**TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**-----□□&□□-----**

Icon

Description automatically generated

**Bài tập lớn**

**Học máy**

***65CS2***

|  |  |  |
| --- | --- | --- |
| ***Thành viên:*** | ***Nguyễn Huy Hiệu*** | ***1520065*** |
|  | ***Nguyễn Hồng Quân*** | ***162865*** |
|  | ***Khiếu Trung Hiếu*** | ***71865*** |
|  |  |  |

***Hà Nội, 1/2022***

**A – Naïve Bayes**

**I. Lý thuyết**

Xét bài toán phân loại với các nhãn 1, 2, …, C. Giả sử có một điểm dữ liệu x∈. Hãy tính xác suất để điểm dữ liệu này rơi vào nhãn c. Nói cách khác, hãy tính:

Tức tính xác suất để đầu ra là nhãn c biết rằng đầu vào là một vector x.

Biểu thức này, nếu tính được, sẽ giúp chúng ta xác định được xác suất để điểm dữ liệu rơi vào mỗi nhãn của bài toán. Từ đó có thể giúp xác định nhãn của điểm dữ liệu đó bằng cách chọn ra nhãn có xác suất rơi vào cao nhất:

Biểu thức trên thường khó được tính trực tiếp. Thay vào đó, quy tắc Bayes thường được sử dụng như sau:

Do mẫu số p(x) độc lập với c nên bài toán trở thành như sau:

Tiếp tục xét biểu thức trên, p(c) có thể được hiểu là xác suất để một điểm rơi vào nhãn c. Giá trị này có thể được tính bằng MLE, tức tỉ lệ số điểm dữ liệu trong tập training rơi vào class này chia cho tổng số lượng dữ liệu trong tập training; hoặc cũng có thể được đánh giá bằng MAP. Trường hợp thứ nhất thường được sử dụng nhiều hơn.

Thành phần còn lại p(x|c), tức phân phối của các điểm dữ liệu trong nhãn c, thường rất khó tính toán vì x là một biến ngẫu nhiên nhiều chiều, cần rất rất nhiều dữ liệu training để có thể xây dựng được phân phối đó. Để giúp cho việc tính toán được đơn giản, người ta thường giả sử một cách đơn giản nhất rằng các thành phần của biến ngẫu nhiên x là độc lập với nhau nếu biết nhãn c. Tức là:

Giả thiết các chiều của dữ liệu độc lập với nhau, nếu biết c, là quá chặt và ít khi tìm được dữ liệu mà các thành phần hoàn toàn độc lập với nhau. Tuy nhiên, giả thiết ngây ngô này lại mang lại những kết quả tốt bất ngờ. Giả thiết về sự độc lập của các chiều dữ liệu này được gọi là Naive Bayes, Cách xác định nhãn của dữ liệu dựa trên giả thiết này có tên là Naive Bayes Classifier (NBC).

Ở bước test, với mỗi điểm dữ liệu x, nhãn của nó sẽ được xác định bằng cách:

Khi d lớn và các xác suất nhỏ, biểu thức ở vế phải sẽ là một số rất nhỏ, khi đó tính toán có thể gặp sai số. Để giải quyết vấn đề này, biểu thức trên thường được viết dưới dạng tương đương như sau:

Việc này không ảnh hưởng tới kết quả vì log là một hàm đồng biến trên tập các số dương.

**Kỹ thuật Laplace Smoothing:**

trong đó: là tổng số lần giá trị thứ i xuất hiện trong nhãn c

là tổng số giá trị kể cả lặp xuất hiện trong nhãn c

Cách tính này có một hạn chế là nếu có một từ mới chưa bao giờ xuất hiện trong class c thì biểu thức sẽ bằng 0, điều này dẫn đến vế phải bằng 0 bất kể các giá trị còn lại có lớn thế nào. Việc này sẽ dẫn đến kết quả không chính xác. Để giải quyết vấn đề này kỹ thuật Laplace Smoothing được áp dụng như sau:

Với là một số dương (thường bằng 0) để tránh trường hợp tử số bằng 0. Mẫu số được cộng với d. để đảm bảo tổng xác suất bằng 1.

Như vậy, mỗi nhãn c sẽ được mô tả bởi bộ các số dương có tổng bằng 1.

**II. Bài toán gắn nhãn chữ số sử dụng Naïve Bayes**

**1. Tóm tắt**

Cho bộ dữ liệu gồm có:

- File testimages bao gồm 1000 ảnh để test

- File testlabel bao gồm 1000 nhãn là nhãn đúng của tập ảnh test

- File trainingimages bao gồm 5000 ảnh để training

- File traininglabel bao gồm 5000 nhãn tương ứng cho tập ảnh training

Mỗi ảnh sẽ được biểu diễn bằng một ma trận 28x28 tương ứng cho một ảnh xám 28x28 pixel với mỗi phần tử là một pixel. Trong đó ký tự ‘ ’ là điểm ảnh trắng, ký tự ‘#’ là điểm ảnh xám còn ký tự ‘+’ là điểm ảnh đen. Các ảnh biểu diễn các số từ 0-9, sử dụng thuật toán Naïve Bayes để huấn luyện mô hình phân lớp nhận diện số từ ảnh.

**2. Kết quả huấn luyện**

**Sử dụng hai đặc trưng:**

Thuật toán sử dụng kỹ thuật Laplace Smoothing với giá trị alpha = 1 ta có kết quả như sau:

- Accuracy: 75.4%

- Ma trận nhầm lẫn sau khi được chuẩn hóa

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | 0.87 | 0. | 0.01 | 0. | 0. | 0.01 | 0.03 | 0. | 0.08 | 0. |
| **1** | 0. | 0.94 | 0. | 0. | 0. | 0. | 0.01 | 0. | 0.05 | 0. |
| **2** | 0.01 | 0.02 | 0.75 | 0.03 | 0.02 | 0.01 | 0.06 | 0.01 | 0.1 | 0. |
| **3** | 0. | 0 | 0 | 0.84 | 0. | 0.01 | 0.02 | 0.04 | 0.02 | 0.07 |
| **4** | 0. | 0 | 0.02 | 0 | 0.67 | 0.01 | 0.04 | 0 | 0.03 | 0.23 |
| **5** | 0.02 | 0.01 | 0.01 | 0.18 | 0.05 | 0.5 | 0.02 | 0.01 | 0.1 | 0.09 |
| **6** | 0.02 | 0.05 | 0.02 | 0 | 0.03 | 0.03 | 0.79 | 0. | 0.04 | 0. |
| **7** | 0 | 0.05 | 0.03 | 0 | 0.04 | 0 | 0. | 0.71 | 0.04 | 0.14 |
| **8** | 0.01 | 0.02 | 0.03 | 0.12 | 0.03 | 0.05 | 0. | 0.01 | 0.64 | 0.1 |
| **9** | 0.01 | 0. | 0. | 0.03 | 0.08 | 0 | 0. | 0.02 | 0.04 | 0.82 |

Sử dụng các giá trị alpha khác nhau ta có kết quả như sau:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **alpha** | 0.2 | 0.4 | 0.6 | 0.8 | 1 | 1.2 | 1.4 | 1.6 | 1.8 | 2 |
| **Acc** | 75.5% | 75.4% | 75.5% | 75.5% | 75.4% | 75.6% | 75.6% | 75.5% | 75.3% | 75.3% |

Thử trộn tập training và chạy lại thuật toán ta có kết quả như sau:

- Accuracy: 75.4%

- Ma trận nhầm lẫn sau khi chuẩn hóa:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | 0.87 | 0. | 0.01 | 0. | 0. | 0.01 | 0.03 | 0. | 0.08 | 0. |
| **1** | 0. | 0.94 | 0. | 0. | 0. | 0. | 0.01 | 0. | 0.05 | 0. |
| **2** | 0.01 | 0.02 | 0.75 | 0.03 | 0.02 | 0.01 | 0.06 | 0.01 | 0.1 | 0. |
| **3** | 0. | 0 | 0 | 0.84 | 0. | 0.01 | 0.02 | 0.04 | 0.02 | 0.07 |
| **4** | 0. | 0 | 0.02 | 0 | 0.67 | 0.01 | 0.04 | 0 | 0.03 | 0.23 |
| **5** | 0.02 | 0.01 | 0.01 | 0.18 | 0.05 | 0.5 | 0.02 | 0.01 | 0.1 | 0.09 |
| **6** | 0.02 | 0.05 | 0.02 | 0 | 0.03 | 0.03 | 0.79 | 0. | 0.04 | 0. |
| **7** | 0 | 0.05 | 0.03 | 0 | 0.04 | 0 | 0. | 0.71 | 0.04 | 0.14 |
| **8** | 0.01 | 0.02 | 0.03 | 0.12 | 0.03 | 0.05 | 0. | 0.01 | 0.64 | 0.1 |
| **9** | 0.01 | 0. | 0. | 0.03 | 0.08 | 0 | 0. | 0.02 | 0.04 | 0.82 |

Nhận xét:

- Mô hình sử dụng Naïve Bayes có độ chính xác khoảng 75% là mô hình có độ chính xác chưa quá cao nhưng có thể chấp nhận được.

- Có thể thấy các nhãn 3, 8 và 9 là các nhãn dễ bị nhầm lẫn nhất có lẽ là do có sự tương đồng trong cách viết các số này.

- Việc trộn tập training là huấn luyện lại không cải thiện độ chính xác của thuật toán

- Việc thử các giá trị alpha khác nhau không làm thay đổi độ chính xác của thuật toá

**Sử dụng ba đặc trưng:**

Thuật toán có kết quả như sau:

- Accuracy: 74.3%

- Ma trận nhầm lẫn chuẩn hóa:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | 0.87 | 0. | 0.01 | 0. | 0. | 0.01 | 0.03 | 0. | 0.08 | 0. |
| **1** | 0. | 0.94 | 0. | 0. | 0. | 0. | 0.01 | 0. | 0.05 | 0. |
| **2** | 0.01 | 0.02 | 0.75 | 0.03 | 0.02 | 0.01 | 0.06 | 0.01 | 0.1 | 0. |
| **3** | 0. | 0 | 0 | 0.84 | 0. | 0.01 | 0.02 | 0.04 | 0.02 | 0.07 |
| **4** | 0. | 0 | 0.02 | 0 | 0.67 | 0.01 | 0.04 | 0 | 0.03 | 0.23 |
| **5** | 0.02 | 0.01 | 0.01 | 0.18 | 0.05 | 0.5 | 0.02 | 0.01 | 0.1 | 0.09 |
| **6** | 0.02 | 0.05 | 0.02 | 0 | 0.03 | 0.03 | 0.79 | 0. | 0.04 | 0. |
| **7** | 0 | 0.05 | 0.03 | 0 | 0.04 | 0 | 0. | 0.71 | 0.04 | 0.14 |
| **8** | 0.01 | 0.02 | 0.03 | 0.12 | 0.03 | 0.05 | 0. | 0.01 | 0.64 | 0.1 |
| **9** | 0.01 | 0. | 0. | 0.03 | 0.08 | 0 | 0. | 0.02 | 0.04 | 0.82 |

Nhận xét:

- Việc sử dụng 3 đặc trưng thay vì 2 không giúp cải thiện độ chính xác trong quá trình phân loại

**B – K-means**

**I. Lý thuyết**

Trong thuật toán K-means clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau. Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở gần nhau trong một không gian Euclidean đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn)

**Chart, scatter chart

Description automatically generated**

Giả sử mỗi cluster có một điểm đại diện (center*)* màu vàng. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

Thuật toán K-Means có thể tóm tắt như sau:

**Đầu vào:** Dữ liệu X và số lượng cluster cần tìm K

**Đầu ra:** Các center M và label vector cho từng điểm dữ liệu Y. Thực hiện qua các bước sau:

1. Xác định cụm K phù hợp với bộ dữ liệu
2. Chọn K điểm bất kỳ làm các center ban đầu.
3. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
4. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
5. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
6. Quay lại bước 2.

Ơ bước 1, để xác định số cụm K phù hợp với bộ dữ liệu của bài toán, chúng ta sử dụng một kỹ thuật là phương pháp Elbow. Phương pháp Elbow là một cách giúp ta lựa chọn được số lượng các cụm phù hợp dựa vào đồ thị trực quan hoá bằng cách nhìn vào sự suy giảm của hàm biến dạng và lựa chọn ra điểm khuỷ tay(elbow point)*.* Điểm khuỷ tay là điểm mà ở đó tốc độ suy giảm của hàm biến dạng sẽ thay đổi nhiều nhất. Tức là kể từ sau vị trí này thì gia tăng thêm số lượng cụm cũng không giúp hàm biến dạng giảm đáng kể. Nếu thuật toán phân chia theo số lượng cụm tại vị trí này sẽ đạt được tính chất phân cụm một cách tổng quát nhất mà không gặp các hiện tượng vị khớp (overfitting)

Ở bước 2 có thể thấy thuật toán sẽ dừng khi việc gán dữ liệu vào từng cluster không thay đổi so với vòng lặp trước đó, để xác định điều này chúng ta sẽ sử dụng các cách như sau:

* Tại 1 vòng lặp: có ít các điểm dữ liệu được gán sang cluster khác hoặc
* Điểm trung tâm (centroid) không thay đổi nhiều
* Giá trị hàm mất mát không thay đổi nhiều:

Trong đó hàm mất mát thường được sử dụng là hàm:

Error =

với d(x, là khoảng cách Euclide giữa điểm m và điểm

**II. Bài toán phân cụm khách hàng sử dụng K-Means**

**1. Tóm tắt**

Cho tập dữ liệu bao gồm thông tin của 200 khách hàng, trong đó mỗi khách hàng có các thông tin như giới tính, tuổi, mức chi tiêu, lương hàng năm.

Sử dụng thuật toán K-Means phân tập dữ liệu thành các cụm khác nhau và nhận xét đặc điểm của mỗi cụm.

**2. Kết quả huấn luyện**

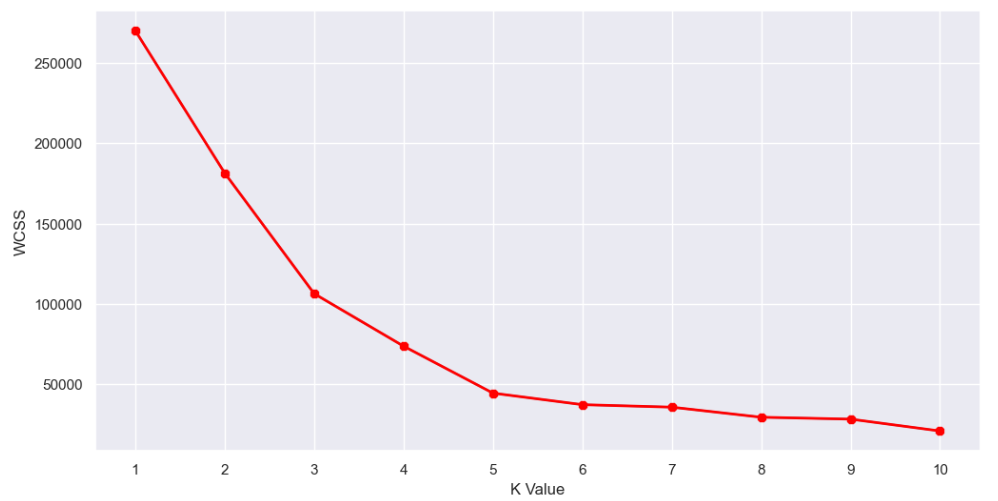
Chọn đặc trưng để huấn luyện:

Cột CustomerID không ảnh hưởng đến mục đích nghiên cứu nên bị loại bỏ

Ta thấy đặc trưng giới tính là một tập phân loại với các giá trị là rời rạc (Male hoặc Female) nên ta không sử dụng đặc trưng này trong thuật toán

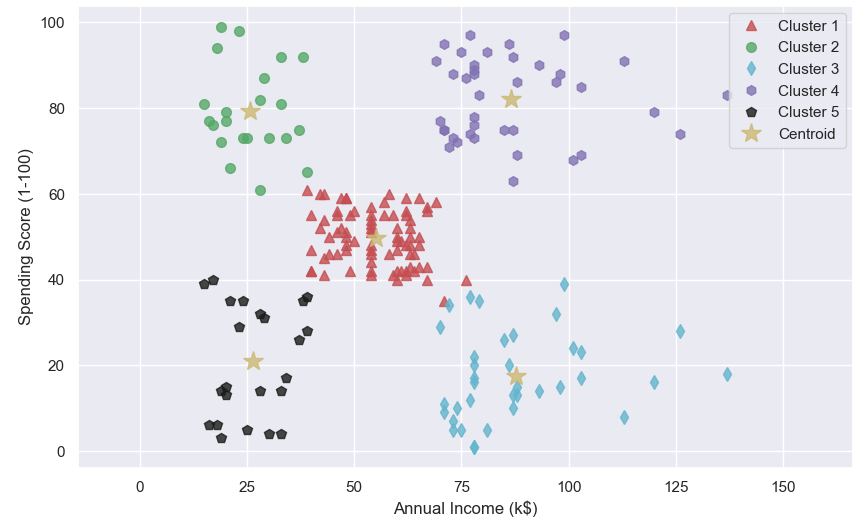
Do vậy ta thực hiện huấn luyện với hai đặc trưng của tập dữ liệu là Annual Income và Spending và sử dụng đặc trưng tuổi để phân tích các cụm

Sử dụng phương pháp Elbow có đồ thị dưới đây:

****

Có thể thấy từ đồ thị rằng 5 là số cụm tối ưu cho bài toán

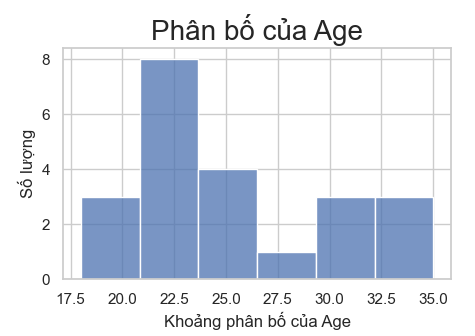
Sau khi thực hiện phân cụm được đồ thị dưới đây:

****

**Phân tích kết quả:**

- Cụm 1 là các khách hành có mức chi tiêu cao mặc dù thu nhập hàng năm lại thấp. Đây là đối tượng rất đáng lưu í và có thể lý giải hành vi này của họ là do họ rất hài lòng với chất lượng thu được từ số tiền bỏ ra hoặc cũng có thể do tính chất công việc hoặc vì một lý do nào đó mà họ mới chi tiêu nhiều.

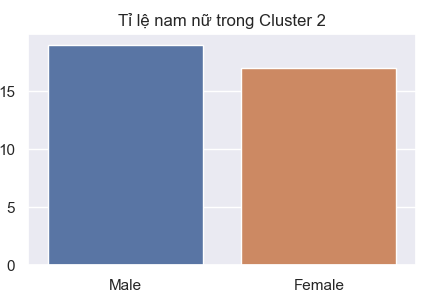
Phân tích kỹ hơn có thể thấy nhóm khách hành có tuổi tương đối trẻ từ 18 đến 35 tuổi



Đây cũng có thể là một cách giải thích cho hành vi của nhóm đối tượng này vì những người trẻ thường có xu hướng chi tiêu phóng khoáng hơn. Đây là đối tượng mà các cửa hàng hay trung tâm thương mại không cần quá để ý tuy nhiên cũng không muốn mất đi.

- Cụm 2 là các khách hàng có mức chi tiêu và mức thu nhập hàng năm ở mức trung bình. Nhìn vào phân bố về độ tuổi của nhóm khách hàng này cũng không thấy điều gì quá đặc biệt nên tuổi không phải một yếu tố quan trọng trong hành vi của nhóm đối tượng này. Đây là đối tượng khách hàng bình thường, họ sẽ chi tiêu cân bằng dựa trên mức thu nhập của bản thân mình. Những khách hàng này không phải là đối tượng tiềm năng khi mua sắm nhưng vẫn đáng để lưu tâm và có thể có một số chương trình để thu hút họ mua nhiều hơn.

Cụm 3 là các khách hàng có mức thu nhập và mức chi tiêu đều cao. Phân bố về tuổi của nhóm đối tượng này có thể thấy nằm trong khoảng từ 28 đến 40 nên có thể thấy rằng đây là độ tuổi mà kinh tế của các khách hàng này ổn định và họ sẵn sàng chi tiêu cho bản thân mình và người xung quanh. Đây cũng là một đối tượng khách hàng tiềm năng do họ là nguồn thu chính của các cửa hàng hay trung tâm thương mại. Những người này thường xuyên ra vào trung tâm thương mại và hài lòng với chất lượng nhận được.

- Cụm 4 là các khách hàng có mức chi tiêu chỉ ở mức thấp tuy nhiên thu nhập hàng năm của họ lại ở mức cao. Phân bố của nhóm này cũng cho thấy rằng độ tuổi không ảnh hưởng quá nhiều đến hành vi của họ. Tuy nhiên xét trong cả 5 nhóm thì đây là nhóm duy nhất có số lượng nam giới lớn hơn nữ giới mặc dù sự chênh lệch là không nhiều. 

Điều này giải thích một phần cho kết quả của đối tượng này do nam giới thường không quan tâm đến việc chi tiêu làm đẹp bản thân bằng nữ giới. Điều này tức là họ có thể chi tiêu nhưng lại quyết định không chi tiêu mua sắm. Đây có thể là đối tượng tiềm năng khi họ hoàn toàn có thể mau sắm. Do vậy các trung tâm thương mại hay của hàng cần phải có nhiều chương trình và tìm cách thu hút nhiều đối tượng khách hàng ở cụm này hơn.

Cụm 5 các khách hàng có mức thu nhập và mức chi tiêu đều thấp. Đây không phải là đối tượng khách hàng tiềm năng do họ không có xu hướng chi tiêu cao và thu nhập của họ cũng không cho phép họ làm thế. Phân bố về độ tuổi của những người này cho thấy nhóm này bao gồm phần lớn là những người cao tuổi trên 40 tuổi.

Chart, histogram

Description automatically generated

Điều này trở nên rất hợp lý do những người ở độ tuổi này có rất ít nhu cầu mua sắm. Những người này có xu hướng tiết kiệm số tiền của mình. Các cửa hàng hay trung tâm thương mại không cần phải để í đến các đối tượng này.