机器学习1030作业

```
利用Linear Regression模型，根据足长和步幅预测身高
```

重要性检验

- 利用t检验判断特征和预测值之间的关系
  - 用0.05作为重要性阈值
  - 得出结论：足长和步幅都和身高相关，可以作为预测的特征

```
# 重要性检验
t_statistic, p_value = ttest_ind(df['足长'], df['身高'])
alpha = 0.05
if p_value < alpha:
    print("足长和身高存在显著差异")
else:
    print("足长和身高无显著差异")
```

数据预处理

- 足长和步幅作为特征X，身高作为预测值y

```
X = df[['足长','步幅']]
y = df[['身高']]
```

- 训练集和测试集按照8:2划分

```
# 按照8：2划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

模型训练

- 初始化Linear Regression模型

```
# 模型训练
lr = LinearRegression()
```

- 在训练集上训练，并输出拟合的参数

```
lr.fit(X_train, y_train)
# 参数
print("权重：", lr.coef_)
print("截距：", lr.intercept_)
```

模型评价

- 用训练好的模型预测测试集上数据的身高
  - 在测试集上预测

```
y_pred = lr.predict(X_test)
```

- 计算不同评价指标
  - 均方误差 (MSE)为1.02，决定系数 (R^2)为0.98，可见预测值和真实值之间的差别不大，模型性能理想

```
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("均方误差 (MSE):", mse)
print("决定系数 (R^2):", r2)
```

- 模型可视化
  - 绘制三维图，同时将真实值和预测值表示在图上，比较两者的区别

```
# 可视化图
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D  # 用Axes3D库画3D模型图

x1_data = X_test.drop(columns=['步幅'], axis=1)
x2_data = X_test.drop(columns=['足长'], axis=1)

# 创建一个3D图形窗口
fig = plt.figure(figsize=(8, 6))

# 创建3D坐标轴
ax3d = fig.add_subplot(111, projection='3d')
```

```
# 绘制散点图
ax3d.scatter(x1_data, x2_data, y_test, color='b', marker='*',
label='actual')
ax3d.scatter(x1_data,x2_data,y_pred,color='r',label='predict')

ax3d.set_xlabel('foot length')  # 设置x轴标签
ax3d.set_ylabel('stride')  # 设置y轴标签
ax3d.set_zlabel('height')  # 设置z轴标签
plt.legend(loc='upper left')
plt.show()
```

- 下图可以看出，真实值和预测值基本在同一平面上，因此模型性能较为理想