Python 数据分析与应用

实验六：使用 scikit-learn 构建模型

丁烨
dingye@dgut.edu.cn
计算机科学与技术学院

2023-12-13

東莞理工學院
DONGGUAN UNIVERSITY OF TECHNOLOGY

机器学习简介

- 机器学习（Machine Learning）
- 人工智能（Artificial Intelligence，AI）的一个分支
- 机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题
- 机器学习在近 30 多年已发展为一门多领域交叉学科
- 涉及概率论、统计学、逼近论、凸分析、计算复杂性理论等多门学科

- 机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法
- 机器学习算法从数据中自动分析获得规律，并利用规律对未知数据进行预测

机器学习简介

- 机器学习的核心问题是：
- 通过学习 n 个数据样本及其标注的结果，预测其它 k 个数据样本的结果

- 大部分机器学习的算法实际上都在解决这个问题，例如：
- AlphaGo 通过学习已存在的围棋选手的下棋方法，预测战胜其他围棋选手的下棋方法
- Tesla 的自动驾驶系统通过学习已存在的驾驶方法，预测不同路况的驾驶方法
- 人脸识别系统通过学习已知的人脸图片，预测一张图片是否为人脸

scikit-learn 简介

- scikit-learn
- https://scikit-learn.org/
- https://github.com/scikit-learn/scikit-learn

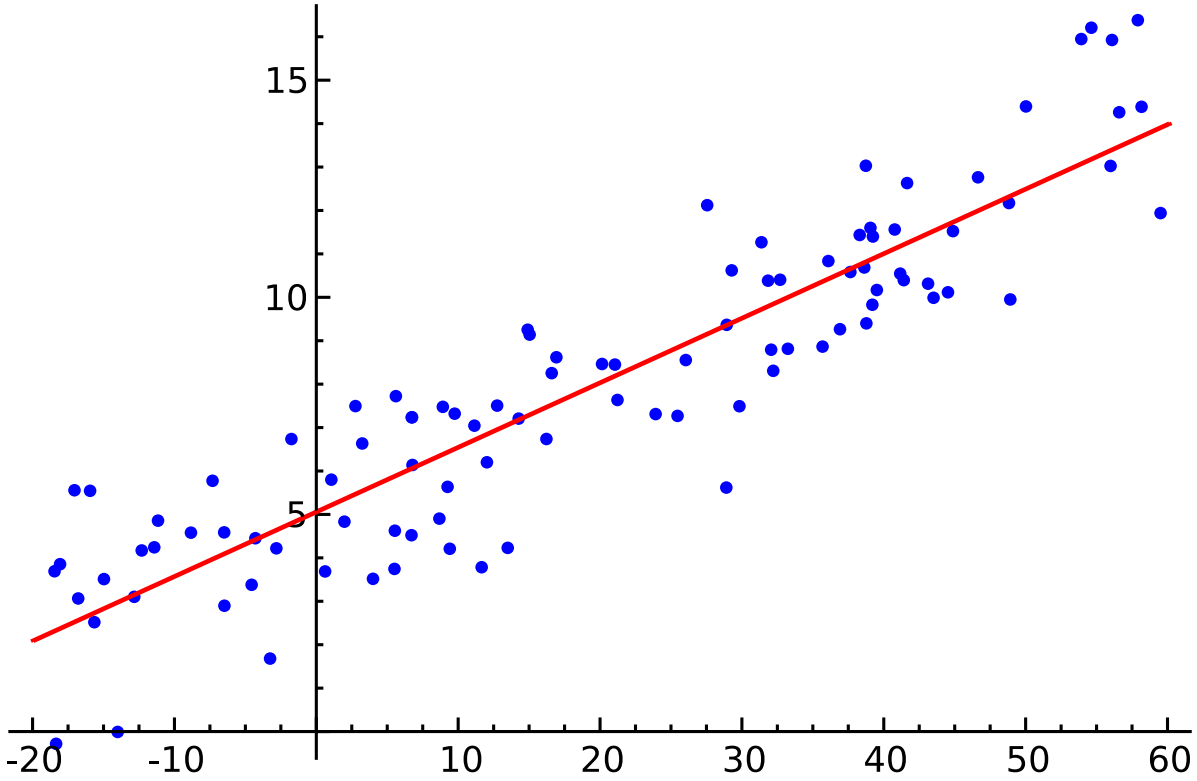
- 一个开源的基于 Python 的科学计算及机器学习工具包
- 属于 SciPy 项目的一部分
- scikit-learn 是一个非常基础、简单的机器学习工具包
- scikit-learn 本身不支持深度学习
- scikit-learn 不支持 GPU 加速
- scikit-learn 只提供了经过广泛验证的算法

scikit-learn 简介

- 使用 pip 安装 scikit-learn：
- `pip3 install --user -U scikit-learn`

- 如果安装不成功，可尝试使用 apt 安装：
- `sudo apt install python3-sklearn`

- 线性回归（Linear Regression）
- 找到一个线性方程尽可能的表达原始样本数据的分布
- 找到这个线性方程之后，每给定一个样本的特征，就能预测对应的标签



线性回归

lr.py

```
1  import numpy
2  from sklearn.linear_model import LinearRegression
3
4  X = numpy.array([[1, 1], [2, 2], [3, 3], [4, 4]])
5  y = numpy.array([1, 2, 3, 4])
6  m = LinearRegression().fit(X, y)
7  print(m.score(X, y))
8  print(m.predict(numpy.array([[5, 5]])))
9
```

线性回归

```
----------------------------------------
~/Workspace/test » python3 lr.py
1.0
[5.]
----------------------------------------
```

线性回归

lr.py

```
1  import numpy
2  from sklearn.linear_model import LinearRegression
3
4  X = numpy.array([[1, 1], [2, 2], [3, 3], [4, 4]])
5  y = numpy.array([1, 2, 3, 4])
6  m = LinearRegression().fit(X, y)
7  print(m.score(X, y))
8  print(m.predict(numpy.array([[5, 6]])))
9
```

```
----------------------------------------
~/Workspace/test » python3 lr.py
1.0
[5.5]
----------------------------------------
```

作业

- UCI 葡萄酒质量数据集
- https://archive.ics.uci.edu/ml/datasets/wine+quality

- 两套（红葡萄酒、白葡萄酒）的物理化学性质数据，已量化
- 这些葡萄酒对应的品质等级

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	1992906

作业

- 获取白葡萄酒数据：https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv
- 取后 100 条为测试数据，其他为训练数据
- 用训练数据训练一个线性回归模型
- 在测试数据上测试训练好的线性回归模型
- 使用均方误差（MSE）计算预测准确度

- 提供完整的代码
- 提供完整的实验结果截图

作业

```
~/Workspace/test » python3 lr.py
[[ 7.    0.27  0.36 ...  0.45  8.8   6.  ]
 [ 6.3   0.3   0.34 ...  0.49  9.5   6.  ]
 [ 8.1   0.28  0.4  ...  0.44 10.1   6.  ]
 ...
 [ 6.5   0.24  0.19 ...  0.46  9.4   6.  ]
 [ 5.5   0.29  0.3  ...  0.38 12.8   7.  ]
 [ 6.    0.21  0.38 ...  0.32 11.8   6.  ]]
[5.63144651 6.10115806 5.85674592 ... 5.34351476 6.57567039 6.35192467]
[5. 5. 5. ... 6. 7. 6.]
0.5630446931574506
```

作业

- 在作业系统中下载并完成本实验课对应实验报告
- https://hw.dgut.edu.cn/
- 注意：所有标识为 * 的地方都需要填写
- 截止日期：2023-12-20 23:59:59

课程名称：Python 数据分析与应用　　学期：20

实验名称	Python 语言回顾			实验序号	
姓　名	***	学　号	***	班　级	
实验地点	***	实验日期	***	指导老师	
教师评语	-			实验成绩	
				百分制	
同组同学	无				

四、　实验作业及分析

4.1　实验过程

1) *** 请将详细实验过程填写在此处 ***

4.2　实验结果

*** 请将实验结果截图填写在此处 ***

五、　实验总结

*** 请撰写一段 200 字左右的实验总结 ***

GOOD LUCK!