**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT**

**KHOA CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN MÔN HỌC**

**GIẢI THUẬT NAÏVE BAYES VÀ ỨNG DỤNG**

Giáo viên hướng dẫn: Quách Đình Hoàng

Sinh viên thực hiện : Võ Nhựt Kha

MSSV: 15110061

Bùi Minh Tiên

MSSV: 15110139

**Thành phố Hồ Chí Minh, tháng 12, năm 2017**

**Mục lục**

[**PHẦN I: GIỚI THIỆU** 2](#_Toc502836277)

[**I.1 Tổng quan** 2](#_Toc502836278)

[**I.2 Mục tiêu đề tài** 4](#_Toc502836279)

[**CHƯƠNG II: NỘI DUNG** 5](#_Toc502836280)

[**II.1. Phát biểu định lý Bayes** 5](#_Toc502836281)

[**II.2. Giới thiệu về bài toán phân lớp Naïve Bayes** 5](#_Toc502836282)

[**II.3. Các mô hình xác suất Naive Bayes** 6](#_Toc502836283)

[**II.4. Ước lượng tham số** 8](#_Toc502836284)

[**II.5. Xây dựng một classifier từ mô hình xác suất** 9](#_Toc502836285)

[**II.6. Thuật toán bayes** 10](#_Toc502836286)

[**II.7. Ví dụ bài toán Naïve Bayes** 11](#_Toc502836287)

[**II.8. Thực hiện thuật toán Naïve Bayes trên C#** 12](#_Toc502836288)

[**II.9. Bảng danh mục các lớp được sử dụng trong chương trình** 14](#_Toc502836289)

[**II.10. Bảng miêu tả phương thức trong các lớp** 15](#_Toc502836290)

[**II.11. Bảng phân công công việc** 17](#_Toc502836291)

[**CHƯƠNG III: PHẦN KẾT** 18](#_Toc502836292)

[**PHẦN IV: TÀI LIỆU THAM KHẢO** 19](#_Toc502836293)

# **PHẦN I: GIỚI THIỆU**

## **I.1 Tổng quan**

Khoa học thống kê đóng một vai trò cực kỳ quan trọng, một vai trò không thể thiếu được trong bất cứ công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm như y khoa, sinh học, nông nghiệp, hóa học, và ngay cả xã hội học. Thí nghiệm dựa vào các phương pháp thống kê học có thể cung cấp cho khoa học những câu trả lời khách quan nhất cho những vấn đề khó khăn nhất.

Khoa học thống kê là khoa học về thu thập, phân tích, diễn giải và trình bày các dữ liệu để từ đó tìm ra bản chất và tính quy luật của các hiện tượng kinh tế, xã hội - tự nhiên. Khoa học thống kê dựa vào lý thuyết thống kê, một loại toán học ứng dụng. Trong lý thuyết thống kê, tính chất ngẫu nhiên và sự không chắc chắn có thể làm mô hình dựa vào lý thuyết xác suất. Vì mục đích của khoa học thống kê là để tạo ra thông tin "đúng nhất" theo dữ liệu có sẵn, có nhiều học giả nhìn khoa thống kê như một loại lý thuyết quyết định.

Thống kê là một trong những công cụ quản lý vĩ mô quan trọng, cung cấp các thông tin thống kê trung thực, khách quan, chính xác, đầy đủ, kịp thời trong việc đánh giá, dự báo tình hình, hoạch định chiến lược, chính sách, xây dựng kế hoạch phát triển kinh tế - xã hội và đáp ứng nhu cầu thông tin thống kê của các tổ chức, cá nhân. Trong số những vai trò quan trọng thì dự báo tình hình là một trong những vai trò mang nhiều ý nghĩa, nó có cả một quá trình huấn luyện bên trong và có tính xử lý tự động khi đã được huấn luyện. Hay nói khác hơn là khi đã có tri thức lấy từ các dữ liệu thống kê hay kinh nghiệm của người dùng kết hợp với một phương pháp học (huấn luyện) dựa trên lý thuyết thống kê ta sẽ có được một cỗ máy có tri thức để tự nó có thể đưa ra được những quyết định với độ chính xác khá cao.

Phân tích thống kê là một khâu quan trọng không thể thiếu được trong các công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm. Một công trình nghiên cứu khoa học, cho dù có tốn kém và quan trọng cỡ nào, nếu không được phân tích đúng phương pháp sẽ không bao giờ có cơ hội được xuất hiện trong các tập san khoa học. Ngày nay, chỉ cần nhìn qua tất cả các tập san nghiên cứu khoa học trên thế giới, hầu như bất cứ bài báo y học nào cũng có phần “Statistical Analysis” (Phân tích thống kê), nơi mà tác giả phải mô tả cẩn thận phương pháp phân tích, tính toán như thế nào, và giải thích ngắn gọn tại sao sử dụng những phương pháp đó để hàm ý “bảo kê” hay tăng trọng lượng khoa học cho những phát biểu trong bài báo. Các tập san y học có uy tín càng cao yêu cầu về phân tích thống kê càng nặng. Không có phần phân tích thống kê, bài báo không thể xem là một “bài báo khoa học”. Không có phân tích thống kê, công trình nghiên cứu chưa được xem là hoàn tất.

Trong khoa học thống kê, có hai trường phái “cạnh tranh” song song với nhau, đó là trường phái tần số (frequentist school) và trường phái Bayes (Bayesian school). Phần lớn các phương pháp thống kê đang sử dụng ngày nay được phát triển từ trường phái tần số, nhưng hiện nay, trường phái Bayes đang trên đà “chinh phục” khoa học bằng một suy nghĩ “mới” về khoa học và suy luận khoa học. Phương pháp thống kê thuộc trường phái tần số thường đơn giản hơn các phương pháp thuộc trường phái Bayes. Có người từng ví von rằng những ai làm thống kê theo trường phái Bayes là những thiên tài!

Để hiểu sự khác biệt cơ bản giữa hai trường phái này, có lẽ cần phải nói đôi qua vài dòng về triết lý khoa học thống kê bằng một ví dụ về nghiên cứu y khoa. Để biết hai thuật điều trị có hiệu quả giống nhau hay không, nhà nghiên cứu phải thu thập dữ liệu trong hai nhóm bệnh nhân (một nhóm được điều trị bằng phương pháp A, và một nhóm được điều trị bằng phương pháp B). Trường phái tần số đặt câu hỏi rằng “nếu hai thuật điều trị có hiệu quả như nhau, xác suất mà dữ liệu quan sát là bao nhiêu”, nhưng trường phái Bayes hỏi khác: “Với dữ liệu quan sát được, xác suất mà thuật điều trị A có hiệu quả cao hơn thuật điều trị B là bao nhiêu”. Tuy hai cách hỏi thoạt đầu mới đọc qua thì chẳng có gì khác nhau, nhưng suy nghĩ kỹ chúng ta sẽ thấy đó là sự khác biệt mang tính triết lý khoa học và ý nghĩa của nó rất quan trọng. Đối với người bác sĩ (hay nhà khoa học nói chung), suy luận theo trường phái Bayes là rất tự nhiên, rất hợp với thực tế. Trong y khoa lâm sàng, người bác sĩ phải sử dụng kết quả xét nghiệm để phán đoán bệnh nhân mắc hay không mắc ung thư (cũng giống như trong nghiên cứu khoa học, chúng ta phải sử dụng số liệu để suy luận về khả năng của một giả thiết).

## **I.2 Mục tiêu đề tài**

Mục tiêu được đặt ra cho đề tài là nghiên cứu một số thuật toán cơ bản để xây dưng chương trình phân tích đánh giá dữ liệu phần nâng cao hiệu quả trong đời sống. Nhăm giúp con người dễ dàng phân loại nhiều tập dữ liệu với số lượng dữ liệu nhiều, mà con người tính sẽ rất nhiều có thể vài ngày, vài tháng hay vài năm cũng không ra kết quả như mong muốn.

Với các mục tiêu đó, đề tài sẽ thực hiện những công việc sau :

* + Tìm hiểu định lý Bayes
  + Giới thiệu về bài toán phân lớp Naïve Bayes
  + Tìm hiểu các mô hình xác suất Naïve Bayes
  + Tìm hiểu ước lượng tham số
  + Xây dựng một classifier từ mô hình xác suất
  + Tìm hiểu Thuật toán Bayes
  + Ví dụ bài toán Naïve Bayes
  + Thuật hiện thuật toán trên Naïve Bayes

# **CHƯƠNG II: NỘI DUNG**

## **II.1. Phát biểu định lý Bayes**

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B), và đọc là "xác suất của A nếu có B". Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

Theo định lí Bayes, xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:

* Xác suất xảy ra A của riêng nó, không quan tâm đến B. Kí hiệu là P(A) và đọc là xác suất của A. Đây được gọi là xác suất biên duyên hay xác suất tiên nghiệm, nó là "tiên nghiệm" theo nghĩa rằng nó không quan tâm đến bất kỳ thông tin nào về B.
* Xác suất xảy ra B của riêng nó, không quan tâm đến A. Kí hiệu là P(B) và đọc là "xác suất của B". Đại lượng này còn gọi là hằng số chuẩn hóa (normalising constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.
* Xác suất xảy ra B khi biết A xảy ra. Kí hiệu là P(B|A) và đọc là "xác suất của B nếu có A". Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra. Chú ý không nhầm lẫn giữa khả năng xảy ra A khi biết B và xác suất xảy ra A khi biết B.

Khi biết ba đại lượng này, xác suất của A khi biết B cho bởi công thức:

Từ đó dẫn tới

## **II.2. Giới thiệu về bài toán phân lớp Naïve Bayes**

Naive Bayes classifier là một thuật ngữ trong xử lý số liệu thống kê Bayesian với một phân lớp xác suất dựa trên các ứng dụng định lý Bayes với giả định độc lập bền vững. Một thuật ngữ mô tả chi tiết cho những mô hình xác suất sẽ là “mô hình đặc trưng không phụ thuộc”.

Trong thuật ngữ đơn giản, một naive Bayes classifier giả định rằng sự có mặt (hay không có mặt) của một đặc trưng của một lớp học là không liên quan đến sự hiện diện (hay thiếu vắng) của bất kỳ các đặc trưng. Ví dụ, một trái cây có thể được coi là một quả táo nếu nó có màu đỏ chung quanh, và đường kính khoảng 4 inch. Mặc dù các đặc trưng này phụ thuộc vào sự tồn tại của các đặc trưng khác, naive Bayes classifier xem xét tất cả các đặc tính độc lập góp phần vào khả năng trái cây này là quả táo.

Tùy thuộc vào tính chính xác bản chất của mô hình xác suất, naive Bayes classifiers có thể được đào tạo rất hiệu quả trong một thiết lập học có giám sát. Trong nhiều ứng dụng thực tế, tham số ước lượng cho các mô hình naive Bayes sử dụng các phương pháp maximum likehood; nói cách khác, có thể làm việc với các mô hình naive Bayes mà không tin ở xác suất Bayesian hoặc bằng cách sử dụng bất cứ phương pháp Bayesian.

Mặc dù thiết kế ngây thơ và hình như giả định đơn giản hơn, naive Bayes classifiers thường làm việc trong nhiều tình huống thế giới thực phức tạp tốt hơn có thể mong đợi. Mới đây, xem xét vấn đề phân lớp Bayesian đã có thể thấy có một số lý thuyết giải thích cho tính hiệu quả của naive Bayes classifiers. Một lợi thế của naive Bayes classifier là nó đòi hỏi một số lượng nhỏ dữ liệu đào tạo để ước lượng các tham số (các nghĩa và sự khác nhau của các biến) cần thiết cho việc phân loại. Bởi vì các biến được giả định độc lập, chỉ những khác biệt của các biến cho mỗi lớp học cần phải được xác định và không phải toàn bộ ma trận thống kê.

## **II.3. Các mô hình xác suất Naive Bayes**

Tóm lại, các mô hình xác suất cho một classifier là một mô hình có điều kiện đối với một biến lớp phụ thuộc C với một số lượng nhỏ của các kết quả hay các lớp học, phụ thuộc vài biến đặc trưng F1 cho tới Fn.

Vấn đề là nếu số các đặc trưng n là lớn hay khi một đặc trưng có thể chiếm một số lượng lớn các giá trị, sau đó dựa vào một mô hình trên các bảng xác suất là không thể làm được. Do vậy, chúng ta công thức hóa lại các mô hình để dễ xử lý.

Bằng cách sử dụng định lý Bayes, có được:

Trong thực hành, chỉ cần quan tâm tới tử số của phân số, khi mà mẫu số không phụ thuộc vào C và các giá trị của các đặc trưng của Fi đã cho, nên mẫu số là hằng thực sự.

Từ số tương đương với mô hình xác suất có thể được viết lại như sau, sử dụng định nghĩa của xác suất có điều kiện:

Bây giờ giả định "naive" giả định có điều kiện độc lập đưa vào: giả định rằng mỗi đặc trưng Fi có điều kiện độc lập với tất cả các đặc trưng Fj cho j # i. Điều này có nghĩa là

Do đó có thể được thực hiện như:

Điều này có nghĩa là dưới sự độc lập giả định ở trên, các điều kiện phân phối trên các lớp học biến C có thể được thể hiện:

Ở đây Z là một nhân tố xác định tỷ xích phụ thuộc vào F1, F2, .., Fn, chẳng hạn một hằng số nếu các giá trị của các biến đặc trưng đều được biết.

Nếu có k lớp học và nếu một mô hình cho P(Fi) có thể được thể hiện trong các thuật ngữ của r tham số, sau đó các mô hình naive Bayes tương ứng có (k - 1) + nrk tham số. Trong thực tế, thường k = 2 (phân loại nhị phân) và r = 1 (các biến Bernoulli như là các đặc trưng) được phổ biến, và như vậy tổng số lượng các tham số của mô hình naive Bayes là 2n + 1, ở đây n là số các đặc trưng nhị phân sử dụng cho các dự đoán.

## **II.4. Ước lượng tham số**

Tất cả các tham số mô hình (tức là, lớp học ưu tiên và các đặc trưng phân phối xác suất) có thể được gần đúng với các tần số liên quan từ việc thiết lập đào tạo. Đây là các đánh giá maximum likehood khả năng có thể xảy ra. Các đặc trưng không riêng biệt cần phải được rời rạc đầu tiên. Sự rời rạc có thể không giám sát (các ràng buộc lựa chọn đặc biệt) hoặc giám sát (ràng buộc hướng dẫn bởi thông tin trong dữ liệu đào tạo).

Nếu một lớp học và giá trị đặc trưng không bao giờ xảy ra cùng với nhau trong thiết lập đào tạo sau đó ước tính xác suất dựa tần số sẽ được 0. Đây là vấn đề vì nó sẽ phá hủy tất cả các thông tin trong các xác suất khi chúng được nhân rộng. Vì vậy, mong muốn kết hợp một mẫu nhỏ chỉnh sửa trong tất cả các xác suất ước tính rằng như vậy không bao giờ được thiết lập chính xác 0.

## **II.5. Xây dựng một classifier từ mô hình xác suất**

Các thảo luận cho đến nay đã bắt nguồn những mô hình đặc trưng độc lập, có nghĩa là, mô hình xác suất naive Bayes. Naive Bayes classifier kết hợp mô hình này với một luật quyết định. Là một luật chung để chọn nhiều nhất các giả thuyết có khả năng xảy ra, điều này được biết đến như là maximum a posteriori hay luật quyết định MAP. Classifier tương ứng là chức năng phân lớp được xác định như sau:

Một chú ý rằng giả định độc lập có thể dẫn đến một số kết quả không mong muốn trong tính toán sau xác suất. Trong một số trường hợp khi có một phụ thuộc giữa sự quan sát, xác suất kể trên có thể mâu thuẫn với xác suất tiền đề thứ hai do mọi xác suất luôn nhỏ hơn hoặc bằng một.

Mặc dù rằng sự thật có thể áp dụng rộng rãi, giả định độc lập thường không chính xác, các naive Bayes classifier có vài thuộc tính làm cho nó hữu ích trong thực hành. Đặc biệt thực hành, sự tách riêng của lớp có điều kiện phân loại đặc trưng có nghĩa là mỗi phân loại có thể được ước tính độc lập như là một phân phối một chiều. Toàn bộ classifier là mạnh đủ để bỏ qua các thiếu sót nghiêm trọng của nó trong những mô hình xác suất naive.

## **II.6. Thuật toán bayes**

Bộ phân lớp Bayes có thể dự báo các xác suất là thành viên của lớp, chẳng hạn xác suất mẫu cho trước thuộc về một lớp xác định

X là mẫu dữ liệu chưa biết nhãn lớp

H là giả thuyết sao cho X thuộc về lớp C

Ấn định xác suất hậu nghiệm posterior probability P(H|X) sao cho H đúng khi cho trước quan sát X (H conditioned on X)

Giả sử thế giới các mẫu dữ liệu gồm trái cây, được mô tả bằng màu sắc và hình dáng.

Giả sử X là màu đỏ và tròn

H là giả thuyết mà X là quả táo

Thì P(H|X) phản ánh độ tin cậy X là quả táo khi biết trước X có màu đỏ và tròn

P(X|H) là xác suất tiên nghiệm của X có điều kiện trên H.

Khi có n giả thuyết

**Phân lớp Naive Bayesian (NBC)**

Phân lớp Naïve Bayesian (NBC)

Mỗi mẫu dữ liệu được biểu diễn bằng X = (x1, x2,…., xn) với các thuộc tính A1, A2,…., An

Các lớp C1, C2,…, Cm. Cho trước mẫu chưa biết X. NBC gán X vào Ci nếu P(Ci|X) > P(Cj|X) với 1 ≤ j ≤ m, j ≠ . Do vậy, chúng ta cực đại P(Ci|X). Lớp Ci sao cho P(Ci|X) là cực đại được gọi là giả thuyết hậu nghiệm cực đại (maximum posterior hypothesis).

Theo định lý Bayes ta có công thức sau:

Để phân lớp mẫu chưa biết X, ta tính P(X|Ci)P(Ci) cho từng Ci. Sau đó mẫu X được gán vào Ci nếu P(Ci|X) > P(Cj) P(Ci) là cực đại.

## **II.7. Ví dụ bài toán Naïve Bayes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Income | Student | Credit-rating | computer |
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |

Dự đoán nhãn lớp với phân lớp bayes

Mẫu thữ X

X = (age = “<=30” ,income = “medium”, student = “yes” ,credit\_rating = “fair”)

P(buy\_computer = “yes”) = 9/14 = 0.643

P(buys\_computer = “no”) = 5/14 = 0.357

Để tính P(X|Ci) P(Ci) , cho I = 1,2 chung ta tính :

P(age = “ <=30”|buys\_computer = “yes”) = 2/9 = 0.222

P(age = “ <=30”|buys\_computer = “no”) = 3/5 = 0.600

P(income = “ medium”|buys\_computer = “yes”) = 4/9 = 0.444

P(income = “ medium”|buys\_computer = “no”) = 2/5 = 0.444

P(student = “ yes”|buys\_computer = “yes”) = 6/9 = 0.667

P(student = “ no”|buys\_computer = “no”) = 1/5 = 0.200

P(credit\_rating = “ yes”|buys\_computer = “yes”) = 6/9 = 0.667

P(credit\_rating = “yes”| buys\_computer = “no” = 2/5 = 0.400

P(X|buys\_computer = “yes”) = 0.222\*0.667\*0.667 \*0.044 = 0.044

P(X|buys\_computer = “no”) = 0.600\*0.400\*0.200 \*0.400 = 0.019

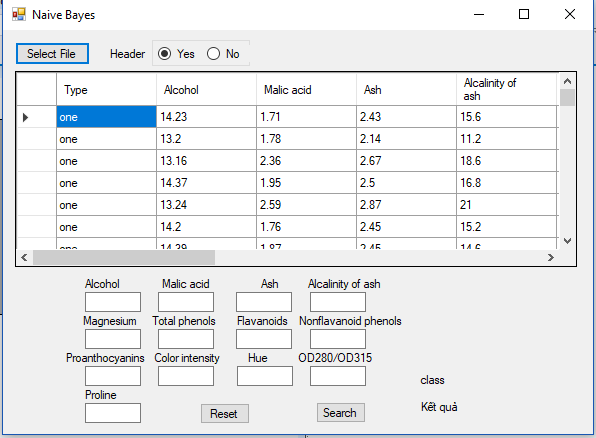
P(X|buys\_computer = “yes”) P(buy\_computer = “yes”) = 0.44\*0.643 = 0.028

P(X|buys\_computer = “no”) P(buy\_computer = “no”) = 0.019\*0.357 = 0.007

Kết quả dự đoán : buys\_computer = “yes”

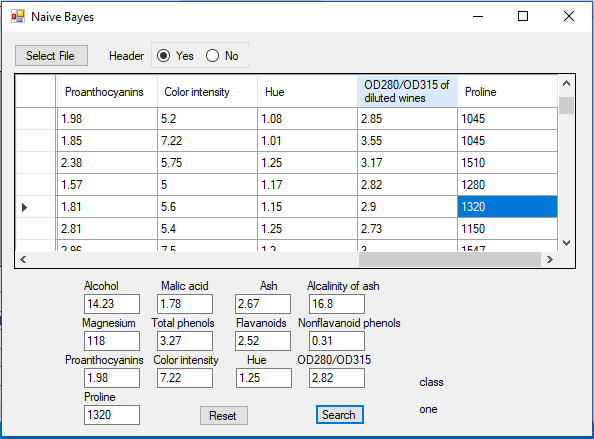
## **II.8. Thực hiện thuật toán Naïve Bayes trên C#**

Trong đề tài này nhóm em lựa chọn bài toán phân loại nhận dạng rượu để áp dụng thuật toán Naïve Bayes. Dữ liệu của bài toán này được sử dụng với nhiều người khác để so sánh các nhà phân loại khác nhau. Các lớp có thể tách rời, mặc dù chỉ RDA đã đạt được phân loại đúng 100%. Ở đây, dữ liệu được sử dụng để minh họa hiệu suất vượt trội của việc sử dụng chức năng đánh giá mới với RDA. Những dữ liệu này là kết quả của một phân tích hóa học của các loại rượu được chế tạo trong cùng một vùng ở Ý nhưng có nguồn gốc từ ba giống khác nhau. Phân tích xác định số lượng 13 thành phần trong mỗi loại rượu vang. Ban đầu tập dữ liệu có khoảng 30 biến cố, nhưng vì một số lý do nên tác giả chỉ có phiên bản 13 chiều. Các thuộc tính: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline.



Hình ảnh thể hiện các thuộc tính của các loại rượu

Tập dữ liệu gồm 3 loại rượu khác nhau. Mỗi loại rượu đều có các thuộc tính riêng biệt của chúng. Khi chúng ta nhập các thuộc tính khác, thì chúng ta nhấn button Search nó sẽ hiện ra loại rượu mà chúng ta đã nhập các thuộc tính trên những textbox trên hình



Hình ảnh nhập thuộc tính và phân lớp

Trên form có 13 thuộc tính của rượu vang. Chúng ta nhập vào 13 thuộc tính vào 13 textbox trên form. Sau đó nhấn vào button Search để cho phần mềm sẽ xử lý và tìm ra loại khi chúng ta nhập vào các thuộc tính. Hình trên là nhóm đã nhập vào những thuộc tính của loại one nên khi nhấn button search thì kết quả ra là loại one. Không nhất thiết là phải nhập trên cùng dòng, chúng ta có thể nhập các thuộc tính khác dòng với nhau. Khi muốn thực hiện phần mềm, lúc chúng ta cần chọn dữ liệu dùng để phân lớp. Sau đó nhập vào các thuộc tính cần phân lớp và nhấn vào button Search để tìm ra loại của các thuộc tính đã nhập.

## **II.9. Bảng danh mục các lớp được sử dụng trong chương trình**

|  |  |  |
| --- | --- | --- |
| **Thứ tự** | **Tên lớp** | **Mục đích chính của lớp trong chương trình** |
| 1 | Helper | Lớp tạo các hàm tính toán để cho thực hiện thuật toán của lớp Classifier |
| 2 | Classifier | Lớp tạo datatable và thực hiện thuật toán |
| 3 | Form1 | Lớp tạo giao diện gọi lại lớp Classifier để thực hiện thuật toán |

## **II.10. Bảng miêu tả phương thức trong các lớp**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Tên hàm** | **Mục đích** | **Input** | **Out** | **Mã giả** |
| 1 | openFileDialog1\_FileOk () | Lấy dự liệu từ file Excel thông qua hộp thoại openFileDialog | File Excel | Data được hiện thị trên GridView theo giống với file Excel | Chuyên dữ liệu của file Excel thành kiểu DataTable và gán lên cho  DataGridView để hiển thị |
| 2 | btntim\_Click () | Lấy dữ liệu của người dùng nhập vào và hiện thị kết quả | Dữ liêu lấy từ các Textbox khi người dùng nhập vào | Đưa ra kết quá dạng chuỗi và được hiện thị trên Lable | Lấy dữ liệu từ các TextBox,  xữ lý và hiện thị ra Lable |
| 3 | btnrest\_Click () | Làm mới các ô nhập dữ liệu và kết quả |  | Đặt lại giá trị của TextBox và Label là rỗng | ResetText của các TextBox và Label |
| 4 | btnSelect\_Click () | Mở hộp thoại openFileDialog | File Excel | Path và dữ liệu của file Execl | Thông qua cửa số nhập Path file Excel |
| 5 | Variance () | Tính độ lệch chuẩn | List dữ liệu kiểu double | Độ lệch chuẩn |  |
| 6 | Mean () | Tính trung bình | List dữ liệu kiểu double | Giá trị trung bình |  |
| 7 | NormalDist () | Tính toán xác xuất theo phân phối Gause của biến liên tục | Giá trị thuộc tính hiện tại,  Giá trị trung bình, độ lệch chuẩn | Xát xuất. |  |
| 8 | NORMDIST () | Trả về kết quả phân bố | Giá trị thuộc tính hiện tại,  Giá trị trung bình, độ lệch chuẩn, | Đưa ra kết quả dữ đoán. |  |
| 9 | SquareRoot () | Tính căn bặc 2 |  |  |  |
| 10 | TrainClassifier () | Tạo các datatable mới với  Tên thuộc tính, trạng thái của thuộc tín và số lần xuất hiện của mỗi trạng thái trong của thuộc tính đó | DataTable | DataTable |  |
| 11 | SelectRows () | Tính toán số lần xuất hiện của trạng thái của mỗi thuộc tính | Data, thứ tự col, giá trị của thuộc tính | Số lần suất hiện của thuộc tính đó |  |
| 12 | Classify () | Thực hiện gọi các hành tính toán để trả về giá trị tìm kiếm | 1 Object chứ dữ liệu người dùng nhập vào | Trả về kết quả tìm kiếm |  |

## **II.11. Bảng phân công công việc**

|  |  |
| --- | --- |
| **Tên SV** | **Mô tả khái quát mảng công việc SV thực hiện trong đồ án** |
| Võ Nhựt Kha | Tìm hiểu thuật toán, các class tính toán, truy xuất file |
| Bùi Minh Tiên | Mô phỏng ví dụ minh họa đơn giản, hiện thực thuật toán, phần giao diện |

# **CHƯƠNG III: PHẦN KẾT**

Sau khi tìm hiểu và cài đặt thuật toán, nhóm chúng em tự đánh giá kết quả như sau :

* Ưu điểm:

Đọc, xuất và xử lý được các file có dữ liệu dạng lớn nhỏ khác nhau

Trả kết quả ra file

Tìm ra lớp sau khi nhập dữ liệu cần tìm.

* Khuyết điểm:

Chưa thử trên nhiều tập dữ liệu khác

Chưa xử lý nhiều vào vấn đề độ chính xác thuật toán

Khó khăn gặp phải:

* Thuật toán naïve bayes vẫn còn là một vấn đề còn mới nên việc tìm hiều có phần gặp khó khăn.
* Tài liệu tiếng anh có nhiều từ chuyên ngành không hiểu.

# **PHẦN IV: TÀI LIỆU THAM KHẢO**

1. Slides in PowerPoint (<http://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm>)
2. <https://www.packtpub.com/books/content/what-na%C3%AFve-bayes-classifier>
3. <https://www.slideshare.net/ashrafmath/naive-bayes-15644818>
4. <https://www.slideshare.net/marinasantini1/lecture-5-bayesian-classification>
5. <https://www.slideshare.net/aorriols/lecture10-nave-bayes>
6. <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>
7. https://viblo.asia/p/ung-dung-thuat-toan-naive-bayes-trong-giai-quyet-bai-toan-chuan-doan-benh-tieu-duong-eW65GYejZDO