**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

Thứ năm, ngày 25, tháng 06, năm 2015

**ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

GV HDLT: NGUYỄN TUẤN ĐĂNG

GV HDTH: NGUYỄN TRỌNG CHỈNH

LỚP: CS324.F21

SVTH: NGÔ KHẮC HÙNG – 11520572

NGUYỄN KHÁNH LIỀN – 11520587

THÁI HỒNG CHÂU - 11520518

**BÁO CÁO**

**MÁY HỌC TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

*ĐỀ TÀI:* *PHÂN LỚP VĂN BẢN THEO MÔ HÌNH NAÏVE BAYES*

**Thuật toán Naïve Bayes**

1. **Giới thiệu bài toán Naïve Bayes**

Naïve Bayes là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học [Mitchell, 1966], [Joachims, 1997], [Jason, 2001], được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961, sau đó thuật toán Naïve Bayes trở nên phổ biến trong nhiều lĩnh vuewcj như công cụ tìm kiếm [Risbergen et al, 1970], các bộ lọc email [Sahami et al, 1998],…

1. **Thuật toán Naïve Bayes**

Từ một tập các mẫu đã phân lớp trước, ta xây dựng mô hình cho từng lớp. Mục đích là gán cho mẫu mới vào các lớp với độ chính xác cao nhất có thể. Phương pháp Naïve Bayes được phân lớp theo mô hình dự đoán xác suất, dự đoán xác suất là thành viên của lớp mẫu mới nào. Nền tảng dựa vào định lý Bayes.

Thuật toán Naïve Bayes dựa trên định lý Bayes được phát biểu như sau:

Áp dụng trong bài toán phân loại, các dữ kiện gồm có:

* D: tập dữ liệu huấn luyện đã được vector hóa dưới dạng
* Ci: phân lớp i, với i = {1, 2, …, m}.
* Các thuộc tính độc lập điều kiện đôi một với nhau.

Theo định lý Bayes:

Theo tính chất độc lập điều kiện:

Trong đó:

* là xác suất thuộc phân lớp i khi biết trước mẫu X.
* xác suất là phân lớp i.
* xác suất thuộc tính thứ k mang giá trị xk khi đã biết X thuộc phân lớp i.

Các bước thực hiện thuật toán Naïve Bayes:

* Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính và
* Bước 2: Phân lớp , ta cần tính xác suất thuộc từng phân lớp khi đã biết trước Xnew. Xnew được gán vào lớp có xác suất lớn nhất theo công thức

1. **Áp dụng cho bài toán phân lớp văn bản**

Ý tưởng: Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Giả định đó làm cho việc tính toán NB hiệu quả và nhanh chóng hơn các phương pháp khác vì không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề. Kết quả dự đoán bị ảnh hưởng bởi kích thước tập dữ liệu, chất lượng của không gian đặc trưng…

Thuật toán gồm 2 giai đoạn huấn luyện và phân lớp:

* 1. *Huấn luyện*: tính và

Đầu vào:

* Tập nhãn/lớp cho từng nhóm của tập huấn luyện.

Đầu ra:

* Các giá trị xác suất và .

Các bước thực hiện:

* Xác định tập từ vựng (vocabulary)
* Tính dựa trên công thức:

Trong đó:

* là số lượng từng doc trong tập dữ liệu.
* là tổng số doc trong tập dữ liệu.
* Tính dựa trên công thức:

Trong đó:

* là số lần xuất hiện của một từ trong doc đang xét.
* là số lượng từ trong một doc đang xét
* là số lượng từ trong tập dữ liệu (những từ trùng nhau chỉ tính là 1)
  1. *Phân lớp*

Đầu vào:

* Các giá trị xác suất và .

Đầu ra:

* Nhãn/lớp văn bản cần phân loại.

Công thức xác suất khi phân lớp:

Áp dụng công thức trên ta tính xác suất thuộc từng phân lớp cho văn bản và chọn ra lớp có xác suất cao nhất.

* **Ví dụ:**

Cho 3 tập ngữ liệu huấn luyện sau:

\* Tập ngữ liệu huấn luyên 1: Nam đã đi đến trường. Nam đang ở trường. Nó đang học bài.

\* Tập ngữ liệu huấn luyên 2: Lan đi chợ. Lan mua nhiều đồ ăn. Cô ấy đang nấu cơm.

\* Tập ngữ liệu huấn luyên 3: Bạn của Nam mới xin được việc. Nó làm ở một công ty tin học.

Cài đặt thuật toán Naive Bayes để phân lớp văn bản sau:

Lan gặp một cô bạn ở chợ. Cô ấy cũng đi chợ. Họ mua đồ để nấu ăn.

**Giải:**

* Huấn luyện

Vocabulary = {nam, đã, đi, đến, trường, đang, ở, nó, học, bài, lan, chợ, mua, nhiều, đồ, ăn, cô, ấy, nấu, cơm, bạn, của, mới, xin, được, việc, làm, một, công, ty, tin} => V=31.

* Xét nhóm 1:

* Tương tự ta xét nhóm 2:
* Xét nhóm 3
* Phân lớp: (chỉ xét các từ có trong tập vocabulary)

“Lan gặp một cô bạn ở chợ. Cô ấy cũng đi chợ. Họ mua đồ để nấu ăn.”

v(1) = nhóm 1

-20

v(2) = nhóm 2

-18

v(3) = nhóm 3

-20

Kết luận: vNB = nhóm 2.

1. **Demo chương trình.**

Hướng dẫn sử dụng chương trình

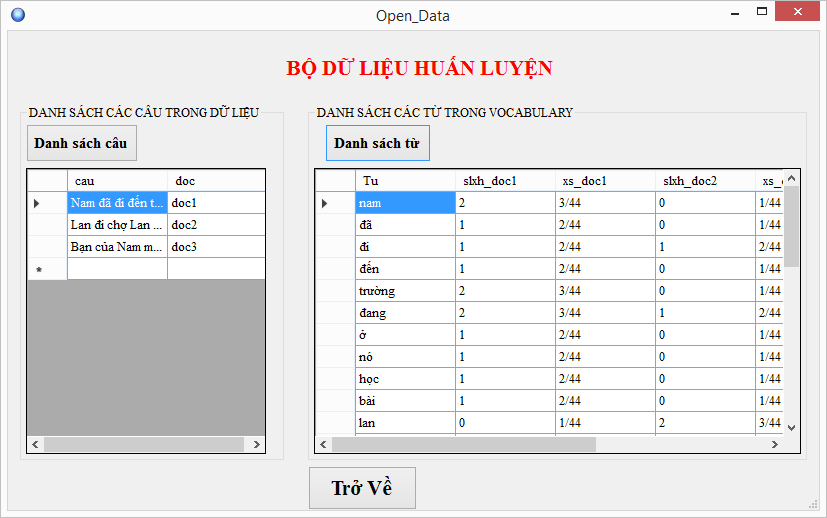
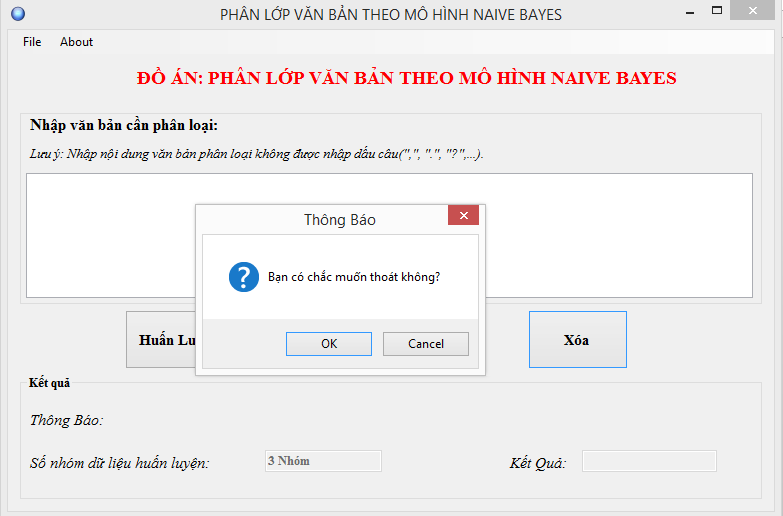
Giao diện chính của chương trình

Chọn nút huấn luyện để huấn luyện dữ liệu, khi huấn luyện xong sẽ hiện thông báo và số nhóm dữ liệu huấn luyện như hình.

Nhập văn bản muốn phân loại vào ô nhập, nhấn nút phân loại. Chương trình sẽ phân loại và kết quả xuất ra như hình.

Nhấn nút xóa sẽ xóa ô nhập văn bản, thông báo, và kết quả để chạy câu tiếp theo.

Các chức năng khác của chương trình

* File->open data: mở danh sách dữ liệu của chương trình từ file dữ liệu (input.data)
* About: Một số thông tin của chương trình
* File->Exit : thoát chương trình

1. **Ưu nhược điểm của thuật toán Naïve Bayes**

* Ưu điểm:
* Dễ dàng cài đặt
* Thời gian thi hành tương tự như cây quyết định
* Đạt kết quả tốt trong phần lớn các trường hợp
* Nhược điểm:
* Giả thiết về tính độc lập điều kiện của các thuộc tính làm giảm độ chính xác