实验三

中国科大2024年春季学期"数据分析及实践"课程 - 实验三说明文档。

任务概述

威斯康辛州乳腺癌数据集（Breast Cancer Wisconsin Dataset）由威斯康辛州医院的Dr. William H. Wolberg收集得到，涵盖了从569个患者收集的乳腺肿瘤特征的测量值及相应肿瘤的诊断标签（良性，恶性）。现欲对该数据集进行探索性分析、可视化展示和简单统计推断，请你按要求编写 Python 代码实现任务列表中的内容。

任务列表

1.（33%）导入`pandas`库，并使用相关方法进行数据集读取、基本信息处理和探索性分析等操作（各子任务请分别使用一行代码完成）。

- Q1.（3%）以`UTF-8`编码格式读取数据集`data.csv`，存储到变量`df`中。
- Q2.（3%）使用`head()`方法展示当前数据集的前10行。
- Q3.（3%）使用`info()`方法展示数据集基本信息：数据集有多少行？有多少列？列名称和类型分别是什么？每一列非空值各自有多少个？
- Q4.（3%）将数据集中含有缺失值的行删除，并将更新后的结果保存到原变量`df`中。
- Q5.（3%）执行上一步后，数据集的行数减少了，然而索引的范围并没有因此改变。请重置该数据集的索引（注意不要新增`index`列到数据集中，只需更新索引），并将更新后的结果保存到原变量`df`中。
- Q6.（3%）将数据集中的`id`列删除，并将更新后的结果保存到原变量`df`中。
- Q7.（3%）数据集中`diagnosis`列代表该患者的肿瘤类型，其中`B`代表良性，`M`代表恶性，请输出这两种类型的计数结果。
- Q8.（4%）为了方便后续处理，请使用`apply()`的相关方法，将`diagnosis`这一列中的值`B`修改为0，值`M`修改为1，并将更新后的结果保存到原变量`df`中。
- Q9.（4%）选取除`diagnosis`以外的五个数值型特征，获取它们的统计信息（平均值、标准差、四分位点、最小值、最大值）。
- Q10.（4%）请使用`groupby()`的相关方法，获取不同`diagnosis`值的各组数据所有特征的变异系数。

2.（17%）导入`numpy`和`matplotlib`库，对数据集`df`进行一定数据可视化分析。

- Q1.（5%）选取除`diagnosis`以外的一个数值型特征，绘制其频率分布直方图。请将条柱数设置为10，条柱填充色为绿色，条柱边框色为黑色，并配上合适的标题。
- Q2.（5%）选取除`diagnosis`以外的两个`mean`类型特征，绘制分布散点图。请将`diagnosis`值为1的点标为红色，值为0的点标为蓝色，散点大小设置为10，并配上合适的标题和图例注记。
- Q3.（7%）选取`diagnosis`及其它任意五个`mean`类型特征，求它们的Pearson相关系数矩阵，并绘制相应的相关系数热力矩阵图。请为每个位置增添对应数值表示（保留三位小数），利用`colorbar`设置数值与颜色的对应关系条，并配上合适的标题和坐标表示。

3.（18%）线性回归是一类经典的统计建模方法。基于响应变量数据 $\mathbf{y} \in \mathbb{R}^n$ 和特征向量数据 $X \in \mathbb{R}^{n\times d}$，使用线性回归模型 $y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_dx_d$ 进行建模时，若 X 为列满秩矩阵，则模型参数向量 $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ 的最小二乘估计有矩阵形式解 $\hat{\mathbf{w}} = (\widetilde{X}^\top\widetilde{X})^{-1}\widetilde{X}^\top\mathbf{y}$，其中 $\widetilde{X} = (\mathbf{1}_n, X)$。

请以`area_mean`为响应变量，`radius_mean`为特征变量，利用`numpy`库的矩阵计算方法和`matplotlib`库的绘图方法，完成以下任务。

- Q1.（5%）利用二次多项式 $y = w_0 + w_1x + w_2x^2$ 进行回归建模，求模型参数 $\mathbf{w}$ 的最小二乘估计。
- Q2.（3%）使用`numpy.polyfit()`方法做二次多项式拟合，求相应模型参数估计，比较与上一步所得估计的结果。
- Q3.（6%）绘制原始数据分布散点图，设置散点颜色为红色，并将拟合曲线添加到散点图中，设置颜色为蓝色，要求图例注记标题坐标等信息清晰美观。
- Q4.（4%）请从特征含义与散点图分布情况阐述线性拟合是否适用于本场景。

4.（18%）数据降维，即选择性地削减数据集的属性维度，可以在牺牲一小部分信息的情况下大幅增加数据处理的效率。其中主成分分析是一种应用非常广泛的数据降维方法。

在本数据集中，`perimeter_mean`和`radius_mean`具有较强的线性性，故而考虑采用主成分分析进行降维，请执行`X = df[["perimeter_mean", "radius_mean"]].to_numpy()`，利用`numpy`库的矩阵计算方法和`matplotlib`库的绘图方法，完成以下任务。

- Q1.（6%）利用`numpy.cov()`和`numpy.linalg.eig()`方法，求`X`的协方差矩阵`corX`，进而求`corX`的特征值`eigV`与特征向量矩阵`eigMat`，通过计算验证`eigMat`的正交性。
- Q2.（7%）计算`X`与`eigMat`的矩阵乘积`Z`，以`X`的第一列为横坐标、第二列为纵坐标，绘制数据分布散点，设置散点颜色为红色，大小为3；同时，以`Z`的第一列为横坐标、第二列为纵坐标，绘制数据分布散点，设置散点颜色为蓝色，大小为3。将以上两部分散点绘制在同一张图中，并简述它们之间的关系。
- Q3.（5%）计算`Z`的协方差矩阵`corZ`，比较`corZ`与`eigV`的结果；基于以上分析，请删除`Z`的一维数据，完成主成分分析降维过程。

5.（14%）假设检验是数理统计学中根据一定假设条件由样本推断总体性质的方法，在统计推断中的地位举足轻重，其中t检验是一类非常重要的假设检验方法。

抽取`concavity_worst`特征数据并根据`diagnose`的值进行分组，现要使用t检验方法对两组样本间的均值差异进行推断，请阅读本文档的内容，导入`scipy`库，完成以下任务。

- Q1.（3%）简述本情景下应使用成组检验还是成对检验。
- Q2.（3%）计算两组数据的平均值，写出单侧检验原假设。
- Q3.（4%）使用`scipy.stats`中的相关方法，执行两样本单侧t检验。
- Q4.（4%）简述你从以上两样本t检验结果中得到的结论。

格式要求

1. 请按具体任务分步编写代码，存储于`.ipynb`格式文件中用于复现，必要时可增加注释。
2. 实验报告必须涵盖任务列表中的所有内容和相应结果，并请存储于`.pdf`格式文件中。

参考资料

以下资料可能会对你顺利完成实验有所帮助。

1. Kaggle 数据挖掘与预测竞赛平台网站：点击这里
2. 数据分析库 pandas 官方网站：点击这里
3. 科学计算库 numpy 官方网站：点击这里
4. 数据可视化库 matplotlib 官方网站：点击这里
5. 科学计算库 scipy 官方网站：点击这里