

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 categorical sử dụng: độ đo thích ứng 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)

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

- $p = \infty$

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

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

a. Độ đo tương đồng:

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

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

với $S(x_i, y_i)$ là sự tương đồng giữa các giá trị thuộc tính x_i, y_i . Lựa chọn đơn giản nhất cho $S(x_i, y_i)$ là

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

- b. Độ đo tần suất xuất hiện ngược:** là sự tổng quát hóa của độ đo tương đồng đơn giản. Độ đo này gắn thêm sự tương đồng giữa các thuộc tính tương đồng của 2 bản ghi bởi 1 hàm nghịch đảo của tần suất của giá trị tương đồng. Do đó, khi $x_i = y_i$ thì sự tương đồng $S(x_i, y_i)$ bằng với tần suất có trọng số nghịch đảo và ngược lại bằng 0. Cho $p_k(x)$ là một tỉ số của các bản ghi mà thuộc tính thứ k lấy giá trị x trong tập dữ liệu. Mặc khác,

$$S(x_i, y_i) = \begin{cases} \frac{1}{p_k(x)^2} & \text{nếu } x_i = y_i \\ 0 & \text{ngược lại} \end{cases}$$

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 includes a search bar, navigation links (Datasets, Contribute Dataset, About Us), and a search results section. The Ionosphere dataset is highlighted with a blue header. It is described as a classification of radar returns from the ionosphere, donated on 12/31/1988. The dataset characteristics are listed: Multivariate, Integer, Real, 351 instances, and 34 attributes. The subject area is Physical, and the associated task is Classification. The creators listed are V. Sigillito, S. Wing, L. Hutton, and K. Baker. There are 47 citations and 33619 views. A 'DOWNLOAD' button is visible in the top right corner.

- 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	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 categorical

- Download the KDD Cup Network Intrusion Data Set for the UCI Machine Learning Repository
(<https://archive.ics.uci.edu/dataset/130/kdd+cup+1999+data>)

archive.ics.uci.edu/dataset/130/kdd+cup+1999+data

UC Irvine Machine Learning Repository

Datasets Contribute Dataset About Us

Search datasets...

KDD Cup 1999 Data

Donated on 12/31/1998

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Computer	Classification

Attribute Type	# Instances	# Attributes
Categorical, Integer	4000000	42

Information

DOWNLOAD

CITE

0 citations
15317 views

Creators

- Salvatore Stolfo
- Wei Fan
- Wenke Lee
- Andreas Prodromidis
- Philip Chan

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

```
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]

- Bỏ cột 1, 2, 3 và 41

```
df= df.drop([1,2,3,41], axis=1)
print(df)
```

	0	4	5	6	7	8	9	...	34	35	36	37	38	39	40
0	0	215	45076	0	0	0	0	...	0.0	0.00	0.00	0.0	0.00	0.0	0.0
1	0	162	4528	0	0	0	0	...	0.0	1.00	0.00	0.0	0.00	0.0	0.0
2	0	236	1228	0	0	0	0	...	0.0	0.50	0.00	0.0	0.00	0.0	0.0
3	0	233	2032	0	0	0	0	...	0.0	0.33	0.00	0.0	0.00	0.0	0.0
4	0	239	486	0	0	0	0	...	0.0	0.25	0.00	0.0	0.00	0.0	0.0
...
4898426	0	212	2288	0	0	0	0	...	0.0	0.33	0.05	0.0	0.01	0.0	0.0
4898427	0	219	236	0	0	0	0	...	0.0	0.25	0.05	0.0	0.01	0.0	0.0
4898428	0	218	3610	0	0	0	0	...	0.0	0.20	0.05	0.0	0.01	0.0	0.0
4898429	0	219	1234	0	0	0	0	...	0.0	0.17	0.05	0.0	0.01	0.0	0.0
4898430	0	219	1098	0	0	0	0	...	0.0	0.14	0.05	0.0	0.01	0.0	0.0

[4898431 rows x 38 columns]

- Chuẩn hóa dữ liệu

```
arr = df.values
scaler = StandardScaler()
arr_scaled = scaler.fit_transform(arr)
```

- Rời rạc hóa dữ liệu để chuyển dữ liệu từ dạng số sang categorical với "uniform" tương ứng với "equi-width":

```
kbins = KBinsDiscretizer(n_bins=10, encode='ordinal', strategy='uniform')
data_equi_width = kbins.fit_transform(arr_scaled)
data_categorical = pd.DataFrame(data_equi_width)
print(data_categorical)
array=data_categorical.values
print(array)
```

```

      0      1      2      3      4      5      6      ...      31      32      33      34      35      36      37
0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0  0.0
1      0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  9.0  0.0  0.0  0.0  0.0  0.0
2      0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  5.0  0.0  0.0  0.0  0.0  0.0
3      0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  3.0  0.0  0.0  0.0  0.0  0.0
4      0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  2.0  0.0  0.0  0.0  0.0  0.0
...
...
4898426 0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  3.0  0.0  0.0  0.0  0.0  0.0
4898427 0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  2.0  0.0  0.0  0.0  0.0  0.0
4898428 0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  2.0  0.0  0.0  0.0  0.0  0.0
4898429 0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  1.0  0.0  0.0  0.0  0.0  0.0
4898430 0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  1.0  0.0  0.0  0.0  0.0  0.0

[4898431 rows x 38 columns]
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

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 cho 100 dòng đầu tiên của array ở mục 2 sử dụng mục 2 với độ đo tương đồng và độ đo tần suất xuất hiện ngược.
- Viết file báo cáo.