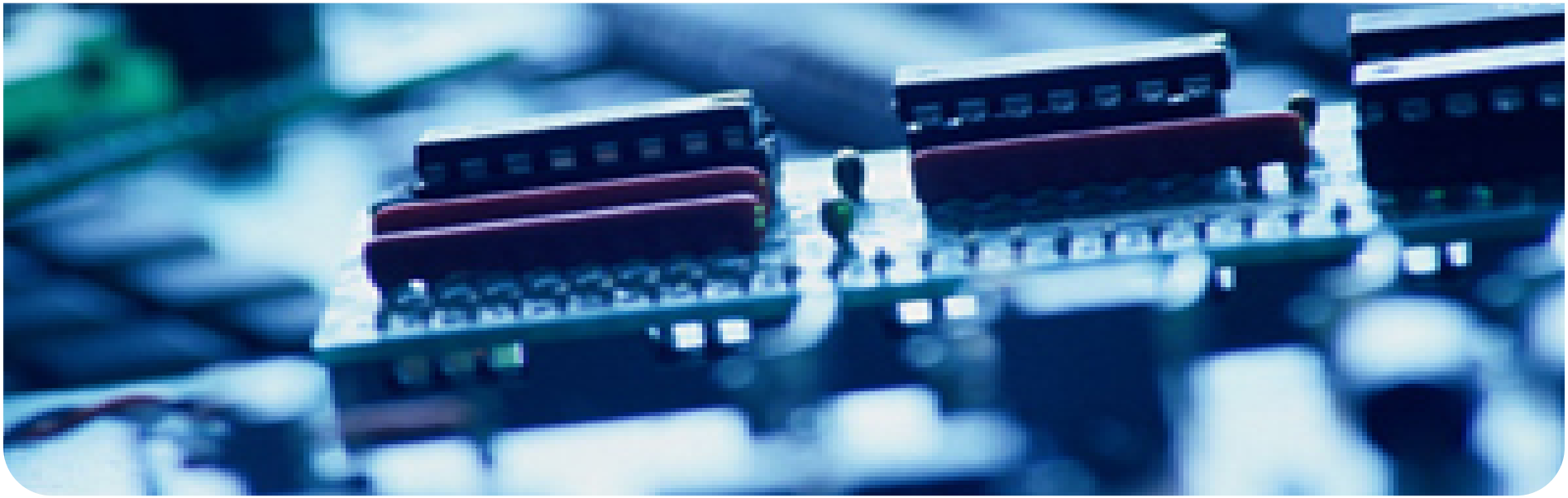
数据挖掘和大数据分析



Outline

① Review

② Big Data

③ Statistic

④ Statistic & DM

主观题 10分

设置

Assignment1 :

1.1 ① What is data mining?

② Is it a simple transformation or application of technology developed from databases, statistics, machine learning, and pattern recognition?

正常使用主观题需2.0以上版本雨课堂

作答

主观题 10分

设置

Assignment2 :

1.4 Present an example where data mining is crucial to the success of a business.

正常使用主观题需2.0以上版本雨课堂

作答

投票　最多可选1项

设置

Did you install Python and Orange3 successfully?

A Yes

B No

>>> import Orange

>>>

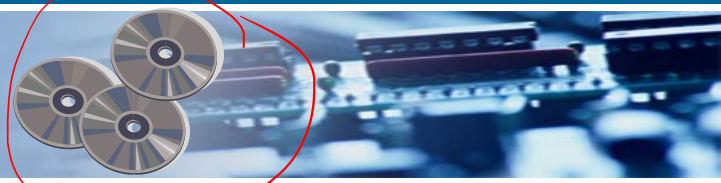
提交

DATA ANALYTICS:

DATA MINING AND BIG DATA

Data



Definition

- "Data are pieces of information that represent the qualitative or quantitative attributes of a variable or set of variables. Data are often viewed as the lowest level of abstraction from which information and knowledge are derived."

Data Types

- Continuous, Binary
- Discrete, String
- Symbolic

Storage

- Physical
- Logical

Major Issues

- Transformation
- Errors and Corruption

单选题 1分

设置

In the following statement, which one is wrong about the unit of computer storage capacity ? ( )

A 1KB < 1MB < 1GB

B Basic Unit is Byte

C A Chinese Character needs a Byte Storage Space

D A Byte can hold a English Character

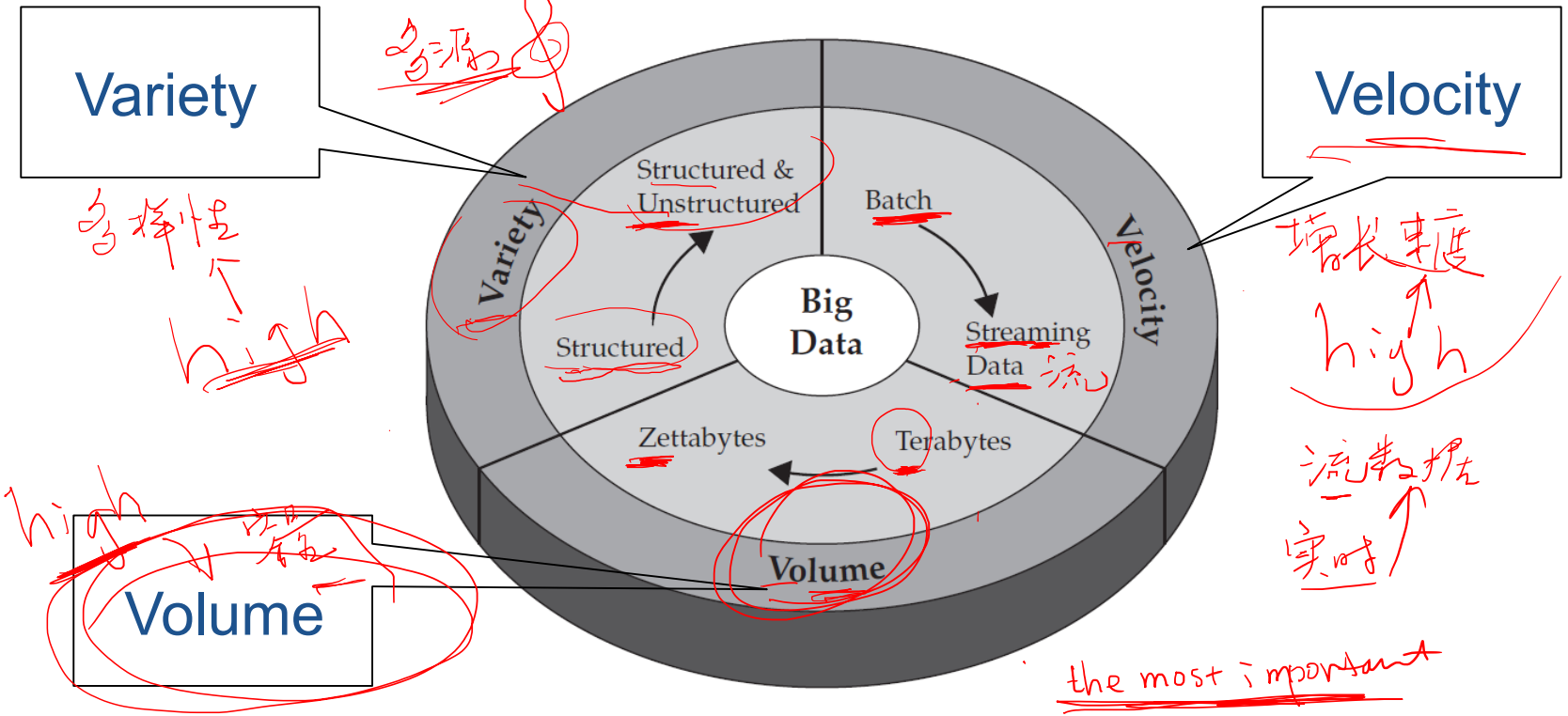
提交

What is Big Data?

- "**Big data** is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." — ***Gartner***

- "**Big data** refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." — ***Mckinsey & Company***

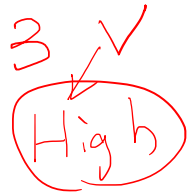
Big Data (3V)


Variety
Velocity
Volume
Structured & Unstructured
Batch
Structured
Streaming Data
Zettabytes
Terabytes
Big Data
Variety
Velocity
Volume

单选题 1分

设置

The Most Significant feature of Big Data is ( )



A Large Data Scale

B Diverse Data Types

C Fast Data Processing

提交

多选题 1分

设置

Big Data Talents need to have ( ) and other core knowledge as a whole.

A Mathematics and Statistics

B Computer related knowledge

C Marxist philosophy knowledge

D Market management knowledge

E Knowledge in specific business areas

提交

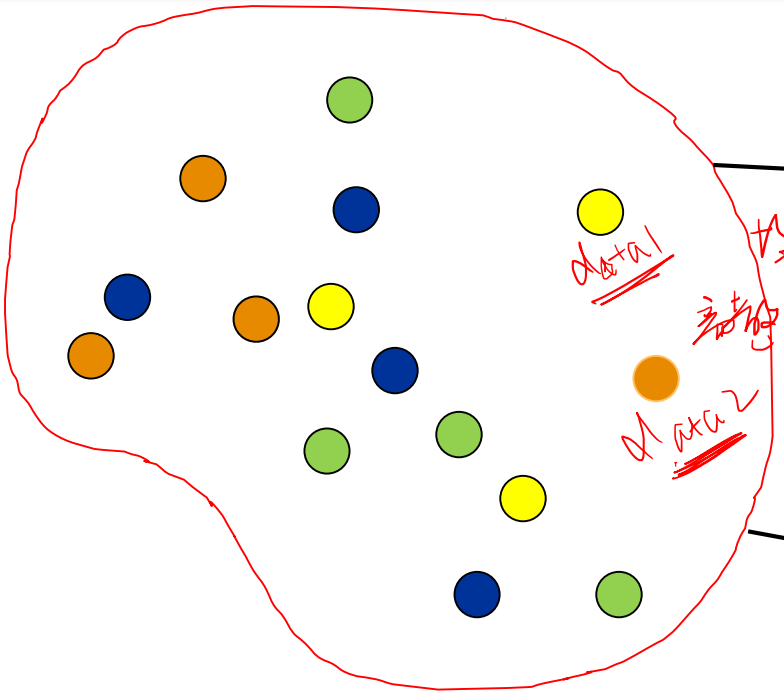
DATA ANALYTICS:

DATA MINING AND BIG DATA

—— Statistic 1

4H

Statics Basic Knowledge



Centralized Features

- Mean number
- Median number
- Mode number

Dispersion Features

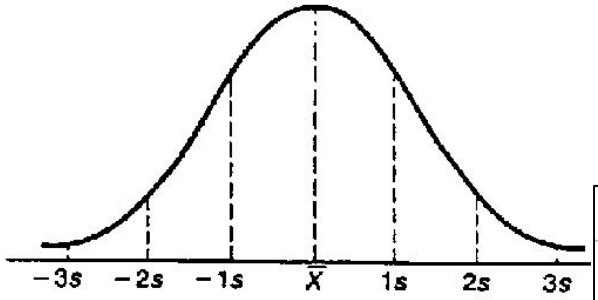
- Standard deviation σ
- variance σ^2

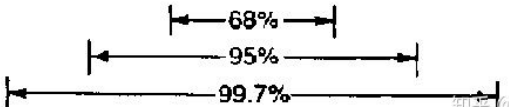
Statics Basic Knowledge

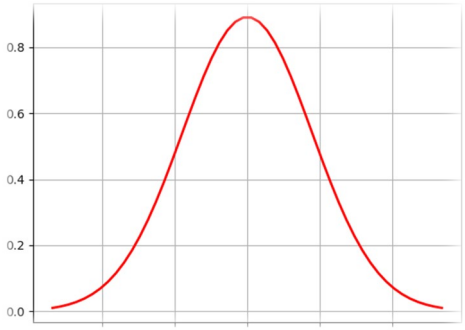
$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The parameter μ in this definition is the ①***mean*** or expectation of the distribution (and also its ②***median*** and ③***mode***).
- The parameter σ is its ① ***standard deviation***; its ②***variance*** is therefore σ^2.

Normal distribution







$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

正态分布.py - C:/Users/鲁凌云/Desktop/数据挖掘课程教案/数据挖掘与数据分析课程/第9周-4月20日/正态分布.py (3.8.2)

File Edit Format Run Options Window Help

```
import numpy as np
import matplotlib.pyplot as plt
import math

u = 0                          # 均值 μ
sig = math.sqrt(0.2)           # 标准差 δ

x = np.linspace(u - 3 * sig, u + 3 * sig, 50)
y_sig = np.exp(-(x - u) ** 2 / (2 * sig ** 2)) / (math.sqrt(2 * math.pi) * sig)
print(x)
print("=" * 20)
print(y_sig)
plt.plot(x, y_sig, "r-", linewidth=2)
plt.grid(True)
plt.show()
```

投票　最多可选1项

设置

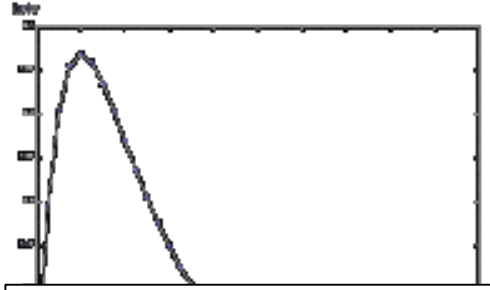
Please coding the task in 6 minutes.

A　Yes

B　No

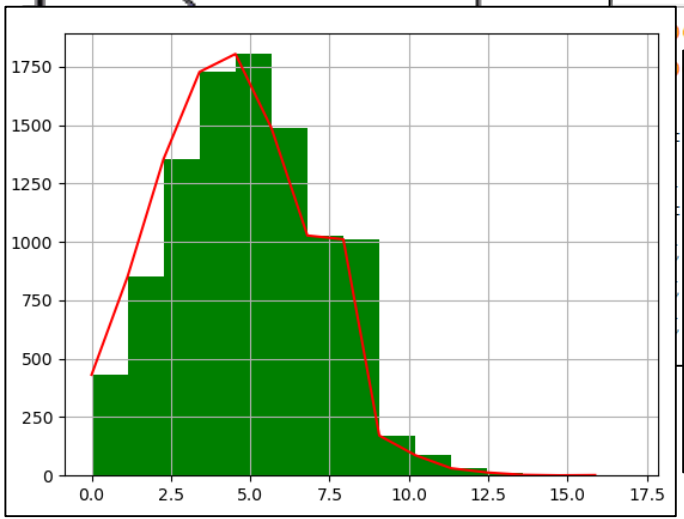
提交

Poisson Distribution



$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},k=0,1,\cdots$$

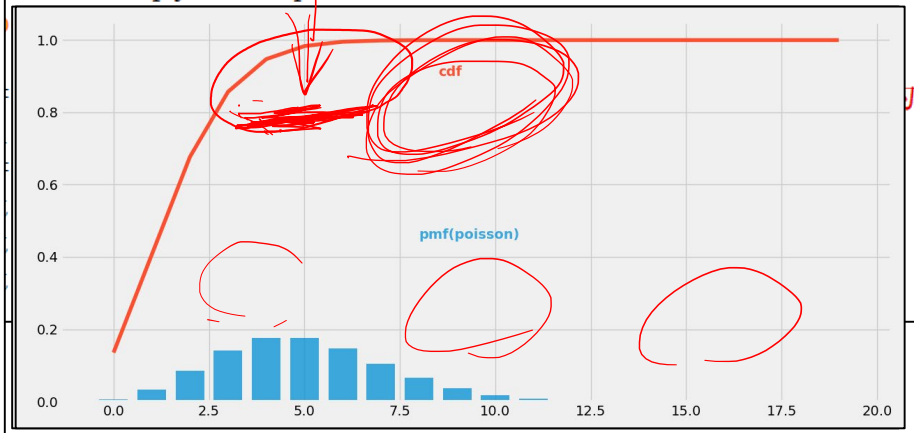
$$EX=\lambda, DX=\lambda$$



泊松分布.py - C:/Users/鲁凌云/Desktop/数据挖掘课程教案/数据挖掘与数据分析课程/第9周-4月20日/泊松分布.py (3.8.2)

File Edit Format Run Options Window Help

ort numpy as np



λ size为k

投票　最多可选1项

设置

Please coding the task in 6 minutes.

A　Yes

B　No

提交

Assignment

- ✓ Two point distribution
- ✓ Binomial distribution
- ✓ Geometric distribution
- ✓ Poisson distribution
- ✓ Uniform distribution
- ✓ Exponential distribution
- ✓ Normal distribution

Formula **Coding** **Figure**

Maximum likelihood estimation

$$f(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

MLE 极大似然估计

Normal distribution is rational, how to estimate the parameters?

$$\mu^{mle}=\arg\max_{\mu} p(x_1,x_2,\ldots x_N\mid\mu,\sigma^2)$$

arg max
arg min

$f(x)\ x\in\{0,1,2\}$ $f(0)=10$; $f(1)=20$ $f(2)=7$ $a=\arg\max f(x)=1$

Maximum likelihood estimation

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_N \mid \mu, \sigma^2)$$

$$\arg\max_{\mu} \prod_{i=1}^{N} p(x_i \mid \mu, \sigma^2)$$

$$\arg\max_{\mu} \sum_{i=1}^{N} \log p(x_i \mid \mu, \sigma^2)$$

$$\arg\max_{\mu} \frac{1}{\sqrt{2\pi}\,\sigma} \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\arg\min_{\mu} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} (x_i - \mu)^2 = -\sum_{i=1}^{N} 2(x_i - \mu)$$

Thus $$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

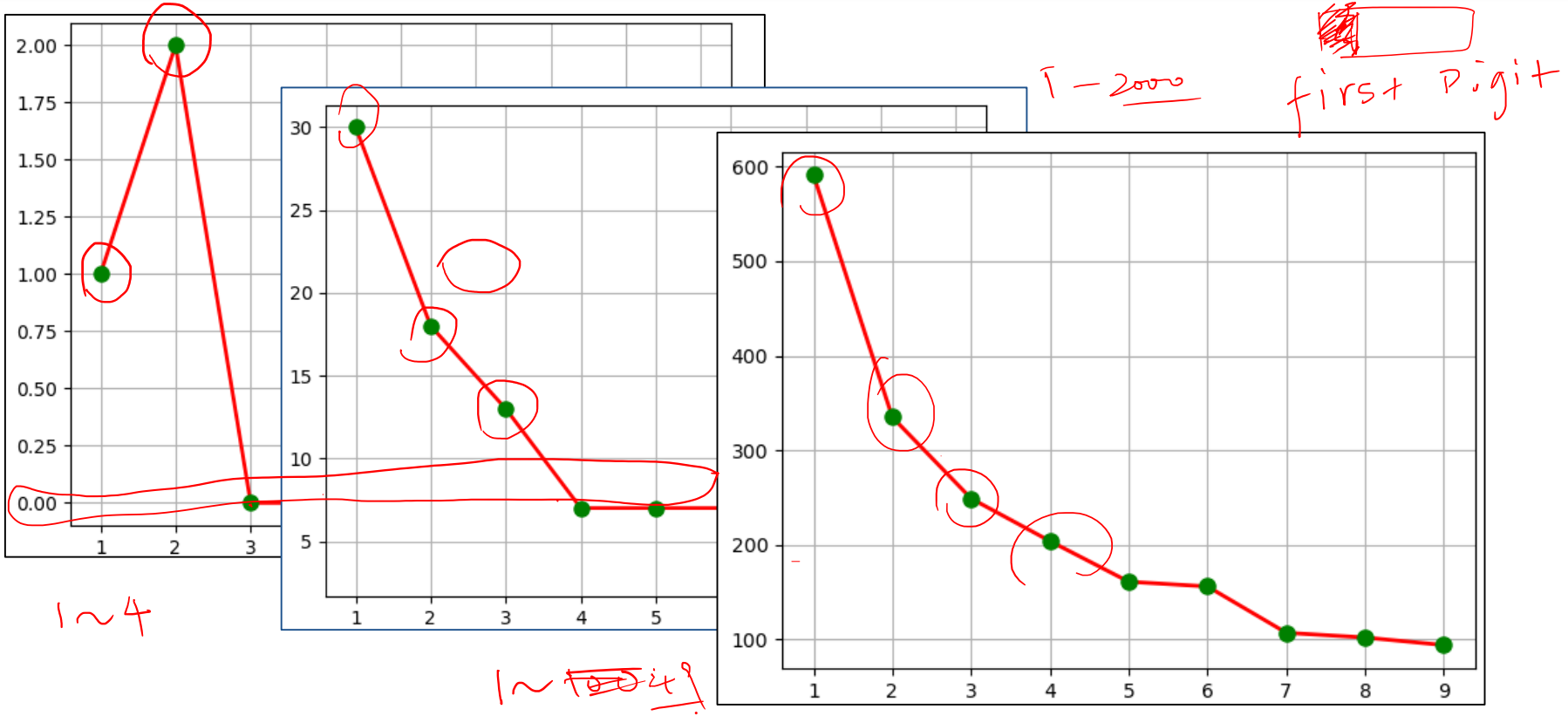
https://mlln.cn/2019/01/24/%E6%95%B0%E5%AD%A6%E5%B0%8F%E7%99%BD%E7%94%A8python%E5%81%9A%E6%9E%81%E5%A4%A7%E4%BC%BC%E7%84%B6%E4%BC%B0%E8%AE%A1MLE/

DM Lab2

Q: Probability of Occurrence of Statistical First Digit

Note: given a positive integer n, the probability of the first digit appearing in the factorial corresponding to all numbers from 1 to n is counted. Then, we can calculate the probability that the first digit is 2, and the probability that the first digit is 3, so we can get a "nine point distribution".

DM Lab2


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
2
3
30
25
20
15
10
5
1
2
3
4
5
600
500
400
300
200
100
1
2
3
4
5
6
7
8
9

投票　最多可选1项

设置

Please coding the task in 15 minutes.

A　Yes

B　No

提交

DM Lab2

test3.py - C:\Users\鲁凌云\Desktop\数据挖掘课程教案\数据挖掘与数据分析课程\第9周-4月20日\test3.py (3.8.2)

File Edit Format Run Options Window Help

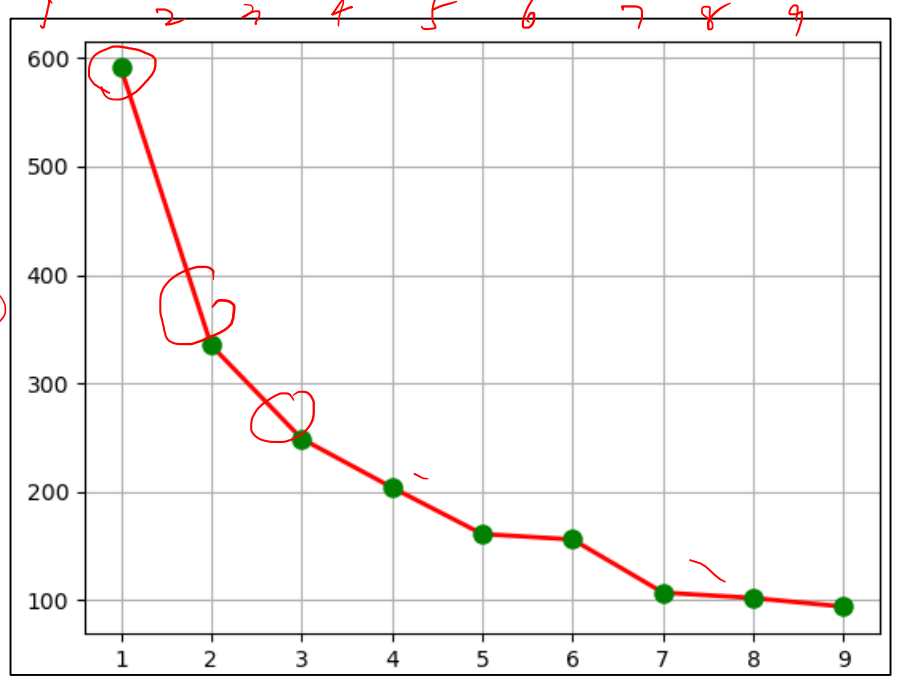
```
import matplotlib.pyplot as plt

def first_num(x):
    while(x>=10):
        x=x//10
    return x

def main():
    n = 1
    frequency = [0]*9
    for i in range(1,2000):
        n*=i
        m = first_num(n)-1
        frequency[m] += 1


    print(frequency)
    x= [1,2,3,4,5,6,7,8,9]
    plt.plot(x,frequency,'r-',lw=2)
    plt.plot(x,frequency,'go',markersize=8)
    plt.xticks(x)
    plt.grid(True)
    plt.show()
```

[591, 335, 249, 204, 161, 156, 107, 102, 94]



Statistic & DM Example – Question

- Suppose we believe that certain groups of evil-doers are meeting occasionally in hotels to plot doing evil.
- We want to find (unrelated) people who have stayed at the same hotel on the same two days.

Statistic & DM Example - Conditions

- 10^9 people being tracked.
- 1000 days.
- 10^5 hotels.
- Each person stays in a hotel 1% of the time (10 days out of 1000).

Will the data mining detect anything suspicious?

Statistic & DM Example – Calculations

p at same hotel *q* at same hotel Same hotel

- **Probability that given persons *p* and *q* will be at the same hotel on given day *d* :**
  - $1/100 \times 1/100 \times 10^{-5} = 10^{-9}$.
- **Probability that *p* and *q* will be at the same hotel on given days d_1 and d_2:**
  - $10^{-9} \times 10^{-9} = 10^{-18}$.
- **Pairs of days:**
  - 5×10^5.

meet

$C_n^m = \frac{n*(n-1)*(n-m+1)}{m!}$

$C_{10^3}^2 = \frac{10^3 * (10^3 - 1)}{2!} = 0.5 \times 10^6 = 5 \times 10^5$

Statistic & DM Example - Calculations

- **Probability that *p* and *q* will be at the same hotel on same two days:**
  - $5\times10^{5} \times 10^{-18} = 5\times10^{-13}$.
- **Pairs of people:**
  - 5×10^{17}.
- **Expected number of "suspicious" pairs of people:**
  - $5\times10^{17} \times 5\times10^{-13} = 250{,}000$.

Statistic & DM Example – Calculations

- **Suppose there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice.**
- **Analysts have to sift through 250,010 candidates to find the 10 real cases.**
  - But how can we improve the scheme?

25010 Scan

Data Mining Moral

- **When looking for a property**

E.g. "two people stayed at the same hotel twice", make sure that the property does not allow so many possibilities that random data will surely produce facts "of interest."

Data Mining Moral

Rhine Paradox

He devised (something like) an experiment where subjects were asked to guess 10 hidden cards – red or blue.

He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right!

Joseph Rhine

Data Mining Moral

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he conclude?
  - ???

Summary

① Review (3 Assignment)

② Big Data (3V)

③ Statistic ★

④ Statistic & DM (DM Lab2/Moral)

贵在坚持！