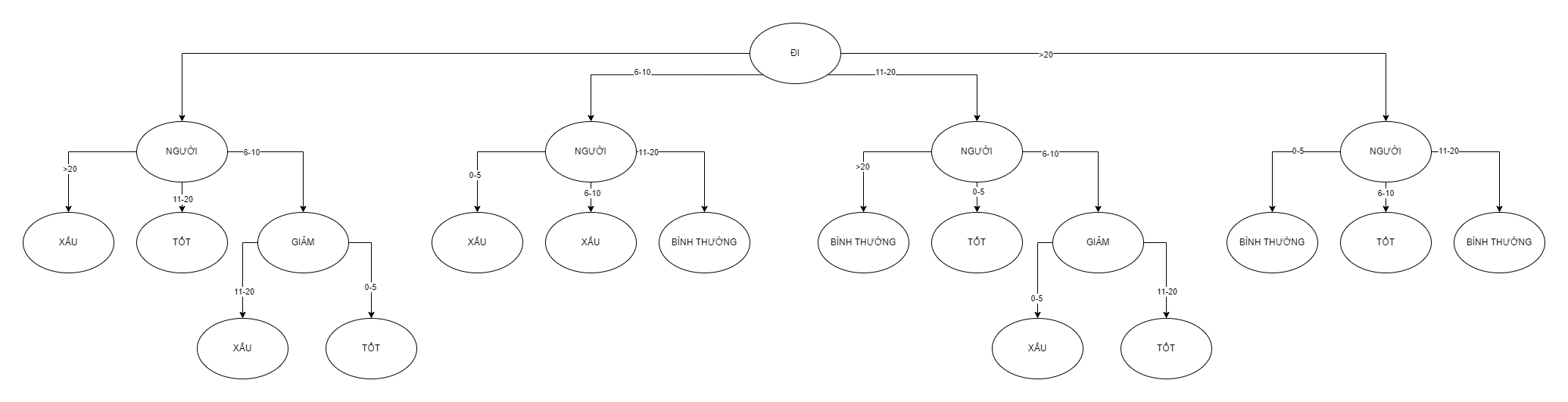
Bài 2\_ Thực hành

(Cơ bản) Phân tích cảm xúc (sentiment analysis) là một lĩnh vực nghiên cứu rất quan trọng và thú vị trong khai thác dữ liệu văn bản (text mining). Sinh viên có thể làm quen với vấn đề này thông qua bài tập sau. Người ta phân tích các trạng thái trên mạng xã hội và thống kê được số lần xuất hiện của các từ khóa (term) được trình bày trong bảng dữ liệu bên dưới, Cảm xúc là thuộc tính phân lớp

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **giảm** | **người** | **chuyển** | **yêu** | **vừa** | **đi** | ***Cảm xúc*** |
| 1 | 0..5 | 11..20 | >20 | 11..20 | >20 | 0..5 | *tốt* |
| 2 | 11..20 | 6..10 | 6..10 | 0..5 | 11..20 | 11..20 | *tốt* |
| 3 | 6..10 | 0..5 | 6..10 | 11..20 | 0..5 | 6..10 | *xấu* |
| 4 | >20 | 0..5 | 11..20 | 6..10 | 0..5 | >20 | *bình thường* |
| 5 | 0..5 | >20 | 11..20 | 0..5 | 6..10 | 0..5 | *xấu* |
| 6 | 0..5 | 6..10 | 0..5 | 0..5 | 11..20 | 11..20 | *xấu* |
| 7 | 0..5 | 6..10 | 11..20 | 0..5 | 6..10 | 0..5 | *tốt* |
| 8 | 11..20 | >20 | 0..5 | 11..20 | 0..5 | 11..20 | *bình thường* |
| 9 | 0..5 | 0..5 | 6..10 | 6..10 | 6..10 | >20 | *tốt* |
| 10 | 11..20 | 0..5 | 11..20 | 11..20 | 0..5 | 11..20 | *tốt* |
| 11 | >20 | 6..10 | 0..5 | 0..5 | 0..5 | 6..10 | *xấu* |
| 12 | 0..5 | 0..5 | 11..20 | 0..5 | 11..20 | >20 | *bình thường* |
| 13 | 6..10 | 11..20 | 6..10 | >20 | 0..5 | 6..10 | *bình thường* |
| 14 | 11..20 | 6..10 | >20 | 11..20 | 0..5 | 0..5 | *xấu* |

* + 1. Xác định tất cả những mâu thuẫn có thể có trong dữ liệu.
* Sau khi tiền xử lý dữ liệu cho phù hợp với thuật toán phân lớp, người ta thường rà soát lại để tìm ra những mâu thuẫn. Mâu thuẫn trong tập dữ liệu phân lớp là những dòng dữ liệu có giá trị thuộc tính giống nhau nhưng lại thuộc các phân lớp khác nhau. Trong dữ liệu bài tập trên không xảy ra trường hợp này.
  + 1. Tính giá trị chỉ số Gini của các thuộc tính và vẽ cây quyết định theo thuật toán CART cho dữ liệu trên.
* Chỉ số Gini thuộc tính **giảm**
* Chỉ số Gini thuộc tính **người**
  + - 5
* Chỉ số Gini thuộc tính **chuyển**
* Chỉ số Gini thuộc tính **Yêu**
* Chỉ số Gini thuộc tính **Vừa**
* Chỉ số Gini thuộc tính **Đi**
* Trong 6 thuộc tính ta xét, thuộc tính ***Đi*** có chỉ số Gini nhỏ nhất. Do đó, ta chọn thuộc tính này làm phép chia nhánh cho cây tại nút gốc.
* **Xét trường hợp 1:** Nhánh Đi = “0…5”
* Chỉ số Gini thuộc tính **giảm** với điều kiện Đi = “0…5”
* Chỉ số Gini thuộc tính **Người** với điều kiện Đi = “0…5”
* Chỉ số Gini thuộc tính **Chuyển** với điều kiện Đi = “0…5”
* Chỉ số Gini thuộc tính **Yêu** với điều kiện Đi = “0…5”
* Chỉ số Gini thuộc tính **Vừa** với điều kiện Đi = “0…5”
* Vậy nhánh Đi = “0…5” có 2 thuộc tính ***Người, vừa*** có chỉ số Gini nhỏ nhất nên ta lấy thuộc tính ***Người*** tiếp tục chia nhánh.
* **Xét trường hợp 1.1:** Nhánh Đi = “0…5” và Người = “11…20”, với nhánh này luôn có phân lớp là Cảm xúc = “Tốt”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 1.2:** Nhánh Đi = “0…5” và Người = “>20”, với nhánh này luôn có phân lớp là Cảm xúc = “Xấu”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 1.3:** Nhánh Đi = “0…5” và Người= “6…10”
* Chỉ số Gini thuộc tính **giảm** với điều kiện Đi = “0…5” và Người= “6…10”
* Chỉ số Gini thuộc tính **chuyển** với điều kiện Đi = “0…5” và Người= “6…10”
* Chỉ số Gini thuộc tính **yêu** với điều kiện Đi = “0…5” và Người= “6…10”
* Chỉ số Gini thuộc tính **vừa** với điều kiện Đi = “0…5” và Người= “6…10”
* Vậy nhánh Đi = “0…5” và Người = 6..10 có 4 thuộc tính ***Giảm, chuyển, yêu, vừa*** có chỉ số Gini nhỏ nhất nên ta lấy thuộc tính***Giảm*** tiếp tục chia nhánh.
* **Xét trường hợp 1.3.1:** Nhánh Đi = “0…5”, Người = “11…20”,Giảm =”0…5”, với nhánh này luôn có phân lớp là Cảm xúc = “Tốt”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 1.3.2:** Nhánh Đi = “0…5”, Người = “11…20”,Giảm =”11..20”, với nhánh này luôn có phân lớp là Cảm xúc = “Xấu”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 2:** Nhánh Đi = “6..10”
* Chỉ số Gini thuộc tính **giảm** với điều kiện Đi = “6..10”
* Chỉ số Gini thuộc tính **Người** với điều kiện Đi = “6..10”
* Chỉ số Gini thuộc tính **Chuyển** với điều kiện Đi = “6..10”
* Chỉ số Gini thuộc tính **Yêu** với điều kiện Đi = “6..10”
* Chỉ số Gini thuộc tính **Vừa** với điều kiện Đi = “6…10”
* Vậy nhánh Đi = “6…10” có 2 thuộc tính ***Người, Yêu*** có chỉ số Gini nhỏ nhất nên ta lấy thuộc tính ***Người*** tiếp tục chia nhánh.
* **Xét trường hợp 2.1:** Nhánh Đi = “6…10” và Người = “0…5”, với nhánh này luôn có phân lớp là Cảm xúc = “Xấu”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 2.2:** Nhánh Đi = “6…10” và Người = “6..10”, với nhánh này luôn có phân lớp là Cảm xúc = “Xấu”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 2.3:** Nhánh Đi = “6…10” và Người = “11..20”, với nhánh này luôn có phân lớp là Cảm xúc = “Bình thường”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 3:** Nhánh Đi = “11..20”
* Chỉ số Gini thuộc tính **giảm** với điều kiện Đi = “11..20”
* Chỉ số Gini thuộc tính **Người** với điều kiện Đi = “11..20”
* Chỉ số Gini thuộc tính **Chuyển** với điều kiện Đi = “11..20”
* Chỉ số Gini thuộc tính **Yêu** với điều kiện Đi = “11..20”
* Chỉ số Gini thuộc tính **Vừa** với điều kiện Đi = “11..20”
* Vậy nhánh Đi = “11..20” có 2 thuộc tính ***Người, chuyển*** có chỉ số Gini nhỏ nhất nên ta lấy thuộc tính ***Người*** tiếp tục chia nhánh.
* **Xét trường hợp 3.1:** Nhánh Đi = “11…20” và Người = “0…5”, với nhánh này luôn có phân lớp là Cảm xúc = “Tốt”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 3.2:** Nhánh Đi = “11…20” và Người = “>20”, với nhánh này luôn có phân lớp là Cảm xúc = “Bình thường”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 3.3:** Nhánh Đi = “11…20” và Người = “6..10”
* Chỉ số Gini thuộc tính **giảm** với điều kiện Đi = “11…20” và Người= “6…10”
* Chỉ số Gini thuộc tính **chuyển** với điều kiện Đi = “11…20” và Người= “6…10”
* Chỉ số Gini thuộc tính **yêu** với điều kiện Đi = “11…20” và Người= “6…10”
* Chỉ số Gini thuộc tính **vừa** với điều kiện Đi = “11…20” và Người= “6…10”
* Vậy nhánh Đi = “11…20” và Người = 6..10 có 4 thuộc tính ***Giảm, chuyển,*** có chỉ số Gini nhỏ nhất nên ta lấy thuộc tính***Giảm*** tiếp tục chia nhánh.
* **Xét trường hợp 3.3.1:** Nhánh Đi = “11…20”, Người = “6…10”,Giảm =”0…5”, với nhánh này luôn có phân lớp là Cảm xúc = “Xấu”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 3.3.2:** Nhánh Đi = “11…20”, Người = “6…10”,Giảm =”11..20”, với nhánh này luôn có phân lớp là Cảm xúc = “Tốt”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 4:** Nhánh Đi = “>20”
* Chỉ số Gini thuộc tính **giảm** với điều kiện Đi = “>20”
* Chỉ số Gini thuộc tính **người** với điều kiện Đi = “>20”
  + - Gini0…5|SĐi=">20"=1−*3*2−*3*2GiniNgườiĐi=">20"= 33
* Chỉ số Gini thuộc tính **Chuyển** với điều kiện Đi = “>20”
* Chỉ số Gini thuộc tính **Yêu** với điều kiện Đi = “>20”
* Chỉ số Gini thuộc tính **Vừa** với điều kiện Đi = “>20”
* Vậy nhánh Đi = “>20” có thuộc tính ***Vừa, chuyển*** có chỉ số Gini nhỏ nhất nên ta lấy thuộc tính ***Vừa*** tiếp tục chia nhánh.
* **Xét trường hợp 4.1:** Nhánh Đi = “>20” và Vừa= “0…5”, với nhánh này luôn có phân lớp là Cảm xúc = “Bình thường”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 4.2:** Nhánh Đi = “>20” và Vừa= “6..10”, với nhánh này luôn có phân lớp là Cảm xúc = “Tốt”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Xét trường hợp 4.3:** Nhánh Đi = “>20” và Vừa= “11..20”, với nhánh này luôn có phân lớp là Cảm xúc = “Bình thường”, vì vậy nhánh này nút lá không cần xét tiếp.
* **Cây quyết định:**



* + 1. Sử dụng cây quyết định và thuật toán Naïve Bayes để dự đoán cảm xúc của những trạng thái sau:
* Sử dụng cây quyết định CART

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **giảm** | **người** | **chuyển** | **yêu** | **vừa** | **đi** | **Cảm xúc** |
| 0..5 | 6..10 | 0..5 | 11..20 | 6..10 | 0..5 | Tốt |
| 0..5 | 0..5 | 6..10 | 0..5 | 11..20 | >20 | Bình thường |
| 6..10 | 0..5 | 11..20 | >20 | 6..10 | 6..10 | Xấu |
| 6..10 | 11..20 | 6..10 | 6..10 | >20 | 0..5 | Tốt |

* Sử dụng thuật toán Naïve Bayes để dự đoán cảm xúc của những trạng thái
* Áp dụng qui tắc làm trơn Laplace ta có
* Ta có P(Cảm xúc = ‘Tốt’) = = 6/17
* Ta có P(Cảm xúc = ‘Xấu’) = = 6/17 0,3529
* Ta có P(Cảm xúc = ‘Bình thường’) = = 5/17 0,2941
* Với tập hồ sơ X1 = {Giảm: 0..5, Người: 6..10, Chuyển: 0..5Yêu: 11..20, Vừa: 6..10, Đi 0..5}
* Xét trường hợp Cảm xúc = ‘Tốt’
* P(Cảm xúc = ‘Tốt’)\*P(X1|Cảm xúc = ‘Tốt’) = 2,14.10-4
* Xét trường hợp Cảm xúc = ‘Xấu’
* P(Cảm xúc = ‘Xấu’)\*P(X1|Cảm xúc = ‘Xấu’)= 4,28.10-4
* Xét trường hợp Cảm xúc = ‘Bình thường’
* P(Cảm xúc = ‘Bình thường’)\*P(X1|Cảm xúc = ‘Bình thường’) = 8,975.10-6
* Như vậy hồ sơ X1 có xác suất xảy ra Cảm xúc = ‘Xấu’ lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán vào phân lớp Cảm xúc = ‘Xấu’
* Với tập hồ sơ X2= { Giảm: 0..5, Người: 0..5, Chuyển: 6..10, Yêu: 0..5, Vừa: 11..20, Đi >20}
* Xét trường hợp Cảm xúc = ‘Tốt’
* P(Cảm xúc = ‘Tốt’)\*P(X2|Cảm xúc = ‘Tốt’) = 2,85.10-4
  + Xét trường hợp Cảm xúc = ‘Xấu’
* P(Cảm xúc = ‘Xấu’)\*P(X2|Cảm xúc = ‘Xấu’) = 6,34.10-5
  + Xét trường hợp Cảm xúc = ‘Bình thường’
* P(Cảm xúc = ‘Bình thường’)\*P(X2|Cảm xúc = ‘Bình thường’) = 1,6155.10-4
* Như vậy hồ sơ X2 có xác suất xảy ra Cảm xúc = ‘Tốt’ lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán vào phân lớp Cảm xúc = ‘Tốt’
* Với tập hồ sơ X3{ Giảm: 6..10, Người: 0..5, Chuyển: 11..20, Yêu: >20, Vừa: 6..10, Đi 6..10}
  + Xét trường hợp Cảm xúc = ‘Tốt’
* P(Cảm xúc = ‘Tốt’)\*P(X3|Cảm xúc = ‘Tốt’) = 1,782.10-5
  + Xét trường hợp Cảm xúc = ‘Xấu’
* P(Cảm xúc = ‘Xấu’)\*P(X3|Cảm xúc = ‘Xấu’) = 3,168.10-5
  + Xét trường hợp Cảm xúc = ‘Bình thường’
* P(Cảm xúc = ‘Bình thường’)\*P(X3|Cảm xúc = ‘Bình thường’) = 8,0777.10-5
* Như vậy hồ sơ X3 có xác suất xảy ra Cảm xúc = ‘Bình thường’ lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán vào phân lớp Cảm xúc = ‘Bình thường’
* Với tập hồ sơ X4{ Giảm: 6..10, Người: 11..20, Chuyển: 6..10, Yêu: 6..10, Vừa: 6..10, Đi 0..5}
  + Xét trường hợp Cảm xúc = ‘Tốt’
* P(Cảm xúc = ‘Tốt’)\*P(X4|Cảm xúc = ‘Tốt’) = 4,753.10-5
  + Xét trường hợp Cảm xúc = ‘Xấu’
* P(Cảm xúc = ‘Xấu’)\*P(X4|Cảm xúc = ‘Xấu’) = 7,9208.10-6
  + Xét trường hợp Cảm xúc = ‘Bình thường’
* P(Cảm xúc = ‘Bình thường’)\*P(X4|Cảm xúc = ‘Bình thường’) = 1,795.10-5
* Như vậy hồ sơ X4 có xác suất xảy ra Cảm xúc = ‘Tốt’ lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán vào phân lớp Cảm xúc = ‘Tốt’
  + 1. Trên thực tế những trạng thái này lần lượt có cảm xúc là: *xấu, tốt, bình thường, tốt*. Hãy lập ma trận nhầm lẫn, sau đó tính giá trị độ chính xác, độ phủ của cả hai phương pháp trên rồi so sánh chúng với nhau. Sinh viên có kết luận gì về kết quả này?
* Thuật toán cây quyết định CART

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lớp dự đoán mô hình** | | | | |
| **Lớp trên thực tế** |  | Tốt | Bình thường | Xấu |
| Tốt | 1 | 1 | 0 |
| Bình thường | 0 | 0 | 1 |
| Xấu | 1 | 0 | 0 |

Precion(M) = = 50%

recall(M) = = 50%

* Thuật toán Naïve Bayes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lớp dự đoán mô hình** | | | | |
| **Lớp trên thực tế** |  | Tốt | Bình thường | Xấu |
| Tốt | 2 | 0 | 0 |
| Bình thường | 0 | 1 | 0 |
| Xấu | 0 | 0 | 1 |

Precion(M) = = 100%

recall(M) = = 100%

* Thuật toán Naïve Bayes có độ chính xác cao hơn CART nên
  + 1. Nếu nắm bắt được cảm xúc của người dùng mạng xã hội thì sinh viên sẽ sử dụng chúng như thế nào?
* Người dung có cảm xúc tốt và bình thường: bình thường.
* Người dung có cảm xúc xấu: hạn chế các bài đăng mang tính tiêu cực.