## Phân loại/lớp bằng Naïve Bayse

## 3.2. Phân loại dựa trên xác suất

Kỹ thuật này có thể hiểu đơn giản như sau: với một mẫu dữ liệu cần phân lớp, ta tính xác suất có điều kiện để mẫu dữ liệu đó rơi vào từng lớp trong tập các lớp đã biết trước. Mẫu dữ liệu sẽ được phân vào lớp nào có xác suất cao nhất.

### 3.2.1. Một số khái niệm ban đầu

**Hiện tượng tất yếu:** là những hiện tượng nếu được thực hiện ở điều kiện giống nhau thì cho kết quả giống nhau. Chẳng hạn khi đun nước đến 1000C thì nước sôi. Hiện tượng tất yếu là đối tượng nghiên cứu của Vật lý, Hóa học.

**Hiện tượng ngẫu nhiên:** là những hiện tượng dù đã được quan sát ở điều kiện giống nhau, nhưng kết quả có thể khác nhau. Ví dụ: tung đồng xu, và quan sát xem đồng xu là “*sấp*” hay “*ngửa*”. Hiện tượng ngẫu nhiên là đối tượng nghiên cứu của xác suất học.

Trong một hiện tượng ngẫu nhiên ta không thể biết được chắc chắn kết quả xảy ra như thế nào, nhưng có thể hình dung ra được các khả năng mà kết quả có thể xảy ra. Tập hợp tất cả các kết quả có thể xảy ra được gọi là không gian mẫu, ký hiệu là *Ω.* Ví dụ: tung một đồng xu, *Ω* = {*sấp, ngửa*}; tung một con xúc sắc và tính điểm, *Ω* = {1, 2, 3, 4, 5, 6} ...

**Biến cố:** là một tập con của không gian mẫu, ký hiệu là: *A, B, C*... Ví dụ: **t**ung một con xúc sắc, gọi *A* là biến cố được số điểm chẵn và *B* là biến cố được số điểm lẻ thì *A* = {2, 4, 6}, *B* = {1, 3, 5}. Vì các biến cố là các tập hợp, nên ta thường sử dụng các phép tính trên tập hợp cho biến cố:

* Phép hội : *A* ∪ *B* (*A* hay *B* xảy ra).
* Phép giao: *A* ∩ *B* = *AB* (*A* và *B* xảy ra).
* Phép bù: (*A* không xảy ra).

Khi quan sát các hiện tượng, ta thấy có những hiện tượng thường xuyên xảy ra, có những hiện tượng ít xảy ra. Xác suất là một con số đo lường mức độ thường xuyên xảy ra của một biến cố. Xác suất xảy ra một biến cố *A* (hay xác suất của *A*), ký hiệu là *P(A)* là tỷ lệ giữa số lần biến cố *A* xảy ra và số lượng tất cả các biến cố:

P(A) = 

(3.1)

**Tính chất cơ bản của xác suất:**

0 ≤ *P(A)* ≤ 1;

*P*(*true*) = 1;

*P*(*false*) = 0;

*P(A ∪ B) = P(A) + P(B) - P(A ∩ B) .*

**Xác suất có điều kiện:**

*P(A|B)* là phần của không gian mà trong đó *A* là đúng, với điều kiện (đã biết) là *B* đúng. Nói cách khác, P(*A|B*) là xác suất xảy ra biến cố *A* với điều kiện là có xảy ra biến cố *B*, thường được gọi là “*xác suất của A nếu có B*”. Ví dụ:

*A*: *Tôi sẽ đi đá bóng vào ngày mai*,

*B: Trời sẽ không mưa vào ngày mai,*

*P(A|B):* Xác suất của việc tôi sẽ đi đá bóng vào ngày mai nếu (đã biết rằng) trời sẽ không mưa vào ngày mai.

Gọi **P(A, B)** là xác suất xảy ra đồng thời hai sự kiện *A* và *B* và *P(B)* là xác suất xảy ra sự kiện *B*. Dễ dàng thấy rằng:

(3.2)



**Công thức xác suất toàn phần:**

Nếu *B1 + B2 + …+ Bn = Ω* và *BiBj* = ∅ ∀*i* ≠ *j*, khi đó với biến cố *A* liên quan được tính theo công thức:

(3.3)

*P(A)* =

### 3.2.2. Định lý Bayes

Cho *h* là một giả thiết và *x* là tập các giá trị quan sát được. Khi đó, xác suất để giả thiết *h* là đúng khi biết *x* được tính như sau:



(3.4)

trong đó:

***P(h)*:** xác suất tiên nghiệm của giả thiết *h*. Đây là xác suất để giả thiết *h* là đúng mà không liên quan gì tới *X*. Nó được gọi là “*tiên nghiệm*” với hàm ý rằng nó không quan tâm tới bất kỳ thông tin nào của *X*.

***P(X):***xác suất tiên nghiệm của việc quan sát được *X*. Đây là xác suất xảy ra *X* mà không quan tâm tới *h*.

***P(X|h)*:** xác suất xảy ra *X*, nếu biết giả thiết *h* là đúng. Nói cách khác, đây là xác suất xảy ra *X* khi biết giả thiết *h* đã xảy ra.

### 3.3.3. Phân lớp bằng kỹ thuật Naïve Bayes

Trước tiên, ta xét bài toán phân lớp. Cho một tập dữ liệu huấn luyện *X* ∈ *Rn×(m+1)* gồm *n* mẫu dữ liệu, mỗi mẫu có *m* thuộc tính và một thuộc tính lớp. Mỗi mẫu huấn luyện *x∈X* được biểu diễn là một vectơ *m*+1 chiều *x*(*x1, x2, ..., xm, y*) trong đó gồm *m* thành phần dữ liệu và *y* là nhãn lớp. Cho một tập xác định các nhãn lớp *C* = {*c1, c2,..., cq*} gồm *q* lớp. Dễ thấy *y*∈*C*. Cho một mẫu dữ liệu mới *z∈Rm* và *z* được biểu diễn bằng: *z(z1, z2,...,zm)*. Hãy xác định lớp của *z*.

Để xác định lớp của *z*, một cách đơn giản là ta tính xác suất xảy ra khả năng *z* được phân vào từng lớp *ci*, *i*=1..*q*, tức khả năng xảy ra *ci*. Mẫu *z* sẽ được phân vào lớp nào có xác suất xảy ra cao nhất.

Tuy nhiên, mẫu *z* là xác định với các thành phần quan sát được là *z1, z2,...zm*. Do đó, xác suất để *z* thuộc vào lớp *ci* phải là xác suất có điều kiện và được ký hiệu là *P(ci|z).* Theo định lý Bayes, xác suất này được tính như sau:

.

(3.5)

Trong phương pháp *Naïve Bayes*, từ *Naïve* có hàm ý giả sử rằng các thuộc tính là độc lập có điều kiện đối với các thuộc tính khác. Do đó



(3.6)

và (3.5) trở thành:

.

(3.7)

Mẫu dữ liệu *z* sẽ được phân vào lớp *ck* nếu *P(ck|z)* là lớn nhất tức:



(3.8)

Vì *P(z)* là hằng số đối với các *ci* khác nhau, do vậy (3.8) tương đương với:



(3.9)

Một cách đơn giản hơn, để xác định lớp cho mẫu dữ liệu *z*, ta lần lượt tính các giá trị của biểu thức  với từng lớp *ci* ∈{*c1, c2,...,cq}*. Lớp *ci* nào cho giá trị của biểu thức lớn nhất sẽ là lớp của *z*. Quá trình phân lớp sử dụng phương pháp *Naïve Bayes* bao gồm hai bước:

**Bước 1:** Đối với mỗi lớp *ci* ∈ *C*, tính giá trị của:

* Xác suất tiên nghiệm *P(ci)*. Xác suất này được tính xấp xỉ bằng tổng số mẫu thuộc lớp *ci* trên tổng số mẫu của bộ dữ liệu huấn luyện.
* Đối với mỗi giá trị thuộc tính *zj*, tính *P(zj|ci)* là xác suất xảy ra của giá trị đó trong lớp *ci*. Giá trị này cũng được tính xấp xỉ bằng tỷ lệ các mẫu có giá trị trên thuộc tính thứ *j* là *zj* trong số các mẫu thuộc lớp *ci*.

**Bước 2:** Cần xác định lớp cho một mẫu dữ liệu mới *z*, ta thực hiện:

* Đối với mỗi lớp *ci* ∈*C*, tính giá trị của biểu thức:

,

(3.10)

* Xác định lớp của *z* là *ck*:

(3.11)



**Ví dụ 3.1.** Cho bảng dữ liệu huấn luyện gồm 14 mẫu về quyết định (có hay không) mua máy tính như trong Bảng 3.1, dựa vào các quan sát về tuổi (*Age*), thu nhập (*Income*), có là sinh viên hay không (*Student*) và tình hình tín dụng (*Credit*).

Bảng 3.1. *Số liệu quan sát về điều kiện mua máy tính*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Age** | **Income** | **Student** | **Credit** | **Buy** |
| 1 | Young | High | No | Fair | no |
| 2 | Young | High | No | Excellent | no |
| 3 | Medium | High | No | Fair | yes |
| 4 | Old | Medium | No | Fair | yes |
| 5 | Old | Low | Yes | Fair | yes |
| 6 | Old | Low | Yes | Excellent | no |
| 7 | Medium | Low | Yes | Excellent | yes |
| 8 | Young | Medium | No | Fair | yes |
| 9 | Young | Low | Yes | Fair | yes |
| 10 | Old | Medium | Yes | Fair | yes |
| 11 | Young | Medium | Yes | Excellent | yes |
| 12 | Medium | Medium | No | Excellent | yes |
| 13 | Medium | High | Yes | Fair | yes |
| 14 | Old | Medium | No | Excellent | no |

Cho một mẫu dữ liệu cần phân lớp *x*(*Youth, Medium, Yes, Fair*), tức xác định xem một sinh viên trẻ với thu nhập trung bình và mức đánh giá tín dụng bình thường sẽ có quyết định mua một chiếc máy tính hay không.

Dễ dàng thấy số mẫu dữ liệu *n*=14; số thuộc tính dữ liệu *m*=4 (do không xem xét thuộc tính *ID*); thuộc tính lớp là *Buy* với tập các lớp *C*={*yes, no*} gồm 2 lớp. Quá trình xác định lớp cho mẫu dữ liệu *x* trải qua hai bước:

**Bước 1:** với mỗi lớp *ci*∈*C*:

* Xét *c1=yes*: dễ dàng tình được *P(yes)* = 10/14. Ta tiếp tục tính *P(xj|c1*):

*P(Age=Young | Buy=yes)*  = 3/10;

*P(Income=Medium | Buy=yes)* = 5/10;

*P(Student = Yes | Buy = yes)* = 6/10;

*P(Credit = Fair | Buy = yes)* = 7/10.

* Xét *c2=no*: dễ dàng tính được *P(no)* = 4/14. Ta tiếp tục tính *P(xj|c2)*:

*P(Age=Young | Buy=no)* = 2/4;

*P(Income=Medium | Buy=no)* = 1/4;

*P(Student = Yes | Buy = no)* = 1/4;

*P(Credit = Fair | Buy = no)* = 1/10.

**Bước 2:** Sử dụng các kết quả vừa tính, ta được:

*  = *P(Age = Youth | Buy = Yes)* ×

*P(Income = Medium | Buy = Yes)* ×

*P(Student = Yes | Buy = Yes)* ×

*P(Credit = Fair | Buy = Yes)*

=  = 0,063.

.P(c1) = 0.063\*10/14 = **0.045**.

*  = *P(Age = Youth | Buy = no)* ×

*P(Income = Medium | Buy = no)* ×

*P(Student = Yes | Buy = no)* ×

*P(Credit = Fair | Buy = no)*

= = 0,0078.

P(c2) = 0.0078\*4/14 = **0.0022**.

Vậy mẫu dữ liệu *x* được phân vào lớp *c1* hay lớp của *x* là “*yes*”.

**Phương pháp Naïve Bayes trong trường hợp dữ liệu liên tục**

Các thuộc tính trong Bảng 3.1 đều có giá trị rời rạc. Trong trường hợp thuộc tính có giá trị liên tục, ta có thể áp dụng các phương pháp rời rạc hóa. Nếu không rời rạc hóa dữ liệu, thay vì tính xác suất, ta sử dụng hàm mật độ xác suất. Thông thường, ta hay giả thiết là dữ liệu trong mỗi lớp *ci* của các thuộc tính liên tục tuân theo phân bố *Gauss* và phương pháp lúc này được gọi là *Gauss Naïve Bayes*.

Xét thuộc tính *A* với các giá trị liên tục. Khi đó, ta phân đoạn các giá trị của *A* theo từng lớp. Với mỗi lớp *ci*, ta tính *µi* là giá trị trung bình và *σi2* là phương sai của các giá trị của *A* trong lớp *ci* (với *Ni* là số mẫu thuộc lớp *ci* và *yi* là lớp của mẫu *xi*):



(3.12)

Giá trị *P(x|ci)* khi đó gọi là phân bố xác suất của *x* vào lớp *ci* và được tính bằng:

.

(3.13)

**Ví dụ 3.2**. Xét bảng dữ liệu 3.2 được xây dựng bằng cách thay thế cột *Income* (thu nhập) bằng các giá trị thực và dữ liệu được sắp lại trên cột *Buy* như sau:

**Bảng 3.2.** *Dữ liệu quan sát với trường Income có giá trị liên tục*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Age** | **Income** | **Student** | **Credit** | **Buy** |
| 1 | Young | 3.1 | No | Fair | no |
| 2 | Young | 2.8 | No | Excellent | no |
| 3 | Old | 3 | Yes | Excellent | no |
| 4 | Old | 3.7 | No | Excellent | no |
| 5 | Medium | 5.9 | No | Fair | yes |
| 6 | Old | 6 | No | Fair | yes |
| 7 | Old | 6.1 | Yes | Fair | yes |
| 8 | Medium | 3.1 | Yes | Excellent | yes |
| 9 | Young | 7 | No | Fair | yes |
| 10 | Young | 2.5 | Yes | Fair | yes |
| 11 | Old | 3.1 | Yes | Fair | yes |
| 12 | Young | 3.9 | Yes | Excellent | yes |
| 13 | Medium | 7.5 | No | Excellent | yes |
| 14 | Medium | 4.6 | Yes | Fair | yes |

Giả sử mẫu dữ liệu *x(Youth, 5.2, Yes, Fair)* cần phân lớp. Các giá trị *P(Income=5.2 | Buy=yes)* và *P(Income = 5.2 | Buy = no)* cần phải tính lại.

Xét lớp *c1 = yes*, ta dễ dàng tính được *µ1* = 4.97 là giá trị trung bình trên cột *Income* của các mẫu thuộc lớp *yes* và *σ12* = 3.12 là phương sai tương ứng. Tương tự với lớp *c2=no*, ta tính được *µ2* = 3.15 và *σ22* = 0.15. Ta có bảng các giá trị trung bình và phương sai trên từng lớp của cột *Income* như sau:

|  |  |  |
| --- | --- | --- |
| Giá trị Lớp | ***µ*** | ***σ2*** |
| **yes** | 4.97 | 3.12 |
| **no** | 3.15 | 0.15 |

Vậy *P(Income=5.2 | Buy=yes)* = ≈ 0.191 và

*P(Income = 5.2 | Buy = no)* = ≈ 1.7966.10-12.

Và do đó, *P(x|c1)* =  và

*P(x|c2)* =  .

Mẫu dữ liệu *x* được phân vào lớp *yes*.

**Trường hợp xuất hiện xác suất bằng không**

Xét một mẫu dữ liệu cần phân lớp *x(x1, x2,...,xm)*. Xét giá trị *xj* trên thuộc tính *j*. Nếu không có mẫu dữ liệu nào trong lớp *ci* có giá trị trên thuộc tính *j* là *xj* thì hiển nhiên *P(xj|ci)* = 0. Điều này kéo theo  = 0.

Giải pháp đưa ra là sử dụng ước lượng *Laplace* để ước lượng *P(xj|ci)* thay cho giá trị 0 đã tính được ở trên.

Giả sử rằng ta có bộ dữ liệu với các thuộc tính giống như Bảng 3.1 và trong lớp *c1*=“*yes*” có 1000 mẫu dữ liệu. Xét thuộc tính *Income* của 1000 mẫu trên với 0 mẫu có giá trị *Income=”Low”;* 990 mẫu dữ liệu có *Income=”Medium”* và 10 mẫu có *Income=”High”*. Khi đó, các xác suất *P(Income=”Low” | Buy=yes)*, *P(Income=”Medium” | Buy=yes)* và *P(Income=”High” | Buy=yes)* lần lượt được xấp xỉ bằng 0, 990/1000 và 10/1000. Do đó, với một mẫu *x* cần phân lớp có *Income=”Low”,* ví dụ *x(Youth, Low, Yes, Fair)* ta tính được

= *P(Income=”Low” | Buy=yes)\*………………..* = 0

Để tránh trường hợp này, ta giả sử rằng số mẫu dữ liệu trong lớp “*yes*” là lớn và do đó, nếu ta bổ sung 01 mẫu dữ liệu cho mỗi tập có *Income=”Low”*, *Income=”Medium”* và *Income=” High”* thì việc này không ảnh hưởng nhiều tới các xác suất đã tính. Nhưng khi đó, các xác suất *P(Income=”Low” | Buy=yes), P(Income=”Medium” | Buy=yes)* và *P(Income=”High” | Buy=yes)* sẽ thay đổi và lần lượt bằng 1/1003, 991/1003 và 11/1003.

## 3.3. Phân lớp dựa trên láng giềng gần nhất (KNN)

Một trong những phương pháp phân lớp đơn giản nhất là dựa trên lớp của các láng giềng gần nhất của mẫu cần phân lớp. Cho một mẫu *x* biểu diễn một đối tượng cần phân lớp. Mẫu *x* sẽ được phân vào lớp xuất hiện phổ biến trong số các lớp của các láng giềng của *x*. Hiển nhiên, các láng giềng của *x* đều đã được phân lớp từ trước. Phương pháp này dựa trên ý tưởng giả định rằng một đối tượng *x* sẽ có các đặc điểm, hành vi tương tự như các láng giềng của nó, và do vậy sẽ là hợp lý nếu nó thuộc cùng một lớp với các láng giềng gần nhất.

Kỹ thuật phân lớp dựa trên láng giềng gần nhất được sử dụng rộng rãi trong các hệ thống nhận dạng mẫu, nhận dạng đối tượng, nhận dạng sự kiện, phân loại dữ liệu văn bản… Khái niệm “láng giềng” dùng để chỉ các đối tượng có khoảng cách hoặc độ tương đồng “gần” với đối tượng *x*. Từ đây, ta cần phải định nghĩa một độ đo khoảng cách hoặc độ đo sự khác biệt giữa các đối tượng.

### 3.3.1. Độ đo khoảng cách

Để xác định độ “gần nhau” giữa hai mẫu dữ liệu, người ta sử dụng một độ đo khoảng cách được định nghĩa trước. Tùy theo kiểu dữ liệu của mẫu và đặc điểm của đối tượng nhận dạng mà ta sử dụng một độ đo phù hợp. Có rất nhiều độ đo khoảng cách (hay độ khác biệt) đã được định nghĩa.

Xét một mẫu dữ liệu *x* gồm *m* thuộc tính. Khi đó, mẫu *x* được xem là một véc tơ trong không gian *m* chiều (*x* có *m* thành phần). Gọi *x(x1, x2,…, xm)* và *y(y1, y2,…, ym)* là hai mẫu dữ liệu. Để tính khoảng cách giữa *x* và *y*, ký hiệu *d(x, y)* ta thường sử dụng một số độ đo sau:

Với dữ liệu kiểu số

**- Khoảng cách Euclidean:**

(3.14)



Dễ thấy, khoảng cách này chính là chuẩn 2 của *x-y*: *d(x, y)* = ||*x-y*||.

**- Khoảng cách Square Euclidean:**

(3.15)



Dễ thấy, khoảng cách này chính là bình phương của khoảng cách Euclidean: *d(x,y)*=||*x-y*||2.

**- Khoảng cách Manhattan:**



(3.16)

Khoảng cách này chính là chuẩn 1 của *x-y*: *d(x, y)* = ||*x-y*||1.

**- Khoảng cách Chebyshev:**

(3.17)



Khoảng cách này chính là chuẩn vô cùng của *x-y:* *d(x, y)* = ||*x-y*||∞.

**- Khoảng cách Cosin:**



(3.18)



Khoảng cách này được hiểu là:

Với dữ liệu kiểu *boolean*, người ta thường sử dụng các độ đo dành riêng như: *Hamming Distance, Jaccard Dissimilarity, Matching Dissimilarity, Dice Dissimilarity, Rogers Tanimoto Dissimilarity, Russell Rao Dissimilarity, Sokal**Sneath Dissimilarity, Yule Dissimilarity*…Với dữ liệu kiểu xâu ký tự, các độ đo thường được sử dụng là: *Edit Distance, Damerau**Levenshtein Distance, Hamming Distance, Smith Waterman Similarity, Needleman Wunsch Similarity*…

Trong một số trường hợp, dữ liệu có thể là một hỗn hợp giữa kiểu số và phi số. Khi đó, một độ đo hỗn hợp có thể là một lựa chọn.

### 3.3.2. Phương pháp k-láng giềng gần nhất

Phương pháp phân lớp *k*-láng giềng gần nhất (*k*-NN) ban đầu dựa trên khoảng cách Euclidean để xác định đâu là các láng giềng gần nhất của mẫu dữ liệu cần phân lớp. Do vậy, dữ liệu thường là kiểu số. Tuy nhiên, việc lựa chọn một độ đo khoảng cách phù hợp có thể mở rộng phương pháp này cho các dữ liệu có kiểu phi số hoặc hỗn hợp.

Cho một bộ dữ liệu huấn luyện *X* gồm *n* mẫu: *X(x1, x2, …, xn)*. Mỗi mẫu dữ liệu *xi (i=1,...,n)* gồm *d* thuộc tính dữ liệu và một thuộc tính lớp.

Một mẫu dữ liệu *y(y1, y2,…,yd) chưa* được xác định lớp. Để phân lớp mẫu *y*, ta tiến hành như sau:

**Bước 1:** Tính khoảng cách từ *y* tới *n* mẫu dữ liệu trong bộ dữ liệu huấn luyện, tức là tính *n* khoảng cách *d(y, xi), i*=1,…,*n*.

**Bước 2:** Lựa chọn ra *k* mẫu dữ liệu trong bộ dữ liệu huấn luyện “*gần*” với *y* nhất, tức có khoảng cách tới *y* là nhỏ nhất.

**Bước 3:** Xác định lớp cho *y*. Lớp của mẫu *y* được xác định là lớp xuất hiện nhiều nhất trong số các giá trị lớp của *k* mẫu gần với *y* nhất đã tìm được ở Bước 2.

Khi sử dụng phương pháp này, ta cần lưu ý một số điểm sau:

- Số láng giềng gần nhất (*k*) được coi như đầu vào của thuật toán, tức là cần chọn một giá trị cho *k* trước khi thực hiện thuật toán. Hiển nhiên là số *k* được chọn sẽ có thể ảnh hưởng tới chất lượng của giải thuật. Một câu hỏi đặt ra là: với một bộ dữ liệu cho trước, số *k* được chọn như thế nào để cho kết quả thuật toán tốt nhất? Câu hỏi này mở ra một lĩnh vực nghiên cứu sôi động và các kết quả hầu như đều chưa đưa ra được câu trả lời thỏa đáng. Nói chung, giá trị của *k* phụ thuộc mạnh vào bộ dữ liệu huấn luyện mà ta có. Một số phương pháp phổ biến để xác định số *k* có thể kể tới như *cross-validation* hoặc *bootstrapping*…Chúng thực chất là phương pháp thử với nhiều giá trị *k* khác nhau và lựa chọn ra một số *k* tối ưu đối với một bộ dữ liệu huấn luyện. Trong một số ứng dụng, người ta sử dụng số *k* mặc định bằng 1.

- Độ đo khoảng cách cần phải phù hợp với dữ liệu hiện có. Hiển nhiên, độ đo này phải phù hợp với kiểu của dữ liệu. Ngoài ra, với một kiểu dữ liệu cụ thể, lại có thể có nhiều độ đo khoảng cách cho kiểu dữ liệu đó. Lựa chọn độ đo nào trong số các độ đo này là tùy thuộc vào bài toán cụ thể.

**Ví dụ 3.3:** Cho bộ dữ liệu huấn luyện gồm 9 mẫu dữ liệu *R1, R2, …, R9*, mỗi mẫu gồm 4 thuộc tính *A, B, C, D* và thuộc một trong 3 lớp (thuộc tính *CLASS*) như Bảng 3.3 sau:

Bảng 3.3. *Dữ liệu huấn luyện cho giải thuật k-NN*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **CLASS** |
| R1 | 5.1 | 3.5 | 1.4 | 0.2 | 1 |
| R2 | 4.9 | 3.0 | 1.4 | 0.2 | 1 |
| R3 | 4.7 | 3.2 | 1.3 | 0.2 | 1 |
| R4 | 7.0 | 3.2 | 4.7 | 1.4 | 2 |
| R5 | 6.4 | 3.2 | 4.5 | 1.5 | 2 |
| R6 | 6.9 | 3.1 | 4.9 | 1.5 | 2 |
| R7 | 6.3 | 3.3 | 6.0 | 2.5 | 3 |
| R8 | 5.8 | 2.7 | 5.1 | 1.9 | 3 |
| R9 | 7.1 | 3.0 | 5.9 | 2.1 | 3 |

*y* là một mẫu dữ liệucần được phân lớp. Hãy xác định lớp của *y* bằng giải thuật *k*-NN với *k*=3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| y: | 5.5 | 3.5 | 5.4 | 0.2 | ? |

**Bước 1:** tính khoảng cách từ *y* tới 9 mẫu dữ liệu của bảng dữ liệu huấn luyện. Giả sử độ đo Euclidean được sử dụng:

|  |  |
| --- | --- |
| d(y, R1) = | 4.02 |
| d(y, R2) = | 4.08 |
| d(y, R3) = | 4.19 |
| d(y, R4) = | 2.07 |
| d(y, R5) = | 1.84 |
| d(y, R6) = | 2.01 |
| d(y, R7) = | 2.52 |
| d(y, R8) = | 1.93 |
| d(y, R9) = | 2.58 |

**Bước 2:** Chọn ra *k* láng giềng gần *y* nhất: *R5, R8, R6*.

**Bước 3:** Dễ dàng xác định được lớp của *y* là 2, là lớp của hai trong số ba láng giềng gần nhất của *y*.

Trong trường hợp ta lựa chọn *k*=6, dễ dàng tìm thấy *k* láng giềng gần với *y* nhất là: *R4, R5, R6, R7, R8, R9* với 3 mẫu thuộc lớp 2 và 3 mẫu thuộc lớp 3. Rõ ràng việc xác định lớp của *y* lúc này là không chắc chắn. Đây là một ví dụ về sự ảnh hưởng của số *k* tới kết quả phân lớp.

Bài tập

1. Sử dụng KNN phân lớp cho mẫu cuối cùng

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** | **Class** |
| 5.1 | 3.5 | 1.4 | 0.2 | 1 |
| 4.9 | 3.0 | 1.4 | 0.2 | 1 |
| 4.7 | 3.2 | 1.3 | 0.2 | 1 |
| 4.6 | 3.1 | 1.5 | 0.2 | 1 |
| 7.0 | 3.2 | 4.7 | 1.4 | 2 |
| 6.4 | 3.2 | 4.5 | 1.5 | 2 |
| 6.9 | 3.1 | 4.9 | 1.5 | 2 |
| 5.5 | 2.3 | 4.0 | 1.3 | 2 |
| 6.3 | 3.3 | 6.0 | 2.5 | 3 |
| 5.8 | 2.7 | 5.1 | 1.9 | 3 |
| 7.1 | 3.0 | 5.9 | 2.1 | 3 |
| 6.3 | 2.9 | 5.6 | 1.8 | 3 |
| 5.9 | 3.0 | 5.1 | 1.8 | ? |

1. Sử dụng Naïve Bayse, ID3 cho tập dữ liệu sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mã số** | **Màu da** | **Chiều cao** | **Chế độ ăn** | **Hoạt động** | **Loại** |
| 1 | Nâu | Cao | Nhiều | Cầu lông | B |
| 2 | Vàng | Thấp | Ít | Bơi lội | A |
| 3 | Nâu | Thấp | TB | Bơi lội | B |
| 4 | Trắng | Thấp | Ít | Cầu lông | A |
| 5 | Vàng | Cao | Ít | Bóng rổ | A |
| 6 | Vàng | Thấp | TB | Cầu lông | B |
| 7 | Trắng | Cao | Nhiều | Bơi lội | A |
| 8 | Vàng | TB | TB | Bơi lội | B |
| 9 | Nâu | TB | Nhiều | Cầu lông | B |
| 10 | Trắng | Cao | Ít | Bóng rổ | ? |