

# 数据挖掘与数据融合编程实验

助课讲师：潘 戈

指导老师：荣 冈

# Github: The State of the Octoverse 2017



## Changing the topic

In January, we released topics: repository tags that let you explore projects by technology, industry, and more. Here are the top topics you used for your repositories since the feature launched, not including frameworks or languages.

GitHub

这节课

非常

重要

01  
02  
03  
04

game

machine-  
learning

database

website

ios

API

deep-learning

blog

助教哥哥真  
是太帅了！

arduino

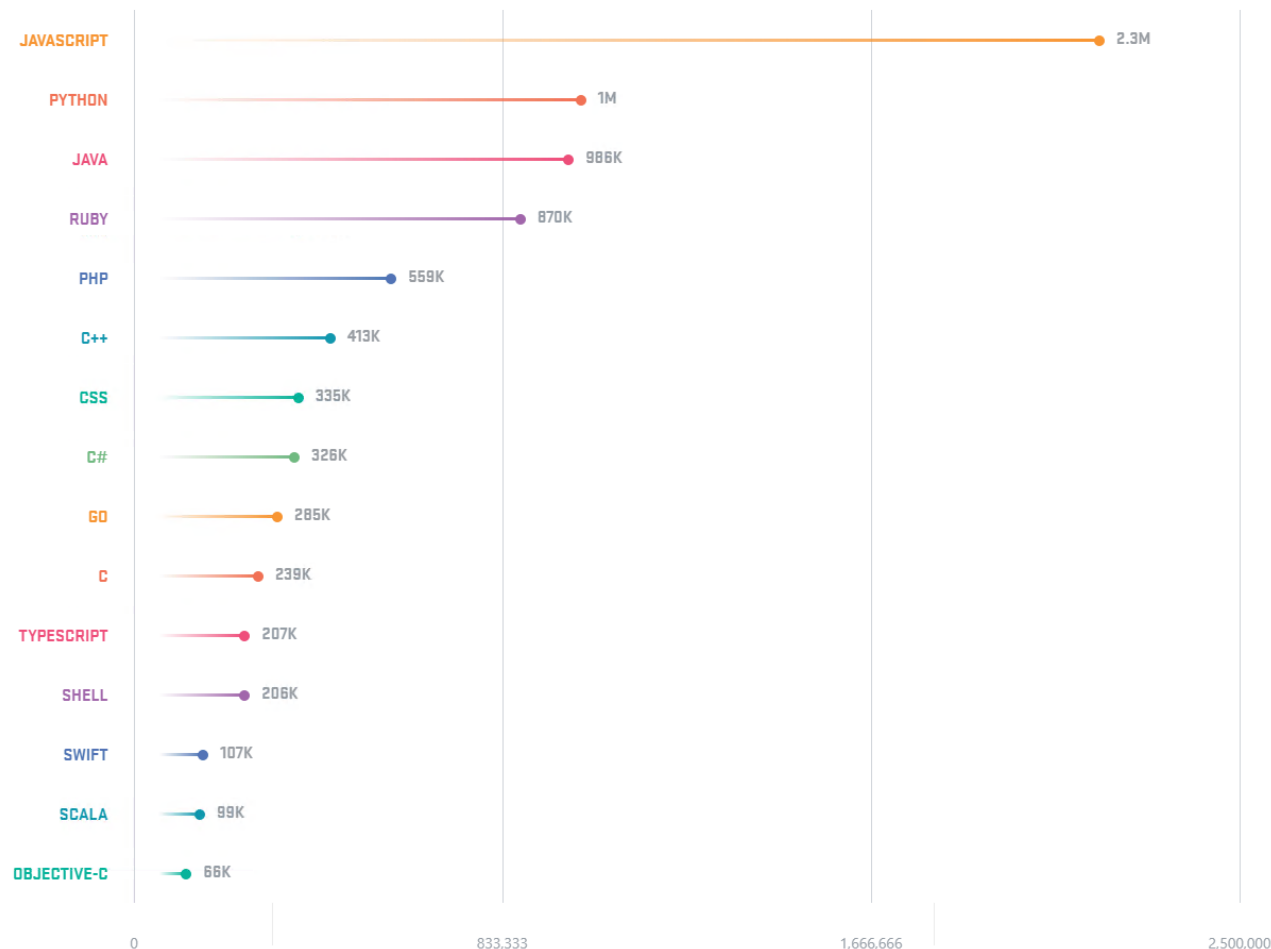
plugin

library

bot

# 第一阶段：Python基础

# Github: The State of the Octoverse 2017



The fifteen most popular languages on GitHub by opened pull request.

Python replaced Java as the second-most popular language on GitHub, with 40 percent more pull requests opened this year than last.

大势所趋 众望所归

# Getting Started

- <https://www.python.org/>
- <https://wiki.python.org/moin/BeginnersGuide/Download>
- <https://docs.python.org/2/tutorial/index.html>

# Getting Started

```
>>> 1/2
```

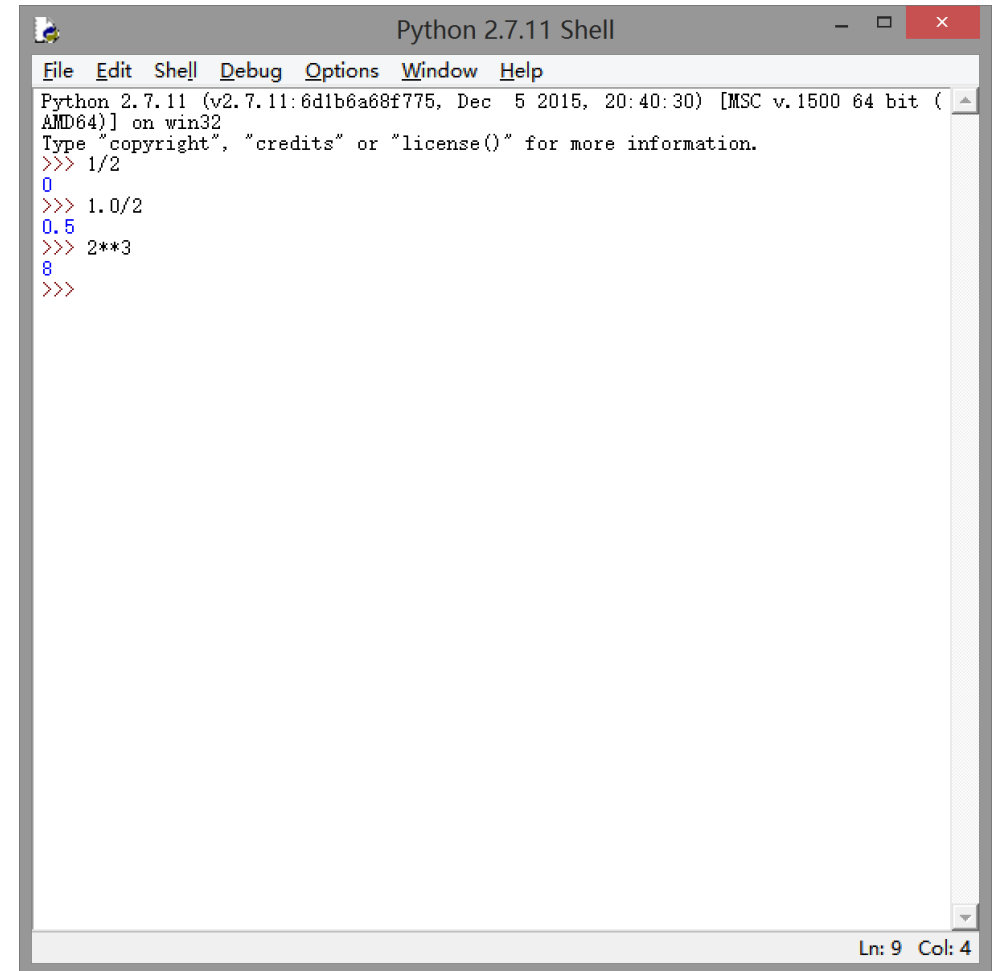
```
0
```

```
>>> 1.0/2
```

```
0.5
```

```
>>> 2**3
```

```
8
```



```
Python 2.7.11 Shell
File Edit Shell Debug Options Window Help
Python 2.7.11 (v2.7.11:6d1b6a68f775, Dec 5 2015, 20:40:30) [MSC v.1500 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> 1/2
0
>>> 1.0/2
0.5
>>> 2**3
8
>>>
```

Ln: 9 Col: 4

# Getting Started

```
>>> print("Hello, I'm Python!")
```

Hello, I'm Python!

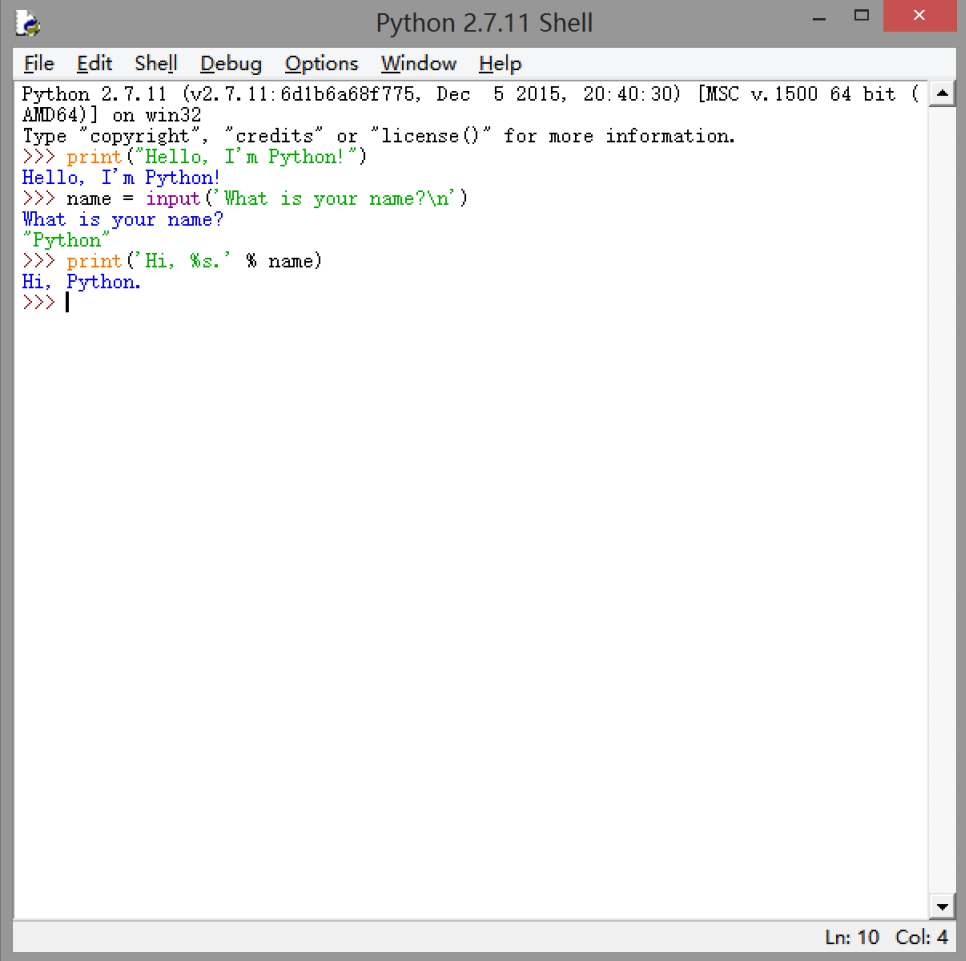
```
>>> name = input('What is your name?\n')
```

What is your name?

"Python"

```
print('Hi, %s.' % name)
```

Hi, Python.



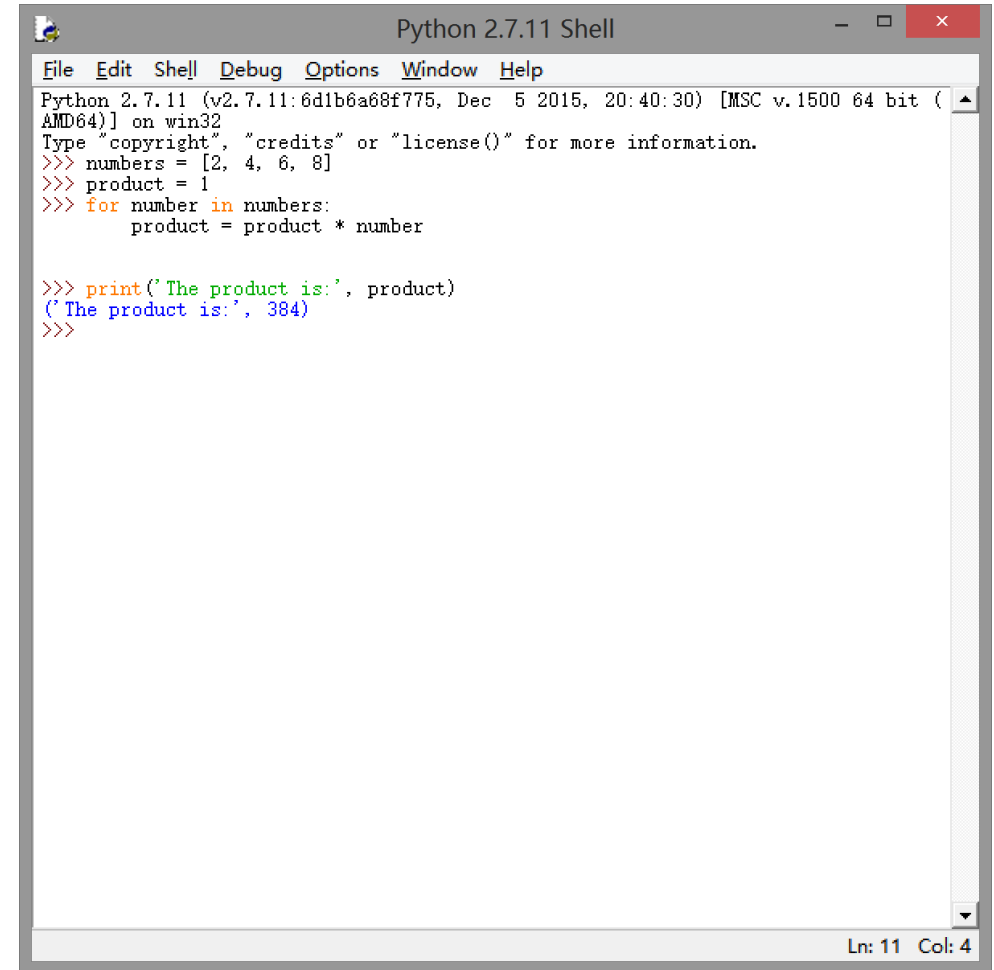
The screenshot shows a 'Python 2.7.11 Shell' window. The title bar includes standard window controls. The menu bar contains 'File', 'Edit', 'Shell', 'Debug', 'Options', 'Window', and 'Help'. The main text area displays the following output and input:

```
Python 2.7.11 (v2.7.11:6d1b6a68f775, Dec 5 2015, 20:40:30) [MSC v.1500 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> print("Hello, I'm Python!")
Hello, I'm Python!
>>> name = input('What is your name?\n')
What is your name?
"Python"
>>> print('Hi, %s.' % name)
Hi, Python.
>>> |
```

The status bar at the bottom right indicates 'Ln: 10 Col: 4'.

# Getting Started

```
>>> numbers = [2, 4, 6, 8]
>>> product = 1
>>> for number in numbers:
...     product = product * number
...
>>> print('The product is:', product)
('The product is:', 384)
```



The screenshot shows a window titled "Python 2.7.11 Shell". The window has a menu bar with "File", "Edit", "Shell", "Debug", "Options", "Window", and "Help". The main text area contains the following code and its output:

```
Python 2.7.11 (v2.7.11:6d1b6a68f775, Dec 5 2015, 20:40:30) [MSC v.1500 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> numbers = [2, 4, 6, 8]
>>> product = 1
>>> for number in numbers:
...     product = product * number
...
>>> print('The product is:', product)
('The product is:', 384)
>>>
```

The status bar at the bottom right indicates "Ln: 11 Col: 4".

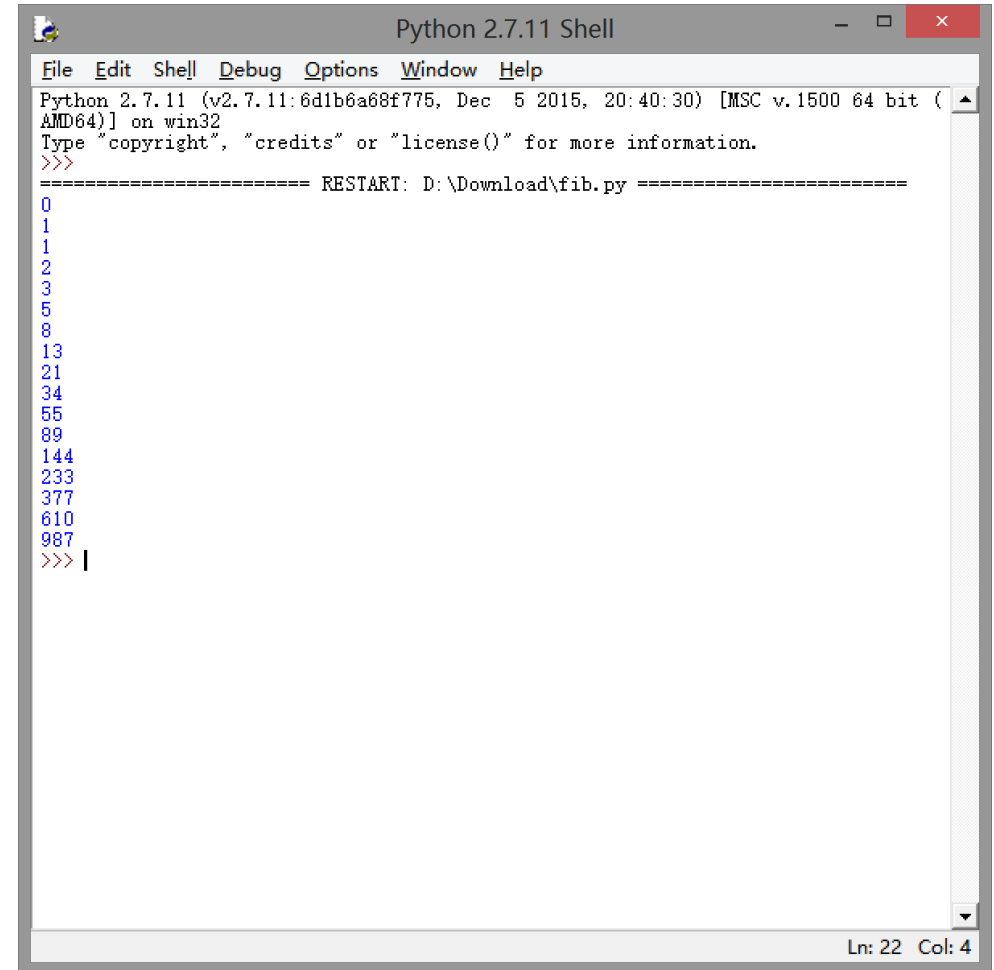


# Getting Started

fib.py

---

```
def fib(n):  
    a, b = 0, 1  
    while a < n:  
        print(a)  
        a, b = b, a+b  
fib(1000)
```

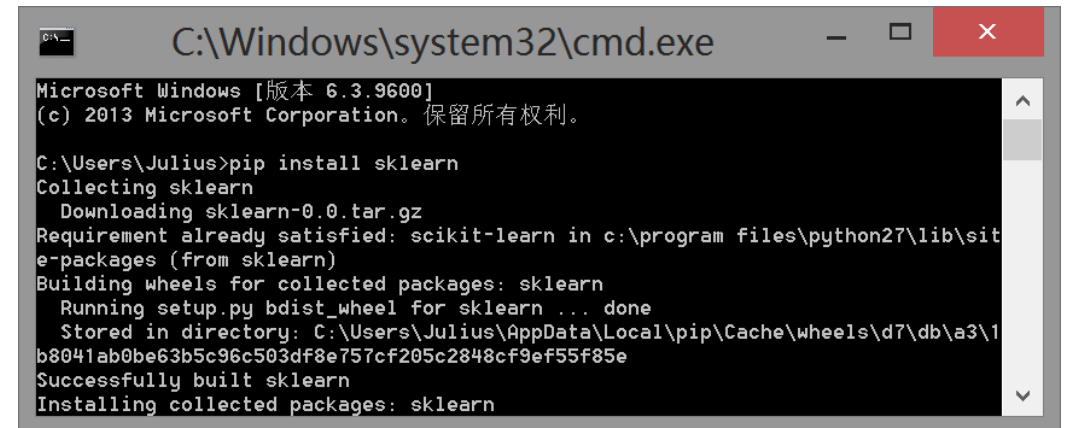


```
Python 2.7.11 Shell  
File Edit Shell Debug Options Window Help  
Python 2.7.11 (v2.7.11:6d1b6a68f775, Dec 5 2015, 20:40:30) [MSC v.1500 64 bit (AMD64)] on win32  
Type "copyright", "credits" or "license()" for more information.  
>>>  
===== RESTART: D:\Download\fib.py =====  
0  
1  
1  
2  
3  
5  
8  
13  
21  
34  
55  
89  
144  
233  
377  
610  
987  
>>> |  
Ln: 22 Col: 4
```

# 论调包侠的自我修养

- numpy
- <http://www.numpy.org/>
- pandas
- <http://pandas.pydata.org>
- sklearn
- <http://scikit-learn.org/>

- Windows: 命令提示符
- Unix/Linux: Shell
- pip install sklearn



```
C:\Windows\system32\cmd.exe
Microsoft Windows [版本 6.3.9600]
(c) 2013 Microsoft Corporation。保留所有权利。

C:\Users\Julius>pip install sklearn
Collecting sklearn
  Downloading sklearn-0.0.tar.gz
Requirement already satisfied: scikit-learn in c:\program files\python27\lib\site-packages (from sklearn)
Building wheels for collected packages: sklearn
  Running setup.py bdist_wheel for sklearn ... done
  Stored in directory: C:\Users\Julius\AppData\Local\pip\Cache\wheels\d7\db\ab\1b8041ab0be63b5c96c503df8e757cf205c2848cf9ef55f85e
Successfully built sklearn
Installing collected packages: sklearn
```

# 伟大的防火墙

```
File "C:\Python27\lib\site-packages\pip\_vendor\requests\packages\urllib3\response.py", line 227, in read
```

```
    raise ReadTimeoutError(self._pool, None, 'Read timed out.')
```

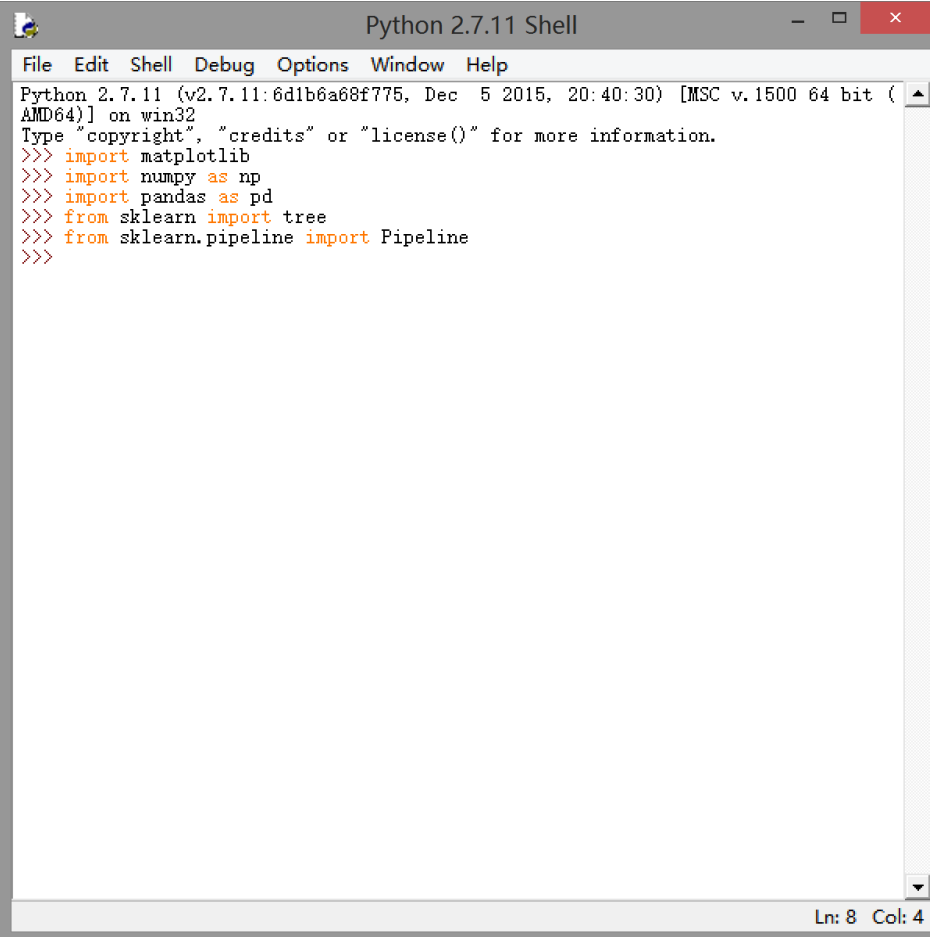
```
ReadTimeoutError: HTTPSConnectionPool(host='pypi.python.org', port=443): Read timed out.
```

# 伟大的防火墙

- 方案一：下载安装包
  - `pip install /path/package`
- 方案二：更换Python默认软件镜像源
  - 豆瓣：<http://pypi.douban.com/simple/>
  - 清华：<https://pypi.tuna.tsinghua.edu.cn/simple/>
- 方案三：吾知子之所以距我， 吾不言。——《墨子》

# 论调包侠的自我修养

```
>>> import matplotlib
>>> import numpy as np
>>> import pandas as pd
>>> from sklearn import tree
>>> from sklearn.pipeline import Pipeline
```



```
Python 2.7.11 Shell
File Edit Shell Debug Options Window Help
Python 2.7.11 (v2.7.11:6d1b6a68f775, Dec 5 2015, 20:40:30) [MSC v.1500 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> import matplotlib
>>> import numpy as np
>>> import pandas as pd
>>> from sklearn import tree
>>> from sklearn.pipeline import Pipeline
>>>
```

Ln: 8 Col: 4

# 实验任务

- 安装python 2.7
- 熟悉pip的基本使用方法
- 学习python的基本语法
- 简单使用几个常用库



## 第二阶段：机器学习

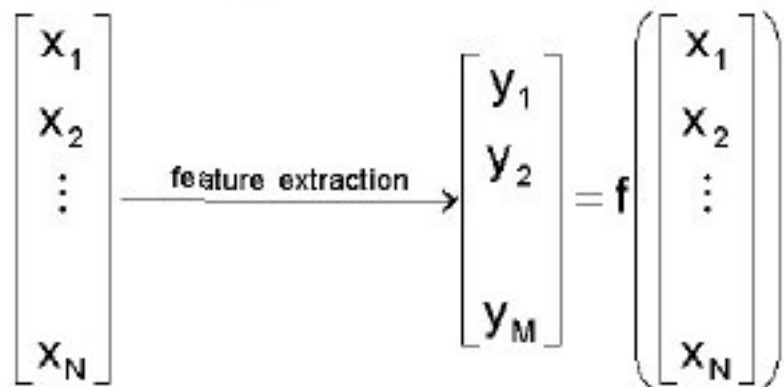
# 鸢尾花数据集

- 鸢尾花的三个分类
  - Iris-setosa
  - Iris-versicolor
  - Iris-virginica
- 每类提供50个数据集
- 共150个数据集
- 取80%做训练， 20%做测试
- 数据集描述的四个属性
  - sepal length, 花萼长度
  - sepal width, 花萼宽度
  - petal length, 花瓣长度
  - petal width, 花瓣宽度
- 根据四大属性  
预测鸢尾花的分类

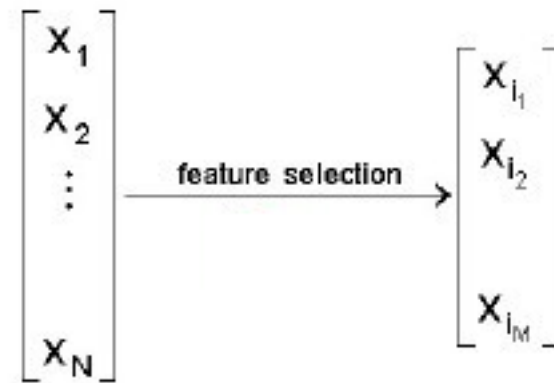


# 特征抽取与特征选择 (降维)

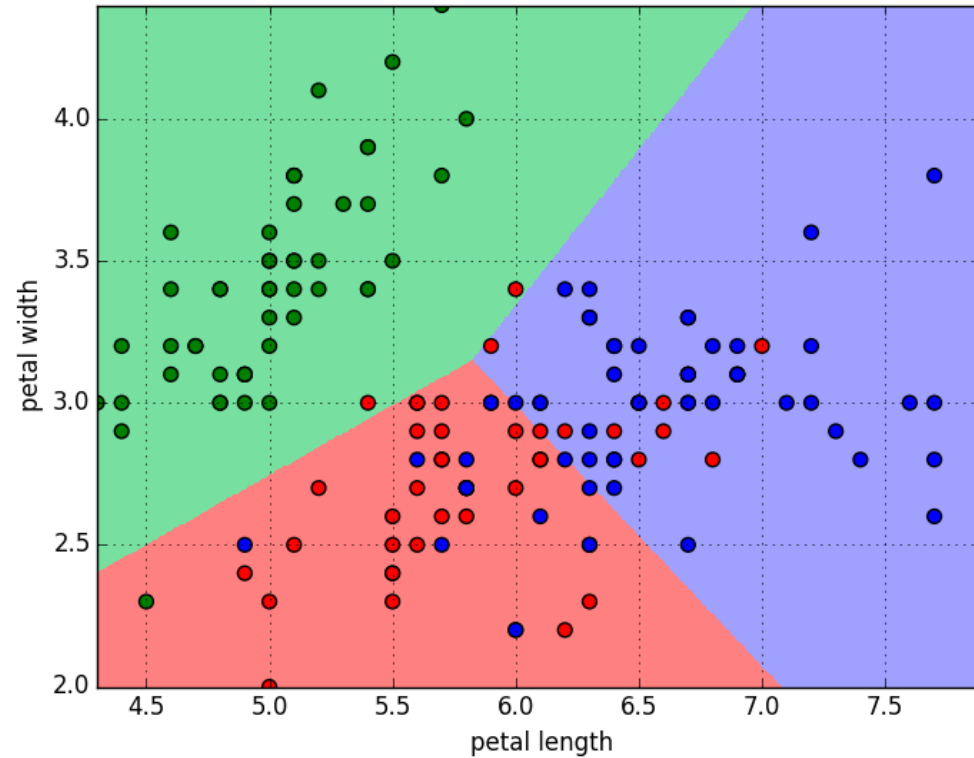
- 特征抽取 (Feature Extraction)
- Creating a subset of new features by combinations of the existing features.
- 特征抽取后的新特征是原来特征的一个映射。



- 特征选择 (Feature Selection)
- choosing a subset of all the features (the ones more informative)
- 特征选择后的特征是原来特征的一个子集。



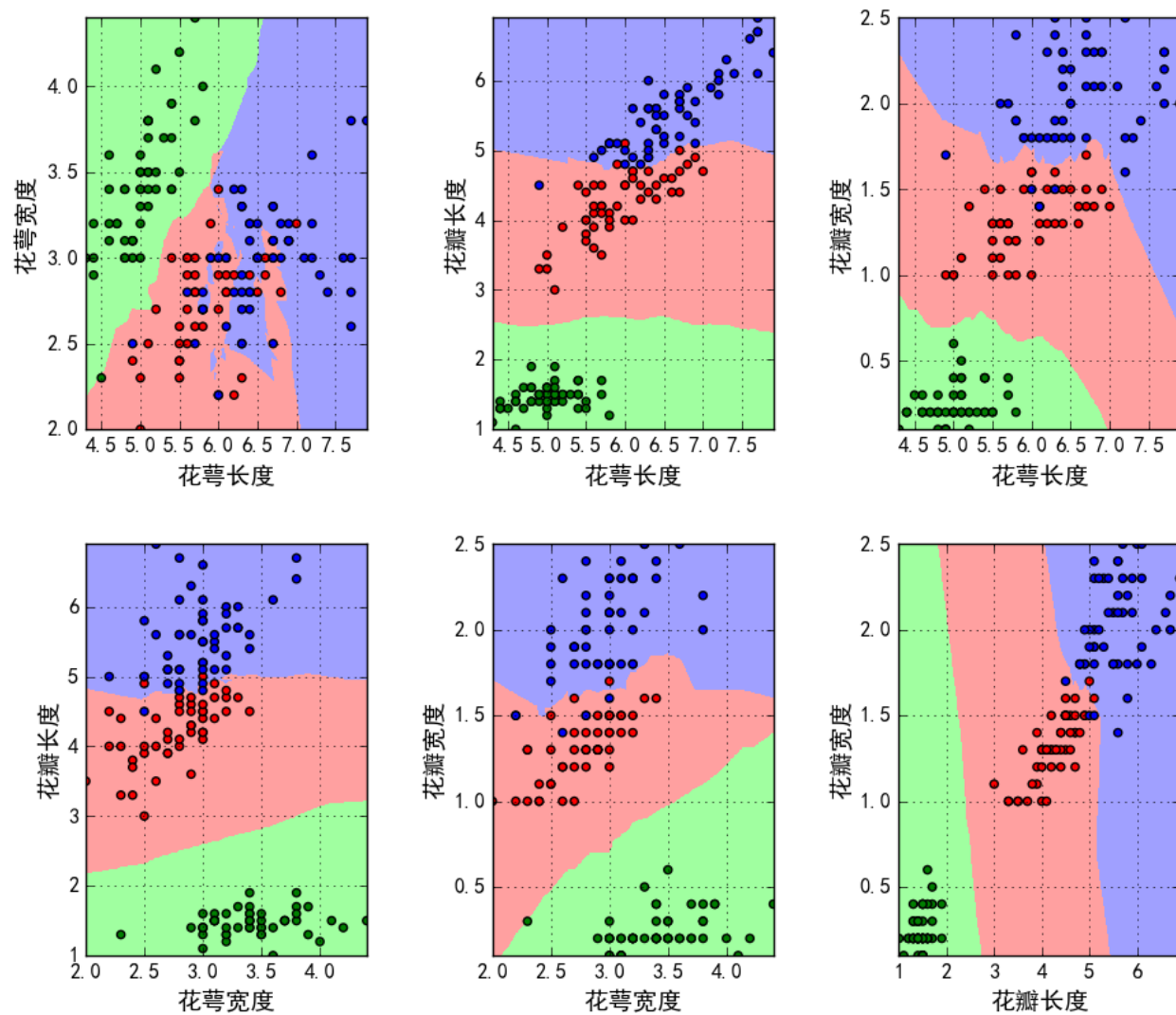
# Logistic Regression



准确率 (4个特征) = 预测正确的数目/测试集样本总数 = 86.67%

# KNN

K最近邻对鸢尾花数据的两特征组合的分类结果



准确率 (4个特征) = 96.67%

# Decision Tree

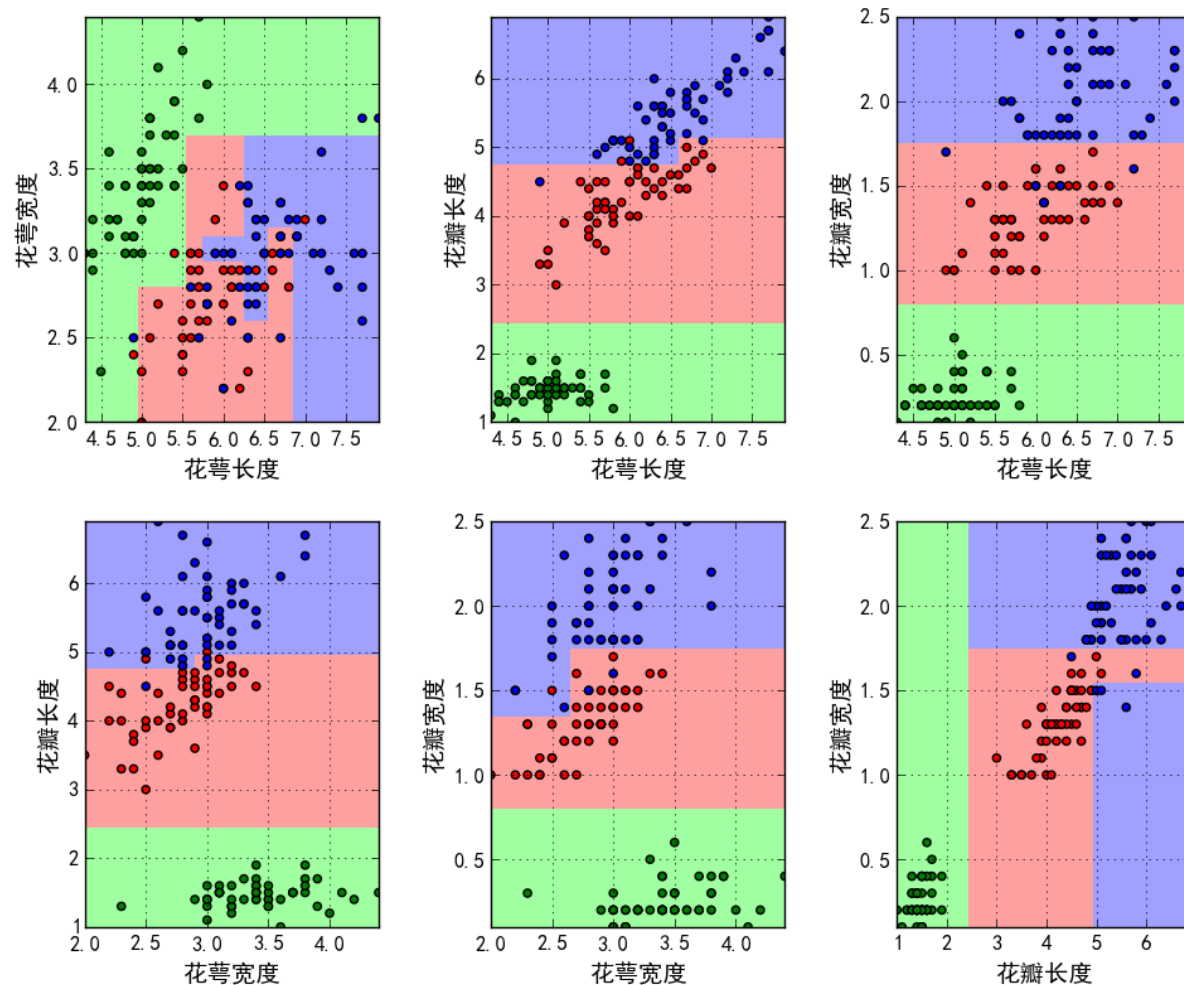
- 特征选择后的训练集准确率

- 花萼长度 + 花萼宽度 = 82.00%
- 花萼长度 + 花瓣长度 = 96.67%
- 花萼长度 + 花瓣宽度 = 96.00%
- 花萼宽度 + 花瓣长度 = 95.33%
- 花萼宽度 + 花瓣宽度 = 96.67%
- 花瓣长度 + 花瓣宽度 = 98.00%

- 4个特征下的测试集准确率

- 96.67%

决策树对鸢尾花数据的两特征组合的分类结果



# Random Forest

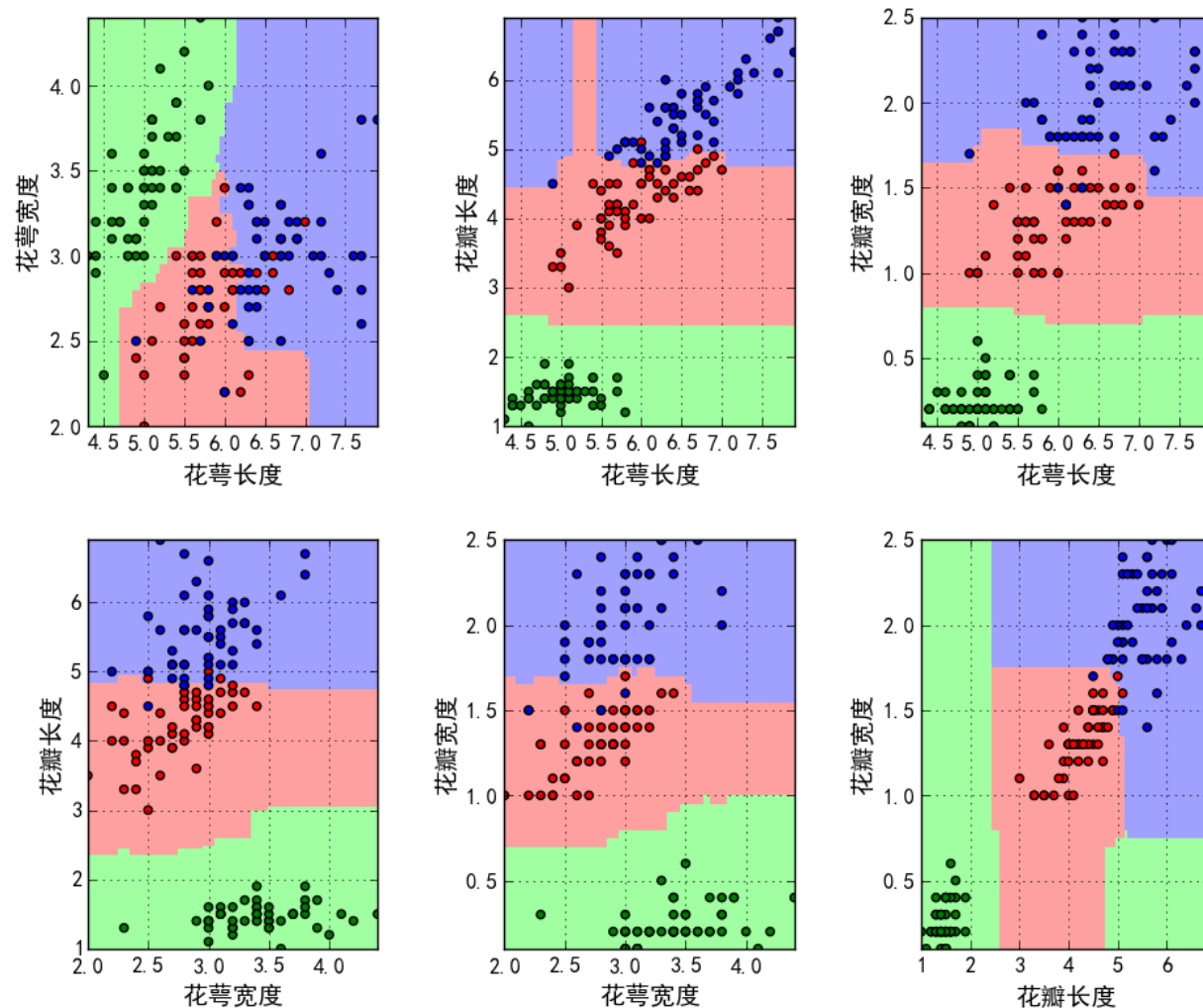
- 特征选择后的训练集准确率

- 花萼长度 + 花萼宽度 = 84.00%
- 花萼长度 + 花瓣长度 = 97.33%
- 花萼长度 + 花瓣宽度 = 96.67%
- 花萼宽度 + 花瓣长度 = 96.00%
- 花萼宽度 + 花瓣宽度 = 96.67%
- 花瓣长度 + 花瓣宽度 = 96.67%

- 4个特征下的测试集准确率

- 100%

随机森林对鸢尾花数据的两特征组合的分类结果



# K-Means

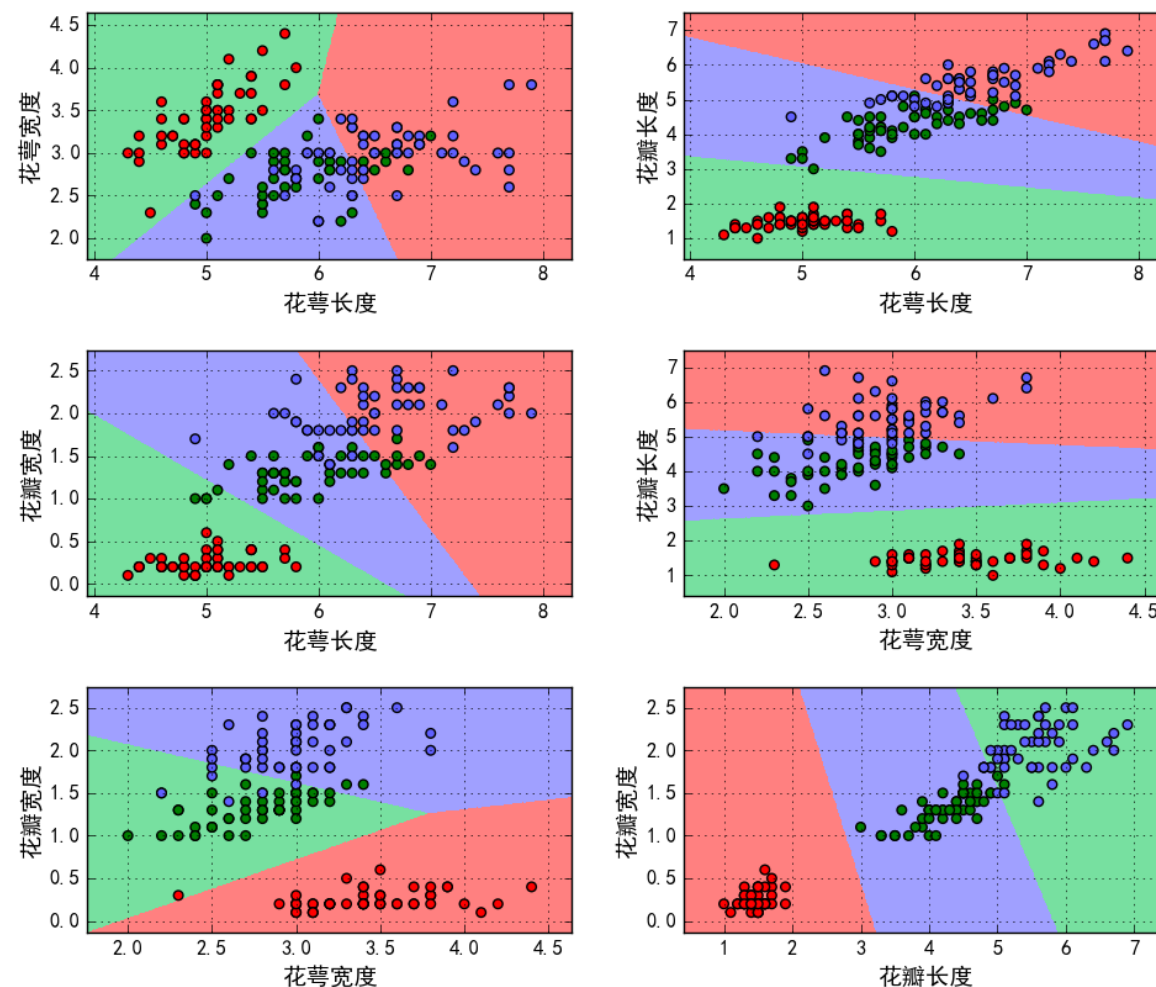
- 特征选择后的训练集准确率

- 花萼长度 + 花萼宽度 = 82.00%
- 花萼长度 + 花瓣长度 = 88.00%
- 花萼长度 + 花瓣宽度 = 82.67%
- 花萼宽度 + 花瓣长度 = 92.67%
- 花萼宽度 + 花瓣宽度 = 92.67%
- 花瓣长度 + 花瓣宽度 = 96.00%

- 4个特征下的测试集准确率

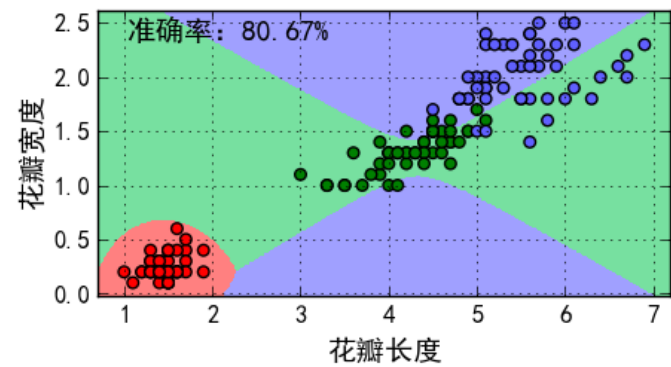
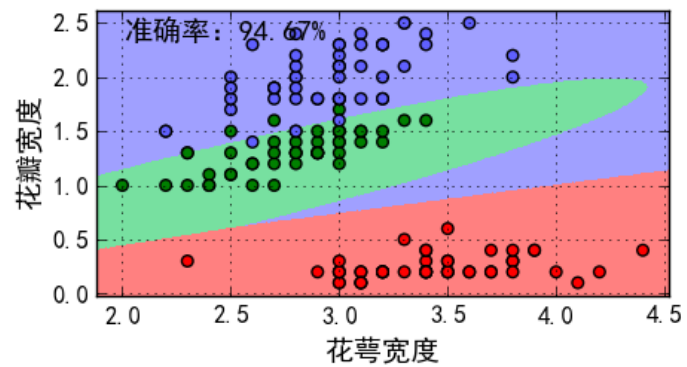
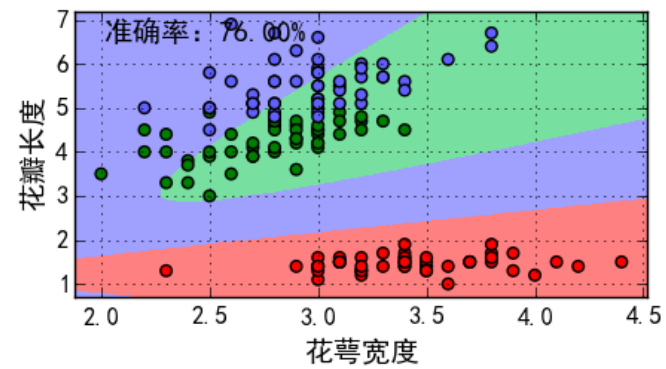
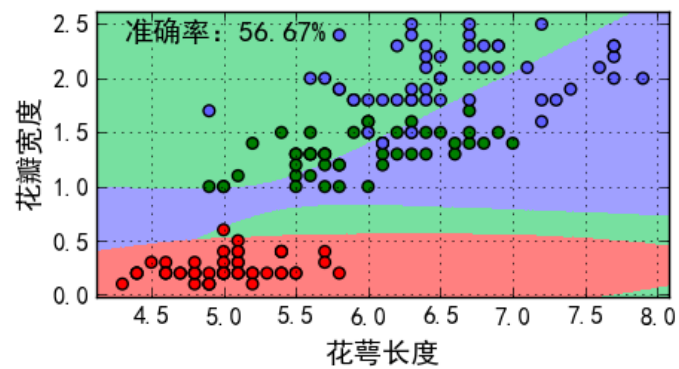
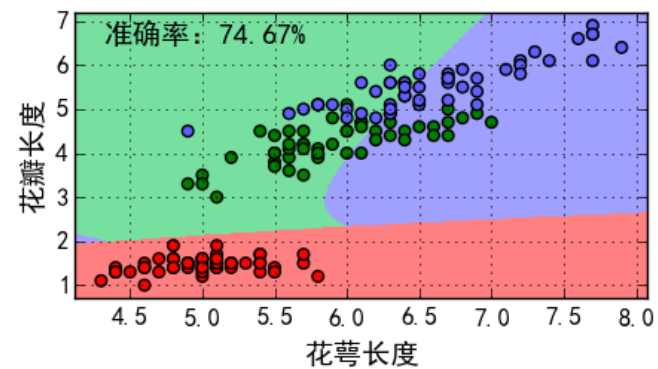
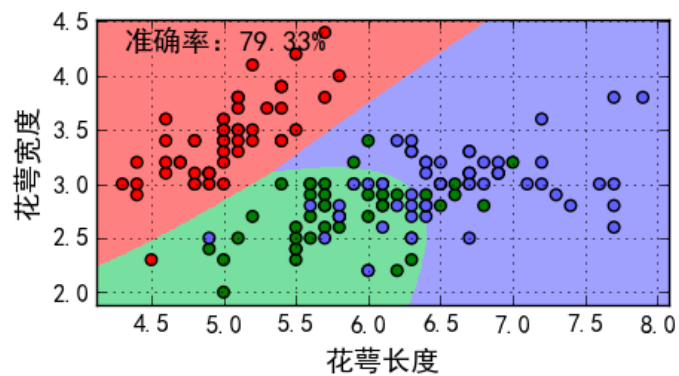
- 90%

kmean++无监督分类鸢尾花数据



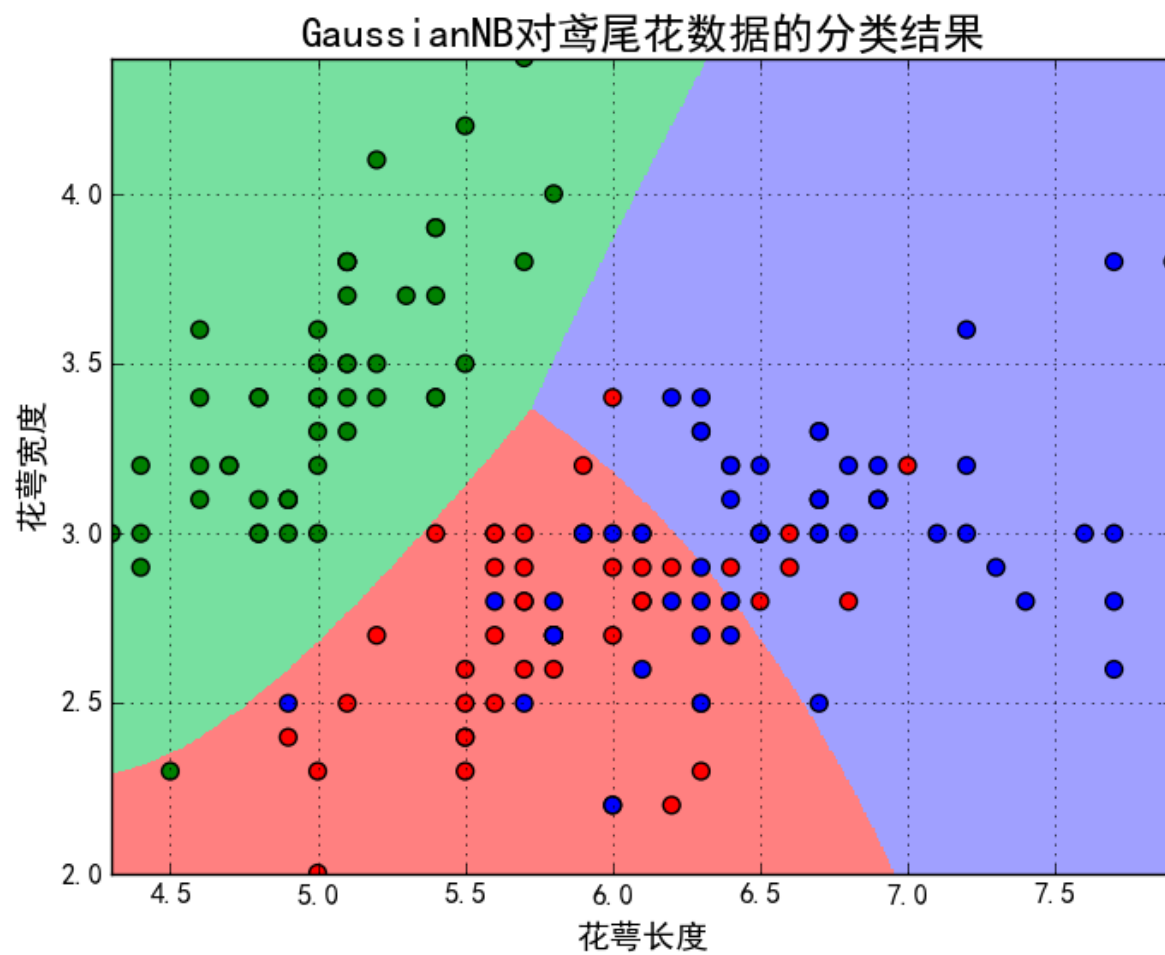
# EM

EM算法无监督分类鸢尾花数据





# Naive Bayes





# 编程学习

- 阅读理解并运行Iris文件夹下的算法测试代码
  - Iris\_LogisticRegression.py
  - Iris\_KNN.py
  - Iris\_DecisionTree.py
  - Iris\_RandomForest.py
  - Iris\_kMeans.py
  - Iris\_DBSCAN.py

# 进阶任务

- 阅读理解algorithms文件夹下的算法实现代码
  - linear\_model/logistic.py
  - neighbors/regression.py
  - tree/tree.py
  - ensemble/forest.py
  - cluster/k\_means\_.py
  - cluster/dbscan\_.py