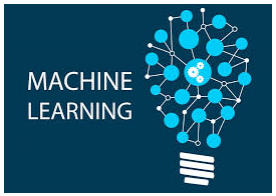
Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

R for Data Science

Bài 16 (bonus): Unsupervised Learning - ***Apriori***

Ngành LT & CSDL

https://csc.edu.vn/lap-trinh-va-csdl/R-Programming-Language-for-Data-Science_190



2021

Nội dung

1. **Association Rule Mining**
2. Giới thiệu Apriori
3. Các ứng dụng
4. Thuật toán
5. Ưu/ khuyết điểm
6. Xây dựng Apriori sử dụng arules

Association Rule Mining

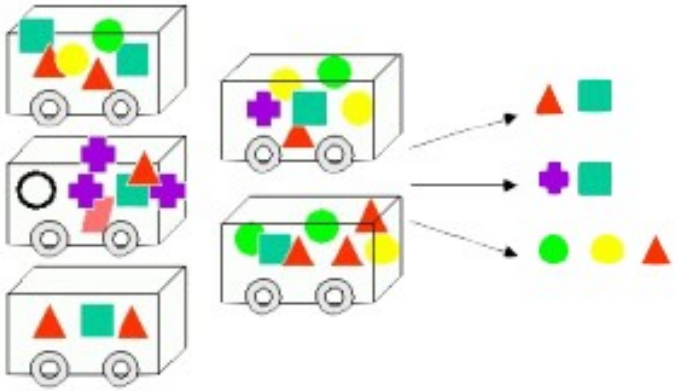
- Association rule learning được sử dụng trong Machine Learning để khám phá mối quan hệ thú vị giữa các biến.

- Association rule mining (khai phá luật kết hợp) là một kỹ thuật để xác định mối quan hệ cơ bản giữa các item khác nhau.

Association Rule Mining

- Ví dụ: Siêu thị là nơi khách hàng có thể mua nhiều mặt hàng. Thông thường, có một mô hình trong những gì khách hàng mua, chẳng hạn như:
  - Các bà mẹ có con nhỏ mua sản phẩm em bé như sữa và tã.
  - Các cô gái có thể mua các mặt hàng trang điểm
  - Các cử nhân vừa tốt nghiệp có thể mua bia, coca và khoai tây chiên
  - …

Association Rule Mining



Association Rule Mining

- Trong thời gian ngắn, các giao dịch liên quan đến một mô hình. Lợi nhuận nhiều hơn có thể được tạo ra nếu mối quan hệ giữa các mặt hàng được mua trong các giao dịch khác nhau được xác định.

- Quá trình xác định mối liên hệ giữa các sản phẩm được gọi là association rule mining, khai phá quy luật kết hợp.

Association Rule Mining

- Ví dụ: Nếu sản phẩm A và sản phẩm B được mua cùng nhau thường xuyên hơn thì có thể thực hiện một số bước để tăng lợi nhuận như sau:
  - A và B có thể được đặt lại gần nhau để khi khách hàng mua một sản phẩm, người ta không phải đi xa để mua sản phẩm khác.
  - Những người mua một trong các sản phẩm có thể được nhắm tới mục tiêu thông qua một chiến dịch quảng cáo để mua một sản phẩm khác.
  - Giảm giá tập thể có thể áp dụng trên các sản phẩm này nếu khách hàng mua cả hai sản phẩm.
  - Cả A và B có thể được đóng gói cùng nhau.

Nội dung

1. Association Rule Mining
2. **Giới thiệu Apriori**
3. Các ứng dụng
4. Thuật toán
5. Ưu/ khuyết điểm
6. Xây dựng Apriori sử dụng arules

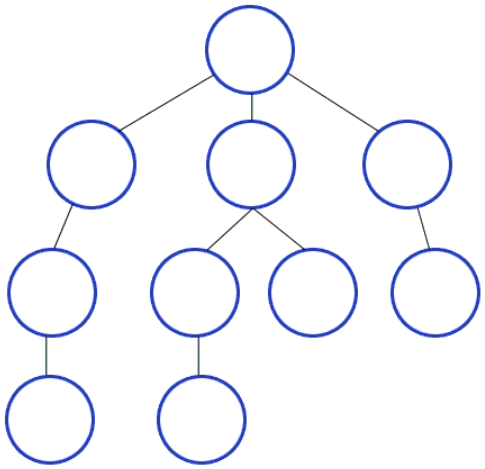
Giới thiệu Apriori

- Apriori nằm trong nhóm **Unsupervised Learning**
- Thuật toán Apriori là một thuật toán phổ biến cho khai thác quy tắc kết hợp và trích xuất các tập hợp thường xuyên với các ứng dụng trong việc học quy tắc kết hợp.
- Nó được thiết kế để hoạt động trên các cơ sở dữ liệu chứa các giao dịch

Giới thiệu Apriori

Đặc điểm

- Sử dụng BFS (Breadth First Search) và Hash tree structure để đếm các bộ item ứng viên một cách hiệu quả
- Trình bày dữ liệu theo định dạng ngang (horizontal data format)



Nội dung

1. Association Rule Mining
2. Giới thiệu Apriori
3. **Các ứng dụng**
4. Thuật toán
5. Ưu/ khuyết điểm
6. Xây dựng Apriori sử dụng arules, efficient-apriori

Các ứng dụng

- Phân tích giỏ hàng của khách hàng mua hàng tại một cửa hàng
- Tìm kiếm các quy tắc trong trong miền điều hướng người dùng web (ví dụ: khách hàng đã truy cập trang web A và trang B cũng đã truy cập trang C)
- Khai thác dữ liệu cháy rừng: được áp dụng để phân tích xác suất và cường độ cháy rừng một cách hiệu quả với dữ liệu cháy rừng. Giúp cho các trang trại rừng có thể dự đoán cháy rừng khi áp dụng các dữ liệu thời tiết.

Các ứng dụng

- Phát hiện phản ứng bất lợi của thuốc trong dữ liệu chăm sóc sức khỏe
- Khám phá tình trạng xã hội của bệnh nhân tiểu đường hoặc một số bệnh lý khác
- Phân tích các yếu tố của sinh viên được nhận vào trường ĐH, CĐ
- Hệ thống đề xuất thương mại điện tử (a mobile e-commerce recommendation system)

Nội dung

1. Association Rule Mining
2. Giới thiệu Apriori
3. Các ứng dụng
4. **Thuật toán**
5. Ưu/ khuyết điểm
6. Xây dựng Apriori sử dụng arules

Thuật toán

Lý thuyết của thuật toán Apriori

- Có 3 thành phần chính của thuật toán Apriori
  - Support
  - Confidence
  - Lift

Thuật toán

❑Xét ví dụ dưới đây:

- Giả sử có 9 giao dịch khách hàng và ta muốn tìm Support, Confidence, Lift cho 2 mục là I1 và I2.

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Thuật toán

- Trong 9 giao dịch có **6** giao dịch có **I1** và **7** giao dịch có **I2**, trong 6 giao dịch có **I1** thì có **4** giao dịch chứa **I2**

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Thuật toán

- **Support:** đề cập đến mức độ phổ biến mặc định của một mục (A) và có thể được tính toán bằng cách tìm số lượng giao dịch có chứa một mục cụ thể đó (A) chia cho tổng số giao dịch.

- Giả sử chúng ta muốn tìm support mục A, tính như sau:

```
Support(A) = (Transactions containing (A))/(Total Transactions)
```

Miền giá trị: [0,1]

- Vậy:
  - Support(I1) = 6/9 = 0.67
  - Support(I2) = 7/9 = 0.78

Thuật toán

- **Confidence:** đề cập đến khả năng một mặt hàng B cũng được mua nếu mặt hàng A được mua. Nó có thể được tính toán bằng cách tìm số lượng giao dịch trong đó A và B được mua lại với nhau, chia cho tổng số giao dịch mà A được mua. Về mặt toán học, nó có thể được biểu diễn như sau:

Confidence(A→B) = (Transactions containing both (A and B))/(Transactions containing A)

Miền giá trị: [0,1]

- Vậy:
  - Confidence(I1->I2) = 4/6 = 0.67

Thuật toán

- Lift(A -> B): đề cập đến sự gia tăng tỷ lệ bán B khi A được bán. Lift(A -> B) có thể được tính bằng cách chia Confidence(A->B) cho Support(B). Về mặt toán học, nó có thể được biểu diễn như sau:

$$Lift(A \rightarrow B) = (Confidence(A \rightarrow B)) / (Support(B))$$

- Vậy:

Miền giá trị: [0,∞]

  - Lift(I1->I2) = Confidence(I1->I2) / Support(I2) = 0.67/0.78 = 0.86

- Về cơ bản, Lift cho biết rằng khả năng mua I1 và I2 với nhau chỉ bằng 0.86 lần so với khả năng mua I2.

Thuật toán

Chú ý:

- Lift = 1 có nghĩa là không có sự liên kết giữa sản phẩm A và B (A và B độc lập, không ảnh hưởng đến nhau)
- Lift > 1 nghĩa là sản phẩm A và B có nhiều khả năng được mua cùng nhau hơn.
- Lift < 1 đề cập đến trường hợp hai sản phẩm khó có thể được mua cùng nhau.

Thuật toán

- Đối với các bộ dữ liệu lớn, có thể có hàng trăm mục trong hàng trăm nghìn giao dịch. Thuật toán Apriori cố gắng trích xuất các quy tắc cho mỗi kết hợp các mục có thể. Quá trình này có thể rất chậm do số lượng kết hợp.

Thuật toán

- Để tăng tốc quá trình, thực hiện các bước sau:
  - Đặt giá trị tối thiểu cho support và confidence. Điều này có nghĩa là ta chỉ quan tâm đến việc tìm kiếm các quy tắc cho các mục có sự tồn tại mặc định nhất định (support) và có giá trị tối thiểu cho sự xuất hiện cùng với các mục khác (confidence).

Thuật toán

- Trích xuất tất cả các tập con có giá trị support cao hơn ngưỡng tối thiểu.
- Chọn tất cả các quy tắc từ các tập con với giá trị confidence cao hơn ngưỡng tối thiểu.
- Sắp xếp các quy tắc theo thứ tự giảm dần của Lift.

Nội dung

1. Association Rule Mining
2. Giới thiệu Apriori
3. Các ứng dụng
4. Thuật toán
5. **Ưu/ khuyết điểm**
6. Xây dựng Apriori sử dụng arules

Ưu/ khuyết điểm

Ưu điểm

- Thuật toán dễ hiểu, dễ triển khai
- Có thể được sử dụng trên tập dữ liệu lớn

Ưu/ khuyết điểm

Khuyết điểm

- Có thể cần phải tìm một số lượng lớn các quy tắc nên việc tính toán tốn kém
- Tính toán hỗ trợ thực hiện trên toàn bộ dữ liệu và được lặp lại thông qua tất cả các giao dịch mỗi lần
- Yêu cầu không gian vùng nhớ lớn khi thực hiện với số lượng lớn các item tạo itemset
- Thuật toán nặng nề khi có quá nhiều feature

Nội dung

1. Association Rule Mining
2. Giới thiệu Apriori
3. Các ứng dụng
4. Thuật toán
5. Ưu/ khuyết điểm
6. **Xây dựng Apriori sử dụng arules**

Xây dựng Apriori

- **Cài đặt và sử dụng các thư viện**
  - install.packages("arules")
  - install.packages("RColorBrewer")

  =>

  - library("devtools")
  - library("arules")
  - library("RColorBrewer")

Xây dựng Apriori

Các bước thực hiện

- Chọn model sẽ sử dụng là: apriori
- Chuẩn bị dữ liệu, chuẩn hóa dữ liệu
- Áp dụng mô hình, tìm quy luật kết hợp
- Sử dụng mô hình hoàn chỉnh để tìm mối liên hệ của một item cụ thể
- Trực quan hóa

Demo: demo_Apriori.ipynb