Nguyễn Trần Gia Kiệt

MSSV: 21522258

Câu 1: Xác định loại của các thuộc tính trong bảng dữ liệu

|  |  |
| --- | --- |
| Thuộc tính | Loại |
| Survival | Binary Asymmetric |
| Pclass | Ordinal |
| Name | Nominal |
| Sex | Binary Symmetric |
| Age | Ratio-scaled |
| SibSp | Ratio-scaled |
| Parch | Ratio-scaled |
| Embarked | Nominal |
| Fare | Ratio-scaled |

Câu 2: Xử lý dữ liệu bị thiếu trong bảng dữ liệu

* Để xử lý dữ liệu bị thiếu (ở cột Age), đầu tiên ta tính trung bình tuổi của những người còn sống và trung bình tuổi của những người tử nạn
  + Avg\_age[survival = ‘1’] = 32
  + Avg\_age[survival = ‘0’] = 24
* Sau đó ta điền số tuổi trung bình này vào các ô bị thiếu dữ liệu, điền theo thuộc tính Survival, nếu Survival là 1 thì điền là 32, nếu Survival là 0 thì điền là 24.

Câu 3: Thực hiện khử nhiễu trên thuộc tính tuổi (Age) bằng kĩ thuật Binning và làm trơn (smoothing)

* Thực hiện khử nhiễu bằng phương pháp chia giỏ, chia thành 3 giỏ với độ rộng lần lượt là (0, 20], (20, 40], (40, 60]
  + Bin 1 (0, 20]: 2, 2, 4, 14, 14, 20
  + Bin 2 (20, 40]: 22, 24, 26, 27, 31, 32, 32, 35, 35, 38, 39
  + Bin 3 (40, 60]: 54, 55, 58
* Thực hiện làm trơn theo bằng trung vị giỏ
  + Bin 1: 9.3, 9.3, 9.3, 9.3, 9.3, 9.3
  + Bin 2: 31, 31, 31, 31, 31, 31, 31, 31, 31, 31, 31
  + Bin 3: 55.67, 55.67, 55.67

Câu 4: Thực hiện rời rạc hóa dữ liệu trên thuộc tính tuổi (Age) thay thế bởi:

* + Bằng các nhãn khái niệm (youth, senior, adult…)
* Children [0, 14]: 2, 2, 4, 14, 14
* Youth [15, 30]: 20, 22, 24, 26, 27
* Adult [30, 60]: 31, 32, 32, 35, 35, 38, 39, 54, 55, 58

Câu 5: Xét các thuộc tính dạng categorical, nhầm tránh biểu diễn sai giá trị thuộc tính khi sử dụng đối với một số thuật toán khác thác dữ liệu.

|  |  |
| --- | --- |
| Embarked | One-hot |
| Q | 0,0,1 |
| S | 0,1,0 |
| C | 1,0,0 |

Câu 6: Khi sử dụng dữ liệu cho các thuật toán phân lớp hoặc gom cụm (k-NN, neural  
networks, k-Means... sẽ được học ở các chương sau) để tránh tình trạng các thuộc  
tính nằm trong vùng giá trị lớn hơn có xu hướng ảnh hưởng đến mô hình nhiều hơn  
các dữ liệu nằm trong vùng giá trị nhỏ (Ví dụ: Tuổi 20, thu nhập 4.000.000). Ta thực  
hiện việc chuẩn hóa các thuộc tính về một vùng giá trị. Sinh viên thực hiện chuẩn  
hóa dữ liệu trên bằng Min-max normalization.

A table of numbers and digits

Description automatically generated