支持向量机应用：人脸识别

原创 石头 机器学习算法那些事 2018-11-24

前言

人脸识别是当前很火的一个方向，涵盖了数字图像处理，机器学习和深度学习等领域。小编认为，若已经理解了某一个算法理论，可以拿人脸识别来练练手，因为网上关于人脸识别的项目很多，不至于自己一个人瞎弄找不到方向，而且网上的人脸数据库很多，不用担心数据的问题。本文用机器学习模型的设计步骤来描述人脸识别。

> 实验数据集叫做 Labeled Faces in the Wild
> (http://vis-www.cs.umass.edu/lfw/lfw-funneled.tgz)

1. 数据集分类

下载数据集：

```
faces = fetch_lfw_people(data_home=None, min_faces_per_person=70, resize=0.4)
```

下载数据集到特定的目录（C:\Users\Administrator\scikit_learn_data）

data_home可以设置下载路径。min_faces_person表示只下载每个人的图片超过70张的图片，resize对原图进行缩放。

运行后可能会提示如下错误：

```
ImportError: The Python Imaging Library (PI
L) is required to load data from jpeg files
```

解决方法：

（1）下载图像处理框架：PIL库；

（2） import Image库

```
from PIL import Image
```

数据集分类：

数据集有两种分类方法，对应有不同的模型评估方法：

（1） 数据集分为训练集和测试集，训练集用交叉验证的方法选择最优模型参数，然后用测试集来评估模型性能。

（2） 数据集分为训练集、验证集和测试集，训练集构建模型，验证集选择最优模型参数，然后用测试集来评估模型性能。

若数据较少，推荐第一种方法；数据较多，推荐第二种方法。

这里采用第一种方法：

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, random_state=42)
```

将数据集随机分成训练集和测试集，且比例为3:1，random_state=42表示设置种子随机器为常数，因此，每次运行后的训练集和测试集是固定的。

2. 特征预处理

图像的每一个像素是一个特征，本次实验的数据集经过缩放后的图像尺寸为50×37，特征共1850个，大于训练数据集容量1288，因此，需要进行降维处理。这里采用PCA（主成分分析法）进行降维，降维原理：把数据投影到正交的基向量，选择前几个方差较大的基向量（后面的文章详细分析这一原理，请继续关注我吧）

```
pca = PCA(n_components=150, svd_solver='randomized',
          whiten=True).fit(X_train)
X_train_pca = pca.transform(X_train)
X_test_pca = pca.transform(X_test)
```

首先对训练集使用随机奇异矩阵分解构建基向量，然后用测试集的数据投影到基向量，这两种步骤使训练集和测试集都实现了同一种规则降维，n_components指定维度。

3. 最优模型构建

这里采用了第一种模型评估方法，即训练数据集用交叉验证的方法选择最优参数，测试集评估模型性能。

设置模型可选择的参数范围：

```
param_grid = {'C': [1e3, 5e3, 1e4, 5e4, 1e5],
              'gamma': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1], }
```

C：模型误分类的惩罚系数

gamma：核函数参数

参数择优模型：

```
clf = GridSearchCV(SVC(kernel='rbf', class_weight='balanced'),
                   param_grid, cv=5)
```

SVC：选择支持向量机模型进行分类

class_weight = 'balanced' 表示样本的权重相等，若分类正常人和癌症病人的情况，则需要给癌症病人较大的权重。

cv = 5 表示用五折交叉验证的方法去选择最优参数。

构建最优模型：

```
clf = clf.fit(X_train_pca, y_train)
```

4. 评估最优模型

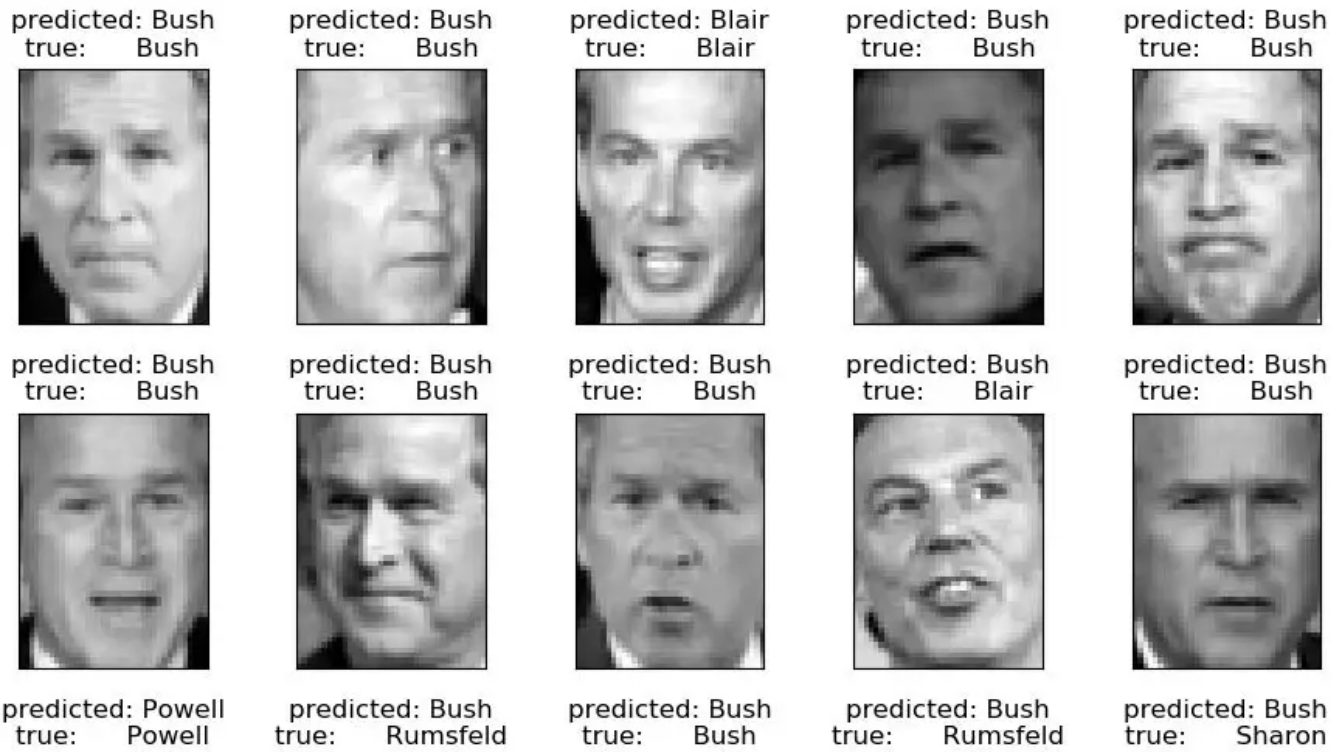
测试数据集评估最优模型：

```
y_pred = clf.predict(X_test_pca)
```

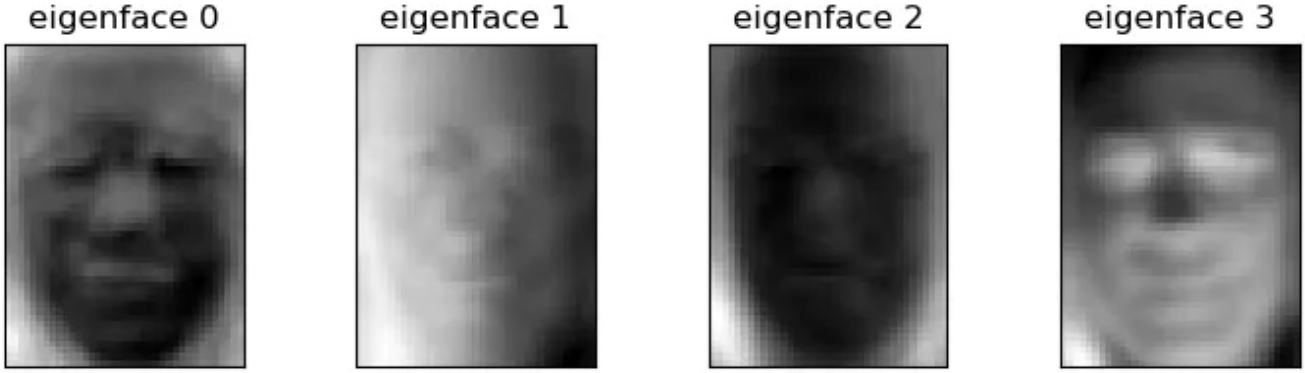
输出混淆矩阵：

```
confusion_matrix(y_test, y_pred, labels=range(n_classes))
```

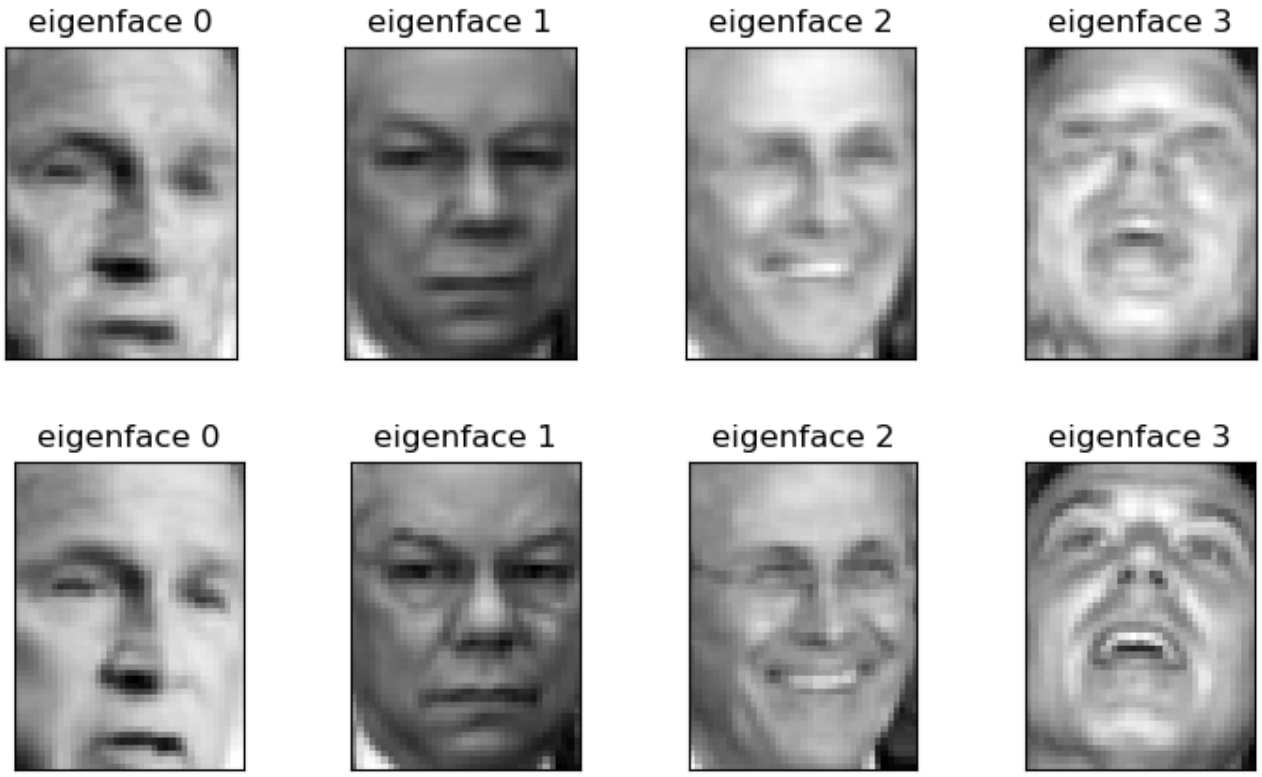
结果：



降维后的特征空间图：



降低后的维度（ndim）越小，丢失的信息越多，如下面的对比图：



第一行降低后的维度是150维，第二行的降低后的维度是900维，然后再逆转换为原始空间，很容易得到第二行更清晰且更接近原图。

参考

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

推荐阅读文章

支持向量机（三）：图解KKT条件和拉格朗日乘子法

机器学习模型评估方法

长按二维码关注

机器学习算法那些事
微信：beautifulife244

砥砺前行 不忘初心