


### 基本建模流程

#### 导入工具包

```
from sklearn import preprocessing(预处理)
from sklearn.model_selection import train_test_split (切分训练集和测试集)
from sklearn.cluster import Kmeans,Birch(聚类)
from sklearn.ensemble import RandomForestClassifier(分类)
from sklearn.naive_bayes import GaussianNB(分类)
from sklearn.ensemble import RandomForestRegressor(回归)
```

103


[Install](#) [User Guide](#) [API](#) [Examples](#) [More](#)

# scikit-learn

Machine Learning in Python

[Getting Started](#)
[Release Highlights for 0.23](#)
[GitHub](#)

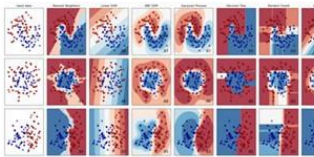
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

### Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, and more...



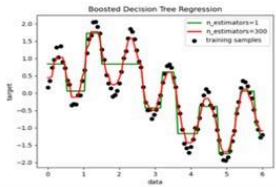
Examples

### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, nearest neighbors, random forest, and more...




Examples

### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, and more...



Examples

### Dimensionality reduction

Reducing the number of random variables to consider.

### Model selection

Comparing, validating and choosing parameters and models.

### Preprocessing

Feature extraction and normalization.

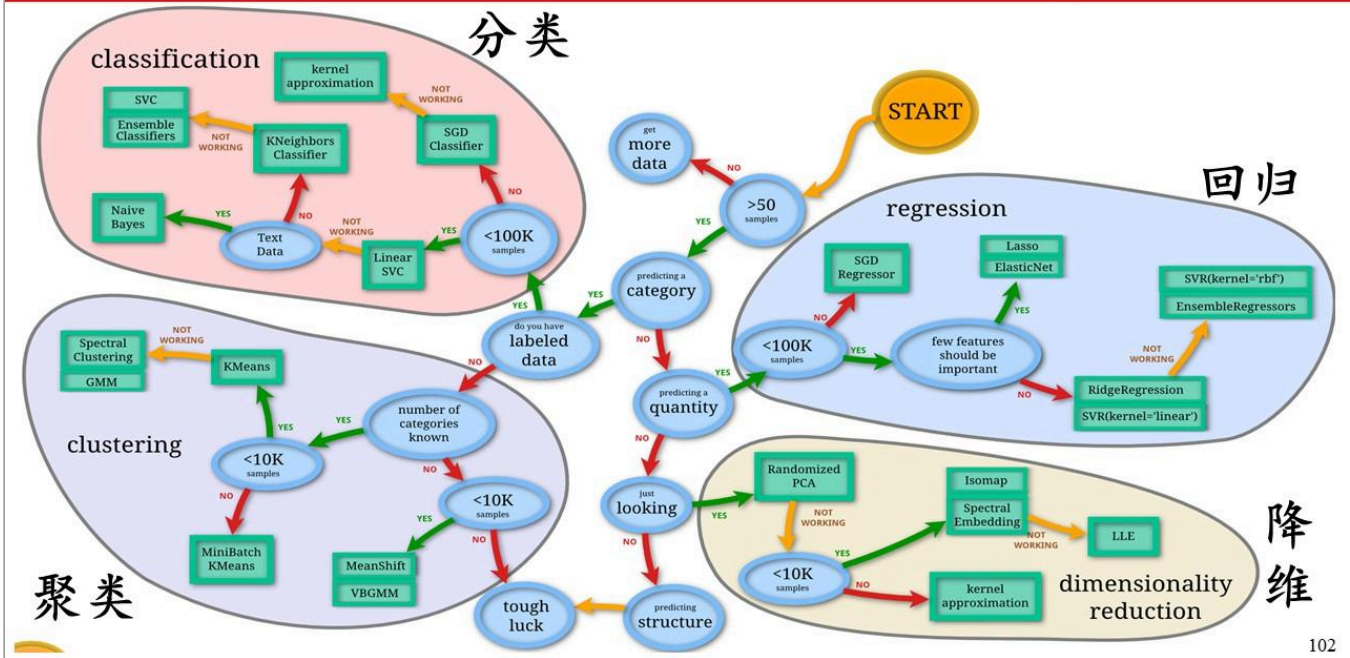
**Applications:** Transforming input data such as text

101

#### 四、Python机器学习常用库sklearn简介



中国石油大学  
CHINA UNIVERSITY OF PETROLEUM



102

#### 五、sklearn——加载数据



中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

数据集：600行数据（CS(粗砂岩)、FS(细砂岩)、M(泥岩)）

GR	AC	DEN	CNL	LLD	LLS	CAL	Type
37.5	241.8	2.5	18.3	25.3	26.6	26.1	CS
38.2	243.1	2.5	19.2	23.7	24.6	26.3	CS
38.4	243.3	2.5	19.6	23.0	24.0	26.5	CS
38.4	242.5	2.5	19.8	22.9	24.0	26.6	CS
39.4	241.0	2.5	20.3	22.6	23.8	26.7	CS
50.3	239.7	2.5	22.6	19.7	21.1	26.7	FS
61.0	241.4	2.5	24.0	16.9	17.9	26.6	FS
71.6	243.5	2.6	25.1	15.7	16.4	26.6	FS
79.4	244.5	2.6	25.7	14.0	14.2	26.7	FS
82.8	243.8	2.6	26.3	13.1	13.0	26.9	FS
78.6	298.9	1.9	38.7	5.4	6.2	32.0	M
81.3	296.7	1.9	38.1	5.5	6.3	31.3	M
82.4	292.0	1.9	37.6	5.7	6.4	32.2	M
82.4	286.6	1.9	37.4	5.8	6.5	33.9	M
81.8	282.4	1.9	37.5	5.8	6.6	35.4	M

加载数据

预处理

劈分训练  
测试样本

训练预测  
模型

预测未知  
样本

104

## 五、sklearn——加载数据



中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

### 加载数据

✓ Scikit-learn 支持以 NumPy 的 arrays 对象、Pandas 对象、SciPy 的稀疏矩阵及其他可转换为数值型 arrays 的数据结构作为其输入，前提是数据必须是数值型的

```
data=pd.read_csv('data.csv')
X = data.iloc[:, 0:7]
y = data.iloc[:,7]
X=np.array(X.values)
y=np.array(y.values)
```

加载数据



预处理



劈分训练  
测试样本



训练预测  
模型



预测未知  
样本

105

## 四、sklearn——预处理



中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

### 数据预处理

Z-Score 标准化  $x^* = \frac{x - \mu}{\sigma}$

处理后的数据均值为0，方差为1

使用 Scikit-learn 进行数据标准化

```
from sklearn import preprocessing
```

构建转换器实例

```
sta=preprocessing.StandardScaler()
```

拟合及转换

```
sta.fit_transform(X)
```

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x^{(i)})^2$$
$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

加载数据



预处理



劈分训练  
测试样本



训练预测  
模型



预测未知  
样本

106



### 数据预处理

$$\text{归一化} \longrightarrow x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

将数据映射到[0,1]区间

使用Scikit-learn进行数据归一化

```
from sklearn import preprocessing
```

构建转换器实例

```
nor = preprocessing.MinMaxScaler()
```

拟合及转换

```
nor.fit_transform(X)
```

加载数据



预处理



劈分训练  
测试样本



训练预测  
模型



预测未知  
样本

107

### 数据划分

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
random_state=12, stratify=y, test_size=0.3)
```



参数 **stratify**: 控制分类问题中的分层抽样，默认为 **None**，即不进行分层抽样，当传入为数组时，则依据该数组进行分层抽样，另外可通过设置 **shuffle=True** 提前打乱数据，再划分数据集。

加载数据



预处理



劈分训练  
测试样本

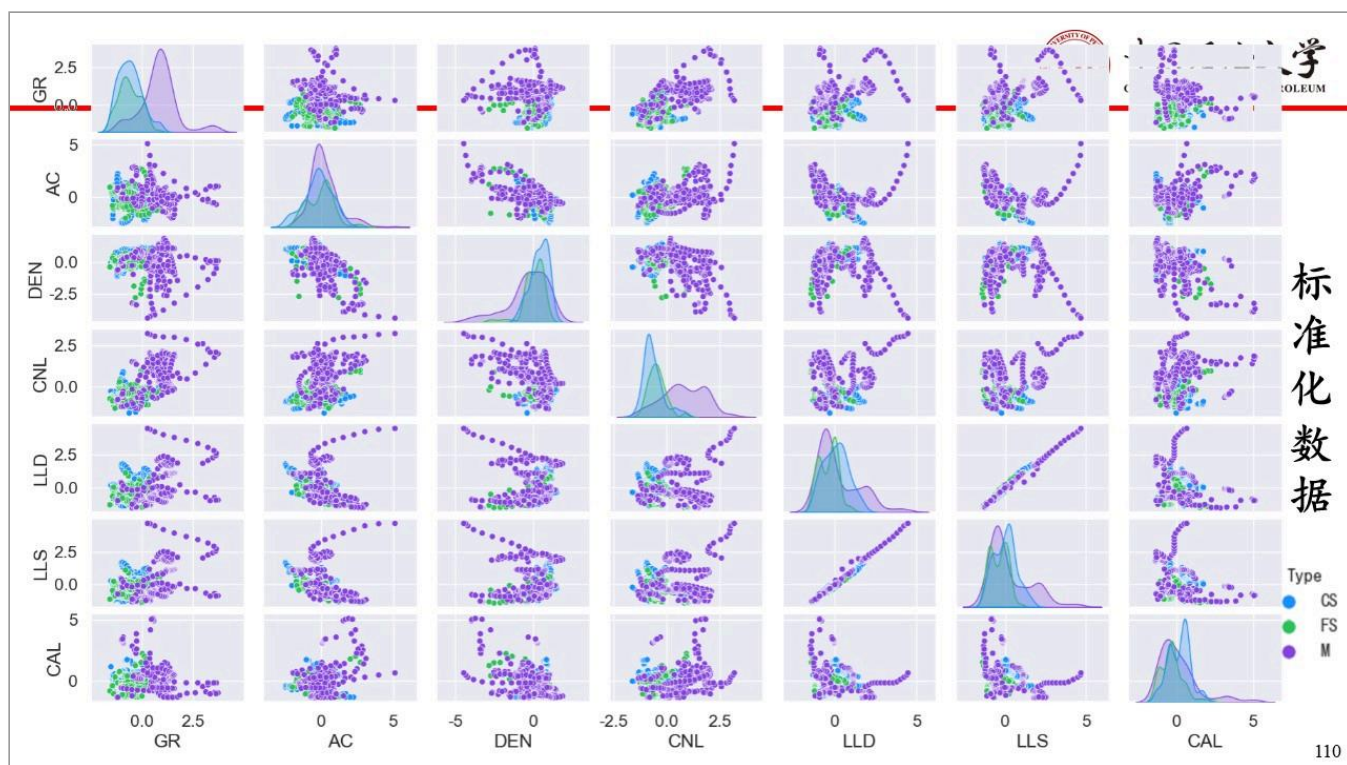
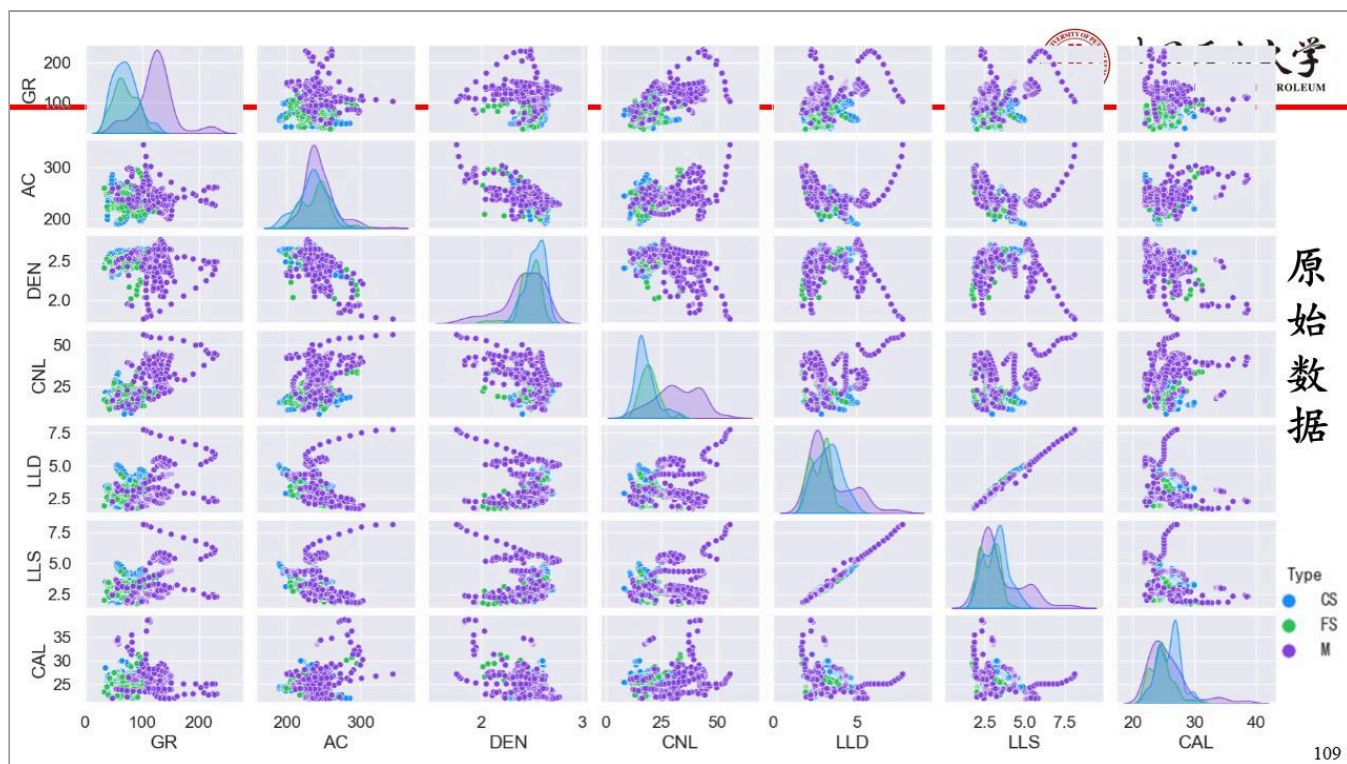


训练预测  
模型

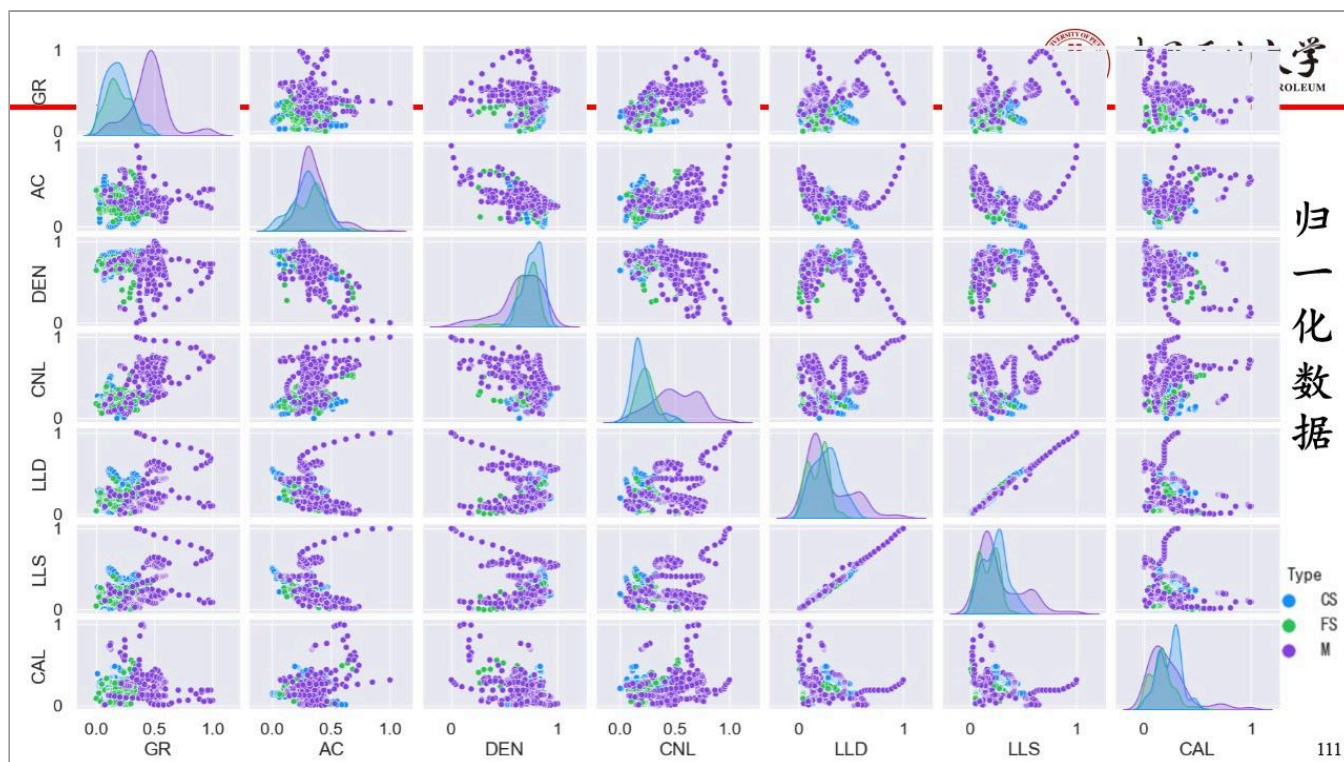


预测未知  
样本

108







111

使用的聚类方法: k-Means, Birch



中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

### k-Means聚类

```
from sklearn.cluster import KMeans
estimator = KMeans(n_clusters=?, max_iter = ? ) # 构造聚类器
estimator.fit(X) # 聚类
```

### Birch聚类

```
from sklearn.cluster import Birch
estimator = Birch(n_clusters=? )
estimator.fit(X)
```

加载数据



预处理



劈分训练  
测试样本



训练聚类  
模型



预测未知  
样本

112

#### 四、sklearn——聚类



中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

使用的聚类方法: k-Means, BIRCH

聚类方法	加载模块
K-means	cluster.KMeans
BIRCH	cluster.Birch

加载数据



预处理



劈分训练  
测试样本



训练聚类  
模型



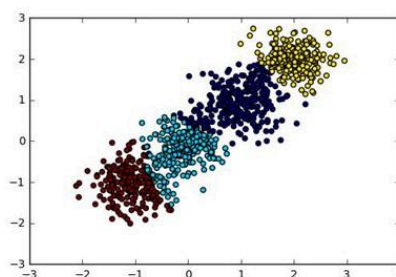
预测未知  
样本

k-Means聚类

K-means clustering on the digits dataset (PCA-reduced data)  
Centroids are marked with white cross



Birch聚类



113

#### 四、sklearn——岩性聚类

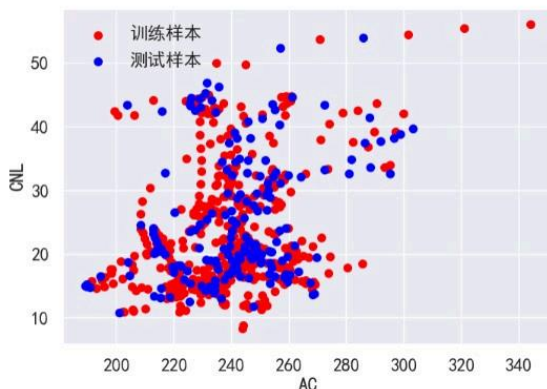


中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

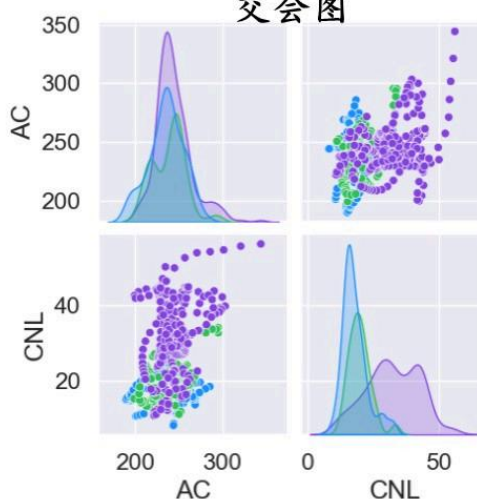
聚类结果

原始数据(训练样本占70%，测试样本占30%)

原始数据



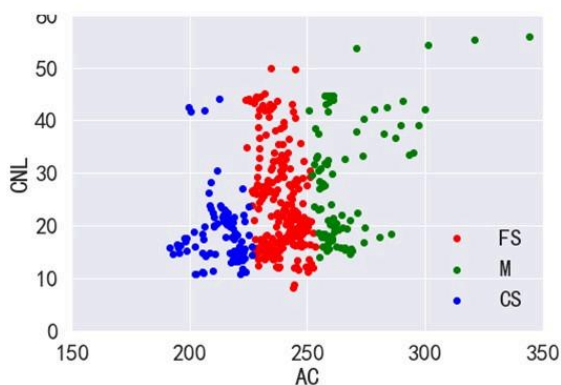
交会图



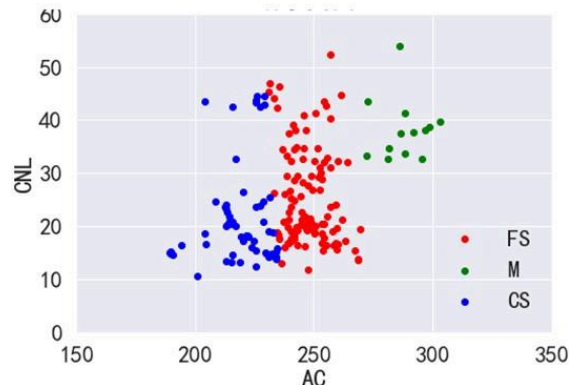
Type  
● CS  
● FS  
● M

114

**K-means聚类结果**  
(训练样本占70%)

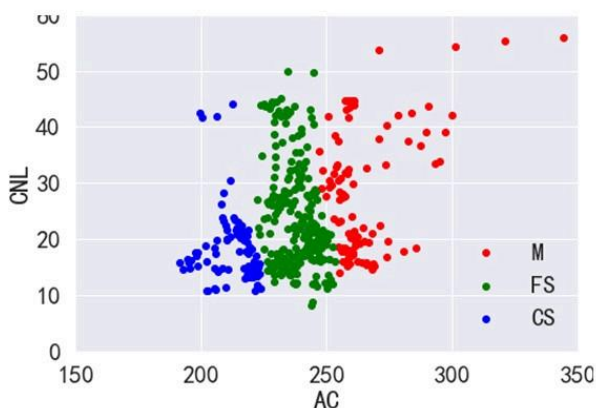


**K-means聚类结果**  
(测试样本占30%)

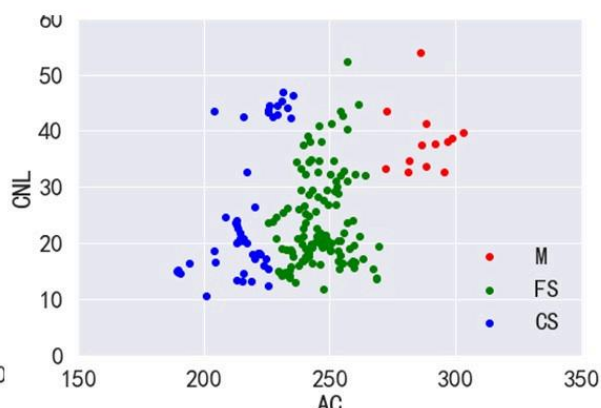


115

**BIRCH聚类结果**  
(训练样本占70%)



**BIRCH聚类结果**  
(测试样本占30%)



116

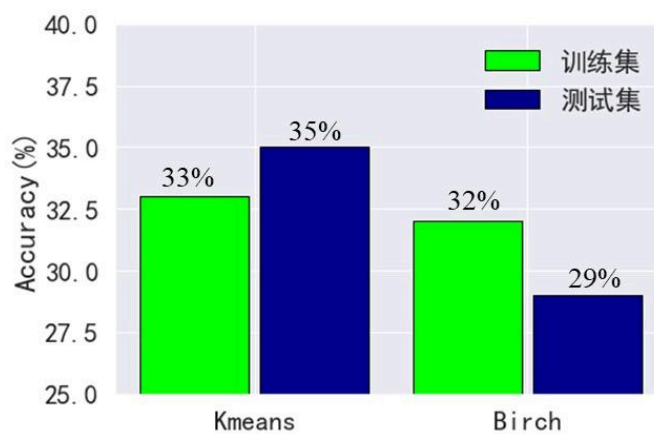


#### 四、sklearn——岩性聚类



中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

两种聚类方法岩性识别准确率



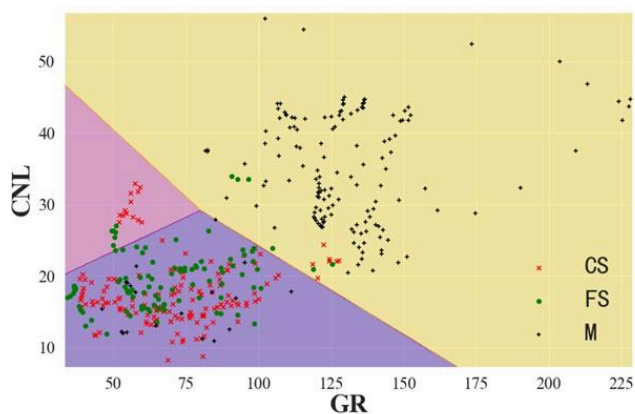
117

#### 四、sklearn——岩性识别



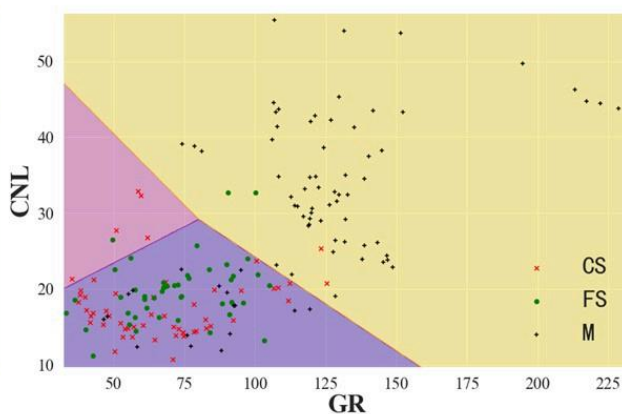
中国石油大学  
CHINA UNIVERSITY OF PETROLEUM

LDA分类结果（训练样本）



正判率67%

LDA分类结果（测试样本）



正判率61%

120