模式识别

期中实验作业

姓名：郝裕玮

班级：计科 1 班

学号：18329015

目录

1 实验环境

Jupyter Notebook (anaconda3) + Python 3.8.5 + scikit-learn 0.23.2

2 PCA+KNN，LDA+KNN，KNN

该部分将分析 PCA+KNN，LDA+KNN，KNN 三种模型的超参数选择和人脸识别性能对比。

本次实验中，大多数代码是具有复用性的：因为主要的代码逻辑是调用 sklearn 库的各种相关函数来进行特征提取（PCA，LDA 等），实现分类器（KNN，SVM 等），超参数选择（调用 GridSearchCV）以及模型评分（sklearn.metrics）

以 PCA + KNN 为例：

（1）需要导入相关库，重点库如下：

```
from sklearn.decomposition import PCA
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import GridSearchCV
```

前两行是 PCA 和 KNN 的相关库，GridSearchCV 则是用于网格化交叉验证寻找最优超参数组合。

（2）为了方便后续数据切片。我的读取逻辑为先按顺序读取每个文件夹的前 8 张（作为训练集），读取结束后，再按顺序读取每个文件夹的最后 2 张（作为测试集），并将这二者存在同二数组中（样本数据 X 和样本标签 Y）。在读取过程中，由于每张图片读取后得到的变量均为 112*92，为了数组方便处理，我将其进行了一维化扁平处理。

```
#对读取到的数据进行一维化扁平处理
def img2vector(filename):
    img = mpimg.imread(filename)
    return img.reshape(1, -1)

# 样本数据和样本标签
X = np.zeros((400,10304),dtype = int)
Y = np.zeros(400,dtype = int)

#cnt 用于计数
cnt = 0
#先读取训练集
for i in range(1,41):
    for j in range(1,9):
        src = 'C:\\Users\\93508\\Desktop\\ORL\\s'
        src += str(i)
        src += '\\'
        src += str(j)
        src += '.bmp'
        X[cnt] = img2vector(src)
        Y[cnt] = i
        cnt = cnt + 1

#再重新读取测试集
for i in range(1,41):
    for j in range(9,11):
        src = 'C:\\Users\\93508\\Desktop\\ORL\\s'
        src += str(i)
        src += '\\'
        src += str(j)
        src += '.bmp'
        X[cnt] = img2vector(src)
        Y[cnt] = i
        cnt = cnt + 1
```

（3）开始进行 PCA 和 KNN 的超参数选择

```
#Max_point 用于保存最优超参数模型的交叉验证评分
Max_point = 0
final_n = 0
final_k = 0

#该循环用于选出 PCA 超参数 n_components 的最佳值
#n_components: 需要保留的特征数量（即降维后的结果）
for n in range(10,410,10):
```

```
    #调用 PCA
    pca = PCA(n_components = n) #实例化
    newX = pca.fit_transform(X) #用已有数据训练 PCA 模型，并返回降维后的数据

    #将降维后的数据拆分为训练集和测试集
    x_train = newX[0:320,:]
    y_train = Y[0:320]
    x_test = newX[320:400,:]
    y_test = Y[320:400]

    #KNN 的超参数：
    #n_neighbors：KNN 用于判别分类的邻居数
    C = np.arange(3,22,2)
    #将需要遍历的超参数定义为字典
    params = {'n_neighbors': C}

    #定义网格搜索中使用的模型和参数
    knn = GridSearchCV(KNeighborsClassifier(), params, scoring =
"f1_macro",cv = 5)
    #使用网格搜索模型拟合数据
    knn.fit(x_train,y_train)

    #存储 PCA 不同超参数下的 KNN 最优超参数模型的交叉验证评分
    cur_point = knn.best_score_
    #选出模型交叉验证评分最高的一组超参数（PCA 和 KNN）
    if cur_point > Max_point:
        Max_point = cur_point
        final_n = n
        final_k = knn.best_params_['n_neighbors']
```

（4）输出结果

```
#输出结果
print("\nPCA 的超参数 n_components 的最优解为: %d\n" %final_n)
print("KNN 的超参数 n_neighbors 的最优解为: %d\n" %final_k)
y_predict = knn.predict(x_test)
accuracy = accuracy_score(y_predict, y_test)
print("测试集预测正确率为：%.2f%%\n" %(accuracy*100))
print("最优超参数模型的评分为: %.2f\n" %Max_point)
print("测试集的预测分类报告如下所示：\n\n")
print(classification_report(y_test, y_predict))
```

对于 LDA + KNN，KNN 以及后续的其他分类器，需要更改的代码仅如下所示：

（1）若特征提取方法改变（如 PCA 变 LDA）：

```
    pca = PCA(n_components = n) #实例化
    newX = pca.fit_transform(X) #用已有数据训练 PCA 模型，并返回降维后的数据
```

该部分需要修改为：

```
    lda = LDA(n_components = n)
    newX = lda.fit_transform(X,Y)
```

（2）若分类器改变（如 KNN 变 SVM 线性核）

```
    #KNN 的超参数：
    #n_neighbors：KNN 用于判别分类的邻居数
    C = np.arange(3,22,2)
    #将需要遍历的超参数定义为字典
    params = {'n_neighbors': C}

    #定义网格搜索中使用的模型和参数
    knn = GridSearchCV(KNeighborsClassifier(), params, scoring =
"f1_macro",cv = 5)
    #使用网格搜索模型拟合数据
    knn.fit(x_train,y_train)
```

该部分需要修改为：

```
    C = np.power(10, np.arange(10))
    params = {'C': C,'kernel':['linear']}
    svc_linear = GridSearchCV(SVC(), params, scoring = "f1_macro",cv =
5)
    svc_linear.fit(x_train,y_train)
```

修改参数列表 params 以及 GridSearchCV 的第一个参数（即分类器类型）

综上所述，下文将不再贴出代码，直接展示结果。

PS:下文中各个结果展示的预测分类报告内容是训练集每个类的内部预测情况（因为共有 40 组不同人脸，每组人脸的最后 2 张作为数据集，所以序号总数为 40，support = 2）

PCA（或 LDA）中的超参数:

n_components: 需要保留的特征数量（即降维后的结果）

KNN 中的超参数：

n_neighbors：KNN 用于判别分类的邻居数

（1）PCA + KNN

PCA的超参数n_components的最优解为: 70

KNN的超参数n_neighbors的最优解为: 3

测试集预测正确率为：93.75%

最优超参数模型的评分为: 0.93

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	0.67	1.00	0.80	2
5	0.50	0.50	0.50	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	0.00	0.00	0.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	0.67	1.00	0.80	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	0.67	1.00	0.80	2
19	1.00	0.50	0.67	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	0.67	1.00	0.80	2
39	1.00	1.00	1.00	2
40	1.00	0.50	0.67	2
accuracy			0.94	80
macro avg	0.93	0.94	0.93	80
weighted avg	0.93	0.94	0.93	80

（2）LDA + KNN

LDA的超参数n_components的最优解为: 13

KNN的超参数n_neighbors的最优解为: 3

测试集预测正确率为：100.00%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	1.00	1.00	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			1.00	80
macro avg	1.00	1.00	1.00	80
weighted avg	1.00	1.00	1.00	80

（3）KNN

KNN的超参数n_neighbors的最优解为：3

测试集预测正确率为：93.75%

最优超参数模型的评分为：0.92

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	0.67	1.00	0.80	2
5	0.50	0.50	0.50	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	0.00	0.00	0.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	0.67	1.00	0.80	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	0.67	1.00	0.80	2
19	1.00	0.50	0.67	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	0.67	1.00	0.80	2
39	1.00	1.00	1.00	2
40	1.00	0.50	0.67	2
accuracy			0.94	80
macro avg	0.93	0.94	0.93	80
weighted avg	0.93	0.94	0.93	80

综合上述结果，人脸识别性能排序为：

LDA + KNN > PCA + KNN > KNN

3 PCA+SVM，LDA+SVM，SVM(线性核)

PCA（或 LDA）中的超参数:

n_components: 需要保留的特征数量（即降维后的结果）

SVM 线性核中的超参数：

C：错误项的惩罚系数。C 越大，对分错样本的惩罚程度越大，因此在训练样本中准确率越高，但是泛化能力降低，也就是对测试数据的分类准确率降低。相反，减小 C 的话，容许训练样本中有一些误分类错误样本，泛化能力会增强。

（1）PCA + SVM 线性核（图见下页）

PCA的超参数n_components的最优解为: 40

SVM线性核的超参数C的最优解为: 1

测试集预测正确率为：97.50%

最优超参数模型的评分为: 0.97

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	0.67	1.00	0.80	2
9	1.00	1.00	1.00	2
10	1.00	0.50	0.67	2
11	0.67	1.00	0.80	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	0.50	0.67	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			0.97	80
macro avg	0.98	0.97	0.97	80
weighted avg	0.98	0.97	0.97	80

（2）LDA + SVM 线性核

LDA的超参数n_components的最优解为: 10

SVM线性核的超参数C的最优解为: 1

测试集预测正确率为：100.00%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	1.00	1.00	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			1.00	80
macro avg	1.00	1.00	1.00	80
weighted avg	1.00	1.00	1.00	80

（3）SVM 线性核

SVM线性核的超参数C的最优解为：1

测试集预测正确率为：97.50%

最优超参数模型的评分为：0.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	0.67	1.00	0.80	2
9	1.00	1.00	1.00	2
10	1.00	0.50	0.67	2
11	0.67	1.00	0.80	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	0.50	0.67	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			0.97	80
macro avg	0.98	0.97	0.97	80
weighted avg	0.98	0.97	0.97	80

综合上述结果，人脸识别性能排序为：

LDA + SVM 线性核 > PCA + SVM 线性核 > SVM 线性核

4 PCA+高斯核，LDA+高斯核，高斯核

PCA（或 LDA）中的超参数:

n_components: 需要保留的特征数量（即降维后的结果）

SVM 高斯核中的超参数：

（1）C：错误项的惩罚系数。C 越大，对分错样本的惩罚程度越大，因此在训练样本中准确率越高，但是泛化能力降低，也就是对测试数据的分类准确率降低。相反，减小 C 的话，容许训练样本中有一些误分类错误样本，泛化能力会增强。

（2）gamma：核函数系数（针对于高斯核，多项式核）

（1）PCA + SVM 高斯核 rbf（图见下页）

PCA的超参数n_components的最优解为: 10

SVM高斯内核rbf的超参数C的最优解为: 10 超参数gamma的最优解为: 0.000001

测试集预测正确率为：51.25%

最优超参数模型的评分为: 0.89

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	0.00	0.00	0.00	2
2	1.00	0.50	0.67	2
3	1.00	0.50	0.67	2
4	1.00	1.00	1.00	2
5	1.00	0.50	0.67	2
6	1.00	0.50	0.67	2
7	1.00	0.50	0.67	2
8	1.00	0.50	0.67	2
9	1.00	0.50	0.67	2
10	0.00	0.00	0.00	2
11	0.00	0.00	0.00	2
12	0.00	0.00	0.00	2
13	1.00	0.50	0.67	2
14	0.00	0.00	0.00	2
15	1.00	1.00	1.00	2
16	0.00	0.00	0.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	0.50	0.67	2
20	1.00	0.50	0.67	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	0.00	0.00	0.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	0.50	0.67	2
28	0.00	0.00	0.00	2
29	1.00	0.50	0.67	2
30	1.00	1.00	1.00	2
31	1.00	0.50	0.67	2
32	0.00	0.00	0.00	2
33	1.00	0.50	0.67	2
34	1.00	0.50	0.67	2
35	0.00	0.00	0.00	2
36	0.00	0.00	0.00	2
37	1.00	1.00	1.00	2
38	1.00	0.50	0.67	2
39	1.00	0.50	0.67	2
40	0.05	1.00	0.09	2
accuracy			0.51	80
macro avg	0.70	0.51	0.56	80
weighted avg	0.70	0.51	0.56	80

（2）LDA + SVM 高斯核 rbf

LDA的超参数n_components的最优解为: 10

SVM高斯内核rbf的超参数C的最优解为: 1 超参数gamma的最优解为: 0.010000

测试集预测正确率为：100.00%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	1.00	1.00	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			1.00	80
macro avg	1.00	1.00	1.00	80
weighted avg	1.00	1.00	1.00	80

（3）SVM 高斯核 rbf

SVM高斯内核rbf的超参数C的最优解为: 1 超参数gamma的最优解为: 0.000100

测试集预测正确率为：51.25%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	0.00	0.00	0.00	2
2	1.00	0.50	0.67	2
3	1.00	0.50	0.67	2
4	1.00	1.00	1.00	2
5	1.00	0.50	0.67	2
6	1.00	0.50	0.67	2
7	1.00	0.50	0.67	2
8	1.00	0.50	0.67	2
9	1.00	0.50	0.67	2
10	0.00	0.00	0.00	2
11	0.00	0.00	0.00	2
12	0.00	0.00	0.00	2
13	1.00	0.50	0.67	2
14	0.00	0.00	0.00	2
15	1.00	1.00	1.00	2
16	0.00	0.00	0.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	0.50	0.67	2
20	1.00	0.50	0.67	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	0.00	0.00	0.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	0.50	0.67	2
28	0.00	0.00	0.00	2
29	1.00	0.50	0.67	2
30	1.00	1.00	1.00	2
31	1.00	0.50	0.67	2
32	0.00	0.00	0.00	2
33	1.00	0.50	0.67	2
34	1.00	0.50	0.67	2
35	0.00	0.00	0.00	2
36	0.00	0.00	0.00	2
37	1.00	1.00	1.00	2
38	1.00	0.50	0.67	2
39	1.00	0.50	0.67	2
40	0.05	1.00	0.09	2
accuracy			0.51	80
macro avg	0.70	0.51	0.56	80
weighted avg	0.70	0.51	0.56	80

综合上述结果，人脸识别性能排序为：

LDA + SVM 高斯核 > SVM 高斯核 > PCA + SVM 高斯核

5 PCA+多项式核，LDA+多项式核，多项式核

PCA（或 LDA）中的超参数:

n_components: 需要保留的特征数量（即降维后的结果）

SVM 多项式核中的超参数：

（1）C：错误项的惩罚系数。C 越大，对分错样本的惩罚程度越大，因此在训练样本中准确率越高，但是泛化能力降低，也就是对测试数据的分类准确率降低。相反，减小 C 的话，容许训练样本中有一些误分类错误样本，泛化能力会增强。

（2）gamma：核函数系数（针对于高斯核，多项式核）

（3）degree：这个参数只对多项式核函数有用，是指多项式核函数的阶数 n

（4）coef0：核函数中的独立项，即常数 c

（1）PCA + SVM 多项式核 poly（图见下页）

PCA的超参数n_components的最优解为: 40

SVM多项式内核poly的超参数C的最优解为: 1 超参数gamma的最优解为: 0.100000 超参数degree的最优解为: 1 超参数coef0的最优解为: 0

测试集预测正确率为：97.50%

最优超参数模型的评分为: 0.97

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	0.67	1.00	0.80	2
9	1.00	1.00	1.00	2
10	1.00	0.50	0.67	2
11	0.67	1.00	0.80	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	0.50	0.67	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			0.97	80
macro avg	0.98	0.97	0.97	80
weighted avg	0.98	0.97	0.97	80

（2）LDA + SVM 多项式核 poly

LDA的超参数n_components的最优解为: 8

SVM多项式内核poly的超参数C的最优解为: 1 超参数gamma的最优解为: 0.100000 超参数degree的最优解为: 1 超参数coef0的最优解为: 0

测试集预测正确率为：100.00%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	1.00	1.00	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			1.00	80
macro avg	1.00	1.00	1.00	80
weighted avg	1.00	1.00	1.00	80

（3）SVM 多项式核 poly

SVM多项式内核poly的超参数C的最优解为: 1 超参数gamma的最优解为: 0.100000 超参数degree的最优解为: 1 超参数coef0的最优解为: 0

测试集预测正确率为：100.00%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	1.00	1.00	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			1.00	80
macro avg	1.00	1.00	1.00	80
weighted avg	1.00	1.00	1.00	80

综合上述结果，人脸识别性能排序为：

LDA + SVM 多项式核 = SVM 多项式核 > PCA + SVM 多项式核

6 逻辑回归，决策树，随机森林，adaboost，神经网络

为了比较不同分类器的性能，我采取的方法是均使用 LDA 进行特征提取，再根据对测试集的准确率和对模型交叉验证的评分来进行排序。

（1）逻辑回归

超参数为：

C：错误项的惩罚系数。C 越大，对分错样本的惩罚程度越大，因此在训练样本中准确率越高，但是泛化能力降低，也就是对测试数据的分类准确率降低。相反，减小 C 的话，容许训练样本中有一些误分类错误样本，泛化能力会增强。

结果见下页：

LDA的超参数n_components的最优解为: 14

逻辑回归的超参数C的最优解为: 1

测试集预测正确率为: 98.75%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	0.67	1.00	0.80	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	0.50	0.67	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			0.99	80
macro avg	0.99	0.99	0.99	80
weighted avg	0.99	0.99	0.99	80

（2）决策树

超参数为：

（1）max_depth：决策树最大深度

（2）min_samples_split：子数据集再切分需要的最小样本量

（3）min_samples_leaf：叶节点（子数据集）最小样本数

LDA的超参数n_components的最优解为: 15

决策树的超参数max_depth的最优解为: 40 超参数min_samples_split的最优解为: 4 超参数min_samples_leaf的最优解为: 1

测试集预测正确率为: 85.00%

最优超参数模型的评分为: 0.81

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	0.50	0.67	2
3	1.00	1.00	1.00	2
4	0.67	1.00	0.80	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	0.67	1.00	0.80	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	0.00	0.00	0.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	0.50	0.67	2
19	0.33	0.50	0.40	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	0.50	0.67	2
28	0.50	0.50	0.50	2
29	1.00	0.50	0.67	2
30	0.67	1.00	0.80	2
31	1.00	1.00	1.00	2
32	0.50	0.50	0.50	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	0.67	1.00	0.80	2
36	1.00	0.50	0.67	2
37	0.67	1.00	0.80	2
38	1.00	1.00	1.00	2
39	0.67	1.00	0.80	2
40	0.00	0.00	0.00	2
accuracy			0.85	80
macro avg	0.86	0.85	0.84	80
weighted avg	0.86	0.85	0.84	80

（3）随机森林

超参数为：

（1）n_estimators：森林中树木的数量

（2）max_depth：树的最大深度

LDA的超参数n_components的最优解为: 12

随机森林的超参数n_estimators的最优解为: 300 超参数max_depth的最优解为: 20

测试集预测正确率为: 100.00%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	1.00	1.00	2
19	1.00	1.00	1.00	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			1.00	80
macro avg	1.00	1.00	1.00	80
weighted avg	1.00	1.00	1.00	80

（4）adaboost

超参数为：

（1）n_estimators：弱学习器的最大迭代次数

（2）learning_rate：每个弱学习器的权重缩减系数

LDA的超参数n_components的最优解为: 12

adaboost的超参数n_estimators的最优解为: 500 超参数learning_rate的最优解为: 0.3

测试集预测正确率为: 71.25%

最优超参数模型的评分为: 0.64

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	0.50	0.67	2
3	1.00	0.50	0.67	2
4	0.67	1.00	0.80	2
5	0.00	0.00	0.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	0.67	1.00	0.80	2
10	1.00	0.50	0.67	2
11	1.00	0.50	0.67	2
12	1.00	1.00	1.00	2
13	0.33	1.00	0.50	2
14	0.00	0.00	0.00	2
15	1.00	0.50	0.67	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	0.00	0.00	0.00	2
19	1.00	0.50	0.67	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	0.40	1.00	0.57	2
23	0.00	0.00	0.00	2
24	1.00	1.00	1.00	2
25	0.22	1.00	0.36	2
26	1.00	1.00	1.00	2
27	0.67	1.00	0.80	2
28	0.33	0.50	0.40	2
29	1.00	1.00	1.00	2
30	1.00	0.50	0.67	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	0.50	0.67	2
35	1.00	1.00	1.00	2
36	1.00	0.50	0.67	2
37	0.50	0.50	0.50	2
38	0.50	1.00	0.67	2
39	0.00	0.00	0.00	2
40	0.00	0.00	0.00	2
accuracy			0.71	80
macro avg	0.73	0.71	0.69	80
weighted avg	0.73	0.71	0.69	80

（5）神经网络

超参数为：

（1）hidden_layer_sizes：默认值（100,）第 i 个元素表示第 i 个隐藏层中的神经元数量

（2）alpha：L2 惩罚（正则化项）参数

（3）max_iter：最大迭代次数

LDA的超参数n_components的最优解为: 13

神经网络的超参数hidden_layer_sizes的最优解为: 50 超参数alpha的最优解为: 0.001000 超参数max_iter的最优解为: 500

测试集预测正确率为: 98.75%

最优超参数模型的评分为: 1.00

测试集的预测分类报告如下所示：

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	0.67	1.00	0.80	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	2
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	2
18	1.00	0.50	0.67	2
19	1.00	1.00	1.00	2
20	1.00	1.00	1.00	2
21	1.00	1.00	1.00	2
22	1.00	1.00	1.00	2
23	1.00	1.00	1.00	2
24	1.00	1.00	1.00	2
25	1.00	1.00	1.00	2
26	1.00	1.00	1.00	2
27	1.00	1.00	1.00	2
28	1.00	1.00	1.00	2
29	1.00	1.00	1.00	2
30	1.00	1.00	1.00	2
31	1.00	1.00	1.00	2
32	1.00	1.00	1.00	2
33	1.00	1.00	1.00	2
34	1.00	1.00	1.00	2
35	1.00	1.00	1.00	2
36	1.00	1.00	1.00	2
37	1.00	1.00	1.00	2
38	1.00	1.00	1.00	2
39	1.00	1.00	1.00	2
40	1.00	1.00	1.00	2
accuracy			0.99	80
macro avg	0.99	0.99	0.99	80
weighted avg	0.99	0.99	0.99	80

五种模型的结果对比如下表所示：

分类器	测试集准确率	交叉验证评分
逻辑回归	98.75%	1.00
决策树	85%	0.81
随机森林	100%	1.00
adaboost	71.25%	0.64
神经网络	98.75%	1.00

所以综上可知，随机森林这一分类器最适合在 ORL 数据集上进行人脸识别。