《人工智能基础 A》实验报告三

机器学习回归与 sklearn 库的使用

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

```
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)
```

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。

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（R^2）。

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()
```

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。

3.1 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。

3.2 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：

3.3 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。

3.4 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。

3.5 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。

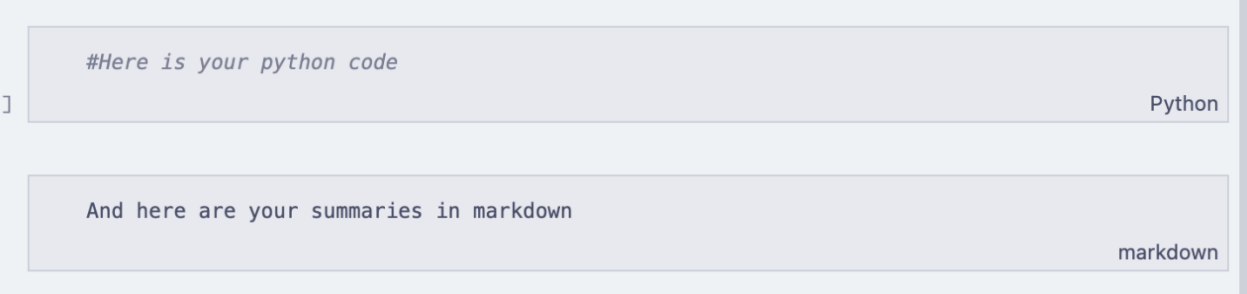
4. 乳腺癌检测——作业

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。