UIOSUN

机 器 学 习 汇 报

主成分分析

BIG DATA EMPOWERS TO CREATE A NEW ERA

PCA（Principal Component Analysis，主成分分析）是一种常用的线性降维方法，它通过寻找数据中的主成分来将高维数据映射到低维空间。

汇报人：游霄童 2023.11

目录

CONTENTS

算法介绍

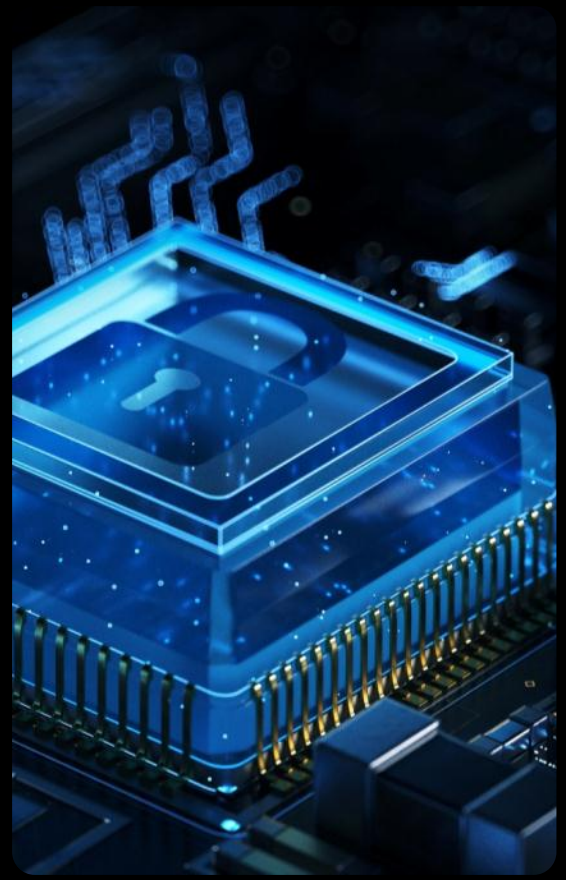
PROGRAM OF ACTIVITIES

PCA（Principal Component Analysis）

主成分是数据中方差最大的方向，通过将数据投影到这些主成分上，可以保留大部分数据的信息。

1.数据标准化
2.计算协方差矩阵
3.特征值分解
4.选择主成分
5.构建投影矩阵

PCA的目标是找到一组新的特征向量，它们是原始特征向量的线性组合，使得数据在新的特征空间中的方差最大。通过保留最重要的主成分，PCA能够在降低数据维度的同时尽量保留数据的信息。



算法介绍

GENERAL IDEA OF THE ACTIVITY

1. **数据标准化：**

   假设我们有n个样本，每个样本有d个特征。首先，计算每个特征的均值（mean）和标准差（standard deviation）：

$$\text{mean}(X_j) = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$$

$$\text{std}(X_j) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \text{mean}(X_j))^2}$$

   然后，将每个特征的值减去均值，然后除以标准差，得到标准化后的数据矩阵Z：

$$Z_{ij} = \frac{X_{ij} - \text{mean}(X_j)}{\text{std}(X_j)}$$

2. **计算协方差矩阵：**

   计算标准化后的数据矩阵Z的协方差矩阵C：

$$C = \frac{1}{n} Z^T Z$$

3. **特征值分解：**

   对协方差矩阵C进行特征值分解，得到特征值（eigenvalues）$\lambda_1, \lambda_2, \ldots, \lambda_d$，和对应的特征向量（eigenvectors）$v_1, v_2, \ldots, v_d$。特征值表示了数据中的方差，特征向量表示了对应特征值的主成分方向。

$$Cv_i = \lambda_i v_i, \quad i = 1, 2, \ldots, d$$

算法介绍

GENERAL IDEA OF THE ACTIVITY

4. **选择主成分：**

   将特征值按从大到小的顺序排列，选择前k个特征值对应的特征向量作为主成分。通常，选择特征值大于某个阈值的主成分，或者选择能够解释总方差的百分比大于某个阈值的主成分。

5. **构建投影矩阵：**

   将选定的k个特征向量按列排列，构成投影矩阵W：

$$W = [v_1, v_2, \ldots, v_k]$$

6. **将数据映射到低维空间：**

   将标准化后的数据矩阵Z乘以投影矩阵W，得到降维后的数据矩阵Y：

$$Y = ZW$$

通过这些步骤，我们将高维数据映射到了低维空间。在降维后的数据矩阵Y中，每一行对应一个样本，每一列对应一个主成分，可以用于后续的分析、可视化等任务。

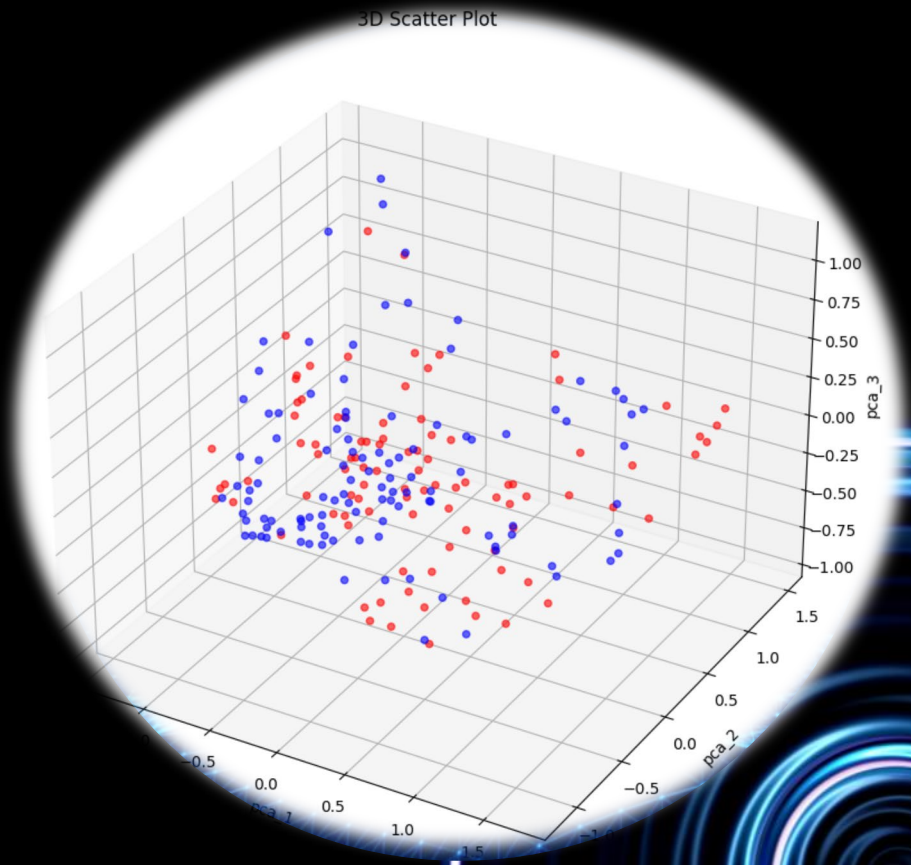
结果展示

绘制三维图

Sonar数据集是通过声纳传感器收集的声纳信号样本，用于区分两种不同类型的目标：岩石（Rock）和金属（Mine）。

1. 样本数量：Sonar数据集包含了208个样本。
2. 特征：每个样本由60个特征组成，这些特征是声纳传感器在不同方向上接收到的信号的幅度。

降维结果



01 先降低为三维

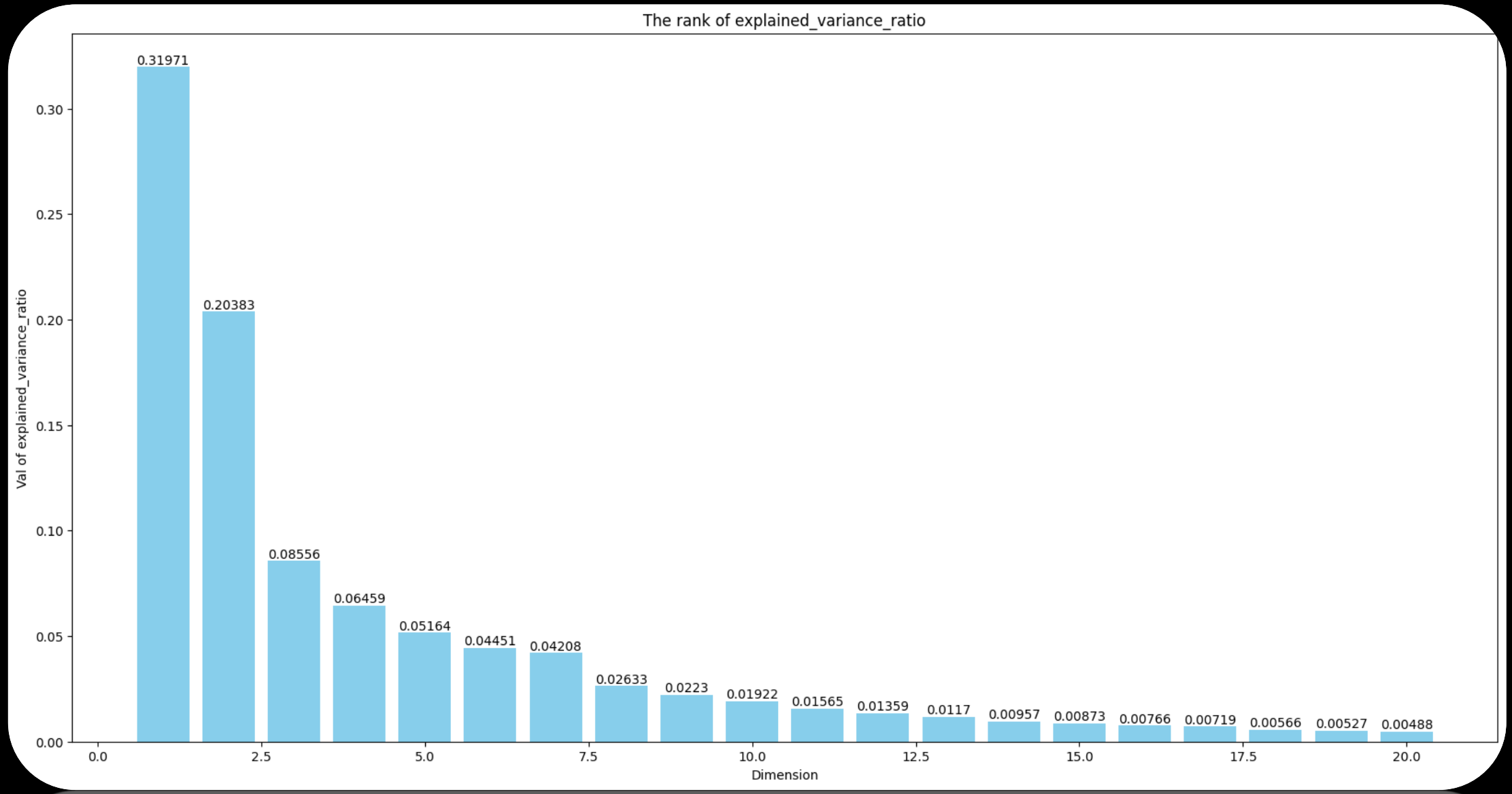
使用logictic 回归 测试得到准确率: 64.285714%

02 观察三维坐标图

并没有表现出明显的聚类现象，后发现前20位解释方差比中，只有前两位较为突出

结果展示

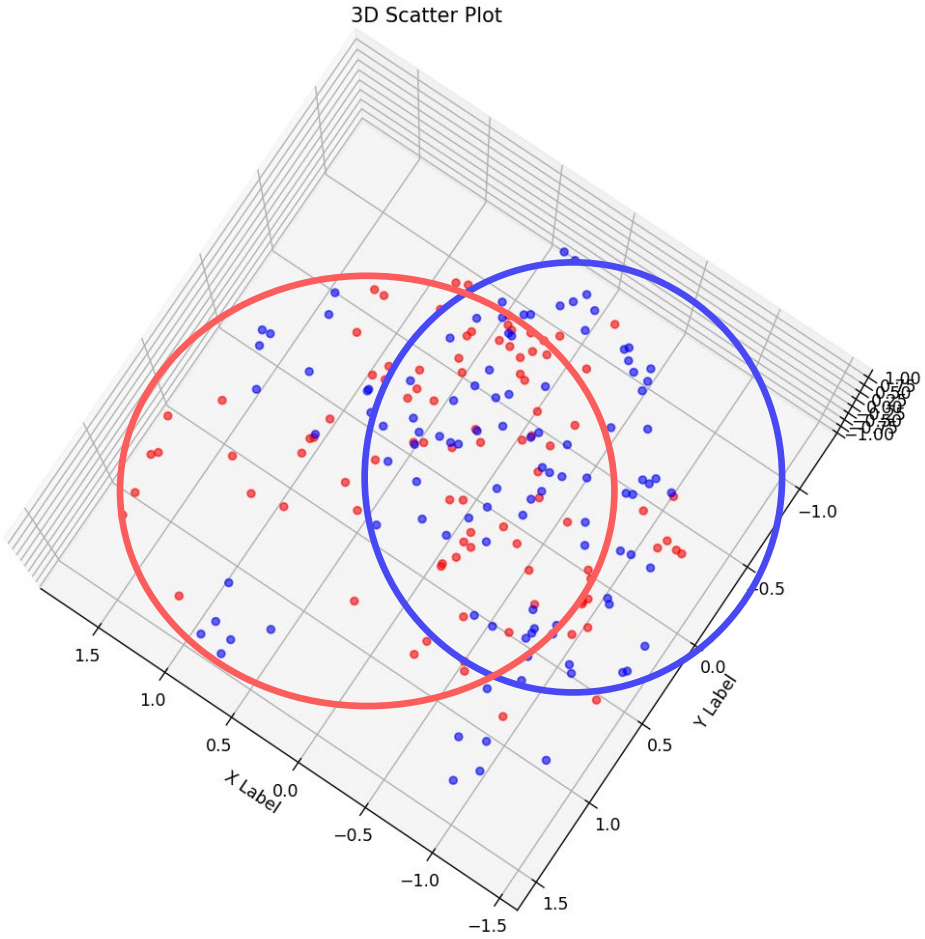
01


The rank of explained_variance_ratio
Val of explained_variance_ratio
Dimension
0.31971
0.20383
0.08556
0.06459
0.05164
0.04451
0.04208
0.02633
0.0223
0.01922
0.01565
0.01359
0.0117
0.00957
0.00873
0.00766
0.00719
0.00566
0.00527
0.00488
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

结果展示

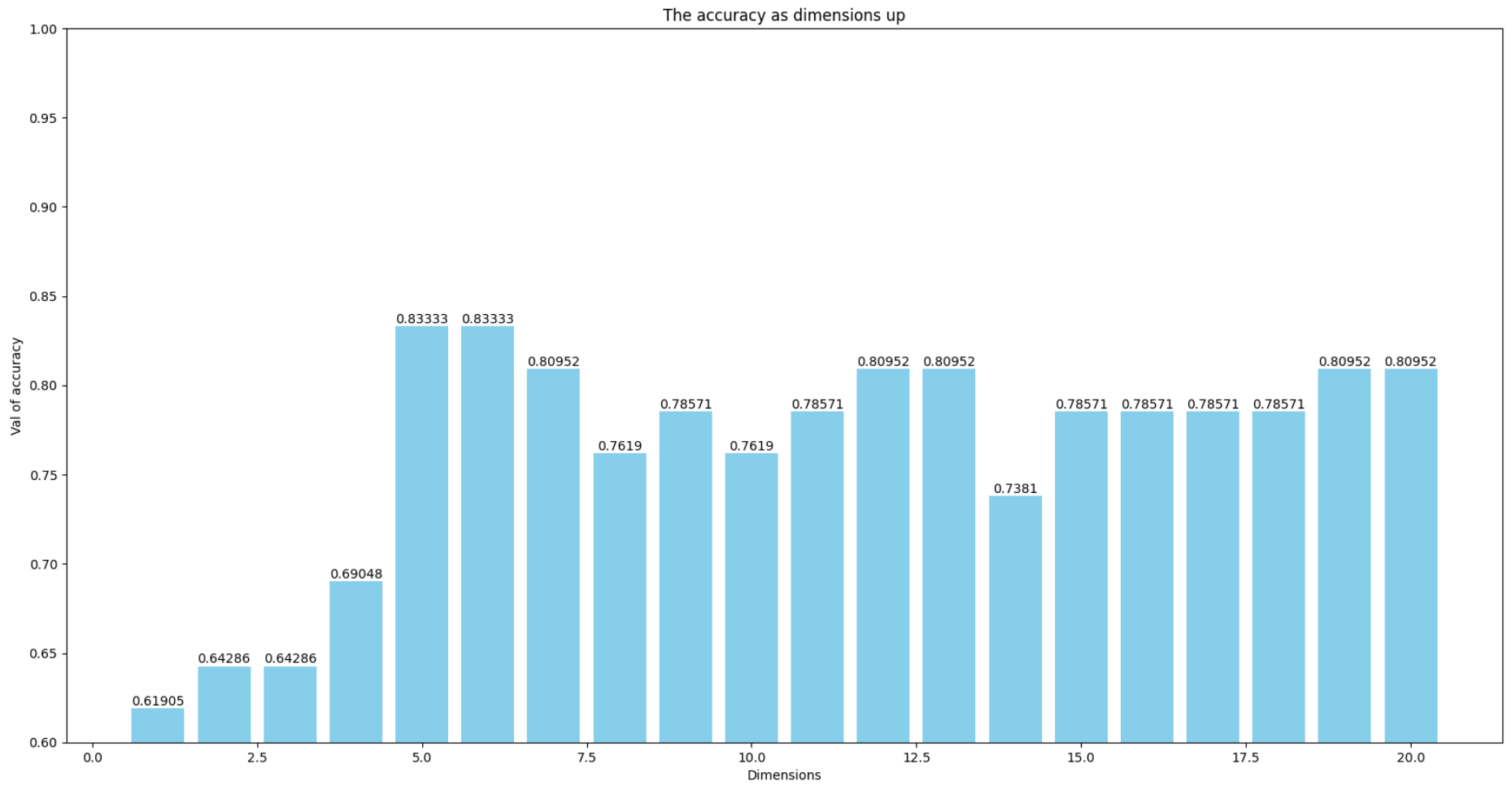
02

根据方差解释比，前两个最大，在pca_1，pca_2即X,Y平面上有了较为明显的聚类现象，也符合结论



结果展示

03


The accuracy as dimensions up
Val of accuracy
Dimensions
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0.61905
0.64286
0.64286
0.69048
0.83333
0.83333
0.80952
0.7619
0.78571
0.7619
0.78571
0.80952
0.80952
0.7381
0.78571
0.78571
0.78571
0.78571
0.80952
0.80952

降维评估

PROGRAM OF ACTIVITIES

可视化

使用二维或三维散点图、热力图等可视化方法，将降维后的数据在低维空间中呈现出来。这可以帮助你直观地观察数据的分布和结构，从而判断降维效果。

累计解释方差比

累计解释方差比表示前k个主成分解释的总方差比例。通常，我们希望保留能够解释大部分数据方差的主成分。可以绘制累计解释方差比的曲线，帮助选择合适的主成分数量。

特征值

特征值表示每个主成分解释的方差。特征值较大的主成分通常包含较多的信息。可以通过观察特征值的大小来评估主成分的重要性。

重构误差

将降维后的数据映射回原始高维空间，然后计算重构误差，即降维后数据与原始数据之间的差异。重构误差可以通过各种距离度量（如欧氏距离）来计算。

聚类性能

如果你的数据用于聚类任务，可以使用降维后的数据进行聚类，并评估聚类性能指标（如轮廓系数、互信息分数等）。如果聚类性能在降维后保持稳定或提高，说明降维是有效的。

监督学习性能

如果数据用于监督学习任务，可以在降维前后使用同样的分类或回归模型，并比较性能指标（如准确率、均方误差等）。如果性能在降维后保持或提高，说明降维是有效的。

Thanks!