**机器学习与数据挖掘**

**实验报告**

|  |  |
| --- | --- |
| 实验名称 | 降维 |
| 专 业 | 软件工程 |
| 学 号 | 21310385 |
| 姓 名 | 张城伟 |
| 完成日期 | 2023.11.04 |

# 降维：主成分分析

## PCA算法流程

*PCA（principal components analysis）*即主成分分析技术，又称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标。

在统计学中，*PCA*是一项用于简化数据集的技术。它采用线性变换将数据映射到一个新的坐标系统中，以便数据在新坐标系下的投影具有最大的方差。任何数据投影的第一大方差位于第一个坐标(称为第一主成分)上，第二大方差位于第二个坐标(第二主成分)上，依次类推。

主成分分析经常用于减少数据集的维数，通过保留低阶主成分、忽略高阶主成分，保持数据集的对方差贡献最大的特征。这样，低阶成分往往能保留数据的最重要方面。

主成分分析的算法流程如下所示：

|  |
| --- |
| **主成分分析算法流程** |
| ***Input*：**样本集 |
| ***Step 1.*** 对所有样本进行中心化： |
| ***Step 2.*** 计算样本的协方差矩阵或相关系数矩阵 |
| ***Step 3.*** 对协方差矩阵或相关系数矩阵进行特征值分解 |
| ***Step 4.*** 依据特征值计算方差的贡献率，先将所有特征值按照降序进行排序，选择降维后的数据方差占比超过阈值*t*%*的k*个特征值对应的特征向量： |
| **Step 5.** 选择的最大的特征向量组成投影矩阵*W*，得到降维后数据集 |
| ***Output***：降维后数据集*Y* |

具体代码如下：

*# 读取数据集*

train\_data = pd.read\_csv("train\_data.csv")

X = train\_data.values.T

print(X)

*# 计算与存储平均值*

X\_means = np.mean(X, axis=1, keepdims=True)

X\_centered = X - X\_means

*# 求取相关系数矩阵*

cov\_matrix = np.cov(X\_centered)

*# 求特征值与特征向量*

eigenvalues, eigenvectors = np.linalg.eig(cov\_matrix)

*# 字典类型*

eigenpairs = [(eigenvalues[i], eigenvectors[:, i]) for i in range(len(eigenvalues))]

*# 排序*

eigenpairs = sorted(eigenpairs, reverse=True)

*# 选择降维后的维度k*

total\_v = sum(eigenvalues)

v\_ratio=[(i/total\_v) for i in [eigenpair[0] for eigenpair in eigenpairs[:]]]

cumulative\_v = np.cumsum(v\_ratio)

*# 设置阈值，这里为99%*

threshold = 0.99

k = np.argmax(cumulative\_v >= threshold) + 1 *# 本实验中, k值计算结果为2*

*# 组成投影矩阵*

top\_k\_eigenvectors=np.array([eigenpair[1] for eigenpair in eigenpairs[:k]])

W = top\_k\_eigenvectors

*# 得到降维到k维后的数据集Y*

Y = np.dot(W, X\_centered)

*# 选择第一主成分和第二主成分进行绘图*

if k >= 2:

    plt.scatter(Y[0], Y[1])

    plt.xlabel('Principal Component 1')

    plt.ylabel('Principal Component 2')

    plt.title('PCA Reduced Data')

plt.show()

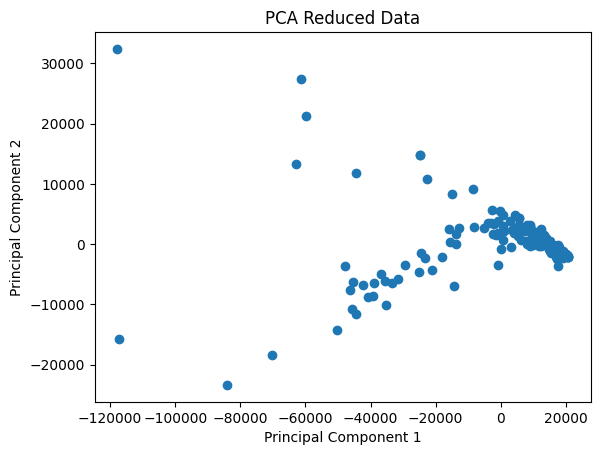
值得注意的是，本代码采用的np.cov函数用于计算相关系数矩阵，而相关系数矩阵为标准后的协方差矩阵，即将协方差值除以对应的两个变量的标准差。

当然，这两种方法求得的矩阵的特征值虽有所不同，但大小排列一致，特征向量也是完全一致的，因此对本实验没有任何影响。

## PCA结果分析

**① 主成分图像分析**

实验所得运行图像如下图所示：



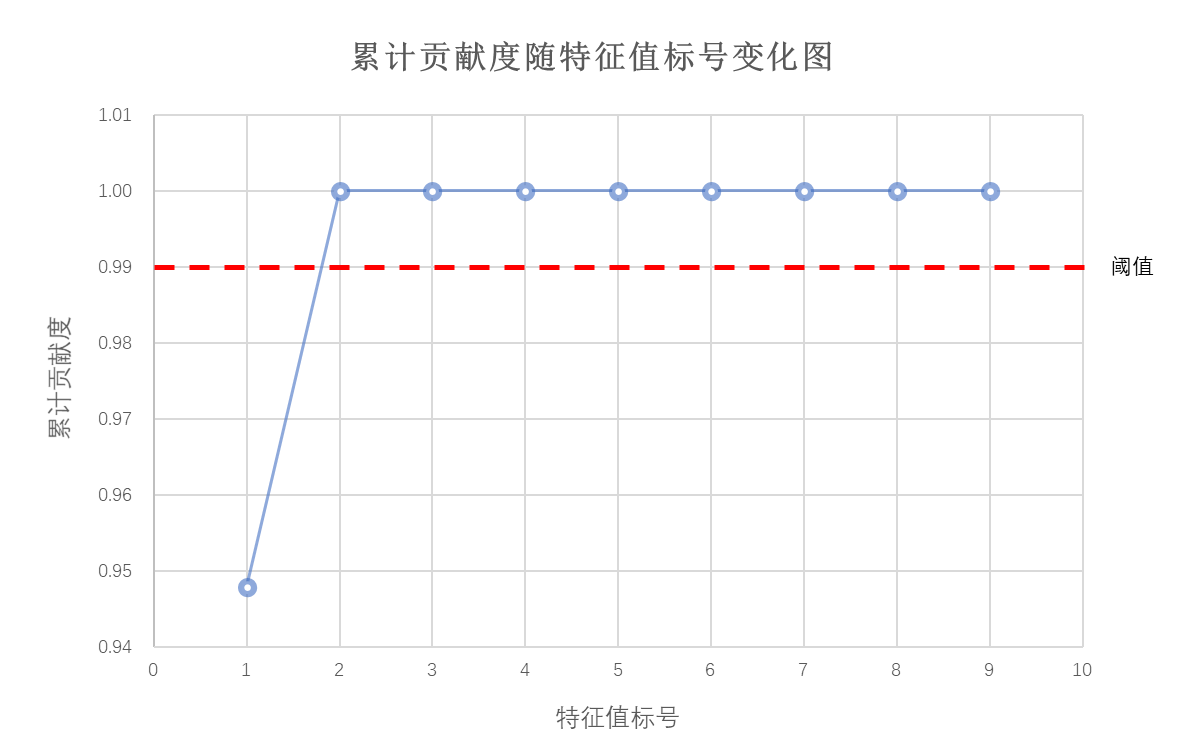
根据主成分分析图，我们可以得知，多数样本集中于图像的右方中间部位，它们的相似度非常高，组内的重复性比较好；而其他样本则较为分散，相似度比较低。

**② 主成分贡献度分析**

作出各成分的贡献度占情况表，如下所示：

|  |  |  |
| --- | --- | --- |
| **特征值** | **贡献度占比** | **累计贡献度** |
| 670739439.70 | 0.94792307 | 0.94792307 |
| 36846580.32 | 0.05207346 | 0.99999653 |
| 1250.98 | 0.00000177 | 0.99999829 |
| 981.35 | 0.00000139 | 0.99999968 |
| 135.15 | 0.00000019 | 0.99999987 |
| 73.49 | 0.00000010 | 0.99999997 |
| 12.88 | 0.00000002 | 0.99999999 |
| 4.24 | 0.00000001 | 1.00000000 |
| 0.61 | 0.00000000 | 1.00000000 |

作出累计贡献度随特征值标号的变化情况图如下所示：



根据上述图表可知，第一主成分占比最高，达到94%，远高于其他成分；第二主成分占比仅5%，但依旧为第三主成分的约三万倍。同时，在选择第一主成分和第二主成分后，累计贡献度超过99%，达到标准，且此后几乎不再发生变化。

因此，选择第一主成分和第二主成分，而舍弃其他成分进行降维，能够在降低数据维度的同时保留绝大部分的信息，是十分合理的。

# 总结

本次实验完整细致地复现了主成分分析（*PCA*）算法，深刻体现了降维的重要性和实用性。作为一个非监督学习的降维方法，*PCA*只需要特征值分解，就可以对数据进行压缩和去噪，故而应用广泛。

在实验中，它的主要优点可体现为：信息量的衡量不受数据集以外的因素影响；各主成分正交，可消除原始数据成分间的影响；计算简单，易于实现。在缺点方面，主要有：主成分的含义具有模糊性，不如原始样本特征的解释性强；非主成分也可能含有区分样本差异的重要信息，丢弃后可能对数据处理产生影响。

总之，*PCA*算法虽然具有一定的缺点，但在有效减少数据的维度、去除冗余信息、提高数据分析和机器学习任务的可行性方面，都具有良好的效果。在实验的过程中，我们能够清晰地认识和理解*PCA*的核心步骤与相关意义，这有助于我们更好地在实际项目中应用降维思想与*PCA*算法。