姓名和学号：李胜志 2210180232

章节名称：P61项目实训

知识目标：理解数据分析的过程；掌握机器学习常用库（NumPy、Pandas、Matplotlib）的使用方法；掌握使用Sklearn训练线性回归模型并进行预测的方法。

能力目标：掌握不同类型的线性回归模型（一元线性回归、多元线性回归）及其应用场景；能够解释模型的参数估计，包括回归系数w、截距b。

素质目标：学习基础知识，提高选择合适方法解决不同问题的能力；养成分析问题、事先做好准本的良好习惯

知识重点： 数据预处理；最小二乘法优化算法的使用。

知识难点：回归系数的统计解释，如斜率w、截距b等。

项目实训

1. 实训目的
2. 理解数据分析的过程。
3. 掌握机器学习常用库（NumPy、Pandas、Matplotlib）的使用方法。
4. 掌握使用Sklearn训练线性回归模型并进行预测的方法。
5. 实训内容

已知某公司员工工龄与平均工资之间有一定的相关关系（见表2-5），要求使用线性回归算法分析平均工资与工龄的关系。

表 2-5工龄与平均工资数据集

|  |  |  |  |
| --- | --- | --- | --- |
| 工龄/（年） | 平均工资/（元） | 工龄/（年） | 平均工资/（元） |
| 1 | 2000 | 6 | 10567 |
| 2 | 2200 | 7 | 9566 |
| 3 | 4900 | 8 | 15678 |
| 4 | 3221 | 9 | 13644 |
| 5 | 6834 | 10 | 15789 |

1. 启动Jupyter Notebook，以Python 3工作方式新建 Jupyter Notebook 文档，并重命名为“item2-sx.ipynb”。
2. 数据准备。
3. 导入NumPy库。
4. 使用NumPy定义两个数组，分别存放工龄数据与平均工资数据。

import numpy as np

work\_age = np.array([[1], [2], [3], [4], [5], [6], [7], [8], [9], [10]])

average\_salary = np.array([[2000], [2200], [4900], [3221], [6834], [10567], [9566], [15678], [13644], [15789]])

1. 训练线性回归模型。
2. 导入线性回归模型LinearRegression。
3. 使用LinearRegression建立基于工龄与平均工资数据集的线性回归模型。
4. 训练线性回归模型。

from sklearn.linear\_model import LinearRegression

lr = LinearRegression()

lr.fit(work\_age, average\_salary)

1. 绘制图像。
2. 导入Matplotlib库。
3. 定义图像的标题为“工龄与平均工资之间的关系图”。
4. 定义图像的横坐标为“工龄/年”，纵坐标为“平均工资/元”。
5. 将步骤②和③中出现的中文字体设置为黑体。
6. 使用scatter()函数绘制原始数据点。

import matplotlib.pyplot as plt

y = lr.predict(work\_age) # 预测

plt.title("工龄与平均工资之间的关系图") # 定义图像的标题为“工龄与平均工资之间的关系图”

plt.xlabel("工龄/年") # 定义图像的横坐标为“工龄/年”

plt.ylabel("平均工资/元") # 定义图像的纵坐标为“平均工资/元”

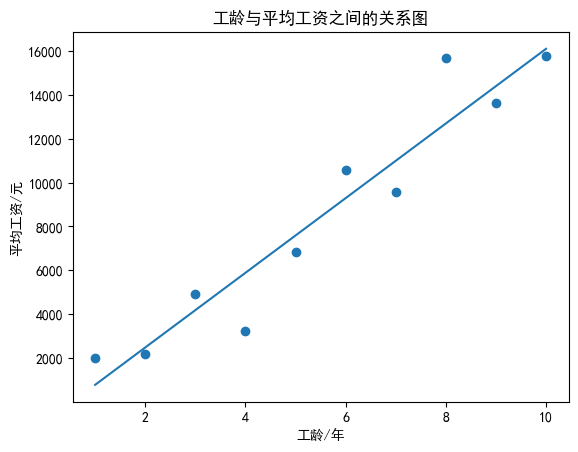
plt.rcParams['font.sans-serif'] = 'Simhei' # ④　中文字体设置为黑体。

plt.scatter(work\_age, average\_salary) # 使用scatter()函数绘制原始数据点。

plt.plot(work\_age, y)

plt.show()

运行结果：



1. 评估模型

计算模型的准确率、斜率和截距

score = lr.score(work\_age, average\_salary)

print("模型的准确率是: ", score)

print("斜率w = ", lr.coef\_, "截距b = ", lr.intercept\_)

运行结果：

模型的准确率是: 0.9122972589588493

斜率w = [[1702.22424242]]

截距b = [-922.33333333]