Ví dụ về tác vụ phân loại

|  |  |  |
| --- | --- | --- |
| Tác vụ | Tập thuộc tính | Nhãn class |
| Lọc thư rác | Tính năng khai thác từ tiêu đề thông báo thư và nội dung | Rác hoặc không rác |
| Xác định các khối u | Tính năng khai thác từ chụp ảnh cộng hưởng từ(quét) | Lành tính hoặc ác tính |
| Phân loại ngân hà | Tính năng khai thác ảnh từ kính thiên văn | Elip ,xoắn ốc ,hoặc hình dạng không đều |

Bảng 3.2 :Một tập dữ liệu đơn giản cho vấn đề phân loại động vật có xương sống .



Bảng 3.1 hiển thị ví dụ về tập thuộc tính và nhãn class đối với các tác vụ phân loại đa dạng.Lọc các thư rác và xác định các khối u là ví dụ về vấn đề phân loại nhị phân (tức là ở đây có 2 nhãn class)trong đó mỗi trường hợp dữ liệu có thể phân loại vào 1 của 2 class.Nếu số lượng class lớn hơn 2 như trong ví dụ phân loại các ngân hà ,sau đó nó có thể đả gọi vấn đề phân loại của nhiều class.

Chúng ta minh họa các khái niệm cơ bản về phân loại trong chương này với 2 ví dụ sau:

Ví dụ 3.2:

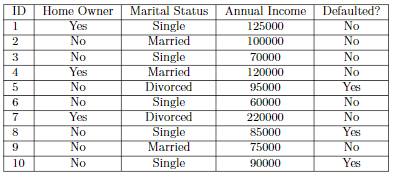
(Phân loại động vật có xương sống).Bảng 3.2 hiển thị 1 tập dữ liệu mẫu cho việc phân loại động vật có xương sống vào trong như động vật có vú ,bò sát ,chim,cá và động vật lưỡng cư.Các tập thuộc tính bao gồm các nét đặc trưng của các động vật có xương sống như nhiệt độ cơ thể của nó ,bìa da,và khả năng để bay.Các tập dữ liệu củng có thể đả được sử dụng cho tác vụ phân loại nhị phân như phân loại động vật có vú ,bởi các nhóm bò sát ,chim ,cá và động vật lưỡng cư vào 1 loại đơn là không phải động vật có vú.

Ví dụ 3.3:

Phân loại người vay.

Xem xét các vấn đề của việc phỏng đoán ,liệu một người vay loan sẽ trả lại tiền. vay hoặc mặc định là phải hoàn trả lại tiền vay.Các tập dữ liệu được sử dụng để xây dựng mô hình phân loại hiển thị ra trong bảng 3.3 .Các tập thuộc tính bao gồm những thông tin cá nhân của người vay như trạng thái hôn nhân,thu nhập hằng năm ,trong khi các nhãn class chỉ ra liệu rằng người vay tiền có mặc định hoàn trả lại tiền .

Bảng 3.3:Một tập dữ liệu mẫu cho các vấn đề phân loại người vay tiền



Một mô hình phân loại phục vụ 2 quy tắc quan trọng trong khai phá dữ liệu.Đầu tiên ,nó được sử dụng như một mô hình phỏng đoán đến phân loại các trường hợp không nhãn trước đây.Một mô hình phân loại tốt ,phải cung cấp phỏng đoán chính xác với 1 thời gian đáp ứng nhanh .Thứ 2 ,Nó được sử dụng như một mô hình mô tả để định danh các đặc điểm ở đó ,trường hợp đặc biệt từ các class khác nhau .Điều này là cụ thể hửu ích cho ứng dụng phán đoán ,giống như chẩn đoán y khoa.,Nơi nó không đủ để có mô hình ở đó làm một sự phỏng đoán mà không cần biện minh,làm cách nào để nó đạt được sự quyết định.

Đối với ví dụ,một mô hình phân loại hướng dẫn từ tập các dữ liệu động vật không có xương sống hiển thị trong bảng 3.2 có thể được sử dụng để phỏng đoán các nhãn class của các động vật không có xương sống như sau:



Ngoài ra ,nó có thể sử dụng như mô hình mô tả để giúp đở xác định các đặc điểm ở đó mà định nghĩa động vật có xương sống như một động vật có vú ,một bò sát ,chim,cá ,hoặc một động vật lưỡng cư .Đối với ví dụ ,các mô hình định danh động vật có vú như động vật có vú máu nóng ở đó sinh ra tuổi trẻ.

Ở đó là một số điểm giá trị lưu ý về các ví dụ trước .Đầu tiên ,mặc dù tất cả thuộc tính hiển thị trong bảng 3.2 là định tính ,ở đó là ko những hạn chế trên các loại của thuộc tính ở đó được sử dụng để phỏng đoán các biến .

Các nhãn class ,mặt khác khác phải là loại thông thường .Điều này phân biệt phân loại từ tác vụ mô hình phỏng đoán khác như hồi quy ,nơi các giá trị phỏng đoán thường là định lượng .Nhiều thông tin về hồi quy có thể tìm thấy trong Appendix D.

Lưu ý giá trị điểm khác là ở đó không phải là tất cả thuộc tính có thể liên quan đến tác vụ phân loại .Đối với ví dụ độ dài hoặc cân nặng trung bình của một động vật không có xương sống không thể hửu ích cho phân loại động vật lưỡng cư ,ở đó thuộc tính có thể hiển thị giá trị mẫu cho cả 2 mammals và non-mammals .Một thuộc tính như vậy thường bị loại bỏ trong quá trình tiền xữ lí.Các thuộc tính còn lại có thể không khả năng để phân biệt các class bởi chúng tôi,và do đó ,phải sử dụng trong các bàn tính với các thuộc tính khác.Đối với trường hợp (instance) các thuộc tính nhiệt độ cơ thể là không đủ để phân biệt động vật có vú từ động vật không có xương sống khác .Trong khi nó được sử dụng cùng nhau để sinh đẻ ,các phân loại động vật có vú là cải thiện đáng kể .Ngoài ra khi bổ sung thuộc tính như skin(da) đảm bảo các mô hình trở nên chỉ định quá mức và không còn bao gồm tất cả các động vật có vú .Tìm các lựa chọn kết hợp của thuộc tính ở đó trường hợp phân biệt đối xử tốt từ class khác nhau là thử thách chính trong xây dựng mô hình phân loại.

3.2.Khung chung cho việc phân loại.

Phân loại là các tác vụ của việc gán nhãn đến không nhãn của các dữ liệu về instances và phân loại là sử dụng giống như thực hiện 1 tác vụ.Một việc phân loại là cùng mô tả trong các mục của một mô hình như mình họa trong các phần trước .Các mô hình là được tạo sử dụng đưa ra 1 tập các instance.biết đến như 1 tập tranning set ,mà chứa giá trị thuộc tính củng như đối với mỗi instance .Các hệ thống tiếp cận cho việc học một mô hình phân loại đưa ra 1 tập cá tranning set được biết đến như việc học 1 thuật toán .Các cái xữ lí của sử dụng một học thuật toán để xây dựng một mô hình phân loại từ các tranning data được biết đến như 1 induction.Xữ lí này củng thường mô tả như học mô hình hoặc xây dựng mô hình .Xữ lí này của việc áp dụng mô hình phân loại trên các tập instance không nhìn thấy để phỏng đoán các nhãn class chúng biết đến như deduction .Như vậy ,các xữ lí về phân loại liên quan 2 bước :

Áp dụng học một thuật toán tới tranning data để học một mô hình ,và sau đó áp dụng các mô hình để gán nhãn tới ko gán nhãn instances .Hình 3.3 mô tả khung cho việc phân loại .

Một kĩ thuật phân loại đề cập đến 1 cái chung tiếp cạnh để phân loại .

Ví dụ như kĩ thuật cây quyết định sẽ đượ học chương này .Kĩ thuật phân loại này giống như phần lớn cái khác .bao gồm về gia đình ,một liên quan mô hình và một số thuật toán cho việc học các mô hình đó .Trong chương 4 ,Chúng ta sẽ học bổ sung thêm kĩ thuật phân loại bao gồm mạng neuron và hổ trợ các máy .

Một vài lưu ý về thuật ngữ .Đầu tiên ,các mục”classifier” và “model” là thường lấy đồng nghĩa.Nếu một kĩ thuật phân loại xây dựng một mô hình đơn ,mô hình toàn cầu,sau đó điều này là tốt ,Tuy nhiên ,mỗi mô hình định nghĩa một phân loại ,ko mỗi mô hình định nghĩa bởi mô hình đơn .Một số phân loại giống như phân loại k-láng giềng gần nhất ,không xây dựng một mô hình rõ ràng (Phần 4.3).Trong khi các phân loại khác như phân loại tập hợp,kết hợp các đầu ra của một sưu tập của mô hình (Phần 4.10).Phần 2,các mục “classifier” là thường sử dụng trong rất nhiều giác quan chung đến tham khảo đến các kĩ thuật phân loại.Như vậy đối với ví dụ phân loại cây quyết đinh có thể tham khảo các kĩ thuật cây quyết định hoặc chỉ định phân loại xây dựng sử dụng kĩ thuật đó .May mắn thay ,các ý nghĩa của “classifier”thường làm rõ từ mức ngữ cảnh .

Trong cái khung chung hiển thị ở hình 3.3,các bước induction(hướng dẫn ) và deduction(suy đoán) nên thực hiện riêng biệt.Trong thực tế ,như sẽ thảo luận sau trong phần 3.6 ,các tranning set và test set nên phải độc lập của mỗi cái khác ,để đảm bảo rằng các mô hình hướng dẫn có thể phỏng đoán các nhãn class là chính xác về trường hợp nó không bao giờ đả gặp trước đây.Mô hình đó cung cấp những hiểu biết gọi nó có 1 hiệu xuất khái quát chung tốt.Hiệu suất của một mô hình có thể được đánh giá bởi việc so sánh các nhãn phỏng đoán với các nhãn true.Những thông tin này có thể tóm tắt lại trong một cái bảng gọi là confusion matrix .Bảng 3.4 miêu tả các confusion matrix(ma trận hổn loạn), for vấn đề phân loại nhị phân .Mỗi nhập cảnh fij biểu thị các số của trường hợp từ class I phỏng đoán đến class j .Đối với ví dụ,f01 là số lượng của trường hợp từ class 0 không chính xác phỏng đoán class 1 .Các số lượng phỏng đoán chính xác được tạo bởi mô hình là (f11+f00) và các số lượng phỏng đoán không chính xác là f10 và f01

Mặc dù ma trận hổn loạn cung cấp các thông tin cần để xác định làm cách nào một mô hình phân loại thực hiện tốt.Tóm tắt lại những thông tin này thành một số đơn ,giúp thuận tiện so sánh các hiệu suất liên quan của mô hình khác nhau.Điều này có thể hoàn thành sử dụng 1 hệ mét đánh giá giống như sự chính xác ,mà thực hiện như cách dưới đây:

Accurancy=số lượng phỏng đoán chính xác/Tổng số các phỏng đoán .

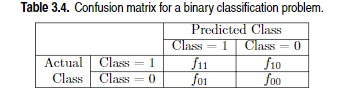
Đối với vấn đề phân loại nhị phân sự chính xác mô hình được đưa ra :

Accurancy=(F11+F00)/(F11+F00+F10+f01)

Tỉ lệ lỗi sai liên quan khác hệ ,mà định nghĩa như sau đối với vấn đề phân loại nhị phân .

Error rate=số lượng phỏng đoán sai/Tổng số phỏng đoán =(F10+F01)/ (F11+F00+F10+f01)

Học các thuật toán về phần lớn các kĩ thuật phân loại thiết kế để học một mô hình ở đó đạt được sư chính xác cao nhất hoặc tương đương .Tỉ lệ lỗi thấp nhất khi áp dụng đến nó đến các tập test set .Chúng tôi sẽ các chủ đề đánh giá mô hình trong phần 3.6.



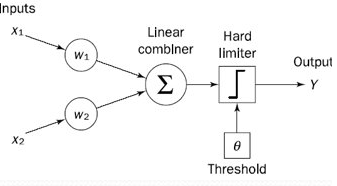
Bài 2:

Cấu trúc mạng đơn tầng .Giải thích các tham số ,Mã giải (Perceptron).

Cấu trúc mạng đơn tầng là gì?

Mạng neuron đơn tầng là có 1 neuron liên kết với các trọng số và mạng neuron đó sẽ học và điều chỉnh các trọng số và với 1 giới hạn cứng nào đó

Giải thích các tham số:



Input:là các dữ liệu vào ,ở đây có 1 tập các thuộc tính x1,..xn.

Với 1 tập thuộc tính vào có gán 1 trọng số w1,w2.

Linear combiner:Kết hợp lại thành 1 tổng x1.w1 +x2.w2+..+xn.wn

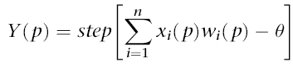
Ouput:các dữ liệu ra mà chúng ta phân loại nó cụ thể trong cấu trúc perceptron thành 2 lớp thông qua bởi 1 hàm phân biệt:  

:giới hạn cứng mà cái mạng nó nó có thể học tới .

Các bước mã giải (Perceptron):

B1:Khởi tạo các trọng số ,các trọng số đó nằm trong khoảng -0.5 đến 0.5

B2: Tính kết xuất thực tại lần lặp thứ p theo công thức:



Trong đó Y(p) là giá trị kết xuất thực .

Y(p)=1 ,nếu cái :>=0

Y(p)=0,nếu cái <0

B3:Cập nhật các trọng số :Trọng số lúc sau bằng trong số lúc đầu cộng với lượng trọng số đưa vào là A\*xi(p)\*e(p) để điều chỉnh trọng lại trọng số .

Cụ thể :với A là tốc độ học ,

xi(p) là giá trị thuôc tính đưa vào .

e(p)lượng sai số trong giữa kết xuất thực và kết xuất mong muốn .

Sau đó quay trở lại bước 2 để tính kết xuất thực để học tiếp.