## 《人工智能基础A》实验报告三

### 机器学习回归与sklearn库的使用

学号 姓名 联系方式 .

一、实验目的

通过实验，学会机器学习回归和机器学习库sklearn的使用。具体目标要求如下：

1. 成功安装sklearn。
2. 掌握scikit-learn的基本用法，包括数据预处理、模型训练和预测。
3. 学习使用sklearn实现并评估一个回归模型。

二、实验内容及要求

**本次实验用Jupyter Notebook完成，可以在mo平台或本地完成，最终提交ipynb文件。**

**1. sklearn安装**

进入之前实验创建的虚拟环境，输入：

pip install scikit-learn

如若提示pip未安装，则需先输入：

conda install pip

**2. sklearn基本用法**

参考资料：<https://scikit-learn.org/0.21/documentation.html>

使用sklearn大致可分为4步：数据预处理、训练模型、预测模型以及模型评估。

**2.1 数据预处理** 这里的数据通常需要从文件读入并做一些预处理，下文示例的数据将直接由sklearn库提供。

import numpy as np

from sklearn.model\_selection import train\_test\_split

np.random.seed(0)

X = 2 \* np.random.rand(100, 1) # 生成100个随机点

y = 3 \* X.flatten() + 2 + np.random.randn(100) \* 0.5 # y = 3x + 2 加入噪声

# 将数据分为训练集和测试集

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

**2.2 模型预测**

from sklearn.linear\_model import LinearRegression

# 创建线性回归模型

model = LinearRegression()

# 训练模型

model.fit(X\_train, y\_train)

**模型预测主要就是创建sklearn库提供的模型实例（传入一定参数），然后直接调用模型实例的fit方法。**

**2.3 模型预测**

# 进行预测

y\_pred = model.predict(X\_test)

模型进行fit训练后便可以调用predict方法来对测试集进行预测。

**2.4 模型评估**

from sklearn.metrics import mean\_squared\_error, r2\_score

# 计算性能指标

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

最后需要对模型的预测结果对模型的性能进行评估。常用的评估函数有均方误差（MSE）与决定系数（R2）。

**2.5 结果可视化**

import matplotlib.pyplot as plt

# 原始数据与预测值的可视化

plt.scatter(X, y, color='blue', label='target', s=10)

X\_line = np.linspace(0, 2, 100).reshape(-1, 1) # 创建用于绘制回归线的X值

y\_line = model.predict(X\_line)

plt.plot(X\_line, y\_line, color='red', linewidth=2, label='predition')

plt.xlabel('X')

plt.ylabel('y')

plt.title("y = 3x + 2")

plt.legend()

plt.grid()

plt.show()

需要的话，可以对模型的输出进行可视化。

**3. 加州房价预测——示例**

下面提供一个示例，实验用到本实验将使用加州房价数据集。该数据集包含不同特征（如房间数、房龄、主人收入等）以及相应的房价信息。目标是构建一个模型来预测房价。

**1. 数据加载与查看**

from sklearn.datasets import fetch\_california\_housing()

import pandas as pd

# 加载数据集

cal = fetch\_california\_housing()

##注，若下载数据集失败，改用下面代码。其中xxx为数据集文件在你本地目录的位置，学在浙大将附上数据集。比如文件位置为"D:\Code\cal\_housing\_py3.pkz"（在文件资源管理器右键点击属性可看到文件路径），那么xxx填"D:\\Code "（所有的\都要写成\\）

cal = fetch\_california\_housing(data\_home="xxx",download\_if\_missing=False)

#得到样本特征与样本目标值

X = pd.DataFrame(cal.data, columns=cal.feature\_names)

y = cal.target

# 数据概览

print(X.head())

print(y[:5])

这里通过库函数直接加载数据集并查看数据集的基本情况。

**2. 数据预处理**

from sklearn.model\_selection import train\_test\_split

# 检查缺失值

print(X.isnull().sum())

# 划分数据集

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

这里检查数据的缺失值（数据集的某些样本可能数据不完整），并决定是否填充或删除这些值（加州房价数据集中没有缺失值），并将数据分为训练集和测试集，使用 80% 的数据用于训练，20% 用于测试：

**3. 模型训练**

from sklearn.linear\_model import LinearRegression

# 创建模型

model = LinearRegression()

# 训练模型

model.fit(X\_train, y\_train)

这里创建线形回归模型，并进行拟合训练。

**4. 模型评估**

from sklearn.metrics import mean\_squared\_error, r2\_score

# 进行预测

y\_pred = model.predict(X\_test)

# 计算性能指标

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

# 打印性能指标

print(f'均方误差 (MSE): {mse}')

print(f'决定系数 (R^2): {r2}')

这里对训练好的模型进行评估。对测试集进行预测，并计算模型性能指标。

**5. 可视化结果**

import matplotlib.pyplot as plt

plt.scatter(y\_test, y\_pred, color='blue', alpha=0.6)

plt.xlabel('target')

plt.ylabel('prediction')

plt.title('target vs prediction')

plt.plot([min(y\_test), max(y\_test)], [min(y\_test), max(y\_test)], color='red', linewidth=2) # 理想预测线

plt.show()

这里通过通过绘制真实值与预测值的散点图，直观展示模型的拟合效果。

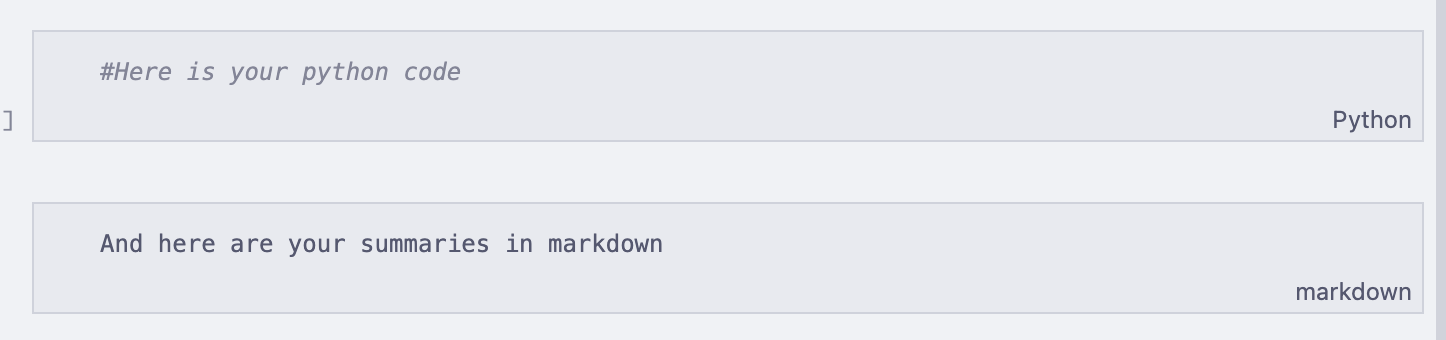
**4. 乳腺癌检测——作业**

**仅提交ipynb文件！！！**

**ipynb文件内不需要加州房价预测的代码！！！**

**文件命名为lab3-姓名-学号.ipynb！！！**

**最后的代码合并在一个notebook的一个cell里。心得体会与思路也附在ipynb文件中，用markdown（不会markdown语法的也可以只输入纯文字）在代码的后面新开一个cell。可以在notebook界面的上面选择cell的属性为Code或者Markdown。**

****

**通过sklearn加载乳腺癌数据集并进行乳腺癌预测，训练一个逻辑回归模型，预测新的样本是良性还是恶性。自行对数据集特征进行观察与分析，决定是否对训练集进行标准化或归一化等预处理，自行设置模型的超参数。**

**分离训练集和数据集时调用train\_test\_split函数时训练集占75%，测试集占25%，random\_state设置为3149。**

**最后的评估标准为对测试样本进行预测的准确率：预测正确的测试样本数/总的测试样本数。**

**乳腺癌数据集不需要从网上额外下载，sklearn已自带。**