

BÀI 2: SỰ TƯƠNG ĐỒNG VÀ CÁC KHOẢNG CÁCH

I. Mục tiêu:

Sau khi thực hành xong, sinh viên nắm được:

- Khoảng cách giữa các điểm trong tập dữ liệu số sử dụng chuẩn L_p với $p = 1, 2, \infty$.
- Sự tương đồng của các điểm trong tập dữ liệu phân loại sử dụng: độ đo Overlap và độ đo tần suất xuất hiện ngược.

II. Tóm tắt lý thuyết:

1. Khoảng cách giữa các điểm trong tập dữ liệu số:

Cho 2 điểm dữ liệu $\bar{X} = (x_1 \dots x_n)$ và $\bar{Y} = (y_1 \dots y_n)$, khoảng cách giữa 2 điểm dữ liệu này dùng chuẩn L_p được xác định như sau:

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

Các trường hợp đặc biệt của chuẩn L_p là

- $p = 1$ (Manhattan)

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^n |x_i - y_i| \right)$$

- $p = 2$ (Euclidean)

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^n |x_i - y_i|^2 \right)^{1/2}$$

- $p = \infty$

$$Dist(\bar{X}, \bar{Y}) = \max_{1 \leq i \leq n} |x_i - y_i|$$

Ví dụ: Cho x, y là tọa độ của 4 điểm như sau:

| | x | y |
|-------|-----|-----|
| p_1 | 3 | 1 |
| p_2 | 1 | 2 |
| p_3 | 2 | 0 |
| p_4 | 2 | 3 |

Ta có các ma trận khoảng cách như sau:

- Với $p = 1$, ta có

$$Dist(p_1, p_2) = |3 - 1| + |1 - 2| = 2 + 1 = 3$$

$$Dist(p_1, p_3) = |3 - 2| + |1 - 0| = 1 + 1 = 2$$

$$Dist(p_1, p_4) = |3 - 2| + |1 - 3| = 1 + 2 = 3$$

$$Dist(p_2, p_3) = |1 - 2| + |2 - 0| = 1 + 2 = 3$$

$$Dist(p_2, p_4) = |1 - 2| + |2 - 3| = 1 + 1 = 2$$

$$Dist(p_3, p_4) = |2 - 2| + |0 - 3| = 0 + 3 = 3$$

Khi đó, ma trận khoảng cách L_1 giữa 4 điểm này là

| | p_1 | p_2 | p_3 | p_4 |
|-------|-------|-------|-------|-------|
| p_1 | 0 | 3 | 2 | 3 |
| p_2 | 3 | 0 | 3 | 2 |
| p_3 | 2 | 3 | 0 | 3 |
| p_4 | 3 | 2 | 3 | 0 |

- Với $p = 2$, ta có

$$Dist(p_1, p_2) = (|3 - 1|^2 + |1 - 2|^2)^{1/2} = \sqrt{5}$$

$$Dist(p_1, p_3) = (|3 - 2|^2 + |1 - 0|^2)^{1/2} = \sqrt{2}$$

$$Dist(p_1, p_4) = (|3 - 2|^2 + |1 - 3|^2)^{1/2} = \sqrt{5}$$

$$Dist(p_2, p_3) = (|1 - 2|^2 + |2 - 0|^2)^{1/2} = \sqrt{5}$$

$$Dist(p_2, p_4) = (|1 - 2|^2 + |2 - 3|^2)^{1/2} = \sqrt{2}$$

$$Dist(p_3, p_4) = (|2 - 2|^2 + |0 - 3|^2)^{1/2} = 3$$

Khi đó, ma trận khoảng cách L_2 (Euclide) giữa 4 điểm này là

| | p_1 | p_2 | p_3 | p_4 |
|-------|------------|------------|------------|------------|
| p_1 | 0 | $\sqrt{5}$ | $\sqrt{2}$ | $\sqrt{5}$ |
| p_2 | $\sqrt{5}$ | 0 | $\sqrt{5}$ | $\sqrt{2}$ |
| p_3 | $\sqrt{2}$ | $\sqrt{5}$ | 0 | 3 |
| p_4 | $\sqrt{5}$ | $\sqrt{2}$ | 3 | 0 |

- Với $p = \infty$, ta có

$$Dist(p_1, p_2) = \max(|3 - 1|, |1 - 2|) = \max(2, 1) = 2$$

$$Dist(p_1, p_3) = \max(|3 - 2|, |1 - 0|) = \max(1, 1) = 1$$

$$Dist(p_1, p_4) = \max(|3 - 2|, |1 - 3|) = \max(1, 2) = 2$$

$$Dist(p_2, p_3) = \max(|1 - 2|, |2 - 0|) = \max(1, 2) = 2$$

$$Dist(p_2, p_4) = \max(|1 - 2|, |2 - 3|) = \max(1, 1) = 1$$

$$Dist(p_3, p_4) = \max(|2 - 2|, |0 - 3|) = \max(0, 3) = 3$$

Khi đó, ma trận khoảng cách L_∞ giữa 4 điểm này là

| | p_1 | p_2 | p_3 | p_4 |
|-------|-------|-------|-------|-------|
| p_1 | 0 | 2 | 1 | 2 |
| p_2 | 2 | 0 | 2 | 1 |
| p_3 | 1 | 2 | 0 | 3 |
| p_4 | 2 | 1 | 3 | 0 |

2. Sự tương đồng giữa các điểm trong tập dữ liệu phân loại:

Cho D là một tập dữ liệu phân loại chứa N đối tượng xác định trên tập d thuộc tính phân loại mà A_K là thuộc tính thứ i . \mathcal{A}_i là thuộc tính A_i lấy n_i giá trị trong tập D .

Xét 2 bản ghi $\bar{X} = (x_1 \dots x_d)$ và $\bar{Y} = (y_1 \dots y_d)$, sự tương đồng đơn giản nhất giữa 2 bản ghi này được xác định như sau

$$Sim(\bar{X}, \bar{Y}) = \sum_{i=1}^d w_i S(x_i, y_i)$$

- a. Độ đo Overlap:** $S(x_i, y_i)$ là sự tương đồng giữa các giá trị thuộc tính x_i, y_i

$$S(x_i, y_i) = \begin{cases} 1 & \text{nếu } x_i = y_i \\ 0 & \text{ngược lại} \end{cases}$$

và $w_i = \frac{1}{d}$ với $i = 1 \dots, d$.

- b. Độ đo tần suất xuất hiện ngược:** Cho $f_i(x)$ là số lần thuộc tính A_i lấy giá trị x trong tập dữ liệu D . Nếu $x \notin \mathcal{A}_i$ thì $f_i(x) = 0$. Cho p_i là xác suất của thuộc tính

A_i lấy giá trị x trong tập dữ liệu D và được cho bởi công thức

$$p_i(x) = \frac{f_i(x)}{N}$$

$$S(x_i, y_i) = \begin{cases} 1 & \text{nếu } x_i = y_i \\ \frac{1}{1 + \log f_i(x_i) \times \log f_i(y_i)} & \text{ngược lại} \end{cases}$$

và $w_i = \frac{1}{d}$ với $i = 1 \dots, d$.

III. Nội dung thực hành:

1. Khoảng cách giữa các điểm trong dữ liệu số

- Download the Ionosphere data set from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/dataset/52/ionosphere>)

The screenshot shows the UCI Machine Learning Repository page for the Ionosphere dataset. The page has a blue header with the dataset name 'Ionosphere' and a 'Donated on 12/31/1988' note. Below the header, there's a description: 'Classification of radar returns from the ionosphere'. A table provides dataset characteristics: Multivariate, Integer, Real, 351 instances, and 34 attributes. On the right, there are buttons for 'DOWNLOAD' and 'CITE', along with citation and view counts. A list of creators is also provided.

| Dataset Characteristics | Subject Area | Associated Tasks |
|-------------------------|--------------|------------------|
| Multivariate | Physical | Classification |

| Attribute Type | # Instances | # Attributes |
|----------------|-------------|--------------|
| Integer, Real | 351 | 34 |

Information

Creators

- V. Sigillito
- S. Wing
- L. Hutton
- K. Baker

- Đọc dữ liệu từ file "ionosphere.data":

```
import pandas as pd
import numpy as np

df = pd.read_csv('D:\\Huynh\\DataMining_Lab\\data\\tuan2\\ionosphere.data', header=None)
print(df)
```

```
   0  1  2      3      4  ...    30    31    32    33    34
0   1  0  0.99539 -0.05889  0.85243  ...  0.42267 -0.54487  0.18641 -0.45300  g
1   1  0  1.00000 -0.18829  0.93035  ... -0.16626 -0.06288 -0.13738 -0.02447  b
2   1  0  1.00000 -0.03365  1.00000  ...  0.60436 -0.24180  0.56045 -0.38238  g
3   1  0  1.00000 -0.45161  1.00000  ...  0.25682  1.00000 -0.32382  1.00000  b
4   1  0  1.00000 -0.02401  0.94140  ... -0.05707 -0.59573 -0.04608 -0.65697  g
..  ..  ..  ..      ..      ..  ...  ..      ..      ..      ..      ..  .
346 1  0  0.83508  0.08298  0.73739  ...  0.86660 -0.10714  0.90546 -0.04307  g
347 1  0  0.95113  0.00419  0.95183  ...  0.94066 -0.00035  0.91483  0.04712  g
348 1  0  0.94701 -0.00034  0.93207  ...  0.92459  0.00442  0.92697 -0.00577  g
349 1  0  0.90608 -0.01657  0.98122  ...  0.96022 -0.03757  0.87403 -0.16243  g
350 1  0  0.84710  0.13533  0.73638  ...  0.75747 -0.06678  0.85764 -0.06151  g

[351 rows x 35 columns]
```

- Xử lý dữ liệu (bỏ cột cuối):

```
df.pop(df.columns[-1])
print(df)
```

```

0      0      1      2      3      ...      30      31      32      33
0      1      0      0.99539 -0.05889 ...      0.42267 -0.54487  0.18641 -0.45300
1      1      0      1.00000 -0.18829 ...     -0.16626 -0.06288 -0.13738 -0.02447
2      1      0      1.00000 -0.03365 ...      0.60436 -0.24180  0.56045 -0.38238
3      1      0      1.00000 -0.45161 ...      0.25682  1.00000 -0.32382  1.00000
4      1      0      1.00000 -0.02401 ...     -0.05707 -0.59573 -0.04608 -0.65697
..     ..     ..     ..     ..     ..     ..     ..     ..     ..
346    1      0      0.83508  0.08298 ...      0.86660 -0.10714  0.90546 -0.04307
347    1      0      0.95113  0.00419 ...      0.94066 -0.00035  0.91483  0.04712
348    1      0      0.94701 -0.00034 ...      0.92459  0.00442  0.92697 -0.00577
349    1      0      0.90608 -0.01657 ...      0.96022 -0.03757  0.87403 -0.16243
350    1      0      0.84710  0.13533 ...      0.75747 -0.06678  0.85764 -0.06151

[351 rows x 34 columns]

```

- Khởi tạo các điểm point1, point2, point3 tương ứng là dòng 0, 1, 2 của array và tính chuẩn $p = 1, 2, \infty$:

```

#array
array = df.values
print(array)

point1 = array[0,:]
point2=array[1,:]
point3 = array[2,:]
#p=1
dist01_2 = np.linalg.norm(point1 - point2,1)
dist01_3 = np.linalg.norm(point1 - point3,1)
#p=2
dist1_2 = np.linalg.norm(point1 - point2)
dist1_3 = np.linalg.norm(point1 - point3)
#p=inf
dist11_2 = np.linalg.norm(point1 - point2,np.inf)
dist11_3 = np.linalg.norm(point1 - point3,np.inf)
#print results
print(dist1_2)
print(dist1_3)
print(dist01_2)
print(dist01_3)
print(dist11_2)
print(dist11_3)

```

```

[[ 1.      0.      0.99539 ... -0.54487  0.18641 -0.453  ]
 [ 1.      0.      1.      ... -0.06288 -0.13738 -0.02447]
 [ 1.      0.      1.      ... -0.2418  0.56045 -0.38238]
 ...
 [ 1.      0.      0.94701 ...  0.00442  0.92697 -0.00577]
 [ 1.      0.      0.90608 ... -0.03757  0.87403 -0.16243]
 [ 1.      0.      0.8471  ... -0.06678  0.85764 -0.06151]]

```

```

2.7763589251571923
1.1697276018372824
13.080950000000001
5.35971
1.12221
0.45772

```

2. Độ đo tương đồng giữa các điểm trong tập dữ liệu phân loại

- Download the KDD Cup Network Intrusion Data Set for the UCI Machine Learning Repository
(<https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>)

- Giải nén file "kddcup.data.gz" và đọc file "kddcup.data.conected"

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import KBinsDiscretizer
from sklearn.preprocessing import StandardScaler

df = pd.read_csv('D:\\Huynh\\DataMining_Lab\\data\\tuan2\\kddcup.data\\kddcup.data.corrected',
                header = None)
print(df)
```

| | 0 | 1 | 2 | 3 | 4 | 5 | ... | 36 | 37 | 38 | 39 | 40 | 41 |
|---------|-----|-----|------|-----|-----|-------|-----|------|-----|------|-----|-----|---------|
| 0 | 0 | tcp | http | SF | 215 | 45076 | ... | 0.00 | 0.0 | 0.00 | 0.0 | 0.0 | normal. |
| 1 | 0 | tcp | http | SF | 162 | 4528 | ... | 0.00 | 0.0 | 0.00 | 0.0 | 0.0 | normal. |
| 2 | 0 | tcp | http | SF | 236 | 1228 | ... | 0.00 | 0.0 | 0.00 | 0.0 | 0.0 | normal. |
| 3 | 0 | tcp | http | SF | 233 | 2032 | ... | 0.00 | 0.0 | 0.00 | 0.0 | 0.0 | normal. |
| 4 | 0 | tcp | http | SF | 239 | 486 | ... | 0.00 | 0.0 | 0.00 | 0.0 | 0.0 | normal. |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4898426 | 0 | tcp | http | SF | 212 | 2288 | ... | 0.05 | 0.0 | 0.01 | 0.0 | 0.0 | normal. |
| 4898427 | 0 | tcp | http | SF | 219 | 236 | ... | 0.05 | 0.0 | 0.01 | 0.0 | 0.0 | normal. |
| 4898428 | 0 | tcp | http | SF | 218 | 3610 | ... | 0.05 | 0.0 | 0.01 | 0.0 | 0.0 | normal. |
| 4898429 | 0 | tcp | http | SF | 219 | 1234 | ... | 0.05 | 0.0 | 0.01 | 0.0 | 0.0 | normal. |
| 4898430 | 0 | tcp | http | SF | 219 | 1098 | ... | 0.05 | 0.0 | 0.01 | 0.0 | 0.0 | normal. |

[4898431 rows x 42 columns]

- Chọn các cột có thuộc tính là 'object' (cột 1, 2, 3 và 41)

```
categorical_columns = df.select_dtypes(include=['object']).columns
print("Categorical attributes:", categorical_columns)
categorical_data = df[categorical_columns]
print("Categorical Data:\n", categorical_data)
```

```
Categorical attributes: Int64Index([1, 2, 3, 41], dtype='int64')
Categorical Data:
   1      2      3      41
0  tcp  http  SF  normal.
1  tcp  http  SF  normal.
2  tcp  http  SF  normal.
3  tcp  http  SF  normal.
4  tcp  http  SF  normal.
...  ...  ...  ..  ...
4898426  tcp  http  SF  normal.
4898427  tcp  http  SF  normal.
4898428  tcp  http  SF  normal.
4898429  tcp  http  SF  normal.
4898430  tcp  http  SF  normal.
[4898431 rows x 4 columns]
```

- Kiểm tra các giá trị giống nhau và loại bỏ những dòng giống nhau này

```

#- Kiểm tra giá trị giống nhau
print("Số lượng dòng giống nhau: ", categorical_data.duplicated().sum())

##loại bỏ các các dòng giống nhau
df1=categorical_data.drop_duplicates()
print("Dữ liệu sau khi xóa những dòng giống nhau: \n",df1)

```

```

Số lượng dòng giống nhau: 4897822
Dữ liệu sau khi xóa những dòng giống nhau:

```

| | 1 | 2 | 3 | 41 |
|---------|-----|----------|------|------------------|
| 0 | tcp | http | SF | normal. |
| 1288 | tcp | http | S2 | normal. |
| 1484 | tcp | http | S1 | normal. |
| 2062 | tcp | smtp | SF | normal. |
| 2067 | udp | domain_u | SF | normal. |
| ... | ... | ... | ... | ... |
| 4574130 | tcp | bgp | REJ | neptune. |
| 4575586 | tcp | z39_50 | REJ | neptune. |
| 4578870 | tcp | smtp | SF | neptune. |
| 4579297 | tcp | other | RSTO | neptune. |
| 4810949 | tcp | telnet | RSTO | buffer_overflow. |

```

[609 rows x 4 columns]

```

3. Yêu cầu:

- Viết hàm tính các chuẩn $p = 1, 2, \infty$ cho 50 dòng đầu tiên của array trong mục 1.
- Tính các láng giềng gần nhất ở mục 2 sử dụng mục 2 với độ đo Overlap và độ đo tần suất xuất hiện ngược.
- Viết file báo cáo.