BTLT - Data Mining - Tuần 1

21280099 - Nguyễn Công Hoài Nam

Ngày 21 tháng 3 năm 2024

1. An analyst collects surveys from different participants about their likes and dislikes. Subsequently, the analyst uploads the data to a database, corrects erroneous or missing entries, and designs a recommendation algorithm on this basis. Which of the following actions represent data collection, data preprocessing, and data analysis?

(a) *Conducting surveys and uploading to database*

Chỉ đạo một cuộc khảo sát và tải lên cơ sở dữ liệu nhằm mục đích thu thập dữ liệu (data collection) nên nó được phân vào bước thu thập dữ liệu

(b) *Correcting missing entries*

Việc sửa lại những giá trị thiếu nằm trong bước làm sạch dữ liệu (data cleaning) và nó là một bước trong tiền xử lý dữ liệu (data preprocessing)

(c) *Designing a recommendation algorithm*

Thiết kế một thuật toán gợi ý là một bước mô hình hóa dữ liệu (data modeling) nhưng ở một khía cạnh rộng hơn, nó có thể được coi là một bước phân tích dữ liệu (data analysis) vì nó liên quan đến việc hiểu sâu hơn về dữ liệu.

2. What is the data type of each of the following kinds of attributes

(a) *Age - Định lượng*

Tuổi là trường giá trị số và có thứ tự tự nhiên giữa chúng nên nó được gọi là dữ liệu số, liên tục hay định lượng

(b) *Salary - Định lượng/Phân loại có thứ tự*

Tương tự như tuổi, lương thường cũng là giá trị số, có thứ tự tự và liên tục hay định lượng.
Ngoài ra, lương cũng có thế là dữ liệu phân loại rời rạc, có thứ tự (ordered catogorical hay oridinal)

(c) *ZIP code - Phân loại không thứ tự*

Mã ZIP là những giá trị rời rạc và không có thứ tự nên nó là kiểu dữ liệu phân loại

(d) *State of residence - Phân loại không thứ tự*

State of residence (nơi/bang cư trú) như mã ZIP là những giá rời rạc, không có thứ bậc nên nó là kiểu dữ liệu phân loại

(e) *Height - Định lượng*

Chiều cao là những giá trị số, có thứ tự, liên tục hay định lượng

(f) *Weight - Định lượng*

Chiều cao là những giá trị số, có thứ tự, liên tục hay định lượng

3. An analysis obtains medical notes from a physician for data mining purposes, and then transforms them into a table containing the medicines prescribed for each patient.

*(**a**) What is the data type of the original data*

Dữ liệu được thu thập từ bác sĩ nhằm mục đích khai phá dữ liệu, nên nó thường là dữ liệu nhiều chiều ở định dạng văn bản. Nó không có bất kỳ mối quan hệ giữa các điểm dữ liệu và thuộc tính của nó. Vì vậy dữ liệu gốc sẽ là kiểu dữ liệu định hướng không phụ thuộc, có thể gồm những đặc trưng cơ bản như tên, giới tính, tuổi, triệu chứng của bệnh nhân,...

*(**b**) What is the data type of the transformed data*

Dữ liệu đã được biến đổi để dùng cho phân tích dữ liệu được dựa vào tổ hợp của mối quan hệ phụ thuộc ngầm và phụ thuộc tường minh giữa các dữ liệu và các thuộc tính tạm thời và thuộc tính không gian. Vì vậy dữ liệu biến đổi là dữ liệu định hướng phụ thuộc vì khi biến đổi các liều thuộc được chỉ định cho mỗi bệnh nhân phải được phân tích bằng cách đánh giá các trường hợp trước cho bộ dữ liệu gốc

*(**c**) What is the process of transforming the data to the new format called*

Quá trình biến đổi dữ liệu sang định dạng mới được gọi là quá trình trích xuất tính năng, làm sạch dữ liệu và chọn lọc tính năng và biến đổi