|  |  |
| --- | --- |
| 班级：数大20-2  学号：2020305267  姓名：彭睿、 | 实验课程：机器学习  实验名称：Diabetes糖尿病数据集PCA重要属性判别  实验时间：2023-4-20 |
| 实验内容 | 一、利用自带数据集Diabetes糖尿病数据集，选取一个特征，使用一元线性回归分析模型来预测糖尿病人的疾病指标。其中，Diabetes包含442个患者10个生理特征和一年以后的疾病级数指标。  要求：  1、测试数据为最后20行，其余数据为训练数据；  2、输出最后预测的均方误差mse，以及可决系数R2；  3、利用PCA分析解释属性中哪些是重要属性，并输出各个特征的可解释方差比pca.explained\_variance\_ratio\_ |
| 实验代码 | *# -\*- coding = utf-8 -\*- # @Time : 2023/4/21 14:06 # @Author : 彭睿 # @File : diabetes.py # @Software : PyCharm* import matplotlib.pyplot as plt from sklearn import linear\_model import numpy as np import pandas as pd from sklearn.model\_selection import train\_test\_split from sklearn.linear\_model import LinearRegression from sklearn.metrics import mean\_squared\_error, r2\_score from sklearn.decomposition import PCA from sklearn import datasets  *# 数据集* diabetes = datasets.load\_diabetes() *# 载入数据  # 获取一个特征* diabetes\_x\_temp = diabetes.data[:, np.newaxis, 2]  diabetes\_x\_train = diabetes\_x\_temp[:-20] *# 训练样本* diabetes\_x\_test = diabetes\_x\_temp[-20:] *# 测试样本 后20行* diabetes\_y\_train = diabetes.target[:-20] *# 训练标记* diabetes\_y\_test = diabetes.target[-20:] *# 预测对比标记  # 回归训练及预测* clf = linear\_model.LinearRegression() clf.fit(diabetes\_x\_train, diabetes\_y\_train) *# 注: 训练数据集  # 系数 残差平法和 方差得分* print('Coefficients :\n', clf.coef\_) print("Residual sum of square: %.2f" % np.mean((clf.predict(diabetes\_x\_test) - diabetes\_y\_test) \*\* 2)) print("variance score: %.2f" % clf.score(diabetes\_x\_test, diabetes\_y\_test))  *# 绘图* plt.title(u'LinearRegression Diabetes of BMI') *# 标题* plt.xlabel(u'Attributes') *# x轴坐标* plt.ylabel(u'Measure of disease') *# y轴坐标 # 点的准确位置* plt.scatter(diabetes\_x\_test, diabetes\_y\_test, color='black') *# 预测结果 直线表示* plt.plot(diabetes\_x\_test, clf.predict(diabetes\_x\_test), color='blue', linewidth=3) plt.show()   print("这是一条分界线-------------------------------------------------------这是一条分界线")   *# 加载数据集* diabetes = datasets.load\_diabetes()  *# 将数据集分为特征和目标变量* X = diabetes.data y = diabetes.target  *# 将数据集分为训练数据和测试数据* X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.045, random\_state=42)  *# 拟合一元线性回归模型* model = LinearRegression() model.fit(X\_train[:, 2].reshape(-1, 1), y\_train)  *# 进行预测* y\_pred = model.predict(X\_test[:, 2].reshape(-1, 1))  *# 计算均方误差* mse = mean\_squared\_error(y\_test, y\_pred) print("均方误差：", mse)  *# 计算可决系数R2* r2 = r2\_score(y\_test, y\_pred) print("可决系数R2：", r2)  *# 进行PCA分析* pca = PCA() pca.fit(X)  *# 获取可解释方差比和贡献度累计比* explained\_var\_ratio = pca.explained\_variance\_ratio\_ cumulative\_var\_ratio = np.cumsum(explained\_var\_ratio)  *# 构建第一个DataFrame* df1 = pd.DataFrame({"属性名": diabetes.feature\_names, "可解释方差比": explained\_var\_ratio}) df1 = df1.sort\_values(by=["可解释方差比"], ascending=False) print("可解释方差比DataFrame：") print(df1)  *# 构建第二个DataFrame* df2 = pd.DataFrame({"属性名": diabetes.feature\_names, "贡献度累计比": cumulative\_var\_ratio}) df2 = df2.sort\_values(by=["贡献度累计比"], ascending=False) print("贡献度累计比DataFrame：") print(df2) |
| 运行结果及分析 | 实验结果如下：   1. 均方误差以及可决系数：     “Coefficients”是对糖尿病数据集中的BMI属性拟合的一元线性回归方程的系数，为938.23786125，  "Residual Sum of Squares"（残差平方和）是回归模型中的一个重要指标，它表示实际观测值与预测值之间的差异。计算方法是将每个观测值的预测值与实际值之差的平方相加，以得到一个总和。在一个好的回归模型中，残差平方和应该尽可能地小，因为这意味着预测值与实际值之间的差异越小。  "Variance score"（方差得分）是另一个回归模型的性能指标，它表示模型对数据的解释程度。方差得分的范围在0到1之间，其中1表示模型完全解释了数据的变化，0表示模型不能解释任何数据的变化。一个好的回归模型的方差得分应该尽可能接近1。在本例中，方差得分为0.47，这意味着回归模型解释了一部分数据的变化，但还有一部分数据的变化没有得到解释。   1. 绘图表示     **图 测试集BMI属性的可解释方差比**   1. PCA分析解释重要属性，可解释方差比:     **图 可解释性方差比**    **图 贡献度累计比**  均方误差(Mean Squared Error, MSE)：它是预测值和实际值之间差异的平方的平均值。在这个数据集中，MSE为5262.625880180969，这意味着模型的预测结果与实际值的差异较大。  可决系数(R-Squared, R2)：它衡量模型对观测值变异性的解释能力。它的值范围从0到1，值越接近1，模型对数据的解释能力越强。在这个数据集中，R2为0.21640939003029047，这意味着模型对数据的解释能力较弱。  可解释方差比(Explained Variance Ratio)DataFrame：它显示了每个属性对总方差的贡献程度。总方差是指所有观测值与其均值之差的平方和。在这个数据集中，年龄(age)属性对总方差的贡献最大，为0.402421，其次是性别(sex)属性和BMI指数(bmi)属性。  贡献度累计比(Cumulative Contribution Ratio)DataFrame：它显示了每个属性对总方差的累计贡献程度。与可解释方差比不同，它按照属性的贡献大小进行排列，并将每个属性的贡献与之前的属性累加。在这个数据集中，S6属性对总方差的贡献最大，其次是S5属性和S4属性。这个结果还显示，所有属性的贡献累计达到100％。 |
| 实验总结 | 本次实验使用Diabetes自带数据集，选择其中的一个特征，使用一元线性回归分析模型来预测糖尿病人的疾病指标。训练数据为前422个患者，测试数据为后20个患者。下面是本次实验的主要结果。  模型的系数为938.24，残差平方和为2548.07，可决系数为0.47。根据可决系数的值来看，该属性（BMI）对糖尿病指标的解释能力较弱。  通过PCA分析，我们可以看到各个特征的可解释方差比和贡献度累计比。其中，前四个特征age、sex、bmi、bp的可解释方差比之和为0.7688，超过了80%的可解释方差比，这表明这些特征对数据的贡献较大，是重要的属性。同时，我们也可以看到s6这个特征的贡献度累计比为1，说明它对数据的解释能力非常弱，可以被忽略。  总体来说，本次实验的回归模型对糖尿病指标的解释能力较弱，需要进一步优化模型或者增加更多的特征来提高预测准确度。同时，通过PCA分析，我们也能够确定哪些特征是重要的属性，可以在后续的特征选择和模型优化中加以利用。 |