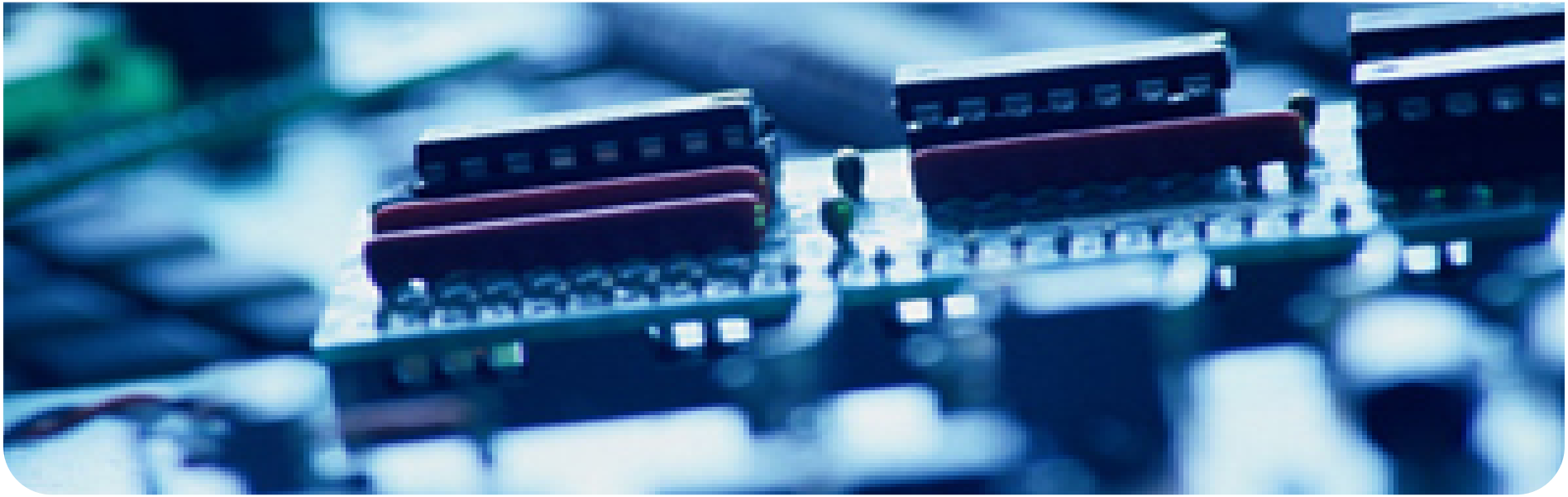
数据挖掘和大数据分析



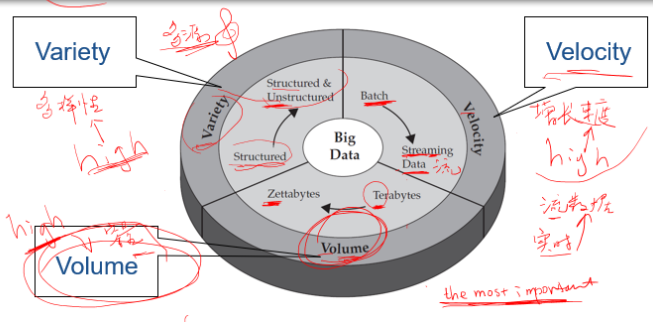
```python
import numpy as np
from sklearn import linear_model
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt

xx, yy = np.meshgrid(np.linspace(0,10,10), np.linspace(0,100,10))
zz = 1.0 * xx + 3.5 * yy + np.random.randint(0,100,(10,10))
X, Z = np.column_stack((xx.flatten(),yy.flatten())), zz.flatten()
regr = linear_model.LinearRegression()
regr.fit(X, Z)
a = regr.coef_
b = regr.intercept_
x = np.array([[5.8, 78.3]])
print(np.sum(a * x) + b)
print(regr.predict(x))
fig = plt.figure()
ax = fig.gca(projection='3d')
ax.scatter(xx, yy, zz)
plt.show()
```

B. What is Big Data? In your answer, address the following:

(a) Describe three features of Big data (**1Points**)

(b) Please explain the Spark, the Hadoop, the MapReduce. (**1Points**)

Big Data (3V)



单选题 1分

The Most Significant feature of Big Data is ( )

3V
High

A. Large Data Scale

B. Diverse Data Types

C. Fast Data Processing

提交

Outline

① Review （Assignments & Process of DM）

② Overview of ML ★

③ Data Cleaning ★

投票　最多可选1项

设置

Do you finish your Homework by yourself?

【1】【2】【3】

A　Yes

B　No

提交

作业清单（4/29、5/4）

【1】 根据下列数据集（数据表存为 csv 格式）建立线性回归模型。

No	square_feet	price
1	150	6450
2	200	7450
3	250	8450
4	300	9450
5	350	11450
6	400	15450
7	600	18450

(1) 预测面积为 1000 平方英尺的房子价格。

要求：完成 2 遍，第 1 遍可以参考课堂笔记、查阅网络资料等方式完成；第 2 遍不参考任何辅助方式，限定 15 分钟内独立编写代码，完成此回归模型。

(2) 建立多元回顾模型。至少增加 2 项房子价格的特征，例如：地段、新旧等因素。

(3) 将（1）和（2）整理成实验报告。5 月 6 日上课检

投票 最多可选1项

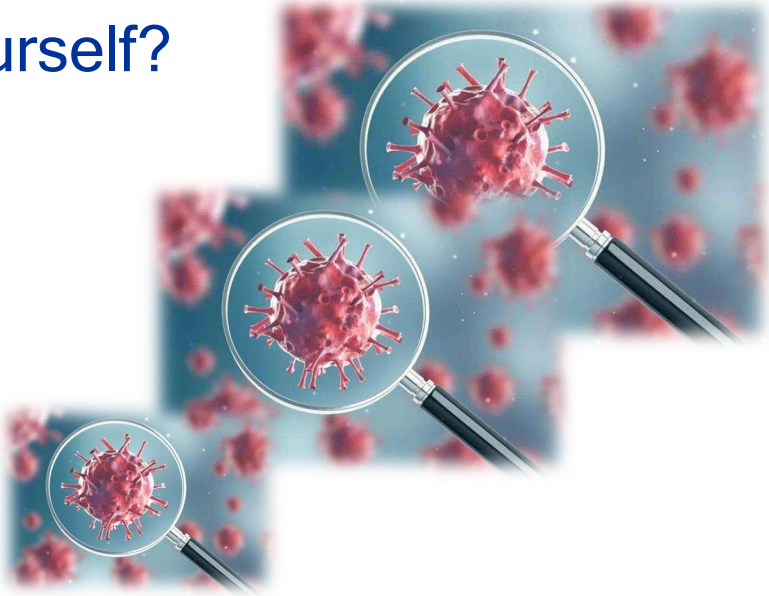
设置

Do you finish your Homework by yourself?

【4】

A Yes

B No



提交

Review Model of Data Mining

① **Data Cleaning** 数据清理

② Data Integration 数据集成

Data Preparation 数据预处理

③ Data Selection 数据选择 — Feature Selection 特征提取

Data Exploration 数据探索

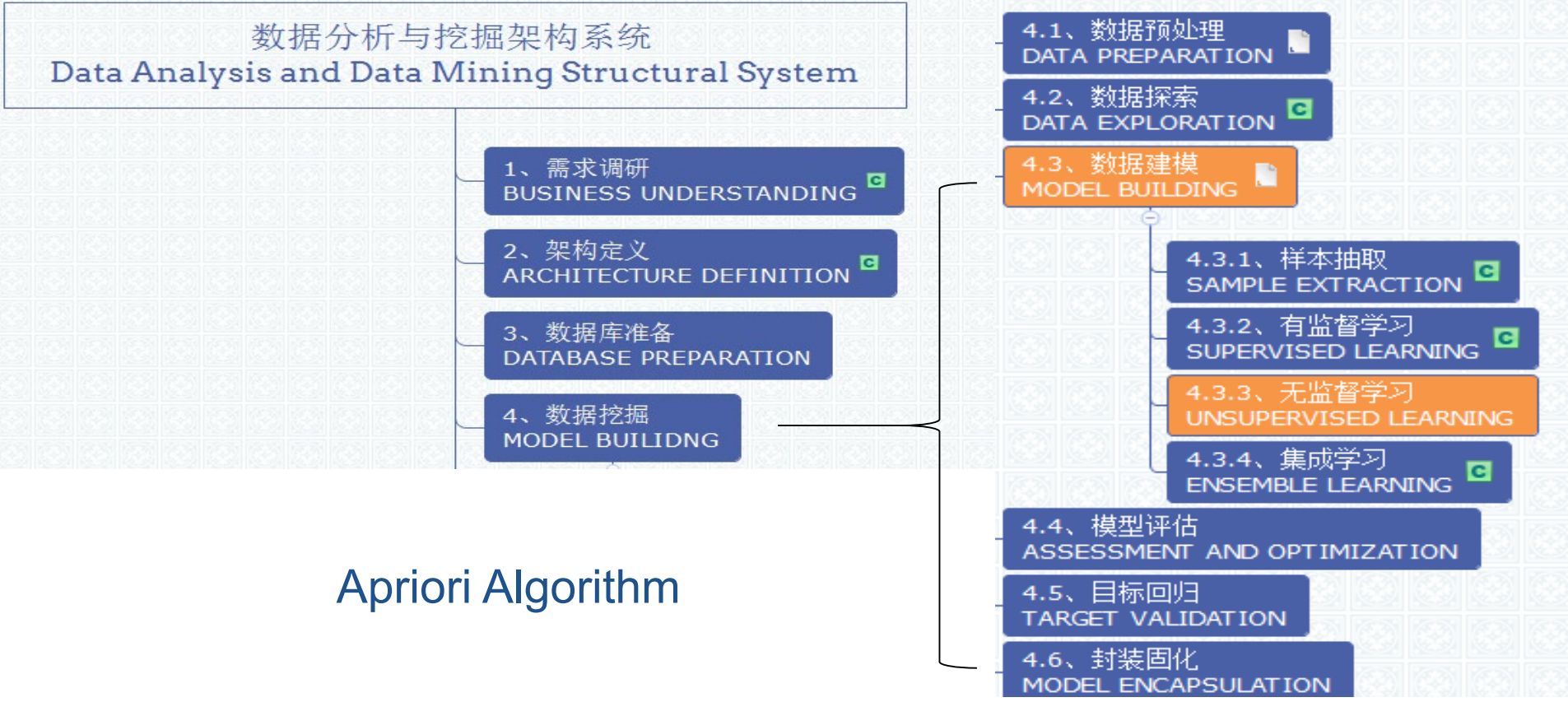
④ Data Transformation 数据变换 — Normalization 归一化

⑤ Data Mining Method 挖掘方法 — Data Model 数据模型

⑥ Pattern Assessment 模式评估

⑦ Knowledge Representation 知识表示

Data Mining



Apriori Algorithm

Data Mining

数据分析与挖掘架构系统
Data Analysis and Data Mining Structural System

- 1、需求调研 BUSINESS UNDERSTANDING
- 2、架构定义 ARCHITECTURE DEFINITION
- 3、数据库准备 DATABASE PREPARATION
- 4、数据挖掘 MODEL BUILIDNG
  - 4.1、数据预处理 DATA PREPARATION
  - 4.2、数据探索 DATA EXPLORATION
  - 4.3、数据建模 MODEL BUILDING
    - 4.3.1、样本抽取 SAMPLE EXTRACTION
    - 4.3.2、有监督学习 SUPERVISED LEARNING
    - 4.3.3、无监督学习 UNSUPERVISED LEARNING
    - 4.3.4、集成学习 ENSEMBLE LEARNING
  - 4.4、模型评估 ASSESSMENT AND OPTIMIZATION
  - 4.5、目标回归 TARGET VALIDATION
  - 4.6、封装固化 MODEL ENCAPSULATION
- 5、测试评审 TESTING
- 6、上线部署 DEPLOYMENT
- 7、监控测评 MONITORING

Data Mining

WAL★MART

Beer←→Paper Diaper

↓

Beer→←Paper Diaper

Data Mining

可以通过对交易数据的分析可能得出"86%买'啤酒'的人同时也买'尿布'"这样一条"啤酒"和"尿布"之间的**关联规则。**

Statistics Machine Learning Database

Apriori Algorithm

投票 最多可选1项

设置

信用卡公司可以将持卡人的信誉度分类为：良好、普通和较差三类。分类分析通过对这些数据类的分析给出一个信誉等级的显式模型：“信誉良好的持卡人是年收入在30000元到50000元之间，年龄在30至45岁之间，居住面积达90M2 左右的人”。这样对于一个新的持卡人，就可以根据他的特征 预测 其信誉度。

Do you think we might build a data mining model?

A Yes

B No

提交

DATA ANALYTICS:

DATA MINING AND BIG DATA

—— Machine Learning 1

Overview of ML

ST Method = Model + Strategy + Algorithm

ML Method = Algorithm ⇨ Model ⇨ Strategy

Supervised Learning

√

Unsupervised learning

x

Annotation /Tagging?

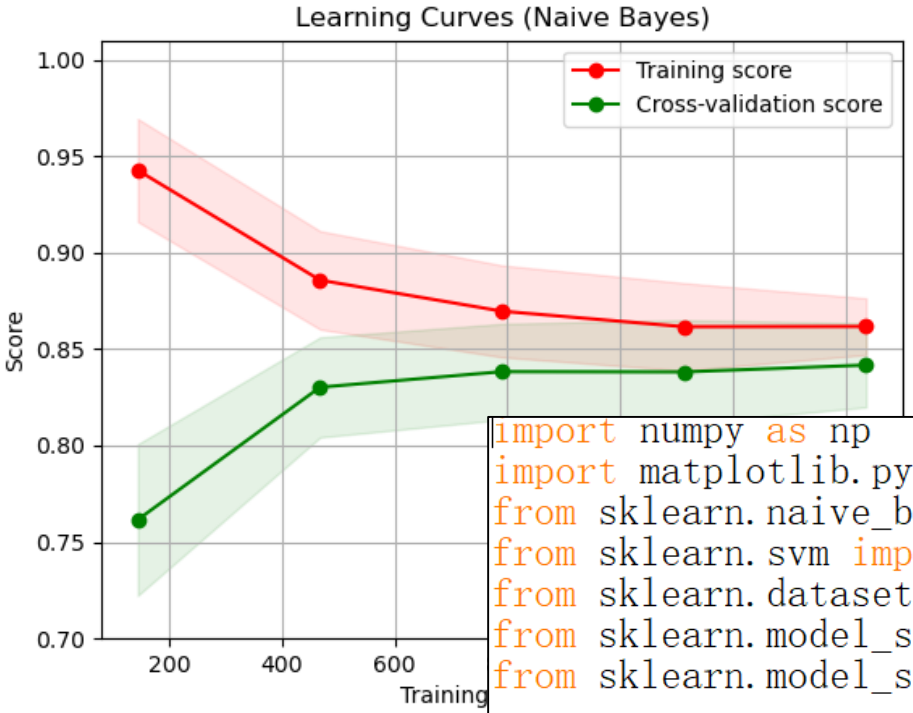
半监督学习

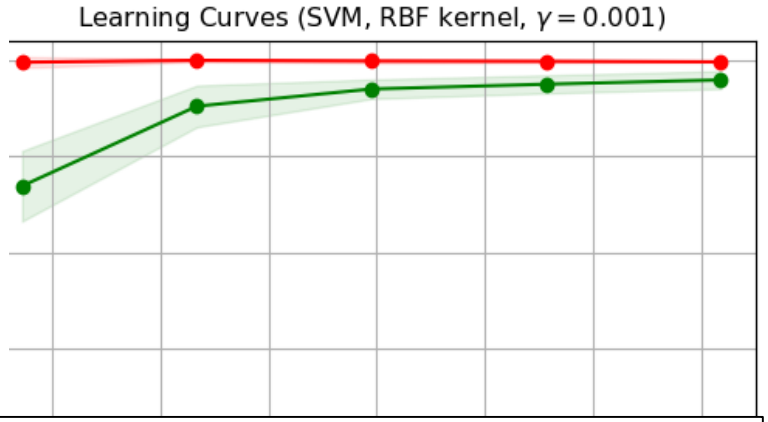
√/ x

强化学习

x

Overview of ML

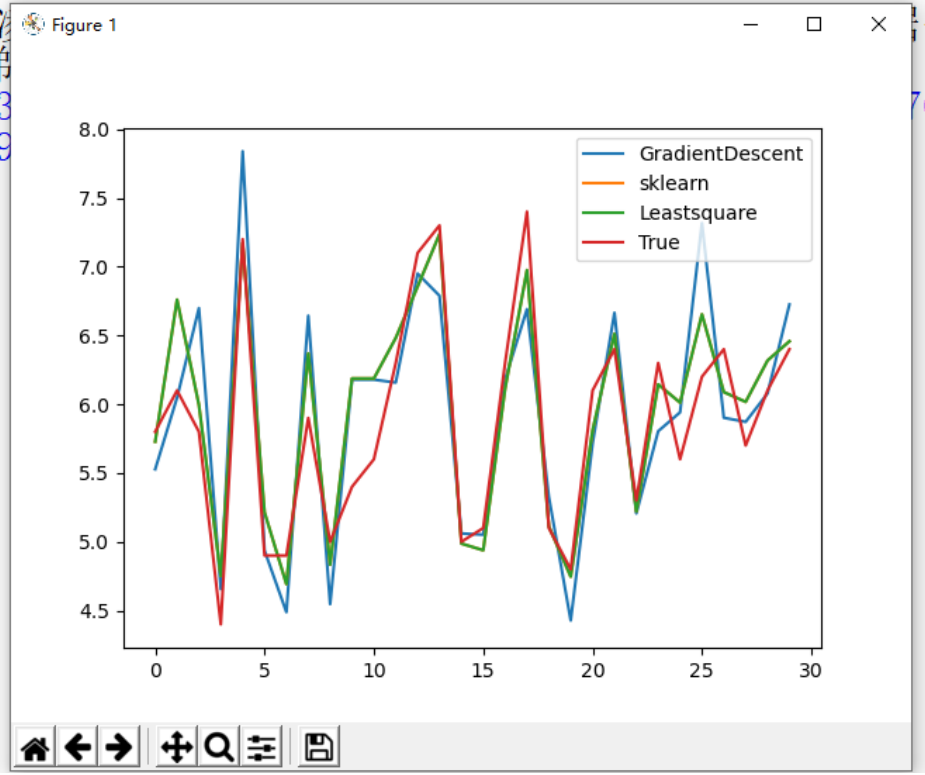





```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.datasets import load_digits
from sklearn.model_selection import learning_curve
from sklearn.model_selection import StratifiedShuffleSplit
```

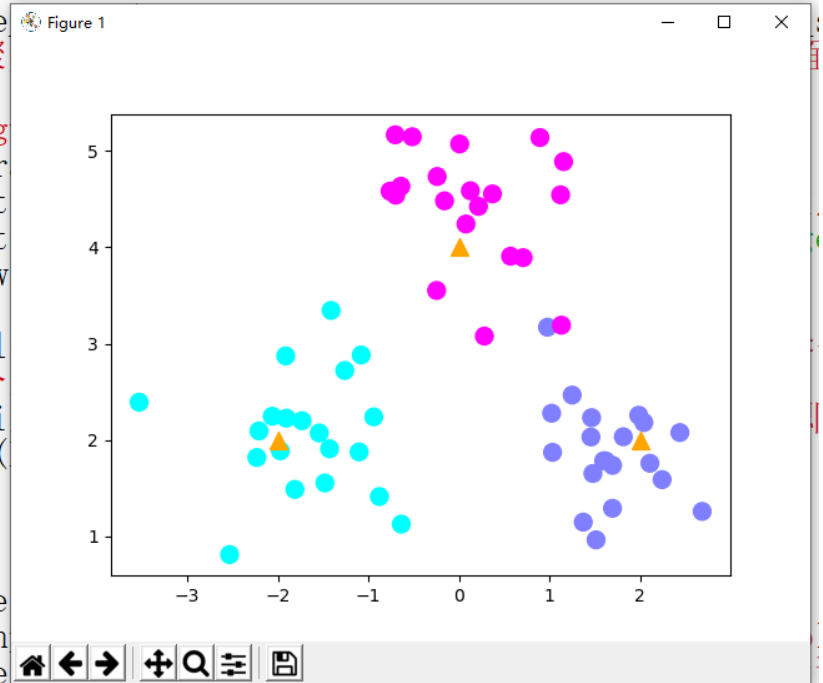
Overview of ML

```
'''
== RESTART: C:\Users\鲁冰
周-5月6日\最小二乘法和梯
[ 0.65368836  0.70955523
193454  1.84603897] [0.9
```

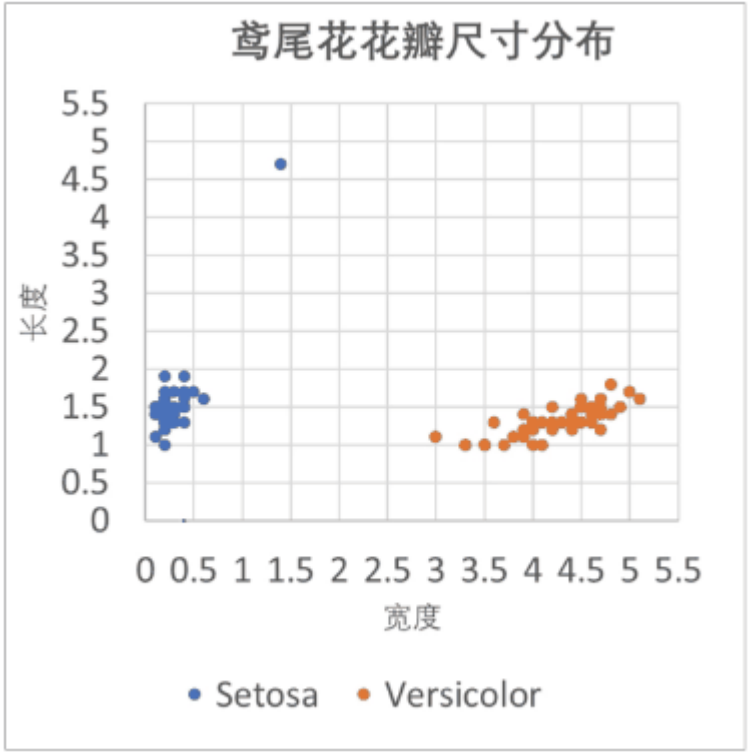


Overview of ML

```
from sklearn.datasets.samples_generator import make_
import numpy as np
import matplotlib.pyplot as plt
centers=[[-2,2],[2,2],[0,4]]
X, y=make
    #为聚
print(X.
#plt.fig
c=np.arr
plt.scat
plt.scat
plt.show

from skl
k=5#这个
clf=KNei
clf.fit(


X_sample
temp = n
y_sample
```



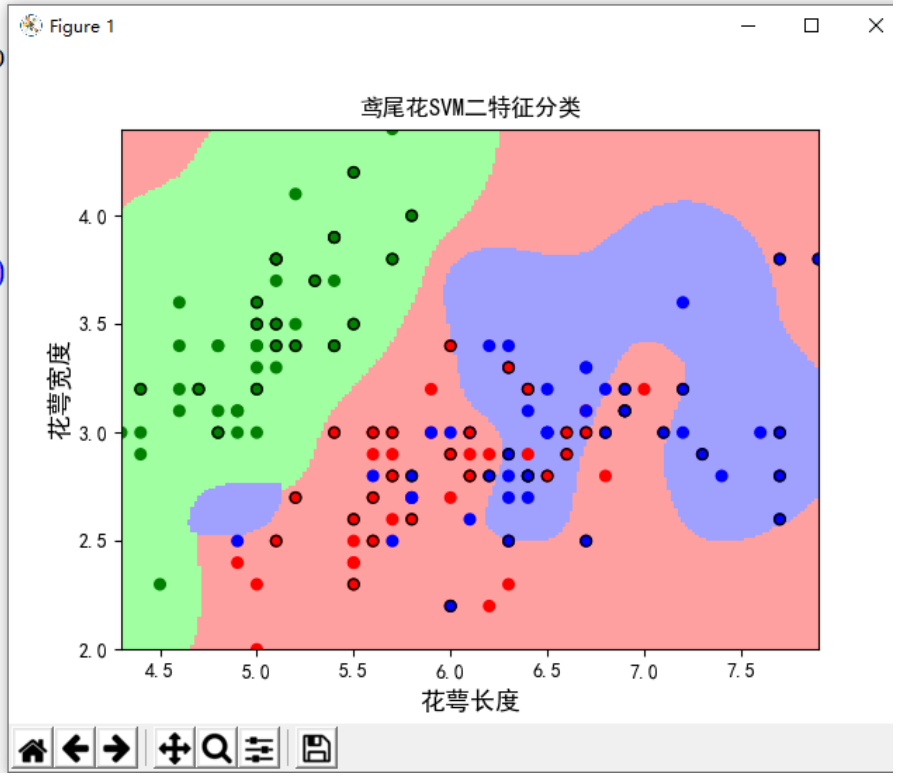
Overview of ML



Overview of ML

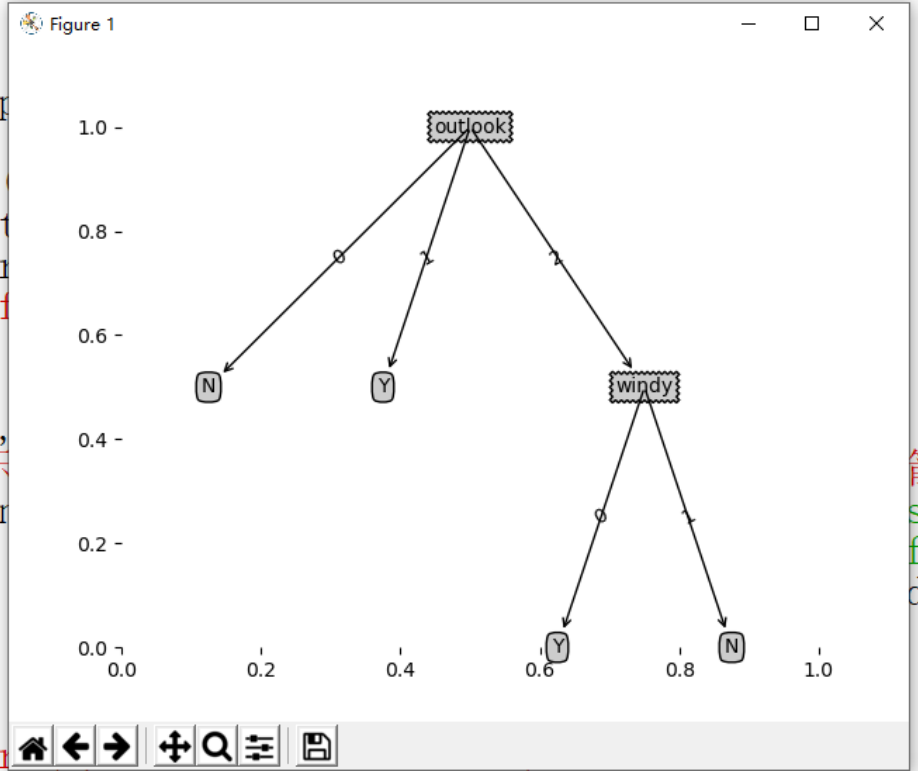
```
>>>
=== RESTART: C:\Users\鲁凌云\Desktop
1周-5月6日\鸢尾花SVM算法.py ===
训练集:  0.8555555555555555
测试集:  0.7
train_decision_function:
predict_result: [0. 1. 2. 0. 0. 2. 0
0. 0. 2.
 0. 1. 2. 1. 0. 0. 1. 0. 2. 1. 2. 2.
 0. 2. 1. 1. 0. 0. 1. 0. 1. 0. 2. 1.
 1. 0. 2. 1. 1. 1. 1. 0. 0. 1. 1. 2.
```

Squeezed text (90 lines).



Overview of ML

```
from math import log
import operator
#import treePlotter
import matplotlib.pyp

descisionNode = dict(
leafNode = dict(boxst
arrow_args = dict(arr
# myTree = { 'no surf                                    'no ', 1:


def plotNode(nodeTxt,
    # nodeTxt为要显示                                    箭头所在的点
    createPlot.ax1.ar                                    s fraction',
                                                         fraction',
                                                         deType, arro
```




```
# def createPlot():
#     fig = plt.figur
```

Lab 4: Data Cleaning

```
import pandas
data = pandas.read_csv("test5-t.csv")
data = data.dropna()
print(data)
```

cs-training.csv

mean 平均值

median 中位数

mode 众数

Lab 4: Data Cleaning

	SeriousDlq	RevolvingU	age	NumberOf	DebtRatio	MonthlyInc	NumberOf	NumberOf	NumberRe	NumberOf	NumberOfDe
1	1	0.766127	45	2	0.802982	9120	13	0	6	0	2
2	0	0.957151	40	0	0.121876	2600	4	0	0	0	1
3	0	0.65818	38	1	0.085113	3042	2	1	0	0	0
4	0	0.23381	30	0	0.03605	3300	5	0	0	0	0
5	0	0.907239	49	1	0.024926	63588	7	0	1	0	0
6	0	0.213179	74	0	0.375607	3500	3	0	1	0	1
7	0	0.305682	57	0	5710	NA	8	0	3	0	0
8	0	0.754464	39	0	0.20994	3500	8	0	0	0	0
9	0	0.116951	27	0	46	NA	2	0	0	0	NA
10	0	0.189169	57	0	0.606291	23684	9	0	4	0	2
11	0	0.644226	30	0	0.309476	2500	5	0	0	0	0
12	0	0.018798	51	0	0.531529	6501	7	0	2	0	2
13	0	0.010352	46	0	0.298354	12454	13	0	2	0	2
14	1	0.964673	40	3	0.382965	13700	9	3	1	1	2
15	0	0.019657	76	0	477	0	6	0	1	0	0
16	0	0.548458	64	0	0.209892	11362	7	0	1	0	2
17	0	0.061086	78	0	2058	NA	10	0	2	0	0

贵在坚持！