Chương 1: Tổng quan về khai phá dữ liệu

1.1. Khái niệm cơ bản

1.2. Các kỹ thuật quan trọng

Chương 2: Bài toán dự đoán tin rác (fake news)

2.1. Tiền xử lí dự liệu

2.1. Cây quyết định (Decision Tree)

2.2. Phân loại Naïve Bayes (Naïve Bayes Classifier)

2.3. Hồi quy logistic (Logistic Regression)

Chương 3: Cài đặt chương trình

3.1. Tiền xử lí dữ liệu

3.2. Cây quyết định

3.3. Phân loại Naïve Bayes

3.4. Hồi quy logistic

Chương 1: Tổng quan về khai phá dữ liệu

1. Khái niệm cơ bản

Data mining – khai phá dữ liệu là quá trình phân loại, sắp xếp các tập hợp dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu. Các MCU khai phá dữ liệu cho phép các doanh nghiệp có thể dự đoán được xu hướng tương lai.

Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.

Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo. Một số tính năng chính của Data Mining:

● Dự đoán các mẫu dựa trên xu hướng trong dữ liệu.

● Tính toán dự đoán kết quả

● Tạo thông tin phản hồi để phân tích

● Tập trung vào cơ sở dữ liệu lớn hơn.

● Phân cụm dữ liệu trực quan

2. Các kỹ thuật quan trọng

Về cơ bản, Data Mining hay khai phá dữ liệu là việc xử lý, nhận biết các xu hướng từ các thông tin dữ liệu để có thể đưa ra quyết định hoặc đánh giá. Thông thường, các bạn sẽ thấy có 6 kỹ thuật cốt lõi, được sử dụng nhiều trong việc khai phá dữ liệu:

a. Kỹ thuật phân tích phân loại (Classification Analysis):

Kỹ thuật khai phá dữ liệu đầu tiên là kỹ thuật phân tích phân loại. Đây là kỹ thuật cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước.

Chúng ta thường sử dụng kỹ thuật khai thác dữ liệu này để lấy các thông tin quan trọng từ dữ liệu và siêu dữ liệu. Vì vậy, trong phân tích 5 phân loại, chúng ta cần áp dụng các thuật toán khác nhau tùy thuộc vào mục tiêu sử dụng.

b. Kỹ thuật Association Rule Learning:

Kỹ thuật Association Rule Learning trong khai phá dữ liệu được sử dụng để xác định mối quan hệ giữa các biến khác nhau trong cơ sở dữ liệu. Ngoài ra, nó còn được sử dụng để “giải nén” các mẫu ẩn trong dữ liệu. Association Rule rất hữu ích để kiểm tra, dự đoán hành vi và thường được áp dụng trong ngành bán lẻ.

c. Kỹ thuật phát hiện bất thường (Anomaly or Outlier Detection):

Về cơ bản, kỹ thuật khai phá dữ liệu (Data Mining) này dùng để nhấn mạnh vào việc quan sát các mục dữ liệu trong bộ dữ liệu để tìm ra các tập dữ liệu không khớp với mẫu dự kiến. Bất thường ở đây có thể đề cập đến độ lệch, sự khác thường, các nhiễu và ngoại lệ.

Sự bất thường được xem là khá quan trọng vì nó có thể cung cấp một số thông tin cần thiết. Nó có thể là một dữ liệu khác biệt so với mức trung bình chung trong một tập dữ liệu. Điều này chỉ ra rằng một cái gì đó khác thường đã xảy ra và các nhà phân tích dữ liệu cần chú ý.

d. Kỹ thuật phân tích theo cụm (Clustering Analysis):

“Cụm” có nghĩa là một nhóm các đối tượng dữ liệu. Các đối tượng tương tự nhau thì sẽ nằm trong một cụm. Kết quả là các đối tượng tương tự nhau trong cùng một nhóm. Về cơ bản, kỹ thuật khai phá dữ liệu này thường được ứng dụng để tạo hồ sơ khách hàng. Hoặc trong lĩnh vực Marketing, đây được xem là việc chia phân khúc khách hàng.

e. Kỹ thuật dự báo (prediction):

Trong khai phá dữ liệu, kỹ thuật dự báo được ứng dụng ở một số trường hợp đặc biệt. Nó được sử dụng để khám phá mối quan hệ giữa các biến độc lập và phụ thuộc.

f. Kỹ thuật Sequential Patterns:

Đây là một kỹ thuật quan trọng trong khai phá dữ liệu. Kỹ thuật này giúp tìm cách khám phá các mẫu tương tự.

g. Kỹ thuật Decision Trees:

Decision Trees là một thuật ngữ rất quan trọng trong khai phá dữ liệu. Nó đóng một vai trò quan trọng trong quá trình khai phá dữ liệu bởi vì mô hình này rất dễ hiểu cho người dùng.

Trong kỹ thuật Decision Trees, gốc cây là một câu hỏi đơn giản có nhiều câu trả lời. Ngoài ra, mỗi câu hỏi dẫn đến bộ câu hỏi khác. Và nó sẽ giúp chúng ta xác định dữ liệu. Vì vậy, chúng ta có thể đưa ra quyết định cuối cùng nhờ vào kỹ thuật này.