

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**- - - - -o0o- - - - -**



**BÀI TẬP LỚN**

**NHẬP MÔN LÝ THUYẾT NHẬN DẠNG**

**Đề tài: NHẬN DẠNG RƯỢU BẰNG NAIVE BAYES**

**Giáo viên hướng dẫn : ThS Trần Hùng Cường**

**Lớp : Khoa Học Máy Tính 2-K13**

**Nhóm : 1**

**Sinh viên thực hiện : Đình Văn Quân**

***Hà Nội, Tháng 11 Năm 2020***

**LỜI NÓI ĐẦU**

Ngày nay với sự phát triển mạnh mẽ của công nghệ thông tin, công nghệ thông tin đang được ứng dụng rất rộng dãi. Một trong số đó là machine learning.

Với nội dung của môn học cho phép, sinh viên có thể hiểu được cơ bản về nhận dạng (một ứng dụng của machine learning).

Nhóm mình sẽ giới thiệu bạn đọc về nhận dạng rượu sử dụng bộ phân lớp Bayes.

Trong bài tập lớn còn nhiều thiếu sót, mong thầy cô và bạn bè góp ý

**Tác giả**

Mục lục

**Chương 1 Tổng quan về nhận dạng**

**1.1 Giới thiệu về nhận dạng**

Nhận dạng mẫu *(pattern recognition)* là một ngành thuộc lĩnh vực học máy *(machine learning)*. Nói cách khác, nó có thể được xem là việc "cần thực hiện một tác động vào dữ liệu thô mà tác động cụ thể là gì sẽ tùy vào loại của dữ liệu đó". Như vậy nó là một tập các phương pháp học có giám sát *(supervised learning).*

Nhận dạng mẫu nhằm mục đích phân loại dữ liệu (là các mẫu) dựa trên: hoặc là kiến thức tiên nghiệm *(a priori)* hoặc dựa vào thông tin thống kế được trích rút từ các mẫu có sẵn. Các mẫu cần phân loại thường được biểu diễn thành các nhóm của các dữ liệu đo đạc hay quan sát được, mỗi nhóm là một điểm ở trong một không gian đa chiều phù hợp. Đó là không gian của các đặc tính để dựa vào đó ta có thể phân loại.

Một hệ thống nhận dạng mẫu hoàn thiện gồm một thiết bị cảm nhận *(sensor)* để thu thập các quan sát cần cho việc phân loại hay miêu tả; một cơ chế trích chọn đặc trưng *(feature extraction)* để tính toán các thông tin dưới dạng số hay dạng tượng trưng *(symbolic)* từ các dữ liệu quan sát được; và một bộ phân loại (hay lược đồ mô tả) nhằm thực hiện công việc phân loại thực sự (hay miêu tả các quan sát đó) dựa vào các đặc tính đã được trích rút.

Việc phân loại (hay lược đồ mô tả) thường dựa vào sự có sẵn của một tập các mẫu mà đã được phân loại (hay miêu tả) sẵn. Tập các mẫu này được gọi là tập huấn luyện và chiến lược học nhằm phân loại mẫu vào một trong các lớp có sẵn được gọi là học có giám sát. Việc học cũng có thể là không có giám sát, theo nghĩa là hệ thống không được cung cấp các mẫu được đánh nhãn (phân loại) *tiên nghiệm*, mà nó phải tự đưa ra các lớp để phân loại dựa vào tính ổn định trong thống kê của các mẫu.

Việc phân loại (hay lược đồ mô tả) thường dùng một trong các hướng tiếp cận sau: thống kê (hay lý thuyết quyết định), cú pháp (hay cấu trúc). Nhận dạng mẫu dùng thống kê là dựa vào các đặc tính thống kê của các mẫu, chẳng hạn rằng các mẫu được tạo mởi một hệ thống xác xuất. Nhận dạng dùng cấu trúc là dựa vào tương quan cấu trúc giữa các mẫu.

Các ứng dụng phổ biến là nhận dạng tiếng nói tự động, phân loại chữ viết (ví dụ: những thư điện tử nào là spam/non-spam), nhận dạng các mã bưu điện viết tay trên các bao thư, hay hệ thống nhận dạng danh tính dựa vào mặt người. Ba ví dụ cuối tạo thành lãnh vực con phân tích ảnh của nhận dạng mẫu với đầu vào là các ảnh số.

**1.2 Quy trình nhận dạng**

Tùy từng đối tượng và mục đích nhận dạng mà mỗi hệ thống nhận dạng có thiết kế khác nhau. Tuy nhiên sơ đồ sau đây thường được sử dụng như một mẫu chung để thiết kế nhận dạng.

Hậu xử lý

Phân lớp, hồi quy, mô tả

Tiền xử lý

Thu nhận mẫu

Các bộ phận chức năng chính bao gồm:

Thu nhận mẫu: có thể dưới nhiều dạng khác nhau như thu nhận tín hiệu, thu thập ảnh hoặc thu thập dữ liệu. Quá trình này được thực hiện bằng cách sử dụng các cảm biến, máy thu (camera, micro,...).

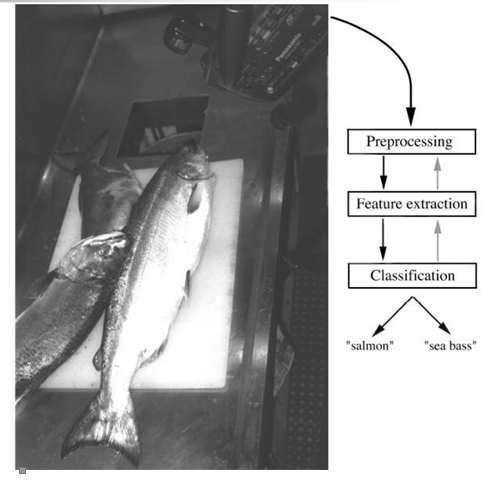
Tiền xử lý: trong một vài trường hợp, các giá trị của các thuộc tính không được đưa trực tiếp vào các công đoạn sau. Thông thường chúng được chuẩn hóa (ví dụ đưa các giá trị về đoạn [0,1]) theo một cách nào đó. Đôi khi quá trình tiền xử lý còn có thể là các công việc làm sạch, làm giàu hoặc chuyển đổi định dạng dữ liệu hoặc trích chọn thuộc tính. Khâu trích chọn thuộc tính thường được áp dụng khi hệ thống phải đối mặt với lượng dữ liệu lớn, có nhiễu. Việc trích chọn những thông tin nào, trích chọn ra sao thường đã được nghiên cứu và thực hiện bởi các giải thuật cụ thể. Đôi khi, việc thực hiện trích chọn thuộc tính còn giúp nâng cao hiệu suất hệ thống.

Phân lớp, hồi quy, mô tả: là những chức năng cơ bản của một hệ thống nhận dạng mà mục đích cuối cùng là phân loại các mẫu và nhận biết những đặc điểm chung của chúng.

Hậu xử lý: đôi khi dữ liệu đầu ra của hệ thống nhận dạng không được sử dụng trực tiếp mà còn trải qua quá trình hậu xử lý, chẳng hạn như các thao tác giải mã.

**1.3 Ví dụ về nhận dạng**

Bài toán phân loại cá trong dây chuyền thành hai loại?



Các bước thực hiện:

Tiền xử lý ảnh cá nhận được từ camera.

Phân vùng đối tượng cá

Trích chọn đặc trưng: độ sáng, chiều dài, chiều rộng

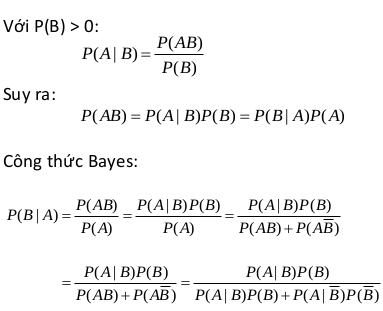
Phân loại đối tượng cá

**Chương 2 Kỹ thuật phân lớp sử dụng bayes**

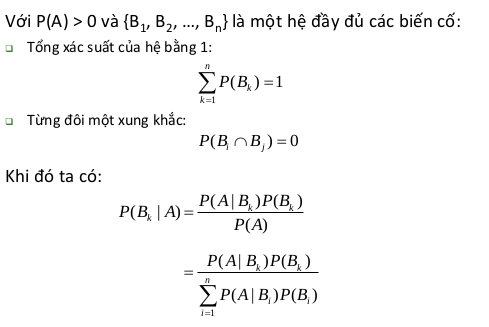
**2.1 Lý thuyết tổng quát**

Bộ phân lớp Bayes là một giải thuật thuộc lớp giải thuật thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu. Phân lớp Bayes được dựa trên định lý Bayes (định lý được đặt theo tên tác giả của nó là Thomas Bayes)

**2.1.1 Định lý Bayes**



Công thức Bayes tổng quát



Trong đó ta gọi A là một chứng cứ (evidence) (trong bài toán phân lớp A sẽ là một phần tử dữ liệu), B là một giả thiết nào để cho A thuộc về một lớp C nào đó. Trong bài toán phân lớp chúng ta muốn xác định giá trị P(B/A) là xác suất để giả thiết B là đúng với chứng cứ A thuộc vào lớp C với điều kiện ra đã biết các thông tin mô tả A. P(B|A) là một xác suất hậu nghiệm (posterior probability hay posteriori probability) của B với điều kiện A.

Giả sử tập dữ liệu liệu khách hàng của chúng ta được mô tả bởi các thuộc tính tuổi và thu nhập, và một khách hàng X có tuổi là 25 và thu nhập là 2000$. Giả sử H là giả thiết khách hàng đõ sẽ mua máy tính, thì P(H|X) phản ánh xác xuất người dùng X sẽ mua máy tính với điều kiện ta biết tuổi và thu nhập của người đó.

Ngược lại P(H) là xác suất tiền nghiệm (prior probability hay priori probability) của H. Trong ví dụ trên, nó là xác suất một khách hàng sẽ mua máy tính mà không cần biết các thông tin về tuổi hay thu nhập của họ. Hay nói cách khác, xác suất này không phụ thuộc vào yếu tố X. Tương tự, P(X|H) là xác suất của X với điều kiện H (likelihood), nó là một xác suất hậu nghiệm. VÍ dụ, nó là xác suất người dùng X (có tuổi là 25 và thu nhập là $200) sẽ mua máy tính với điều kiện ta đã biết người đó sẽ mua máy tính. Cuối cùng P(X) là xác suất tiền nghiệm của X. Trong ví dụ trên, nó se là xác xuất một người trong tập dữ liệu sẽ có tuổi 25 và thu nhập $2000.

**Posterior = Likelihood \* Prior / Evidence**

**2.1.2 Phân lớp Naïve Baye**s

Bộ phân lớp Naive bayes hay bộ phân lớp Bayes (simple byes classifier) hoạt động như sau:

Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính A1, A2,...,An = {x1,x2,...,xn}

Giả sử có m lớp C1, C2,..,Cm. Cho một phần tử dữ liệu X, bộ phân lớp sẽ gán nhãn cho X là lớp có xác suất hậu nghiệm lớn nhất. Cụ thể, bộ phân lớp Bayes sẽ dự đoán X thuộc vào lớp Ci nếu và chỉ nếu:

P(Ci|X) > P(Cj|X) (1<= i, j <=m, i != j)

Giá trị này sẽ tính dựa trên định lý Bayes.

Để tìm xác suất lớn nhất, ta nhận thấy các giá trị P(X) là giống nhau với mọi lớp nên không cần tính. Do đó ta chỉ cần tìm giá trị lớn nhất của P(X|Ci) \* P(Ci). Chú ý rằng P(Ci) được ước lượng bằng |Di|/|D|, trong đó Di là tập các phần tử dữ liệu thuộc lớp Ci. Nếu xác suất tiền nghiệm P(Ci) cũng không xác định được thì ta coi chúng bằng nhau P(C1) = P(C2) = ... = P(Cm), khi đó ta chỉ cần tìm giá trị P(X|Ci) lớn nhất.

Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toàn P(X|Ci) là rất lớn, dó đó có thể giảm độ phức tạp của thuật toán Naive Bayes giả thiết các thuộc tính độc lập nhau. Khi đó ta có thể tính:

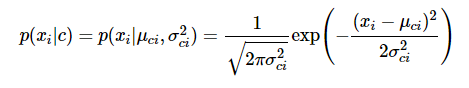
**P(X|Ci) = P(x1|Ci)...P(xn|Ci)**

**2.1.2.1 Các phân phối hay dùng**

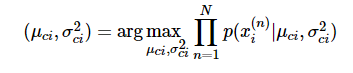
**2.1.2.1.1 Gaussian Naive Bayes**

Mô hình này được sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.

Với mỗi chiều dữ liệu ii và một class cc, xixi tuân theo một phân phối chuẩn có kỳ vọng μciμci và phương sai σ2ciσci2:



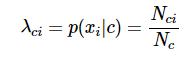
Trong đó, bộ tham số θ={μci,σ2ci}θ={μci,σci2} được xác định bằng Maximum Likelihood:



**2.1.2.1.2 Multinomial Naive Bayes**

Mô hình này chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng Bags of Words. Lúc này, mỗi văn bản được biểu diễn bởi một vector có độ dài dd chính là số từ trong từ điển. Giá trị của thành phần thứ ii trong mỗi vector chính là số lần từ thứ ii xuất hiện trong văn bản đó.

Khi đó, p(xi|c)p(xi|c) tỉ lệ với tần suất từ thứ ii (hay feature thứ ii cho trường hợp tổng quát) xuất hiện trong các văn bản của class cc. Giá trị này có thể được tính bằng cách:

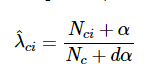


Trong đó:

NciNci là tổng số lần từ thứ ii xuất hiện trong các văn bản của class cc, nó được tính là tổng của tất cả các thành phần thứ ii của các feature vectors ứng với class cc.

NcNc là tổng số từ (kể cả lặp) xuất hiện trong class cc. Nói cách khác, nó bằng tổng độ dài của toàn bộ các văn bản thuộc vào class cc. Có thể suy ra rằng Nc=∑di=1NciNc=∑i=1dNci, từ đó ∑di=1λci=1∑i=1dλci=1.

Cách tính này có một hạn chế là nếu có một từ mới chưa bao giờ xuất hiện trong class cc thì biểu thức (10)(10) sẽ bằng 0, điều này dẫn đến vế phải của (7)(7) bằng 0 bất kể các giá trị còn lại có lớn thế nào. Việc này sẽ dẫn đến kết quả không chính xác (xem thêm ví dụ ở mục sau).

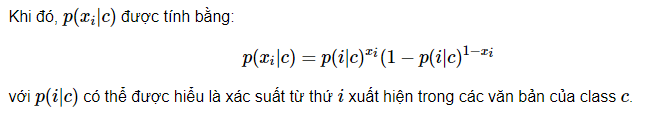


Với αα là một số dương, thường bằng 1, để tránh trường hợp tử số bằng 0. Mẫu số được cộng với dαdα để đảm bảo tổng xác suất ∑di=1^λci=1∑i=1dλ^ci=1.

Như vậy, mỗi class cc sẽ được mô tả bởi bộ các số dương có tổng bằng 1: ^λc={^λc1,…,^λcd}λ^c={λ^c1,…,λ^cd}.

**2.1.2.1.3 Bernoulli Naive Bayes**

Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bẳng 0 hoặc 1. Ví dụ: cũng với loại văn bản nhưng thay vì đếm tổng số lần xuất hiện của 1 từ trong văn bản, ta chỉ cần quan tâm từ đó có xuất hiện hay không.



**2.1.2.2 Ví dụ về phân lớp bayes**

Ví dụ 1:

Phân các bệnh nhân thành 2 lớp ung thư và không ung thư. Giả sử xác suất để một người bị ung thư là 0.008 tức là P(cancer) = 0.008; và P(nocancer) = 0.992. Xác suất để bệnh nhân ung thư có kết quả xét nghiệm dương tính là 0.98 và xác suất để bệnh nhân không ung thư có kết quả dương tính là 0.03 tức là P(+/cancer) = 0.98, P(+/nocancer) = 0.03. Bây giờ giả sử một bệnh nhân có kết quả xét nghiệm dương tính. Ta có:

P(+/canncer)P(cancer) = 0.98 \* 0.008 = 0.0078

P(+/nocancer)P(nocancer) = 0.03 \* 0.992 = 0.0298

Như vậy, P(+/nocancer)P(nocancer) >> P(+/cancer)P(cancer).

Do đó ta xét đoán rằng, bệnh nhân là không ung thư.

Ví dụ 2:

Cơ sở dữ liệu khách hàng:

| **ID** | **Tuổi** | **Thu nhập** | **Sính viên** | **Đánh giá tín dụng** | **Mua máy tính** |
| --- | --- | --- | --- | --- | --- |
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle | low | yes | excellent | yes |
| 8 | youth | medum | no | fair | yes |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle | medium | no | excellent | yes |
| 13 | middle | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

Giả sử ta có một khách hàng mới X có các thuộc tính

X = (age = youth, income = medium, student = yes, credit\_rating = fair)

Bây giớ cần xác định xem khách hàng X có thuộc lớp Cyes (mua máy tính) hay không, ta tính toán như sau:

P(Cyes) = 9/14 = 0.357

Các xác suất thành phần:

P(age = youth|Cyes) = 2/9 = 0.222

P(age = youth|Cno) = 3/5 = 0.6

P(income = medium|Cyes) = 4/9 = 0.444

P(income = medium|Cno) = 2/5 = 0.4

P(student = yes|Cyes) = 6/9 = 0.667

P(student = yes|Cno) = 1/5 = 0.2

P(credit\_rating = fair|Cyes) = 6/9 = 0.667

P(credit\_rating = fair|Cno) = 2/5 = 0.2

Cuối cùng:

P(X|Cyes) = 0.222 \* 0.444 \* 0.667 \* 0.667 = 0.044

P(X|Cno) = 0.60.4 \* 0.2 \* 0.4 = 0.019

P(X|Cyes)\*P(Cyes) = 0.044 \* 0.643

P(X|Cno)\*P(Cno) =0.019 \* 0.357 = 0.007

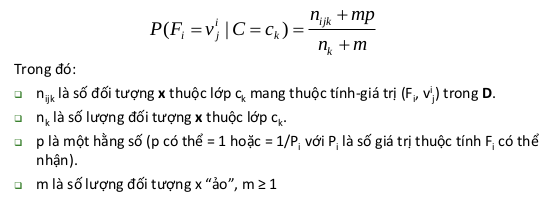
Từ kết quả này ta thấy P(X |Cyes)P(Cyes) có giá trị lớn nhất, do đó thuật toán Bayes sẽ kết luận là khách hàng X sẽ mua máy tính.

**2.1.3Khắc phục vấn đề xác suất điều kiện bằng zero**

Nếu trong dữ liệu huấn luyện không có đối tượng X nào có thuộc tính lớp Ck có thuộc tính Fi nhận một giá trị cụ thể vij, xác suất điều kiện P(Fi = xij | Ck) sẽ bằng 0.

Khi phân lớp, nếu có một đối tượng nào mang thuộc tính này thì xác suất phân vào lớp Ck luôn bằng 0.

Khắc phục bằng cách ước lượng theo công thức sau:



**2.1.4 Ưu điểm**

Giả định độc lập: hoạt động tốt cho nhiều bài toán/miền sữ liệu và ứng dụng.

Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán như phân lớp văn bản, lọc spam,..

Cho phép kết hợp tri thức tiền nghiệm (prior knowledge) và dữ liệu quan sát được (obserwed data).

Tốt khi có sự chệnh lệch số lượng giữa các lớp phân loại.

Huấn luyện mô hình (ước lượng tham số) dễ và nhanh.

**2.1.5 Nhược điểm**

Giả định độc lập (ưu điểm cũng chính là nhược điểm)

hầu hết các trường hợp thực tế trong đó có các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau.

Vấn đề zero (đã nêu cách giải quyết ở phía trên)

Mô hình không được huẩn luyện bằng phượng pháp tối ưu mạnh và chặt chẽ.

Tham số mủa mô hình là các ước lượng xác suất điều kiện đơn lẻ.

Không tính đến sự tương tác giữa các ước lượng này.

**2.1.6 Ứng dụng cụ thể**

\*Phân lớp văn bản (document classification)

Tham khảo cuốn sách sau:

Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008. (Free)

Chương 13. Text classification and Naïve Bayes

Tham khảo thêm:

*http://en.wikipedia.org/wiki/Document\_classification*

*http://en.wikipedia.org/wiki/Naive\_Bayes\_classifier*

\*Lọc spam (Spam filtering)

Tham khảo:

*Bayesian spam filtering*

*http://en.wikipedia.org/wiki/Bayesian\_spam\_filtering*

*http://en.wikipedia.org/wiki/Naive\_Bayes\_classifier*

*http://en.wikipedia.org/wiki/Email\_filtering*

**2.2 Phân loại 1 lớp**

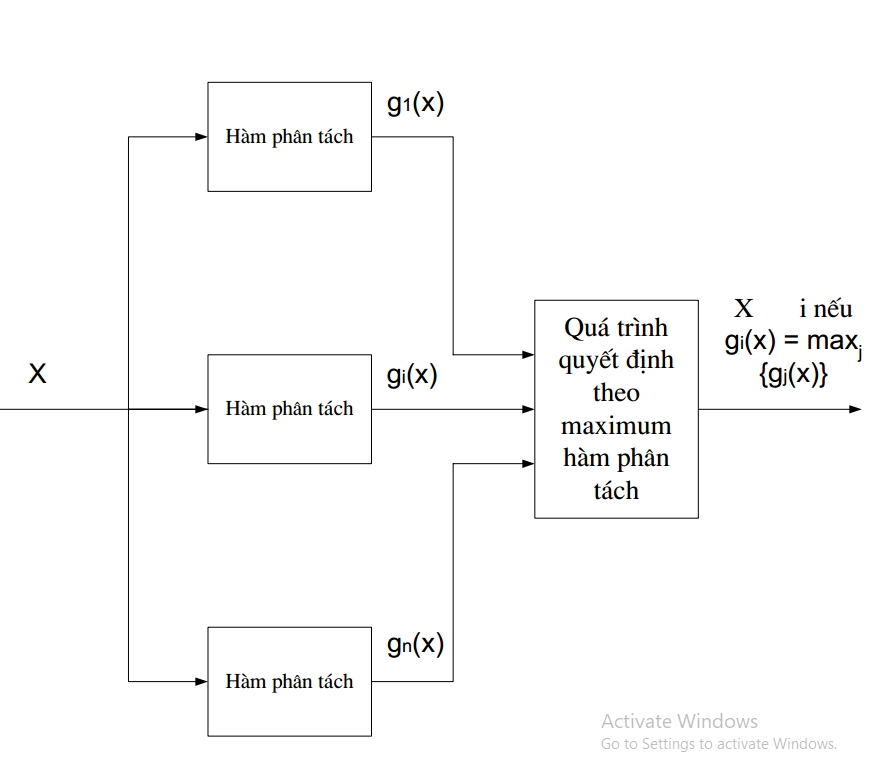
Biểu diễn bộ phân loại bằng hàm phân tách:

gi(x), i=1,2,...,n;

Bộ phân loại gán đối tượng x

vào phân lớp ωi nếu:

gi(x) > gj(x), đối với mọi j ≠ i;



Trường hợp tổng quát, với các rủi ro ta có thể chọn phân tách là rủi ro tương đối:

gi(x) = -R(αi|x)

Trong trường hợp với tỷ lệ sai số cưc tiểu, ta có thể

chọn hàm phân tách là xác suất hậu nghiệm:

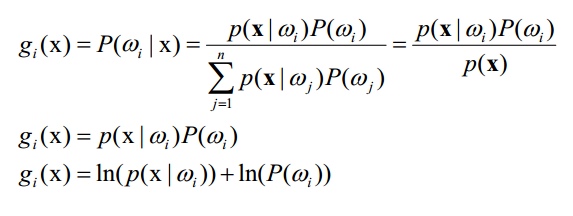
gi(x) = P(ωi|x)

Ta có thể thay thế hàm rủi ro gi(x) bằng f(gi(x)),

trong đó f(.) là hàm đơn điệu tăng thì kết quả không

thay đổi.

Một số dạng hàm phân tách thông dụng:



Tuy các hàm phân tách có thể có những dạng khác nhau,

nhưng luật quyết định vẫn tương đương nhau

Hiệu ứng của các luật quyết định là phân chia

không gian đặc điểm thành n vùng quyết định,

R

1, R2, …, Rn.

Nếu gi(x)>gj(x) ∀ j ≠ i thì x ∈ Ri và luật quyết

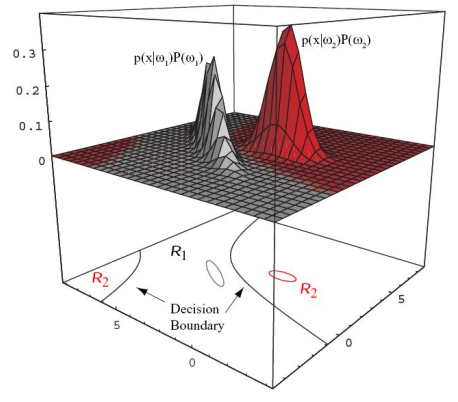
định sẽ gán x vào phân lớp ωi.

Các miền quyết định được phân tách nhau bằng

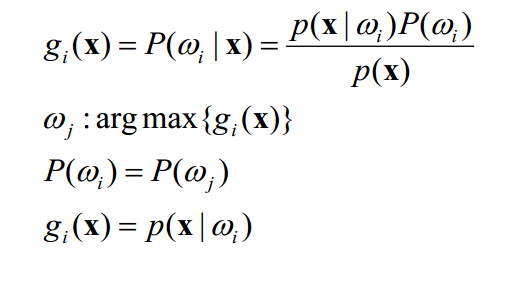
ranh giới quyết định

Miền ranh giới quyết định trong hai trường hợp phân loại hai lớp không gian đặc điểm hai chiều

Trong trường hợp này, các mật độ phân bố xác suất là phân bố Gaus. Miền danh giới có dạng hyperbol và miền quyết định R2 không liên thông đơn



Luật quyết định



**2.3 Phân tách hai lớp**

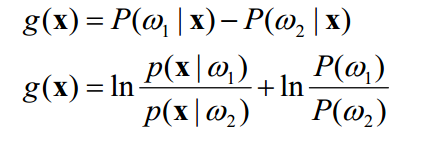
Thay cho việc sử dụng hai hàm phân tách g1(x) và g2(x) và

gán x vào phân lớp ω1 nếu g1(x) > g2(x), sử dụng hàm g(x)

= g1(x) – g2(x)

Luật quyết định: gán x vào phân lớp ω1 nếu g(x) > 0;

Một số dạng hàm phân tách



**Chương 3: Phân loại rượu dựa trên phân lớp naïve Bayes**

**3.1 Giới thiệu dữ liệu**

Rượu vang là một loại đồ uống có cồn được làm từ nho lên men. Men tiêu thụ đường trong nho và chuyển nó thành etanol, carbon dioxide và nhiệt. Nó là một loại đồ uống có cồn có hương vị dễ chịu, được yêu thích. Chắc chắn sẽ rất thú vị khi phân tích các thuộc tính hóa lý của rượu vang và hiểu mối quan hệ và ý nghĩa của chúng với chất lượng và phân loại rượu vang. Để thực hiện việc này, Chúng ta sẽ tiến hành theo các mô hình quy trình công việc Khai thác dữ liệu và Máy học:

Dự đoán chất lượng của từng mẫu rượu, có thể thấp, trung bình hoặc cao.

Bộ dữ liệu liên quan đến các biến thể màu đỏ và trắng của rượu vang "Vinho Verde". Vinho verde là một sản phẩm độc đáo từ vùng Minho (tây bắc) của Bồ Đào Nha. Có độ cồn trung bình, nó đặc biệt được đánh giá cao do sự tươi mát của nó (đặc biệt là vào mùa hè). Bộ dữ liệu này chỉ được cung cấp công khai cho mục đích nghiên cứu, để biết thêm thông tin, hãy đọc Cortez và cộng sự, 2009.. Do các vấn đề về quyền riêng tư và hậu cần, chỉ có các biến số hóa lý (đầu vào) và cảm quan (đầu ra) (ví dụ: không có dữ liệu về loại nho, nhãn hiệu rượu vang, giá bán rượu vang, v.v.).

Thông tin chung về dữ liệu:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 178 | **Area:** | Physical |
| **Attribute Characteristics:** | Integer, Real | **Number of Attributes:** | 13 | **Date Donated** | 1991-07-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 1534914 |

**3.2 Thông tin các thuộc tính**

Input variables (based on physicochemical tests): 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density 9 - pH 10 - sulphates 11 - alcohol Output variable (based on sensory data): 12 - quality (score between 0 and 10)

**Acidity**

Axit là một trong những đặc tính cơ bản của rượu vang và góp phần rất lớn vào hương vị của rượu, Tính axit trong thức ăn và đồ uống có vị chua và nồng. Việc nếm thử độ chua đôi khi cũng bị nhầm lẫn với rượu. Các loại rượu có nồng độ axit cao hơn có cảm giác nhẹ nhàng hơn vì chúng có dạng “cay”. Giảm axit một cách đáng kể có thể dẫn đến rượu vang có vị nhạt. Nếu bạn thích một loại rượu vang đậm đà và tròn vị hơn, bạn nên thưởng thức ít chua hơn một chút.

**volatile acidity**

Cảm giác ngứa ran tập trung ở mặt trước và mặt bên của lưỡi. Cảm giác như đá pop.

Nếu bạn chà lưỡi lên vòm miệng, bạn sẽ cảm thấy có sỏi.

Miệng của bạn có cảm giác ẩm ướt, giống như bạn đã cắn vào một quả táo.

Bạn cảm thấy như bạn có thể lấp lánh.

**Acid Types and Measures**

**fixed acidity**: Các axit cố định bao gồm axit tartaric, malic, citric và succinic được tìm thấy trong nho (ngoại trừ succinic). Biến này thường được biểu thị bằng $ \ frac {g (tartaricacid)} {dm ^ 3} $ trong tập dữ liệu.

**volatile acidity**: Các axit này phải được chưng cất ra khỏi rượu trước khi hoàn thành quá trình sản xuất. Nó chủ yếu được tạo thành từ axit axetic mặc dù các axit khác như axit lactic, axit fomic và axit butyric cũng có thể có mặt. Quá nhiều axit bay hơi là không mong muốn và dẫn đến hương vị khó chịu. Tại Hoa Kỳ, giới hạn pháp lý về độ axit bay hơi là 1,2 g / L đối với rượu vang đỏ và 1,1 g / L đối với rượu vang trắng. Độ axit dễ bay hơi được biểu thị bằng $ \ frac {g (aceticacid)} {dm ^ 3} $ trong tập dữ liệu.

**citric acid**: Đây là một trong những axit cố định mang lại sự tươi mát cho rượu vang. Thông thường hầu hết nó được tiêu thụ trong quá trình lên men và đôi khi nó được thêm riêng để làm cho rượu thêm tươi mới. Nó thường được biểu thị bằng $ \ frac {g} {dm ^ 3} $ trong tập dữ liệu.

**pH**: Còn được gọi là thế của hydro, đây là thang số để xác định độ axit hoặc tính bazơ của rượu. Độ axit cố định đóng góp nhiều nhất vào độ pH của rượu vang. Bạn có thể biết, các dung dịch có độ pH nhỏ hơn 7 có tính axit, trong khi các dung dịch có độ pH lớn hơn 7 là bazơ. Với độ pH bằng 7, nước tinh khiết là trung tính. Hầu hết các loại rượu có độ pH từ 2,9 đến 3,9 và do đó có tính axit.

**Sweetness**

Rượu ngọt hay khô (không ngọt) như thế nào? Nhận thức của con người về vị ngọt bắt đầu từ đầu lưỡi. Thông thường, ấn tượng đầu tiên về một loại rượu là mức độ ngọt ngào của nó. Để thưởng thức vị ngọt, hãy tập trung sự chú ý vào vị giác trên đầu lưỡi. Vị giác của bạn có ngứa ngáy không? –Một chỉ số về độ ngọt. Bạn có tin hay không, nhiều loại rượu vang khô có thể có một chút vị ngọt để khiến chúng trở nên sung mãn hơn.

**Nếm vị ngọt như thế nào?**

Cảm giác ngứa ran trên đầu lưỡi.

Cảm giác hơi nhờn ở giữa lưỡi kéo dài.

Rượu có độ nhớt cao hơn; rượu từ từ rơi lệ trên mặt ly. (cũng là một chỉ báo về ABV cao)

Các loại rượu vang đỏ khô như cabernet sauvignon thường có lượng đường dư lên đến 0,9 g / L (phổ biến với các loại rượu rẻ tiền).

Rượu khô xương thường có thể bị nhầm lẫn với một loại rượu có tanin cao.

**Sweetness Measure**:

**Residual sugar**: Điều này thường đề cập đến đường tự nhiên từ nho còn lại sau khi quá trình lên men ngừng hoặc bị dừng. Nó thường được biểu thị bằng $ \ frac {g} {dm ^ 3} $ trong tập dữ liệu.

**Salty**

Độ mặn không phải là cách mô tả rượu vang thông thường. Đó cũng không phải là điều tích cực có lẽ không cần phải nói. Nhưng thực tế là các quốc gia sản xuất rượu vang có (rất khác nhau) mức tối đa hợp pháp cho natri clorua trong rượu vang sẽ cho bạn biết điều gì đó. Độ mặn là một mối quan tâm ở các địa điểm khô hạn khi việc tưới tiêu thường xuyên làm tăng độ mặn của đất, làm tăng độ mặn của rượu vang. Thành phần đất thường không chuyển dịch theo cách bạn mong đợi thành thành phần nho; muối, thật không may, là một ngoại lệ.

**Salty Measure**:

**Chlorides**: Nồng độ clorua trong rượu vang bị ảnh hưởng bởi khủng bố và nồng độ clorua cao nhất được tìm thấy trong các loại rượu vang đến từ các quốc gia có hệ thống tưới tiêu bằng nước mặn hoặc ở những vùng có địa hình lợ. Đây thường là yếu tố góp phần chính tạo nên độ mặn trong rượu vang. Nó thường được biểu thị bằng $ \ frac {g (sodiumchloride)} {dm ^ 3} $ trong tập dữ liệu.



**Sulfites**

Sulfite trong rượu vang là các hợp chất hóa học (sulfur dioxide, hoặc SO2) xuất hiện tự nhiên, ở mức độ khác nhau, trong tất cả các loại rượu. Sulfur Dioxide được tìm thấy tự nhiên trong rượu vang và là một sản phẩm phụ của quá trình lên men, nhưng hầu hết các nhà sản xuất rượu vang chọn thêm một ít để ngăn chặn sự phát triển của các loại nấm men và vi khuẩn không mong muốn, cũng như bảo vệ chống lại quá trình oxy hóa.

Sulfur dioxide ức chế nấm men, ngăn không cho rượu vang ngọt tham khảo trong chai. Nó là một chất chống oxy hóa, giữ cho rượu vang tươi và không bị ôxy hóa.

Các nền văn hóa cổ đại ở Hy Lạp, La Mã và Ai Cập, đã sử dụng sulfit để khử trùng đồ đựng rượu của họ. Bởi vì sulfit là chất chống vi khuẩn, nó có khả năng tiêu diệt vi khuẩn không mong muốn và nấm men hoang dã trong quá trình nấu rượu.

Những người nếm rất nhạy cảm đã được ghi nhận là ngửi thấy các hợp chất lưu huỳnh trong rượu vang, mặc dù các hợp chất lưu huỳnh phần nào không liên quan đến sulfit. Các hợp chất lưu huỳnh trong rượu vang được gọi là thiols có nhiều loại hương vị từ mùi giống cam quýt đến mùi giống như trứng nấu chín.

Điều thú vị là rượu càng ấm thì lượng lưu huỳnh phân tử càng nhiều. Đây là lý do tại sao một số loại rượu có mùi thơm khó chịu từ trứng nấu chín khi bạn mở chúng ra. Bạn có thể khắc phục vấn đề này bằng cách gạn rượu và ướp lạnh trong khoảng 15-30 phút.

**Cần bao nhiêu sulfur trong rượu?**

Rượu vang có độ axit thấp hơn cần nhiều sulfit hơn rượu vang có độ axit cao hơn. Ở pH từ 3,6 trở lên, rượu vang kém ổn định hơn nhiều và sulfit cần thiết cho thời hạn sử dụng.

Rượu vang có nhiều màu hơn (tức là rượu vang đỏ) có xu hướng cần ít sulfit hơn rượu vang trong (tức là rượu vang trắng). Một loại rượu vang trắng khô điển hình có thể có khoảng 100 mg / L trong khi một loại rượu vang đỏ khô điển hình sẽ có khoảng 50–75 mg / L.

Rượu có hàm lượng đường cao hơn có xu hướng cần nhiều sulfit hơn để ngăn chặn quá trình lên men phụ của phần đường còn lại.

Rượu vang ở nhiệt độ ấm hơn sẽ giải phóng các hợp chất lưu huỳnh tự do (mùi lưu huỳnh khó chịu) và có thể “cố định” đơn giản bằng cách gạn và làm lạnh rượu.

**Sulfites Measure**:

**sulphates**: Là những muối khoáng có chứa lưu huỳnh. Sulphates đối với rượu vang như gluten đối với thực phẩm. Chúng là một phần thường xuyên của việc sản xuất rượu vang trên khắp thế giới và được coi là thiết yếu. Chúng được kết nối với quá trình lên men và ảnh hưởng đến hương thơm và hương vị của rượu. Ở đây, nó được biểu thị bằng $ \ frac {g (potassiumsulphate)} {dm ^ 3} $ trong tập dữ liệu.

**free sulfur dioxide**: Đây là phần lưu huỳnh điôxít khi được thêm vào rượu vang được cho là tự do sau khi phần còn lại liên kết. Các nhà sản xuất rượu sẽ luôn cố gắng để có được tỷ lệ lưu huỳnh tự do cao nhất để kết dính. Chúng còn được gọi là sulfit và quá nhiều là không mong muốn và tạo ra mùi hăng. Biến này được biểu thị bằng $ \ frac {mg} {dm ^ 3} $ trong tập dữ liệu.

**total sulfur dioxide**: Đây là tổng số của liên kết và điôxít lưu huỳnh tự do ($ SO\_2 $). Ở đây, nó được biểu thị bằng $ \ frac {mg} {dm ^ 3} $. Điều này chủ yếu được thêm vào để tiêu diệt vi khuẩn có hại và bảo quản chất lượng và độ tươi. Thường có giới hạn pháp lý đối với mức lưu huỳnh trong rượu vang và vượt quá mức này thậm chí có thể giết chết men tốt và tạo ra mùi không mong muốn.

**Alcohol**

Rượu được hình thành do nấm men chuyển hóa đường trong quá trình lên men. Tỷ lệ phần trăm của rượu có thể thay đổi từ rượu vang. Chúng tôi giải thích rượu bằng nhiều thụ thể vị giác khác nhau, đó là lý do tại sao nó có thể có vị đắng, ngọt, cay và dầu cùng một lúc. Di truyền của bạn thực sự đóng một vai trò trong việc rượu có vị đắng hay ngọt. Dù sao đi nữa, tất cả chúng ta đều có thể cảm nhận được rượu từ phía sau miệng trong cổ họng như một cảm giác nóng lên.

Đặc điểm rượu:

Rượu vang có độ cồn cao hơn có xu hướng có vị đậm hơn và nhiều dầu hơn

Rượu vang có độ cồn thấp hơn có xu hướng có vị nhẹ hơn

**Alcohol Measure**:

Rượu: Nó thường được đo bằng% thể tích hoặc rượu theo thể tích (ABV).

**density**: Điều này có thể được biểu diễn bằng cách so sánh trọng lượng của một thể tích rượu cụ thể với một thể tích nước tương đương. Nó thường được sử dụng như một thước đo để chuyển đổi đường thành rượu. Ở đây, nó được biểu thị bằng $ \ frac {g} {cm ^ 3} $.

Lớp phân loại:

**winetype**: Vì ban đầu chúng tôi có hai tập dữ liệu cho rượu vang đỏ và trắng, chúng tôi đã giới thiệu thuộc tính này trong tập dữ liệu được hợp nhất cuối cùng cho biết loại rượu vang cho mỗi điểm dữ liệu. Rượu có thể là rượu 'đỏ' hoặc 'trắng'. Một trong những mô hình dự đoán mà chúng tôi sẽ xây dựng trong chương này sẽ là để chúng tôi có thể dự đoán loại rượu bằng cách xem xét các thuộc tính rượu khác.

**quality**: Các chuyên gia rượu đã xếp loại chất lượng rượu từ 0 (rất tệ) đến 10 (rất xuất sắc). Điểm chất lượng cuối cùng là điểm trung bình của ít nhất ba đánh giá được thực hiện bởi cùng một chuyên gia rượu vang.

**Quality\_label**: Đây là thuộc tính xuất phát từ thuộc tính chất lượng. Chúng tôi xếp hoặc nhóm điểm chất lượng rượu thành ba nhóm định tính cụ thể là thấp, trung bình và cao. Rượu có điểm chất lượng 3, 4 & 5 là chất lượng thấp, điểm 6 & 7 là chất lượng trung bình và điểm 8 & 9 là rượu chất lượng cao. Chúng tôi cũng sẽ xây dựng một mô hình khác trong chương này để dự đoán nhãn chất lượng rượu này dựa trên các thuộc tính rượu khác.

**3.3 Quy trình xử lý**

Trong phần này, phân tích làm và làm sạch nhóm mình sử dụng R program language và xây dựng bộ phân lớp bằng Python program language.

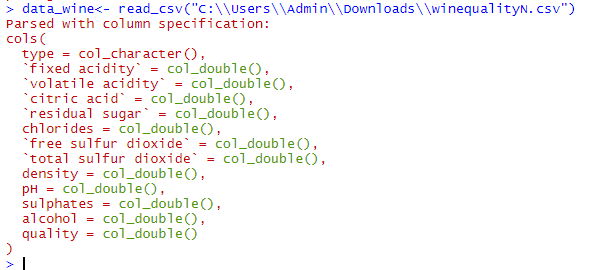
**3.3.1 Thu thập dữ liệu:**

***Dữ liệu được lấy từ: https://archive.ics.uci.edu/ml/datasets/wine***

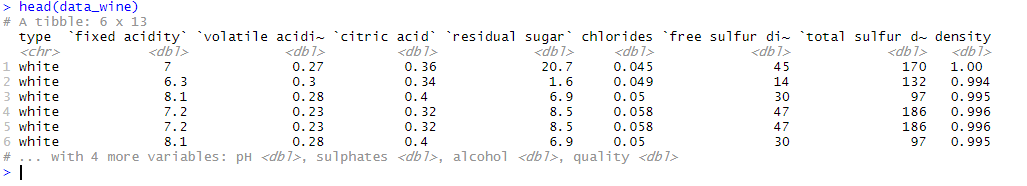
Đầu tiên mình import thư viện cần thiết:



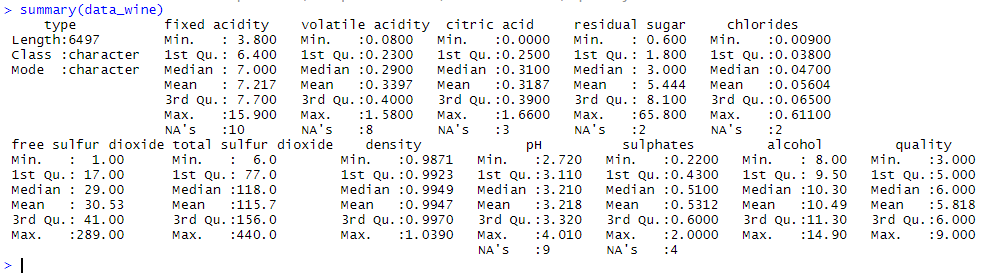
Sau đó mình import dữ liệu:



Phần đầu của dữ liệu :



Mô tả dữ liệu :

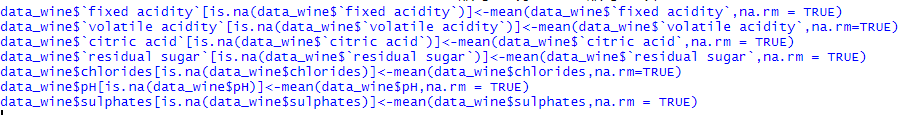


Nhận xét : dữ liệu còn chứa các NA, dữ liệu không chuẩn để xây dựng bộ phân lớp. Bước tiếp theo chúng ta sẽ là tiền sử lý dữ liệu.

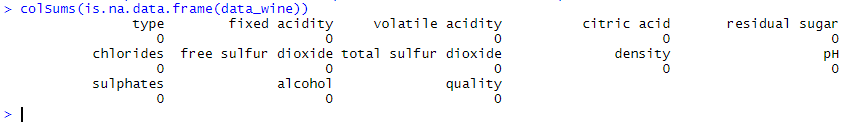
**3.3.2 Tiền xử lý dữ liệu :**

Như ở phần trên :

Bây giờ mình sẽ xử lý các thuộc tính có giá trị NA bằng cách thay chúng bằng tính trung bình :



Check lại :



Như vậy xử lý NA đã xong !

Tuy nhiên thuộc tính quality chúng ta muốn phân loại chúng lại có quá nhiều lớp vì thang điểm đáng giá từ 0->10 : Vậy có 11 nhãn

Bây giờ nhóm mình sẽ thu nhỏ số nhãn lại còn 3 nhãn : +số điểm từ 0->5 là loại rượu thấp

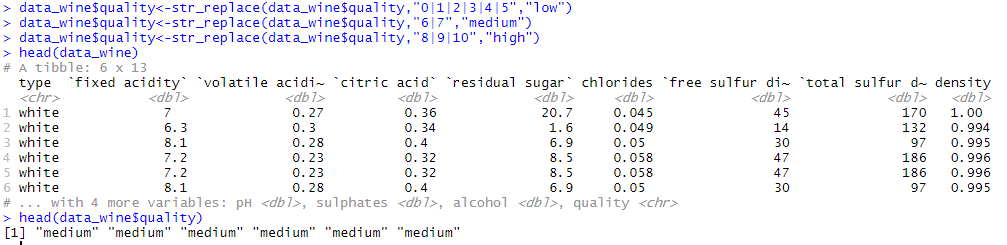
+số điểm 6,7 là loại khá

+số điểm >=8 là loại rượu tốt

Dầu tiên ép kiểu dữ liệu quality thành kiểu : character

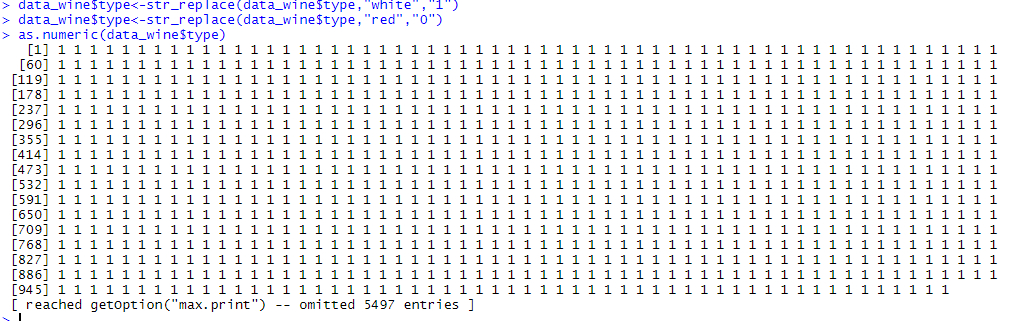


Sau đó thực hiện theo đúng kế hoạch :



Bước tiếp theo chúng ta sẽ thay đổi nhãn cho thuộc tính type loạt rượu trắng là 1, đỏ là 0,

Cách làm tương tự với thay đổi nhãn của thuộc tính quality

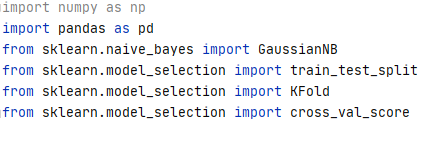


Cuối cùng import dữ liệu ra file .csv để cho bước tiếp theo xây dựng bộ phân lớp :



**3.3.3 Xây dựng bộ phân lớp Naive Bayes và đánh giá bộ phân lớp**

B1, import các thư viện cần thiết :



Thư viện numpy dùng cho việc xử lý tính toán ma trận va mang

Thư viện pandas cho phép xử lý và phân tích dữ liệu

Thư viện GausianNB với phân lớp naive\_bayes là bộ phân lớp chúng ta muốn xây dựng

Thư viện train\_test\_split mục đích để chia tập dữ liệu

Thư viện KFold dùng đề kiểm định bộ phân lớp bằng phương pháp cross\_val\_score

Thư viện cross\_val\_score là kết quả cần dự đoán

B2, chúng ta import dữ liệu đã được xử lý ở trên



B3, chúng ta xáo trộn dữ liệu bởi vì nó tuần tự theo nhãn



B4, chuyển đổi kiểu dữ liệu từ dataframe thành matrix



B5, phân chia các thành phần để học



Nhãn ở đây cần dự đoán là quality là cột Y còn các thuộc tính còn lại thuộc X

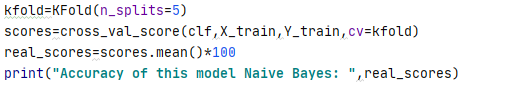
B6,Phân chia dữ liệu thành 2 phần, một phần là test set dùng để kiểm tra bộ phân lớp và phần còn lại là train set dùng để huấn luyện bộ phân lớp



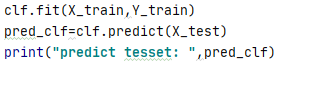
B7,Khởi tạo bộ phân lớp Naive Bayes



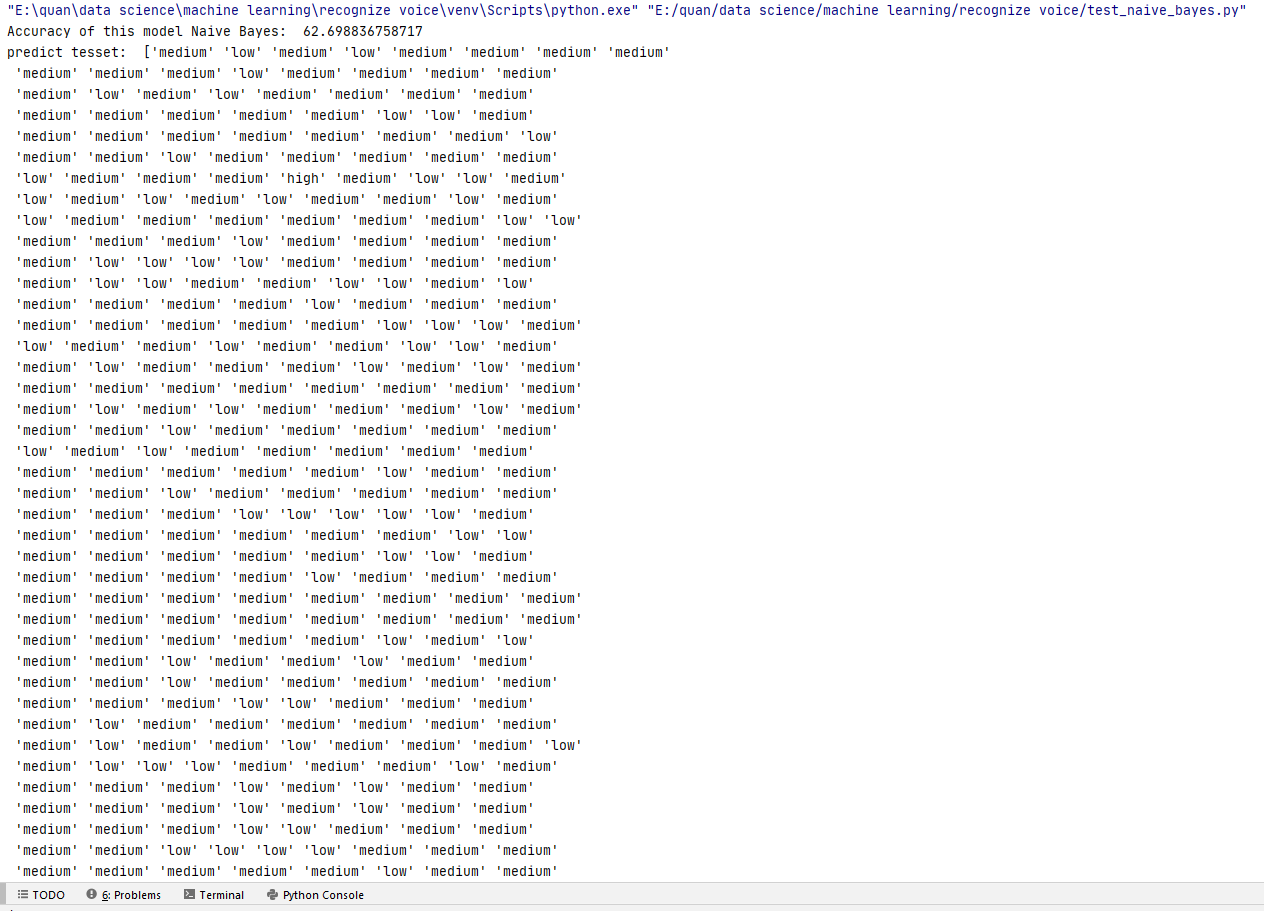
B8, Kiểm tra độ chính xác trong quá trình học bằng phương pháp cross\_validation



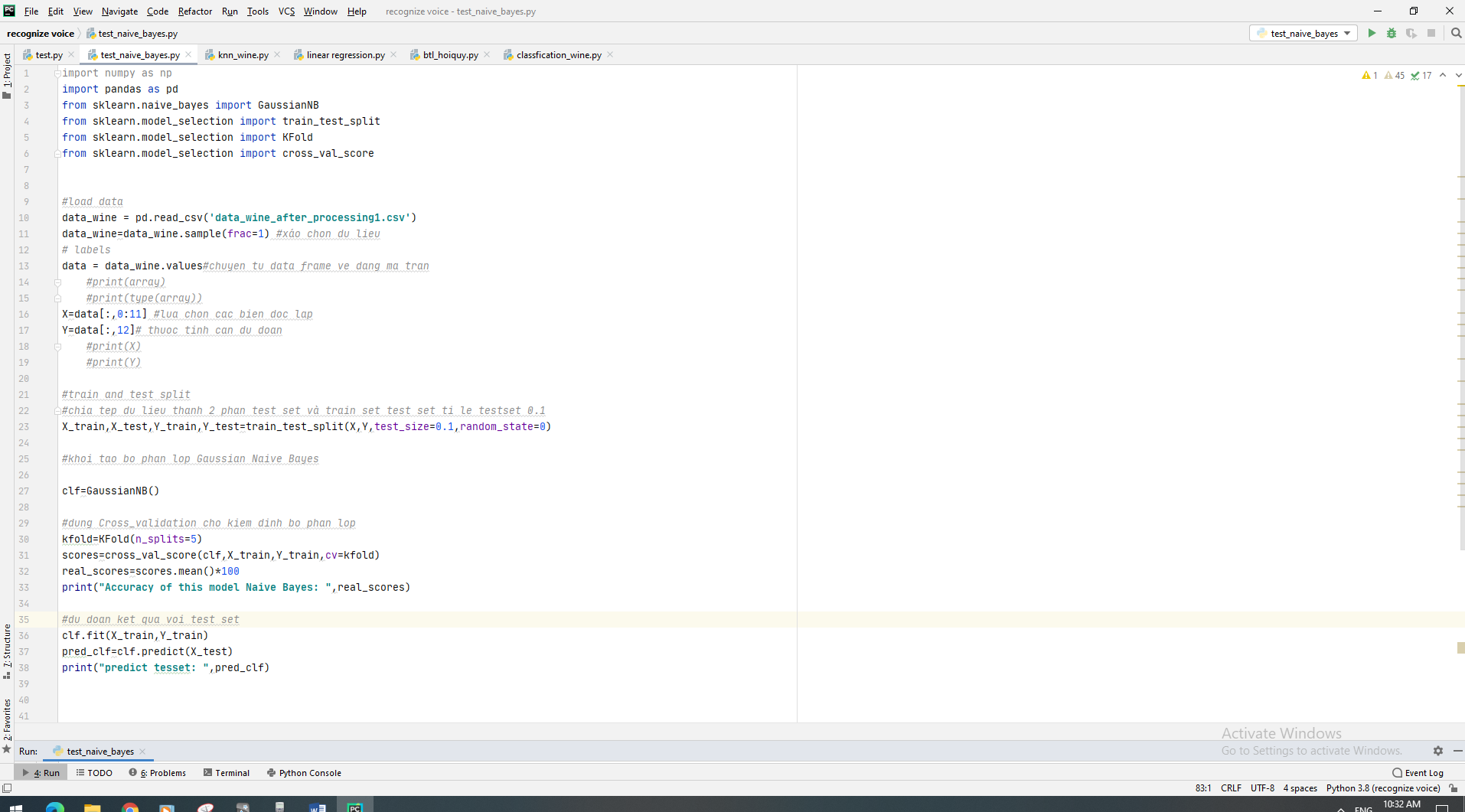
B9, Dự đoán nhãn



Kết quả thu được :



Code đầy đủ của chương trình :



**Kết luận:**

Thông qua cách thí nghiệm nho nhỏ chúng ta đã nhận dạng được chất lượng loại rượu thông qua bộ phân lớp Naive Bayes

**Tài liệu tham khảo:**

*https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn\_d%E1%BA%A1ng\_m%E1%BA%ABu?fbclid=IwAR028kqKKXum1CLqw7VmasWc\_YEuCI5fPmxUicggzilCcXLihjX3Df8RC5I*

*https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM?fbclid=IwAR1Qb2JCICH6rW559kjqK01o575baKe7Vp5Q0Mqvlj\_NnrWONrIK586yhkk*

*Pattern Recognition - Nguyen Linh Giang - 2008*