Big Data Fundamentals and Applications

Data Preprocessing – Numerical Analysis I

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, , Chung Yuan Christian University, Taoyuan, Taiwan

Outlines

1. Introduction
2. General Concept
3. Some Features of Python
4. Coding Part (plz see ipynb)
    - Part I Basic computation
    - Part II Flow control and loop
    - Part III Define a function
5. Part X Assignment (introduce)

Introduction

- In the beginning, we will cover a brief introduction of Python programming as a review.
- We will introduce the importance and common problems in data preprocessing.
- Basically, data preprocessing can be divided into numerical and text (string) data. To achieve the goals of data cleaning, we will also introduce some useful algorithms with both mathematical and programming perspectives.

General Concept

- Data preprocessing
- When we need to do data preprocessing
  - Incorrect format for further analysis
  - Lack of data information
  - Drop null values
  - Duplicated values
  - Data alignment
  - Normalization
  - Standardization
  - Binarization
  - Centralization

Incorrect Format for Further Analysis

- In data analysis, one of the most common errors occurs in incorrect formats. For example, there is string/ list/ other data format appears in the numerical data column. Hence, the data type of this column will be “object”, not “float” or “int”. It will raise an error message that notifies you cannot do numerical calculation, such as mean and standard deviations, etc.
- In general, the size (columns and rows) of dataset is usually very large that you cannot processing by manual. However, if you do not know all possibilities; therefore, it could be take few steps for data cleaning and corrections.

Lack of Data Information

- This is also a common issue in data analysis. Some datasets do not include column name at the top of data, and they usually provide separately. As a result, you need to combine by yourself.
- In some cases, they do provide the column description at the top of data, but not one-by-one.

```
1  * MH (Multifield Hourly Data Format)
2  * 以 '*' 字元開頭的文字為格式說明
3  * 以 '#' 字元開頭的文字為氣象資料欄位標題列7個字元一組
4  ************************************************************
   **************
5  * 格式說明:
6  * 站碼(stno) 時間(yyyymmddhh)  氣象資料欄位
7  * 1~6        8~17              第18個字元開始每個氣象要素(欄位)
   7個字元
8  * 時間為當地時間(LocalTime)
9  * yyyy:西元年
10 * mm:月
11 * dd:日
12 * hh:時
13 ************************************************************
   **************
14 * 欄位標題說明:
15 * CD01 高雲高度(Km)
16 * CD02 中雲高度(Km)
17 * CD03 低雲高度(Km)
18 * CD04 高雲量(10分量)
19 * CD05 中雲量(10分量)
20 * CD06 低雲量(10分量)
21 * CD07 高雲狀
22 * CD08 中雲狀
23 * CD09 低雲狀
24 * CD10 雲幂高(Km)
25 * CD11 總雲量(10分量)
```

Drop Null Values

- Null value is also a common annoying issue in data analysis and it will appear in your dataset with various forms.
- Typically, standard null in the dataset will be empty (i.e., “”), but in reality, there are four common forms, “NaN”, “null”, -999, or -9999.
- To facilitate the data cleaning process for all null values, both numpy and pandas provide several functions to detect and delete the null values.

Question 1: NA Value

If you see a “NA” in your dataset, does it always represent as a null value?

See an example, there is a table “ISO 3166-1 Alpha-2 code” of each country.

Do you see a “NA”?

What will you do with this table?

Moldova (the Republic of)	The Republic of Moldova	UN member state	MD	MDA	498	ISO 3166-2:MD	.md
Monaco	The Principality of Monaco	UN member state	MC	MCO	492	ISO 3166-2:MC	.mc
Mongolia	Mongolia	UN member state	MN	MNG	496	ISO 3166-2:MN	.mn
Montenegro	Montenegro	UN member state	ME	MNE	499	ISO 3166-2:ME	.me
Montserrat	Montserrat	United Kingdom	MS	MSR	500	ISO 3166-2:MS	.ms
Morocco	The Kingdom of Morocco	UN member state	MA	MAR	504	ISO 3166-2:MA	.ma
Mozambique	The Republic of Mozambique	UN member state	MZ	MOZ	508	ISO 3166-2:MZ	.mz
Myanmar[t]	The Republic of the Union of Myanmar	UN member state	MM	MMR	104	ISO 3166-2:MM	.mm
Namibia	The Republic of Namibia	UN member state	NA	NAM	516	ISO 3166-2:NA	.na
Nauru	The Republic of Nauru	UN member state	NR	NRU	520	ISO 3166-2:NR	.nr
Nepal	The Federal Democratic Republic of Nepal	UN member state	NP	NPL	524	ISO 3166-2:NP	.np
Netherlands (the)	The Kingdom of the Netherlands	UN member state	NL	NLD	528	ISO 3166-2:NL	.nl
New Caledonia	New Caledonia	France	NC	NCL	540	ISO 3166-2:NC	.nc
New Zealand	New Zealand	UN member state	NZ	NZL	554	ISO 3166-2:NZ	.nz
Nicaragua	The Republic of Nicaragua	UN member state	NI	NIC	558	ISO 3166-2:NI	.ni
Niger (the)	The Republic of the Niger	UN member state	NE	NER	562	ISO 3166-2:NE	.ne
Nigeria	The Federal Republic of Nigeria	UN member state	NG	NGA	566	ISO 3166-2:NG	.ng
Niue	Niue	New Zealand	NU	NIU	570	ISO 3166-2:NU	.nu
Norfolk Island	The Territory of Norfolk Island	Australia	NF	NFK	574	ISO 3166-2:NF	.nf

https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

Duplicated Values

- In demographic data, datasets are often constructed manually; therefore, duplicate rows appear in datasets.
- Sometimes, we need to merge/select some rows based on their values. In this case, the meaning of replicated values for your data analysis is the same as duplicated values.
- Pandas proffers you several functions to resolve this problem, such as drop_duplicates and groupby.
- When you choose to remove duplicate rows, you need to be careful which row you want to keep.

Data Alignment

- In most of cases, dataset is divided into two parts: attribute and value data. If you have primary key column, and merging table will be very simple.
- In fact, we always need to merge tables from different data sources; therefore, their data formats are usually inconsistent.
- The most common case is datetime data because the sampling rates during data collection are different.

Question 2: Scaling Problem

- Scaling problem in numerical analysis significantly affects the outcome and interpretation. Here, we use a simple case to demonstrate the scaling problem. In the following regression result, What is the most important parameter?

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Coefficients					
	Estimate	Std. Error	t value	Pr (> | t |)	Sig.
(intercept)	-0.1675	0.1356	-1.2302	0.2345	
x1	12567.1234	12354.3584	-2.1245	0.0246	*
x2	23.4582	46.2875	3.5830	0.0019	**
x3	456722.4406	11237.1235	0.7745	0.4451	
x4	1.4872	0.5465	1.1214	0.2356	

Scaling Problem

- Typically, we have five numerical transformation methods.
  1. Normalization
  2. Standardization
  3. Binarization
  4. Centralization
  5. Feature scaling
- Among these five methods, standardization is the most common method in statistical analysis because of its range of transformed value.

Normalization

Let $p \geqq 1$ be a real number.

The $p - norm$ (= $\ell_p - norm$) of vector $x = (x_1, \ldots, x_n)$ is

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Given $p = 1$

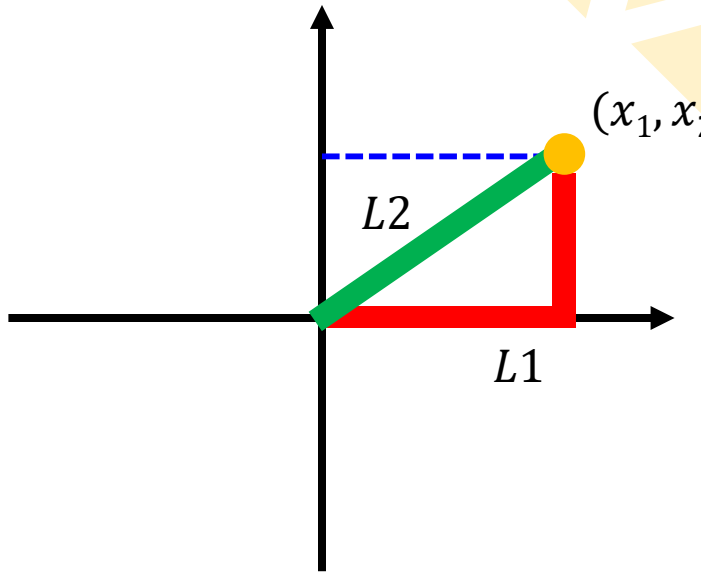
$$L1 = \|x\|_1 = |x_1| + |x_2|$$

Given $p = 2$

$$L2 = \|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2}$$

Given $p = \infty$

$$L\infty = \|x\|_\infty = \max(x)$$



Standardization, Binarization, Centralization, Feature Scaling

Methods	Equations
Standardization	$x' = \frac{x - \mu}{\sigma}$
Binarization	$if\ x \geq threshold; x' = 1$ $else\ x' = 0$
Centralization	$x' = x - \frac{1}{n}\sum_{i=1}^{n} x_i$
Feature scaling	$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$

μ and σ denote mean and standard deviation, respectively.

Question 3: Method Selection

Please give 2-3 examples of each scaling method and explain why they are suitable.

Some Features of Python

- **Python**

1. High-level language
2. Support OOP
3. Row major indexing
4. Start from 0
5. Forced indentation

[#2] Assignment

Part X Assignment: Seismic Risk Map

Plot a seismic risk map based on USGS database (https://earthquake.usgs.gov/earthquakes/search/).
Search conditions:
1) latitude from 21.500 to 25.500 and longitude from 119.200 to 122.200.
2) magnitude from 4.0 to 9.0
3) time range from 2000/01/01 00:00:00 to 2021/12/31 23:59:59.
4) datetime setting with UTC+8.
Download the search results as a csv file.

Here is the requirements of seismic risk map.

1) [25%] Is there any spatial, temporal, or spatiotemporal trend of strong seismic events (M_W > 7.0)?
2) [15%] Is there any regularity of depth from focus to epicenter of all seismic events?
3) [25%] Do shallow seismic events usually have lower energy?
4) [20%] Where are the areas with the most highest seismic risk? (magnitude or energy)
5) [15%] As a data scientist, is there any insightful information that you want to further address?

Please explain your workflow and demonstrate your answers with appropriate visualization tools.

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))