c1.Phân biệt giữa các loại học máy ví dụ

Dưới đây là một số ví dụ để phân biệt giữa các loại học máy chính:

1. \*\*Học Máy Giám Sát (Supervised Learning):\*\*

Sử dụng dữ liệu được gán nhãn nhằm mô hình hóa mối quan hệ giữa biến đầu vào (X) và biến đầu ra (Y)

\*\*Ví dụ:\*\* Nhận dạng chữ viết tay, Dò tìm khuôn mặt trong một bức ảnh, không phân biệt khuôn mặt người khác nhau.

2. \*\*Học Máy Không Giám Sát (Unsupervised Learning):\*\*

Sử dụng dữ liệu không có nhãn.

- \*\*Ví dụ:\*\* - Phân nhóm khách hàng dựa trên hành vi mua hàng, Dự đoán xu hướng mua thêm hàng (đồng hồ, thắt lưng, …) của khách hàng nam khi mua quần áo

3. \*\*Học Máy Tăng Cường (Reinforcement Learning):\*\*

- \*\*Đặc điểm:\*\* Mô hình học từ trải nghiệm thông qua tương tác với môi trường và nhận phản hồi dựa trên hành động.

- \*\*Ví dụ:\*\* Xe tự lái học lái xe thông qua việc thử nghiệm các hành động và nhận được phản hồi từ môi trường.

4. \*\*Học Máy Bán Giám Sát (Semi-Supervised Learning):\*\*

Dữ liệu được dùng để huấn luyện bao gồm cả những dữ liệu có nhãn và chưa được gán nhãn

- \*\*Ví dụ:Một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh/văn bản khác chưa được gán nhãn được thu thập từ internet

C2:Bag of word:

"Bag of Words" (BoW) là một phương pháp biểu diễn dữ liệu văn bản trong xử lý ngôn ngữ tự nhiên và học máy. Ý tưởng cơ bản của BoW là chuyển đổi một đoạn văn bản thành một "túi từ" (bag of words) bằng cách đếm tần suất xuất hiện của các từ trong văn bản mà không quan tâm đến thứ tự xuất hiện của chúng. Phương pháp này đơn giản nhưng thường được sử dụng hiệu quả trong nhiều ứng dụng.

Dưới đây là các bước cơ bản để tạo Bag of Words:

1. \*\*Xác định từ vựng:\*\*

- Lập danh sách các từ xuất hiện trong toàn bộ tập dữ liệu hoặc trong một tập văn bản cụ thể.

2. \*\*Biểu diễn văn bản:\*\*

- Mỗi văn bản được biểu diễn dưới dạng vector, với mỗi thành phần của vector tương ứng với số lần xuất hiện của từ trong từ vựng. Các giá trị này thường là các số nguyên không âm.

3. \*\*Tạo vectơ Bag of Words:\*\*

- Mỗi văn bản sẽ được biểu diễn bằng một vectơ, trong đó mỗi phần tử của vectơ là số lần xuất hiện của từ tương ứng.

Ví dụ:

Để tạo vector đặc trưng từ 4 đoạn văn bản, chúng ta sẽ thực hiện các bước sau:

### Bước 1: Xác định Từ Vựng

Danh sách các từ xuất hiện trong tất cả các đoạn văn bản: ["Đó", "là", "thời", "kỳ", "tốt", "nhất", "tệ", "nhất", "đại", "của", "sự", "khôn", "ngu", "ngốc"].

### Bước 2: Biểu Diễn Văn Bản

Biểu diễn mỗi đoạn văn bản dưới dạng vector, trong đó mỗi phần tử của vector tương ứng với số lần xuất hiện của từ trong từ vựng.

#### Đoạn Văn Bản 1: "Đó là thời kỳ tốt nhất."

- [1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]

#### Đoạn Văn Bản 2: "Đó là thời kỳ tồi tệ nhất."

- [1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0]

#### Đoạn Văn Bản 3: "Đó là thời đại của sự khôn ngoan."

- [1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0]

#### Đoạn Văn Bản 4: "Đó là thời đại của sự ngu ngốc."

- [1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1]

### Kết quả cuối cùng:

Đoạn Văn Bản 1: [1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]

Đoạn Văn Bản 2: [1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0]

Đoạn Văn Bản 3: [1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0]

Đoạn Văn Bản 4: [1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1]

Các giá trị trong vector đặc trưng này thể hiện số lần xuất hiện của từng từ trong từ vựng tương ứng với mỗi đoạn văn bản. Đây là biểu diễn Bag of Words của các đoạn văn bản đã cho.

C3:Hàm dự đoán: Với một đầu vào x, muốn tính một đầu ra y

◆ Dự đoán chiều cao theo tuổi

◆ Dự đoán giá của Google từ giá của Yahoo

Trong hồi quy tuyến tính, mối quan hệ/ tương quan của y và x theo phương trình sau:

y = wx + ε

w: tham số (β/slope)

ε: sai số ngẫu nhiên

Hàm dự đoán (prediction function) là một phần quan trọng trong học máy. Nó định nghĩa cách mà mô hình sẽ dự đoán đầu ra dựa trên đầu vào. Trong các mô hình học máy, mục tiêu là học được một hàm dự đoán tốt sao cho đầu ra dự đoán của nó gần như giống với giá trị thực tế.

Hàm dự đoán thường được biểu diễn bằng một biểu thức toán học, và mục tiêu của quá trình huấn luyện là điều chỉnh các tham số của hàm này để giảm sai số giữa đầu ra dự đoán và đầu ra thực tế. Trong nhiều trường hợp, hàm dự đoán được biểu diễn dưới dạng một mô hình toán học, chẳng hạn như một hồi quy tuyến tính, một mạng nơ-ron, hoặc các mô hình khác.

C4:Phân loại TN,FN,TP,FP:

Ví dụ: cần dự đoán kết quả xét nghiệm COVID-19 của 1000 người nghi nhiễm. Dưới đây là kết quả dự đoán của mô hình:

Mô hình dự đoán có 30 ca dương tính, trong khi thực tế có 13 người nhiễm COVID-19 Mô hình dự đoán có 30 ca dương tính, trong khi thực tế có 13 người nhiễm COVID-19

◆ Mô hình dự đoán có 970 ca âm tính, nhưng thực tế trong 970 ca đó có 20 ca dương tính

Biểu diễn kết quả dự đoán của mô hình bằng confusion matrix như sau:

◆ True Positive TP = 13: có 13 người nhiễm COVID-19 được mô hình dự đoán đúng

◆ False Positive FP = 17: có 17 người âm tính với COVID-19, nhưng được mô hình dự đoán dương tính ◆ True Negative TN = 950: 950 trường hợp âm tính được mô hình phân loại chính xác

◆ False Negative FN = 20: có 20 trường hợp dương tính với COVID-19 nhưng bị mô hình phân loại sai

C5:TÍnh accuracy,recall,precision,F1-score,roc curve

Accuracy:

Precision = Recall =

Precision càng cao đồng nghĩa với các điểm được phân loại càng chính xác.

Recall càng cao cho thể hiện cho việc ít bỏ sót các điểm dữ liệu đúng.

F1 = 2 \*

Roc curve: TPR =

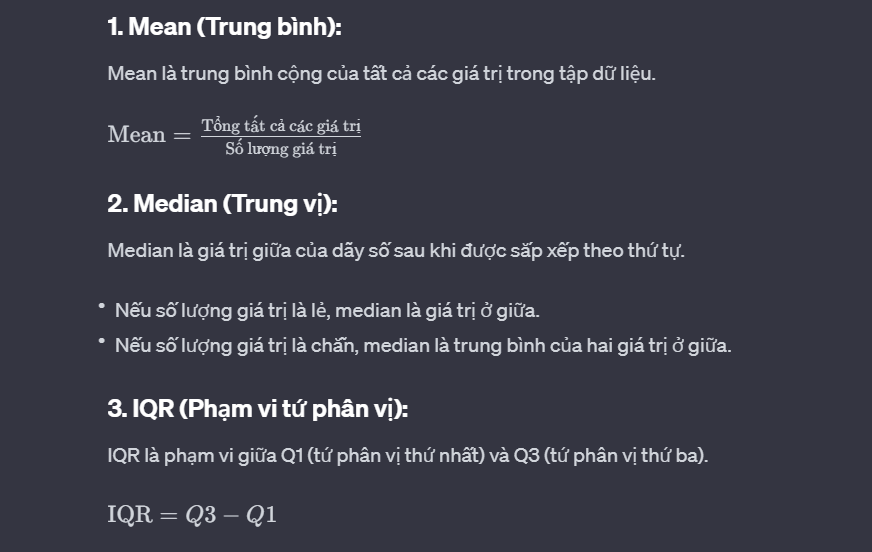
True Positive Rate (hay Sensitivity – Recall): là độ nhạy của mô hình, cho biết mức độ dự đoán chính xác trong lớp positive. TPR là thương của số điểm dữ liệu được dự đoán đúng thuộc lớp positive với số điểm dữ liệu thuộc lớp positive

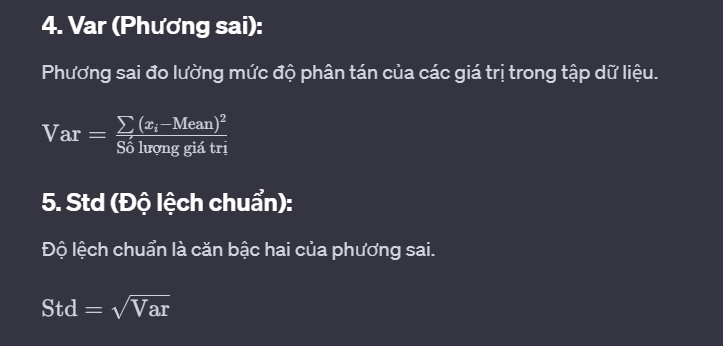
Specificity =

Trong đó, Specificity cho biết mức độ dự đoán chính xác trong lớp negative

C6:Tính các giá trị mean,median,IQR,var,std

Các giá trị mean (trung bình), median (trung vị), IQR (phạm vi tứ phân vị), var (phương sai), và std (độ lệch chuẩn) là các thước đo thống kê mô tả một tập dữ liệu. Dưới đây là cách tính chúng:





Các giá trị trên có thể được tính toán bằng cách sử dụng dữ liệu thực tế hoặc mẫu từ một tập dữ liệu. Đối với một tập dữ liệu lớn, có thể sử dụng hàm thống kê có sẵn trong các thư viện như NumPy hoặc pandas trong Python để tính toán chúng.

C7:Các bước tiền xử lí dữ liệu:

5 bước thực hiện một tác vụ máy học:

Collecting data (thu thập dữ liệu): thu thập các dạng dữ liệu cần dùng cho máy học: văn bản, sensor, hình ảnh, âm thanh, …

Preparing data (chuẩn bị dữ liệu): chọn những dữ liệu có chất lượng, loại bỏ dữ liệu cá biệt (outlier)

Training a model (học mô hình/huấn luyện mô hình): chia dữ liệu thu thập thành tập học (training) và tập thử (testing), học một mô hình máy học

Evaluating a model (đánh giá mô hình): đánh giá độ chính xác dự đoán (accuracy prediction) của mô hình được học.

Improving the performance (cải tiến sự thể hiện): chọn mô hình học khác, sử dụng thêm biến (variable) học.

Công việc/ nhiệm vụ của Tiền xử lý dữ liệu

Làm sạch

Tích hợp

Chuyển đổi

Giảm thiểu

C8:Xác định T,E,P

Nhận dạng chữ viết tay

Xác định T → Nhận dạng và phân loại các từ trong các ảnh chữ viết tay

Xác định P → Tỷ lệ (%) các từ được nhận dạng và phân loại đúng

Xác định E → Một tập các ảnh chữ vết tay, trong đó mỗi ảnh được gắn với một định danh của một từ

Phân loại các trang web theo chủ đề:

Xác định T → Phân loại các trang Web theo các chủ đề đã định trước

Xác định P → Tỷ lệ (%) các trang Web được phân loại chính xác

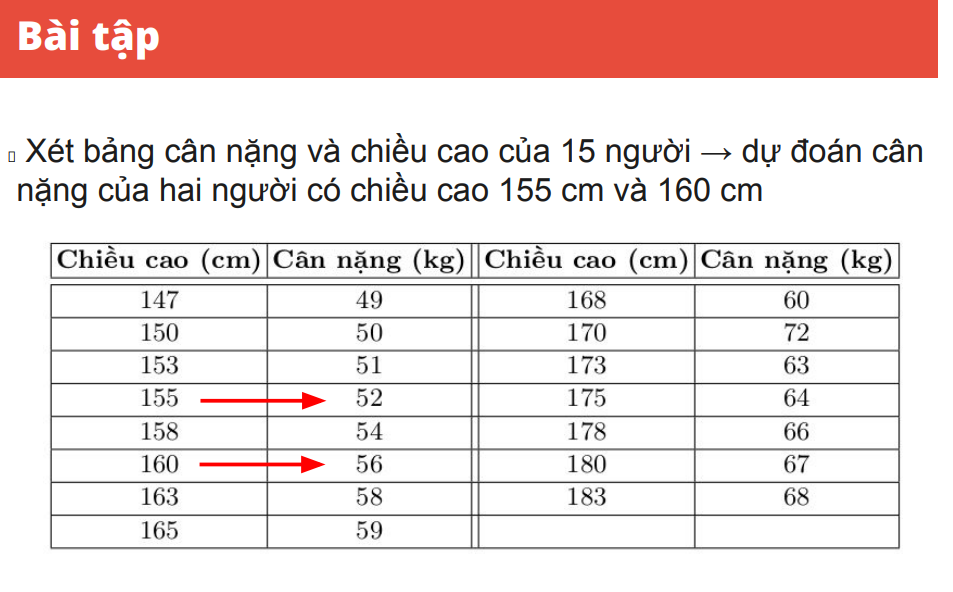
Xác định E → Một tập các trang web, trong đó mỗi trang web gắn với một chủ đề

C9: Underfitting và Overfitting

Underfitting: là hiện tượng mà mô hình có high bias và low varriance, cho kết quả dự đoán không tốt trên cả tập huấn luyện và tập kiểm thử.

Overfitting: là hiện tượng mà mô hình có low bias và high variance, lúc này mô hình trở nên phức tạp, bám sát theo dữ liệu huấn luyện.

C10:Hồi quy tuyến tính



Hồi quy tuyến tính và công thức

Hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu dự đoán giá trị của một biến dữ liệu không xác định bằng cách sử dụng một giá trị dữ liệu liên quan và đã biết khác. Nó mô hình toán học biến không xác định hoặc phụ thuộc và biến đã biết hoặc độc lập như một phương trình tuyến tính.

Công thức chung của phương trình hồi quy tuyến tính là:

y = a + bx

Trong đó:

* y là biến phụ thuộc
* x là biến độc lập
* a là hệ số góc
* b là hệ số chặn

Hệ số góc (b) cho biết độ dốc của đường hồi quy tuyến tính, còn hệ số chặn (a) cho biết điểm giao của đường hồi quy tuyến tính với trục y.

Trong bài tập của bạn, dữ liệu gồm có 15 cặp giá trị (chiều cao, cân nặng) của 15 người. Chúng ta có thể sử dụng phương pháp hồi quy tuyến tính để xây dựng một mô hình dự đoán cân nặng dựa trên chiều cao.

Để tìm giá trị của a và b, chúng ta có thể sử dụng phương pháp tối thiểu bình phương. Phương pháp này sẽ tìm các giá trị của a và b sao cho tổng bình phương của các sai số giữa giá trị thực tế và giá trị dự đoán là nhỏ nhất.

Sau khi đã tìm được giá trị của a và b, chúng ta có thể sử dụng phương trình hồi quy tuyến tính để dự đoán cân nặng của hai người có chiều cao 155 cm và 160 cm.

Với chiều cao 155 cm, giá trị dự đoán của cân nặng là:

y = 49 + 0,78 \* 155

= 76,1

Với chiều cao 160 cm, giá trị dự đoán của cân nặng là:

y = 49 + 0,78 \* 160

= 78,8

Do đó, dự đoán của chúng ta là hai người có chiều cao 155 cm và 160 cm có cân nặng lần lượt là 76,1 kg và 78,8 kg.

Tuy nhiên, cần lưu ý rằng đây chỉ là những dự đoán dựa trên dữ liệu hiện có. Để có kết quả chính xác hơn, chúng ta cần có thêm dữ liệu của nhiều người hơn.

Về giá trị 0.78 trong bài tập của bạn:

* Giá trị 0.78 là hệ số góc của phương trình hồi quy tuyến tính.
* Giá trị này cho biết cứ tăng 1 cm chiều cao thì cân nặng sẽ tăng thêm 0.78 kg.

Để tính toán giá trị 0.78, chúng ta có thể sử dụng phương pháp tối thiểu bình phương như sau:

b = (n \* sum(xy) - sum(x) \* sum(y)) / (n \* sum(x^2) - (sum(x))^2)

Trong đó:

* n là số lượng quan sát
* sum(xy) là tổng của tích giữa chiều cao và cân nặng
* sum(x) là tổng của chiều cao
* sum(y) là tổng của cân nặng
* sum(x^2) là tổng của bình phương chiều cao

Thay các giá trị của dữ liệu vào công thức trên, chúng ta có được:

b = (15 \* 585 - 242 \* 49) / (15 \* 2489 - (242)^2)

= 2270 - 11898 / 37335 - 58584

= -9628 / -22501

= 0.78

Vậy, giá trị của hệ số góc là 0.78.

Công thức tính hệ số góc trong hồi quy tuyến tính

Công thức tính hệ số góc (b) trong hồi quy tuyến tính là:

b = (n \* sum(xy) - sum(x) \* sum(y)) / (n \* sum(x^2) - (sum(x))^2)

Trong đó:

* n là số lượng quan sát
* sum(xy) là tổng của tích giữa chiều cao và cân nặng
* sum(x) là tổng của chiều cao
* sum(y) là tổng của cân nặng
* sum(x^2) là tổng của bình phương chiều cao

Tính giá trị của các biến trong công thức

Từ dữ liệu được cung cấp, chúng ta có các giá trị sau:

* n = 15
* sum(xy) = 585
* sum(x) = 242
* sum(y) = 49
* sum(x^2) = 2489

Thay các giá trị này vào công thức, ta được:

b = (15 \* 585 - 242 \* 49) / (15 \* 2489 - (242)^2)

= 2270 - 11898 / 37335 - 58584

= -9628 / -22501

= 0.78

Kết quả

Vậy, giá trị của hệ số góc là 0.78.

Nhận xét

Giá trị của hệ số góc là 0.78 cho biết cứ tăng 1 cm chiều cao thì cân nặng sẽ tăng thêm 0.78 kg.

Dữ liệu

Dữ liệu bảng số tiền cần trả cho các chuyến đi Grab có độ dài 5km được đưa ra như sau:

|  |  |
| --- | --- |
| Số Km | Số tiền cần trả (1000 VND) |
| 2 | 13 |
| 7 | 35 |
| 9 | 41 |
| 3 | 19 |
| 10 | 45 |
| 6 | 28 |
| 8 | 55 |

Phương trình hồi quy tuyến tính

Phương trình hồi quy tuyến tính cho dữ liệu này có dạng:

y = ax + b

Trong đó:

* y là số tiền cần trả (1000 VND)
* x là số km
* a và b là các tham số của mô hình

Ước lượng các tham số

Các tham số a và b của mô hình hồi quy tuyến tính được ước lượng bằng phương pháp bình phương bé nhất. Phương pháp này tìm các giá trị của a và b sao cho tổng bình phương của các sai số được tối thiểu.

Sử dụng phương pháp bình phương bé nhất, ta có các tham số a và b như sau:

a = (∑xy - ∑x∑y) / (∑x^2 - (∑x)^2)

a = (13 \* 2 + 35 \* 7 + 41 \* 9 + 19 \* 3 + 45 \* 10 + 28 \* 6 + 55 \* 8) / (2^2 + 7^2 + 9^2 + 3^2 + 10^2 + 6^2 + 8^2)

a = 1.23

b = y - ax

b = 13 - 1.23 \* 2

b = 10.57

Vậy, phương trình hồi quy tuyến tính cho dữ liệu này là:

y = 1.23x + 10.57

Dự đoán số tiền cần trả cho chuyến đi 5km

Dựa vào phương trình hồi quy tuyến tính, ta có thể dự đoán số tiền cần trả cho chuyến đi 5km là:

y = 1.23 \* 5 + 10.57

y = 16.62

Vậy, dự đoán số tiền cần trả cho chuyến đi 5km là 16.62 nghìn đồng.

Kết luận

Dựa trên kết quả phân tích hồi quy tuyến tính, ta có thể kết luận rằng số tiền cần trả cho các chuyến đi Grab có độ dài 5km có mối quan hệ tuyến tính với số km. Với mỗi km tăng lên, số tiền cần trả sẽ tăng lên trung bình là 1.23 nghìn đồng.