TRƯỜNG ĐẠI HỌC CNTT **TÀI LIỆU HƯỚNG DẪN THỰC HÀNH**

KHOA HỆ THỐNG THÔNG TIN **Môn: Khai thác dữ liệu**

Bài thực hành 1
TIỀN XỬ LÍ DỮ LIỆU

I. Mục tiêu

1. Nhận diện được các các loại dữ liệu sử dụng cho bài toán khai thác dữ liệu.
2. Hiểu và vận dụng các phương pháp tiền xử lí dữ liệu trong các trường hợp: dữ liệu bị thiếu, nhiễu, dữ liệu đặc biệt...
3. Tìm hiểu một số kĩ thuật tiền xử lý dữ liệu đối với các dữ liệu dạng văn bản(text) hoặc dữ liệu theo thời gian

II. Thời gian

1. Thực hành: 5 tiết
2. Bài tập làm thêm: 8 tiết

III. Hướng dẫn chung

Bài 1: Số liệu doanh thu và chi phí quảng cáo của 5 cửa hàng giới thiệu sản phẩm A như sau

Cửa hàng	Doanh thu *(triệu đồng)* y_i	Chi phí quảng cáo *(triệu đồng)* x_i	x_iy_i	x_i^2	y_i^2
A	850	2	1700	4	722500
B	870	5	4350	25	756900
C	880	6	5280	36	774400
D	900	9	8100	81	810000
E	910	13	11830	169	828100
Tổng	**4410**	**35**	**31260**	**315**	**3891900**

Yêu cầu:

Tính hệ số tương quan giữa 2 thuộc tính "Doanh thu" và "Chi phí quảng cáo"

Hướng dẫn: Công thức tính hệ số tương quan giữa 2 thuộc tính

$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y} = b_1 \frac{\sigma_x}{\sigma_y}$$

- Tính:

$$\sigma_x^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \overline{x^2} - (\overline{x})^2$$

$$= 315/5 - (\mathbf{35/5})^2 = 14$$

- Tính

$$b_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x^2}$$

= ((31260/5)-(4410/5*35/5))/14= 5.57

- Tính

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{\overline{y^2} - \overline{y}^2}$$

= (3891900/5) - (4410/5)* (4410/5) = 21.35

- r = $\boldsymbol{b_1} \times \frac{\boldsymbol{\sigma x}}{\boldsymbol{\delta y}} = 5.57 \times \frac{\sqrt{\mathbf{14}}}{\mathbf{21.35}} = 0.976$
- Như vậy mối mối liên hệ giữa chi phí quảng cáo tới doanh thu là mối liên hệ tương quan tuyến tính thuận và rất chặt chẽ

Bài 2:

Cho bảng dữ liệu về các hành khách trên tàu Titanic[1] như sau:

Survival	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Embared
0	3	Braund, Mr. Owen Harris	male	22	1	0	7.25	S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	71.2833	C
1	3	Heikkinen, Miss. Laina	female	26	0	0	7.925	S
1	1	Futrelle, Mrs. JacquesHeath (Lily May Peel)	female	35	1	0	53.1	S
0	3	Allen, Mr. William Henry	male	35	0	0	8.05	S
0	3	Moran, Mr. James	male		0	0	8.4583	Q
0	1	McCarthy, Mr. Timothy J	male	54	0	0	51.8625	S
0	3	Palsson, Master. GostaLeonard	male	2	3	1	21.075	S
1	3	Johnson, Mrs. Oscar W (Elisabeth VilhelminaBerg)	female	27	0	2	11.1333	S
1	2	Nasser, Mrs. Nicholas(Adele Achem)	female	14	1	0	30.0708	C
1	3	Sandstrom, Miss.Marguerite Rut	female	4	1	1	16.7	S
1	1	Bonnell, Miss. Elizabeth	female	58	0	0	26.55	S
0	3	Saundercock, Mr. WilliamHenry	male	20	0	0	8.05	S
0	3	Andersson, Mr. Anders Johan	male	39	1	5	31.275	S
0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	7.8542	S
1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	16	S
0	3	Rice, Master. Eugene	male	2	4	1	29.125	Q
1	2	Williams, Mr. Charles Eugene	male		0	0	13	S
0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	18	S
1	3	Masselmani, Mrs. Fatima	female		0	0	7.225	C

- **Mô tả các thuộc tính:**
  - ✓ Survival: Hành khách sống sót hay không (1: Yes, 0: No). Survival là thuộc tính quyết định.
  - ✓ Pclass: Loại vé (1: 1st, 2: 2nd, 3: 3rd).
  - ✓ Name: Tên.
  - ✓ Sex: Giới tính.
  - ✓ Age: Tuổi.
  - ✓ SibSp: Số vợ hoặc chồng / anh chị em trên tàu.
  - ✓ Parch: Số lượng cha mẹ / con cái trên tàu.
  - ✓ Fare: Giá vé
  - ✓ Embarked: Cổng lên tàu.

Yêu cầu:

1. Xác định loại của các thuộc tính trong bảng dữ liệu (nominal, categorical, binary...)
2. Xử lý dữ liệu bị thiếu trong bảng dữ liệu:

- Đối với dữ liệu số:
  - ✓ Bằng giá trị trung bình hoặc trung vị

  Hướng dẫn: Ví dụ thuộc tính tuổi (Age): xử lý dữ liệu thiếu bằng giá trị trung bình trên thuộc tính tuổi.

  - Tính giá trị trung bình trên tất cả các giá trị có được của thuộc tính tuổi

$$\boldsymbol{trungbinh} = \frac{22 + 38 + 26 + \cdots + 31}{17}$$

  - Dùng giá trị ***trungbinh*** này để điền vào dữ liệu thiếu và làm tương tự với cácthuộc tính dữ liệu **số** còn lại

  - ✓ Bằng giá trị **trung bình** hoặc **trung vị** của các mẫu dữ liệu thuộc cùng một thuộc tính quyết định.

  Hướng dẫn: Tương tự yêu cầu trên nhưng thay vì tính **trung bình** trên toàn bộ dữ liệu có được của thuộc tính tuổi, ta tính trung bình thuộc tính tuổi thuộc

cùng một thuộc tính quyết định, xét trường hợp thuộc tính quyết định **survival** là 1

$$\boldsymbol{trungbinh} = \frac{38 + 26 + 35 + 27 + 14 + 4 + 58 + 55}{8}$$

- Dùng giá trị này điền vào vị trí dữ liệu còn thiếu ở thuộc tính tuổi có **survival** là 1.
- Làm tương tự với dữ liệu thiếu ở thuộc tính tuổi có survival là 0

- Đối với các loại thuộc tính khác (categorical, binary, ...) ta có thể có một số cách xử lý như sau:
  - ✓ Loại bỏ mẫu có giá trị **thiếu** ra khỏi tập dữ liệu.
  - ✓ Điền giá trị **thiếu** bằng giá trị xuất hiện nhiều nhất trong thuộc tính (mode).
  - ✓ Xem thuộc tính thiếu là một loại mới (đối với categorical)
  - ✓ Sử dụng các thuật toán máy học để dự đoán giá trị của dữ liệu

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

   Hướng dẫn:

   - Để **binning** (chia giỏ) ta thấy giá trị thấp nhất là 2 tuổi, lớn nhất là 58 tuổi nên ta chọn 0 và 60 là giá trị bắt đầu và kết thúc cho mỗi giỏ. Chọn độ rộng mỗi giỏ là 20 tađược các giỏ sau: (0, 20], (20, 40], (40, 60]. Tiếp theo, đem dữ liệu phân vào các giỏ đã chia.
   - Để smoothing (làm trơn) ta tính giá trị **trung bình** của các giá trị trong cùng một giỏ, sau đó thay thế giá trị tuổi bằng giá trị trung bình vừa tính theo mỗi giỏ

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

   - Khoảng giá trị (10 – 20, 0 – 10…)
   - Bằng các nhãn khái niệm (youth, senior, adult…)

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

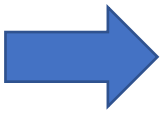
***Ví dụ**: Cổng lên tàu (Embarked) trong bảng dữ liệu trên có 3 giá trị C, Q, S, nếu ta biểudiễn các thuộc tính này là 1, 2, 3 sẽ sai tính chất vì cách biểu diễn này sẽ chứa quan hệcấp bậc 3 > 2 > 1.*

Sinh viên hãy tìm một kiểu biểu diễn khác của dạng dữ liệu này để tránh trường hợptrên.

Hướng dẫn: Để đảm bảo công bằng cho các thuộc tính **categorical**, ta có thể biểu diễn thuộc tính này thành dạng **One-hot** Encoding. Ở dạng biểu diễn này, mỗi giá trị của thuộc tính được biểu diễn bằng một vector với một thành phần có giá trị là 1 và cácthành phần còn lại có giá trị 0. Số lượng thành phần của vector chính là số loại của thuộc tính.

***Ví dụ**: Xét thuộc tính Level trong một mẫu dữ liệu có các loại sau: Easy, Medium, Hardtương ứng theo thứ tự, ta có biểu diễn One-hot như sau*

Level
Easy
Medium
Hard



One-hot
1,0,0
0,1,0
0,0,1

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.**

Hướng dẫn: Đối với Min-max normalization, ta chuyển dữ liệu về khoảng giá trị thuộc vùng 0 – 1 bằng công thức sau

$$v' = \frac{v - minA}{maxA - minA}(new\_maxA - new\_minA) + new\_minA$$

Trong đó

- v' = giá trị sau khi chuyển đổi.
- v = giá trị đang cần chuyển đổi.
- Miền giá trị cũ: v ∈ [minA, maxA]
- Miền giá trị mới: v’ ∈ [new_minA, new_maxA]

Ví dụ*: xét giá trị 7.25* ***của thuộc*** *tính giá vé (****Fare****) áp dụng công thức trên ta tính được giá trị mới*

$$\text{new-value} = \frac{7.25 - 7.225}{71.2833 - 7.225} = 0.00039$$

IV. Thực hành

Bài 1: Tính giá trị **trung bình** và **trung vị** của dãy số sau

12, 17, 15, 19, 20, 51, 22, 21, 24, 18, 23, 16, 33, 36

Bài 2: Xét tuổi của các học viên trong lớp như sau:

36, 27, 21, 22, 24, 32, 15, 16, 19, 20, 33, 27, 22, 35

Hãy khử nhiễu bằng kỹ thuật bôi trơn Binning

Bài 3: Xét dữ liệu về chiều cao và cân nặng của 11 phụ nữ trong độ tuổi 18 – 24 được lựa chọn ngẫu nhiêu như sau:

Chiều cao (cm)	Cân nặng (kg)
175	65
133	67
185	71
163	71
126	66
198	75
153	67
163	70
159	71
151	69
155	69

Yêu cầu: Hãy tính độ tương quan giữa 2 thuộc tính "***Chiều cao***" với "***Cân năng***"

Bài 4:

Để thực hiện chính dịch marketing cho chương trình gửi tiền có kì hạn, một ngân hàng dựa trên bộ dữ liệu người dùng được cho trong mẫu sau để tìm hiểu khả năng người dùng sẽ đăng kí chương trình gửi tiền có kì hạn ở ngân hàng hay không (mẫu được trích từ tập dữ liệu bank marketing[2]):

Job	Marial	Education	Default	Loan	Duration	Cons.price.idx	Subscribe
housemaid	Married	basic.4y	no	no	261	93.994	no
services	Married	high.school	unknown	no	149	93.994	no
blue-collar	Divorced	basic.4y	unknown	no	1575	93.994	yes
admin.	Single	high.school	no	no	338	93.994	no
blue-collar	Single	basic.9y	no	no	179	93.994	no
services	Married	high.school	no	no	1030	93.994	yes
management	Married	high.school	unknown	no	149		no
unemployed		university.degree	unknown	no	424	93.994	no
technician	Married	professional.course	no	no	1623	93.994	yes
services	Divorced	high.school	unknown	no	568	93.994	no
blue-collar	Married	high.school	no	no	1297	93.994	yes
self-employed	Married	basic.9y	no	no	376	93.918	no
entrepreneur	Married	professional.course	no	yes	576	93.994	no
services	Single	high.school	no	no	1059	93.2	yes
technician	Married	basic.9y	no	no	705	93.2	yes
Retired	Married	high.school	no	no	532	93.2	no

Mô tả thuộc tính

- Job: Nghề nghiệp.
- Marial: Tình trạng hôn nhân
- Education: Trình độ học vấn.

- Default: Đã có tính dụng mặc định hay chưa.
- Loan: Đã có khoản vay nào chưa.
- Duration: Thời lượng lần contact cuối cùng, tính trên đơn vị giây
- Cons.price.idx: Chi số giá tiêu dùng

Yêu cầu

Sinh viên thực hiện lại bài tập 1 -> 5 ở phần hướng dẫn chung theo Dữ liệu cho trên. Ở câu 3 và 4 thay thuộc tính tuổi (Age) bằng thuộc tính Duration

V. Bài tập thêm

1. Download hai tập dữ liệu đầy đủ của hai data set trên (Titanic data và Bank Marketing data) được cho ở phần Tài liệu tham khảo. Sau đó, thực hiện giảm các thuộc tính và dòngdữ liệu dư thừa trên tập dữ liệu đầy đủ.
2. Cho mẫu dữ liệu là những bài báo được crawl từ website vnexpress.net được đính kèmtrong tập tin. Thực hiện các thao tác tiền xử lý sau:
   a) Loại bỏ các tag html trong file văn bản
   b) Loại bỏ các dấu cách dư thừa trong câu và cắt các từ trong văn bản theo dấu cách
   c) Loại bỏ các dấu chấm câu trong dữ liệu.
3. Dữ liệu môi trường của một nhà xưởng sản xuất được đo và trình bày trong tập tin đínhkèm *tb_tracking.xlsx*. Thực hiện các thao tác tiền xử lý sau:
   a) Thêm vào tiêu đề các cột theo thứ tự: *id* (định danh của dòng dữ liệu), *device_id* (địnhdanh của thiết bị đo), *co_level* (nồng độ khí CO), *humidity* (độ ẩm), *temperature* (nhiệt độ), *time* (thời điểm đo đạc).
   b) Do điều chỉnh cảm biến nên dữ liệu đo được bắt đầu từ ngày *03/01/2018* mới chínhxác. Những dữ liệu còn lại không có giá trị, yêu cầu sinh viên loại bỏ.
   c) Từ cột *time* sinh viên bóc tách dữ liệu thành thời điểm đo đạc trong ngày tính theo phút, đặt tên là *minutes*.

d) Vẽ đồ thị phân tán (scatter-plots) lần lượt biểu diễn dữ liệu nồng độ khí CO, độ ẩm và nhiệt độ theo cột *minutes* vừa tạo ở yêu cầu trên. Sinh viên tham khảo hướng dẫnvẽ đồ thị phân tán trên R trong phần tài liệu tham khảo

VI. Tài liệu tham khảo

1. *Titanic dataset*, https://data.world/nrippner/titanic-disaster-dataset
2. *Bank marketing dataset*, [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision SupportSystems, Elsevier, 62:22-31, June 2014; https://archive.ics.uci.edu/ml/datasets/bank+marketing
3. *Bài giảng G1: Biểu đồ khoa học với R*, Nguyễn Văn Tuấn; https://www.youtube.com/watch?v=VNExPOfIeLc
4. *Link Google Drive tới các tập dữ liệu*: https://drive.google.com/drive/folders/1rJEK-nBu7VuaJvd7l- iuDm5sumWGdHYD?usp=sharing