Báo cáo môn

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM THÀNH PHỐ HỒ CHÍ MINH**



Khai thác dữ liệu

**Áp dụng Bayes vào Chẩn đoán**

**Bệnh Phụ Khoa**

*TPHCM, tháng 12 năm 2019*

Tên thành viên:

Nguyễn Tú Nguyên – 2001170114

Đào Trần Thanh Đông – 2001170025

Nguyễn Võ Tuấn Thành – 2001170172

Diệp Chấn Đạt – 2001170017

Trang Tấn Nhựt – 2001170393

Lê Thanh Bình – 2001170310

Phan Đức Tài - 2001170155

GV hướng dẫn: Trần Như Ý

1. **Tìm hiểu về định lý Bayes**

*Định lý Bayes* (Bayes’ Theorem) là một định lý toán học để tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra.

*Naive Bayes Classification (NBC)* là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes. Thuật toán này thuộc nhóm Học có giám sát.

*Công thức tổng quát:*

P(Ci|X)=

Trong công thức này, P(Ci) được gọi là xác suất tiền nghiệm mà ta biết về Ci, trước khi tiếp cận dữ liệu X. Một thí dụ thường gặp về P(Ci) trong y học là tỉ suất mắc bệnh trong quần thể. Trong bài toán phân loại, P(Ci) được cung cấp từ chính tập dữ liệu ta dùng để huấn luyện mô hình, vì ta biết tỉ lệ phân bố của mỗi nhãn Ci trên toàn bộ mẫu.

Ở mẫu số, P(X) được hiểu như xác suất quan sát được (những) giá trị của vector dữ liệu X, trong toàn bộ khả năng có thể của chúng trên thực tế. Việc tính giá trị chính xác của P(X) gần như bất khả thi, nhưng may mắn thay, điều này không thực sự cần thiết, vì P(X) là mẫu số chung cho tất cả nhãn Ci.

https://viblo.asia/uploads/a468626e-0831-4efb-b4be-537f5329f050.png

P(x­i|c) là xác suất quan sát được giá trị dữ liệu x khi biết nhãn phân loại c. Việc tính xác suất điều kiện này cũng gần như bất khả thi trên thực tế, vì X là một không gian dữ liệu đa chiều gồm giá trị các biến ngẫu nhiên và tất cả những khả năng tổ hợp có thể trong số này một khi có liên hệ, tương tác giữa chúng.

Để giải quyết khó khăn này, người ta áp dụng một giả định quan trọng, đó là các yếu tố (biến số) xj trong dữ liệu X hoàn toàn độc lập với nhau.

Nếu A và B là 2 sự kiện liên quan đến nhau, và xác suất xảy ra sự kiện B lớn hơn 0, ta có thể định nghĩa xác suất xảy ra A khi biết B xảy ra như sau:

P(A|B)=

Ta có thể viết lại thành:

P (A, B) = P(A|B) P(B)

Theo công thức, mọi xác suất điều kiện đều phải khác 0. Trên thực tế, cụ thể là trong bài tiểu luận, có một số bệnh một khi đã có biểu hiện này thì sẽ không có biểu hiện kia, hoặc không thể có biểu hiện nào đó. Nói cách khác, hoàn toàn có khả năng xảy ra những xác suất bằng 0. Điều này sẽ gây nhiễu hoặc thậm chí làm sai lệch kết quả dự đoán.

Để giải quyết vấn đề này, ta sử dụng kĩ thuật *làm trơn Laplace:*

P(Ci)=

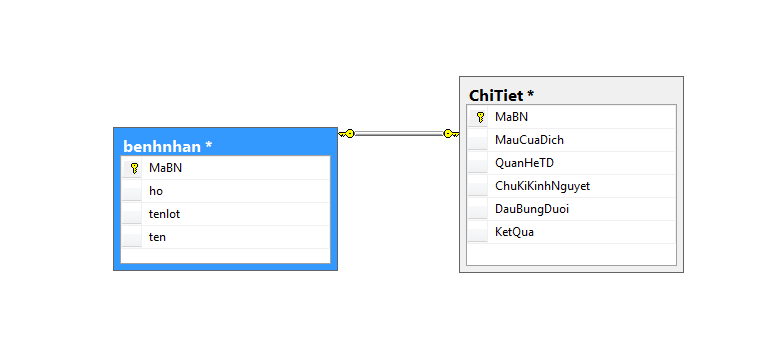
Trong đó, D là tổng số bộ dữ liệu, m là số lớp.

P(xk|Ci)=

Trong đó, r là số giá trị của thuộc tính.

NBC có mối liên hệ mật thiết với ngành Thống kê vì cơ chế của nó dựa vào định lý Bayes và khi vận dụng vào Y học thì Naive Bayes cũng tương đồng với quy trình biện luận lâm sàng của người bác sĩ. Tuy cổ xưa và quá đơn giản, nhưng Naive Bayes vẫn còn chỗ đứng ở thời đại ngày nay; nó vẫn được nhắc đến trong mọi giáo trình về Machine learning bên cạnh những giải thuật phức tạp khác, điều này cho thấy Naive Bayes có hiệu quả thực sự, và đáng để ta tìm hiểu.

1. **Ứng dụng thuật toán: Xây dựng ứng dụng cơ bản Hỗ trợ Chẩn đoán Bệnh Phụ khoa qua các đặc điểm khảo sát**
2. **Cơ sở dữ liệu**

Dữ liệu: Gồm 7500 dòng dữ liệu với cấu trúc bên dưới

Các thuộc tính cần giải thích:

* MauCuaDich: Thuộc tính thể hiện tính trạng dịch âm đạo bệnh nhân, cụ thể là màu sắc. Thuộc tính gồm các giá trị {Không màu, Trắng đục, Nâu đen}
* QuanHeTD: Thể hiện tình trạng Quan hệ tình dục của bệnh nhân, trả lời câu hỏi “Bệnh nhân đã từng quan hệ tình dục hay chưa?”, gồm hai giá trị {Chưa, Rồi}
* ChuKiKinhNguyet: Thuộc tính theo dõi Chu kì kinh nguyệt của Bệnh nhân, trả lời cho câu hỏi “Kinh nguyệt của bệnh nhân đều hay không?”, gồm hai giá trị {Đều, Không đều}
* DauBungDuoi: Thuộc tính theo dõi tình trạng đau bụng dưới của bệnh nhân, trả lời cho câu hỏi “Bệnh nhân có đau bụng dưới hay không?”, gồm hai giá trị {Có, Không}
* KetQua: Kết quả chẩn đoán sơ lược tạm thời qua các dấu hiệu – thuộc tính trên. Có 3 lớp (giá trị của KetQua) là {Bình thường, Viêm âm đạo, Viêm buồng trứng}

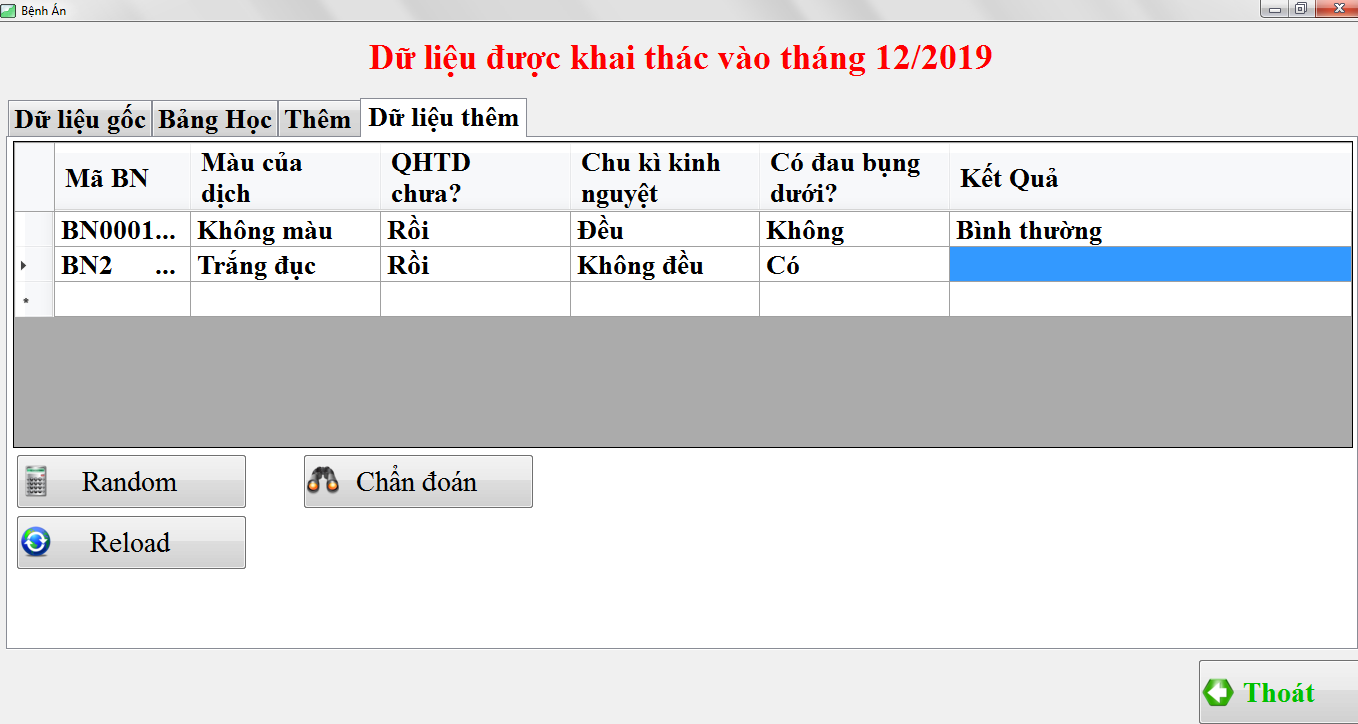
Ứng dụng dựa trên thuật toán Naïve Bayes để đưa ra chẩn đoán tạm thời, dùng để đưa vào khám chuyên sâu (khám tìm ra cấp hoặc mức độ nghiêm trọng của bệnh, hoặc có thể tìm ra các bệnh liên quan) cho bệnh nhân. Kết quả của bảng dữ liệu học mặc định đã qua nghiên cứu và có lời khuyên của chuyên gia nên có thể tin tưởng đưa vào tư vấn.

1. **Giao diện đồ họa**

Bảng dữ liệu gốc gồm 7500 dòng

Bảng học giúp lọc ra số dòng dữ liệu dùng để học, nếu nhập quá khối lượng dữ liệu có trong cơ sở (cụ thể > 7500), hệ thống sẽ mặc định lấy 200 dòng.



Giao diện chính dùng để nhập Bệnh nhân mới và Chẩn đóan. Các xác suất kết quả sẽ hiện ra ở các label tương ứng.

Bảng để Random và kiểm soát dữ liệu Thêm. Sau khi Random có thể nhấn Chẩn đoán để kiểm tra.

1. **Tài liệu tham khảo**

1. “Ứng dụng thuật toán Naive Bayes trong giải quyết bài toán chuẩn đoán bệnh tiểu đường” – TechTalk(2019)

2. “Machine Learning trong NLP” - Vietnamese Natural Language Processing

3. “Naive Bayes ứng dụng cho y học” – Lê Ngọc Khả Nhi (tháng 1 năm 2018)