专业课程实验报告

课程名称：模式识别

开课学期：2022 至 2023 学年 第 1 学期

专　　业：智能科学与技术

年级班级：20 级 3 班

学生姓名：严中圣

学生学号：222020335220177

实验教师：杨颂华

人工智能学院

《模式识别》实验课报告书

实验编号:	3	**实验名称**:	基于主成分分析的人脸识别
姓　　名:	严中圣	**学　　号**:	222020335220177
日　　期:	2022 年 12 月 11 日	**教师打分**:	

1 实验目标

主成分分析（PCA）是一种重要的特征提取方法，也是一种线性的降维技术。通过本实验旨在掌握PCA 的原理和实现算法，将其应用到人脸识别。

2 实验环境

- PyCharm 2022.1.3 (Professional Edition)
- python 3.7.13, numpy 1.21.5
- OS: Windows 11 22H2, CPU:12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz

3 实验原理

在多元统计分析中，主成分分析（Principal components analysis，PCA）是一种统计分析、简化数据集的方法。它利用正交变换来对一系列可能相关的变量的观测值进行线性变换，从而投影为一系列线性不相关变量的值，这些不相关变量称为主成分（Principal Components）。具体地，主成分可以看做一个线性方程，其包含一系列线性系数来指示投影方向。

PCA 的基本思想为：

- 将坐标轴中心移到数据的中心，然后旋转坐标轴，使得数据在 C1 轴上的方差最大，即全部 n 个数据个体在该方向上的投影最为分散。意味着更多的信息被保留下来。C1 成为第一主成分。
- C2 第二主成分：找一个 C2，使得 C2 与 C1 的协方差（相关系数）为 0，以免与 C1 信息重叠，并且使数据在该方向的方差尽量最大。
- 以此类推，找到第三主成分，第四主成分……第 p 个主成分。p 个随机变量可以有 p 个主成分。

具体数学推导如下：假设 $x=(x_1,x_2,...,x_m)^T$ 是 m 维随机变量，其均值向量是 μ

$$\mu = E(x) = (\mu_1,\mu_2,...,\mu_m)^T \tag{1}$$

协方差矩阵是 Σ

$$\Sigma = \operatorname{cov}(\boldsymbol{x},\boldsymbol{x}) = E\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\right] \tag{2}$$

考虑由 m 维随机变量 $\boldsymbol{x}$ 到 m 维随机变量 $\boldsymbol{y}\,(y_1,y_2,\cdots,y_m)^{\mathrm{T}}$ 的线性变换:

$$y_i = \alpha_i^{\mathrm{T}}\boldsymbol{x} = \alpha_{1i}x_1 + \alpha_{2i}x_2 + \cdots + \alpha_{mi}x_m \tag{3}$$

故我们要求解的即使最优的正交变换 α^{T}，使得新特征的方差能够达到极值。正交变换保证了新特征间的不相关性，二新特征的方差越大则可以保留更多的信息分布。同时为了统一新特征的尺度，得到以下定义：

$$\alpha_i^{\mathrm{T}}\alpha_i = 1, \quad cov(y_i, y_j) = 0 (i \neq j) \tag{4}$$

首先求 $\boldsymbol{x}$ 的第一主成分 $y_1 = \alpha_1^{\mathrm{T}}\boldsymbol{x}$, 即求系数向量 α_1 。由定义知, 第一主成分的 α_1 是在 $\alpha_1^{\mathrm{T}}\alpha_1 = 1$ 条件下, $\boldsymbol{x}$ 的所有线性变换中使方差 $\operatorname{var}\left(\alpha_1^{\mathrm{T}}\boldsymbol{x}\right) = \alpha_1^{\mathrm{T}}\Sigma\alpha_1$ 达到最大。

则求解第一主成分转化为求解约束最优化问题：

$$\begin{aligned} \max_{\alpha_1} \quad & \alpha_1^{\mathrm{T}}\Sigma\alpha_1 \\ s.t. \quad & \alpha_1^{\mathrm{T}}\alpha_1 = 1 \end{aligned} \tag{5}$$

利用拉格朗日乘数法，定义拉格朗日函数：

$$\alpha_1^{\mathrm{T}}\Sigma\alpha_1 - \lambda\left(\alpha_1^{\mathrm{T}}\alpha_1 - 1\right) \tag{6}$$

对 α_1 求导，并令其为 0，得：

$$\Sigma\alpha_1 - \lambda\alpha_1 = 0 \tag{7}$$

因此, λ 是 Σ 的特征值, α_1 是对应的单位特征向量。于是, 目标函数

$$\alpha_1^{\mathrm{T}}\Sigma\alpha_1 = \alpha_1^{\mathrm{T}}\lambda\alpha_1 = \lambda\alpha_1^{\mathrm{T}}\alpha_1 = \lambda \tag{8}$$

假设 α_1 是 Σ 的最大特征值 λ_1 对应的单位特征向量, 显然 α_1 与 λ_1 是最优化问题的解。所以, $\alpha_1^{\mathrm{T}}x$ 构成第一主成分, 其方差等于协方差矩阵的最大特征值

$$\operatorname{var}\left(\alpha_1^{\mathrm{T}}\boldsymbol{x}\right) = \alpha_1^{\mathrm{T}}\Sigma\alpha_1 = \lambda_1 \tag{9}$$

接着求 $\boldsymbol{x}$ 的第二主成分 $y_2 = \alpha_2^{\mathrm{T}}\boldsymbol{x}$ 。第二主成分的 α_2 是在 $\alpha_2^{\mathrm{T}}\alpha_2 = 1$, 且 $\alpha_2^{\mathrm{T}}\boldsymbol{x}$ 与 $\alpha_1^{\mathrm{T}}\boldsymbol{x}$ 不相关的条件下, $\boldsymbol{x}$ 的所有线性变换中使方差 $\operatorname{var}\left(\alpha_2^{\mathrm{T}}\boldsymbol{x}\right) = \alpha_2^{\mathrm{T}}\Sigma\alpha_2$ 达到最大的。

则求第二主成分需要求解约束最优化问题

$$\begin{aligned} & \max_{\alpha_2} \alpha_2^{\mathrm{T}}\Sigma\alpha_2 \\ & \text{s.t. } \alpha_1^{\mathrm{T}}\Sigma\alpha_2 = 0, \quad \alpha_2^{\mathrm{T}}\Sigma\alpha_1 = 0 \\ & \alpha_2^{\mathrm{T}}\alpha_2 = 1 \end{aligned} \tag{10}$$

同理利用拉格朗日乘数法可得第二主成分为 Σ 的第二大特征值 λ_2 对应的单位特征向量。按照上述方法求得第一、第二、直到第 m 主成分, 其系数向量 $\alpha_1, \alpha_2, \ldots, \alpha_m$ 分别是 Σ 的第一个、第二个、直到第 m 个单位特征向量, $\lambda_1, \lambda_2, \cdots, \lambda_m$ 分别是对应的特征值。并且, 第 k 主成分的方差等于 Σ 的第 k 个特征值,

$$\operatorname{var}\left(\alpha_k^{\mathrm{T}}\boldsymbol{x}\right) = \alpha_k^{\mathrm{T}}\Sigma\alpha_k = \lambda_k, \quad k = 1, 2, \cdots, m \tag{11}$$

综上所述，主成分分析的具体步骤可概括如下：

(1) 对观测数据进行规范化处理，得到规范化数据矩阵

在实际问题中，不同变量可能有不同的量纲，直接求主成分有时会产生不合理的结果。为了消除这个影响，常常对各个随机变量实施规范化，使其均值为 0 ，方差为 1。

$$X=\frac{x_i-E\left(x_i\right)}{\sqrt{\operatorname{var}\left(x_i\right)}}, \quad i=1,2, \cdots, m \tag{12}$$

(2) 依据规范化数据矩阵，计算样本相关矩阵 R

$$R=\left[r_{ij}\right]_{m \times m}=\frac{1}{n-1} X X^{\mathrm{T}} \tag{13}$$

其中：

$$r_{ij}=\frac{1}{n-1} \sum_{l=1}^n x_{il} x_{lj}, \quad i, j=1,2, \cdots, m \tag{14}$$

(3) 求样本相关矩阵 R 的 k 个特征值和对应的 k 个单位特征向量。

首先求解 R 的特征方程 $|R-\lambda I|=0$ 得 R 的 m 个特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq| \lambda_m$ 再根据方差贡献率 $\sum_{i=1}^k \eta_i$ 达到预定值的主成分个数 k 。最后求得前 k 个特征值对应的单位特征向量 $a_i=\left(a_{1i}, a_{2i}, \cdots, a_{mi}\right)^{\mathrm{T}}, \quad i=1,2, \cdots, k$。

(4) 求 k 个样本主成分

以 k 个单位特征向量为系数进行线性变换,求出 k 个样本主成分

$$y_i=a_i^{\mathrm{T}} \boldsymbol{x}, \quad i=1,2, \cdots, k \tag{15}$$

4 实验步骤

4.1 数据集来源

本次实验采用 ORL 人脸数据库，包含 1992 年 4 月至 1994 年 4 月期间在实验室拍摄的一组人脸图像。该数据库用于与剑桥大学工程系语音、视觉和机器人小组合作进行的人脸识别项目。

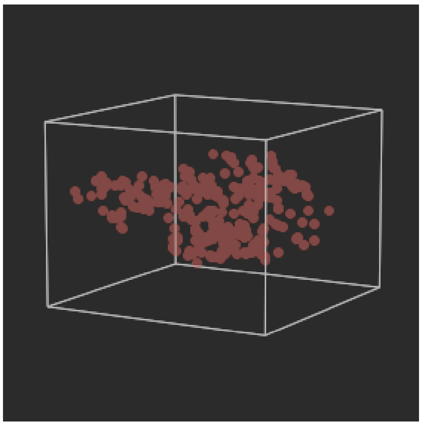
40 个不同的主题中的每一个都有十个不同的图像。对于某些受试者，图像是在不同的时间拍摄的，不同的照明，面部表情（睁开/闭上眼睛，微笑/不微笑）和面部细节（眼镜/不戴眼镜）。所有图像都是在黑暗的均匀背景下拍摄的，拍摄对象处于直立的正面位置（容忍某些侧面运动）。人脸数据库的预览图像可用。

图 1: ORL dataset

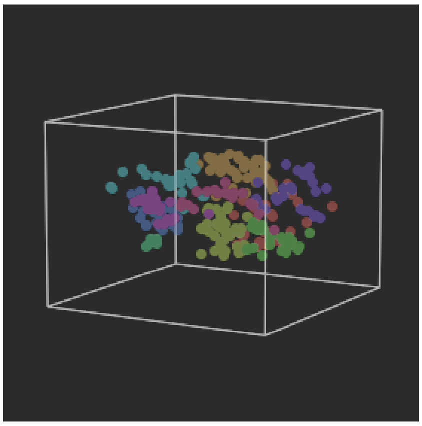
4.2 PCA 特征降维与人脸识别

在加载数据后利用 PCA 进行特征降维，首先对样本数据进行去中心化，其次构建协方差矩阵并进行特征值和特征向量求解。对特征值进行排序后取前 k 个权重较大的对应的特征向量作为主成分，并构建每幅图像投影后的值用以后续人脸识别，降维后的特征分布如下图所示。PCA 代码如下：

```python
def pca(data, k):
    data = np.float32(np.mat(data))
    rows, cols = data.shape
    data_mean = np.mean(data, 0)
    Z = data - np.tile(data_mean, (rows, 1))
    D, V = np.linalg.eig(Z * Z.T)
    V1 = V[:, :k]
    V1 = Z.T * V1
    for i in range(k):
        V1[:, i] /= np.linalg.norm(V1[:, i])
    return np.array(Z * V1), data_mean, V1
```



(a) 降维前数据特征



(b) 降维前数据特征

此后将降维后的主成分特征利用 SVM 分类器进行分类，最终得到识别率为 94.17%。值得一提的是，在未进行降维操作前直接进行分类，识别率仅为 86.42%，可见特征降维对识别精度的提升有较大的帮助。

5 实验分析

本次实验实现了主成分分析的数据降维过程，同时分析了算法的原理和数学推导，更深刻地理解了算法的原理和实际应用方式。试验中通过少量样本亦取得良好效果，说明 PCA 在降维上拥有很好的效果，较好的保留了主要特征。

6 参考文献

[1] 周志华. 机器学习 [M]. 清华大学出版社, 2016.

[2] 李航. 统计学习方法 [M]. 清华大学出版社, 2019

[3] https://zhuanlan.zhihu.com/p/37777074

[4] Christopher M. Bishop: Pattern Recognition and Machine Learning, Chapter 4.3.4