人工智能程序设计


python
import turtle
turtle.setup(650,350,200,200)
turtle.penup()
turtle.fd(-250)
turtle.pendown()
turtle.pensize(25)
turtle.pencolor("purple")
turtle.seth(-40)
for 1 in range(4):
turtle.circle(40, 80)
turtle.circle(-40, 80)
turtle.fd(40)
turtle.circle(40, 80/2)
turtle.fd(40)
turtle.circle(16, 180)
turtle.fd(40 * 2/3)

人工智能程序设计

11.4 机器学习实践案例

北京石油化工学院 人工智能研究院

刘 强

11.4.1 分类任务实践：泰坦尼克号生存预测

泰坦尼克号生存预测是机器学习入门的经典案例：

- **数据量适中**：约800条训练数据
- **特征丰富**：数值型（年龄、票价）和类别型（性别、船舱等级）
- **问题明确**：预测乘客是否生存，典型的二分类问题
- **现实意义**：具有历史背景，能激发学习兴趣

步骤1：数据准备

创建模拟数据集，包含乘客的基本信息和生存标签：

```
## 创建模拟数据集的核心逻辑
data = {
    'Age': np.random.normal(30, 12, n_samples),
    'Sex': np.random.choice(['male', 'female'], n_samples),
    'Pclass': np.random.choice([1, 2, 3], n_samples),
    'Fare': np.random.exponential(30, n_samples),
}

## 根据特征生成生存标签
for i in range(n_samples):
    prob = 0.3  # 基础生存概率
    if data['Sex'][i] == 'female': prob += 0.4  # 女性生存率更高
    if data['Pclass'][i] == 1: prob += 0.3      # 头等舱生存率更高
```

步骤2：探索性数据分析

分析不同特征与生存率的关系，发现数据中的规律：

```
## 分析不同特征与生存率的关系
sex_survival = df.groupby('Sex')['Survived'].mean()
class_survival = df.groupby('Pclass')['Survived'].mean()

## 可视化特征分布和相关性
plt.figure(figsize=(15, 12))
## 生存率分布、特征分布、相关性热力图等
```

步骤3：数据预处理

处理缺失值、编码分类变量、创建新特征：

```
## 处理缺失值
df['Age'].fillna(df['Age'].median(), inplace=True)

## 编码分类变量
le = LabelEncoder()
df['Sex'] = le.fit_transform(df['Sex'])

## 特征工程
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
df['IsAlone'] = (df['FamilySize'] == 1).astype(int)
```

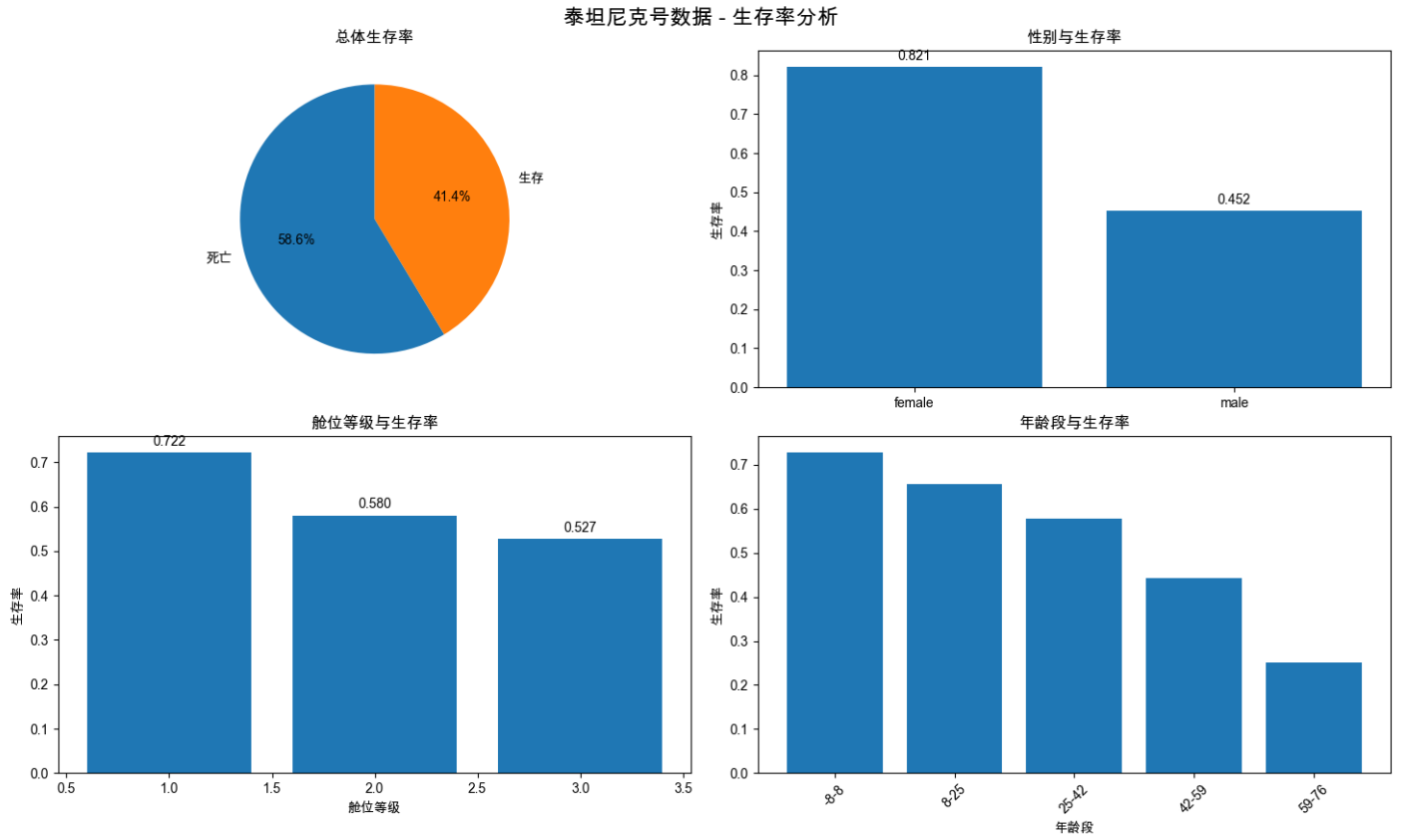
步骤4：模型训练与评估

准备数据并训练多个模型进行比较：

```
## 准备训练数据
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

## 训练多个模型
models = {
    '决策树': DecisionTreeClassifier(max_depth=5, random_state=42),
    '随机森林': RandomForestClassifier(n_estimators=100, random_state=42),
    '逻辑回归': LogisticRegression(random_state=42)
}

## 评估模型性能
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    print("{} 准确率: {:.3f}".format(name, accuracy))
```

泰坦尼克号生存率分析

图 11.4.1 泰坦尼克号生存率分析



11.4.2 回归任务实践：房价预测

房价预测是回归任务的典型应用：

- 目标是根据房屋特征预测其价格
- 展示如何处理连续数值型的目标变量
- 学习评估回归模型的性能指标

数据集特征：房屋面积、卧室数量、浴室数量、房龄、地段等级、是否有车库、是否有花园

步骤1：数据准备

创建模拟房价数据集，根据特征生成合理的房价：

```
## 创建模拟房价数据集
data = {
    'Size': np.random.normal(150, 50, n_samples),     # 房屋面积
    'Bedrooms': np.random.poisson(3, n_samples),      # 卧室数量
    'Location': np.random.choice([1, 2, 3, 4, 5], n_samples),  # 地段等级
    'Age': np.random.exponential(10, n_samples),      # 房龄
}

## 根据特征生成房价
for i in range(n_samples):
    price = base_price + data['Size'][i] * 2000  # 面积影响
    price *= location_multiplier[data['Location'][i]]  # 地段影响
```

步骤2：特征工程

创建新的衍生特征，提升模型预测能力：

```
## 创建新特征
df['PricePerSqm'] = df['Price'] / df['Size']  # 每平米价格
df['TotalRooms'] = df['Bedrooms'] + df['Bathrooms']  # 总房间数
df['RoomRatio'] = df['Bathrooms'] / df['Bedrooms']  # 浴室卧室比

## 分类特征编码
df = pd.get_dummies(df, columns=['SizeCategory', 'AgeCategory'])
```

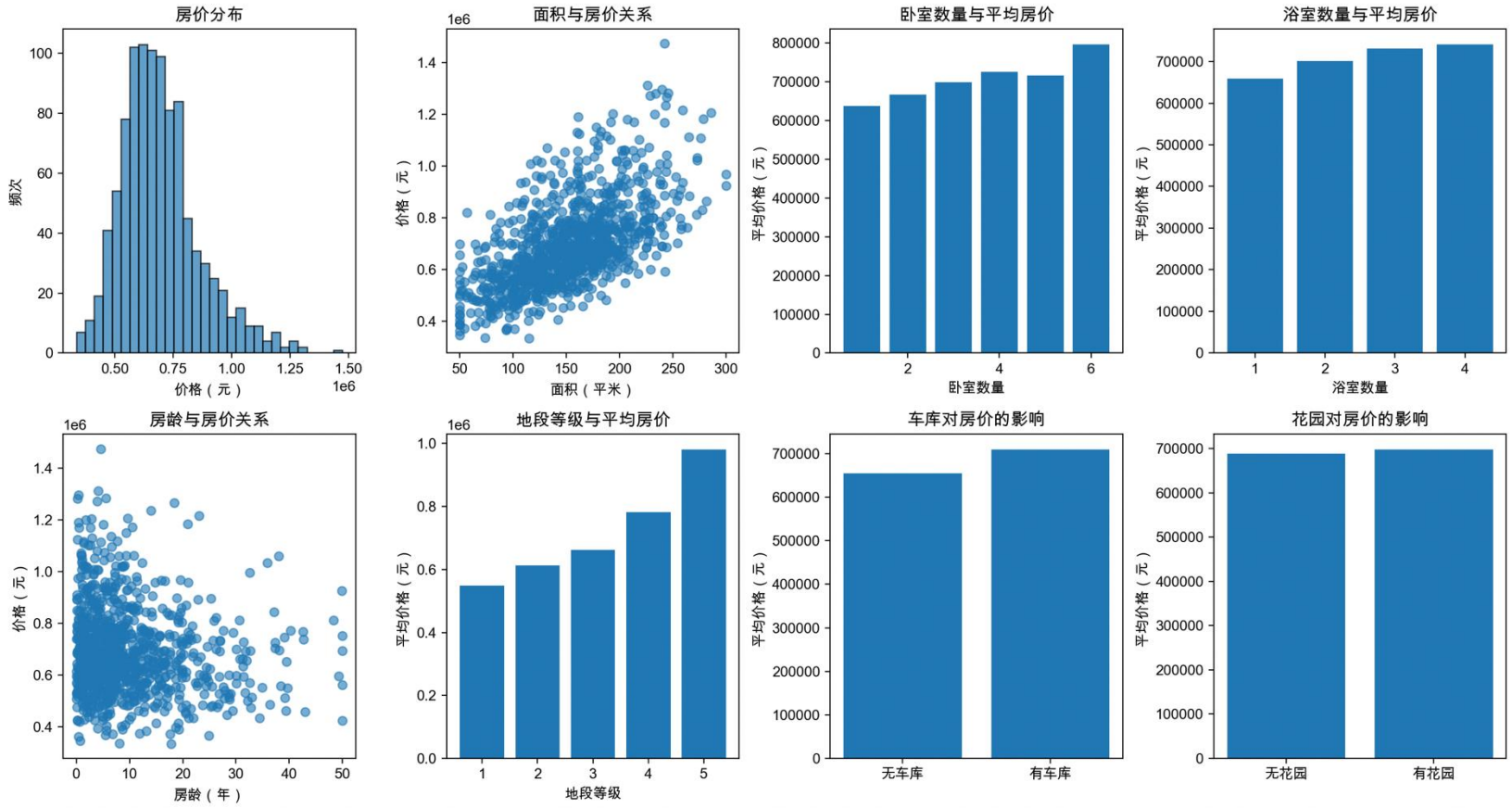
步骤3：模型训练与评估

训练多个回归模型并比较性能：

```
## 训练多个回归模型
regression_models = {
    '线性回归': LinearRegression(),
    '决策树回归': DecisionTreeRegressor(max_depth=10, random_state=42),
    '随机森林回归': RandomForestRegressor(n_estimators=100, random_state=42)
}

## 评估回归性能
for name, model in regression_models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    r2 = r2_score(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    mae = mean_absolute_error(y_test, y_pred)

    print("{} - R²: {:.3f}, RMSE: {:.0f}, MAE: {:.0f}".format(name, r2, rmse, mae))
```

房价回归分析

图 11.4.2 房价回归分析



11.4.3 Ask AI：Kaggle竞赛入门

Kaggle 是全球最大的数据科学竞赛平台：

- 接触真实的商业问题
- 学习最佳实践
- 在竞争中快速提升技能

推荐入门竞赛

泰坦尼克号生存预测

- 最适合初学者的竞赛
- 数据集小巧，问题清晰
- 学习分类任务的绝佳起点

房价预测竞赛

- 学习回归任务和特征工程技巧

手写数字识别

- 理解图像分类的基本方法

实践练习

练习 11.4.1：分类模型优化

1. 在泰坦尼克案例中尝试不同的特征工程方法
2. 比较决策树、随机森林、逻辑回归的性能
3. 分析模型的优缺点和适用场景

实践练习

练习 11.4.2：销量预测模型

1. 构建商品销量预测数据集，包含特征：价格、促销力度、季节、广告投入、库存量等
2. 尝试不同的回归算法（线性回归、决策树回归、随机森林回归）
3. 使用交叉验证评估模型稳定性，分析哪些特征对销量影响最大

实践练习

练习 11.4.3：Kaggle实践

1. 注册 Kaggle 账号，浏览泰坦尼克竞赛
2. 下载数据集，完成基本的数据分析
3. 提交一次预测结果，体验完整流程

本节小结

- **分类任务**：泰坦尼克号生存预测，预测离散类别
- **回归任务**：房价预测，预测连续数值
- **完整流程**：数据准备→探索分析→预处理→模型训练→性能评估
- **Kaggle平台**：提供真实项目实践机会