使用PCA可视化数据

原创 磐怼怼 深度学习与计算机视觉 2020-08-28

收录于话题
#PCA

13个

主成分分析（PCA）是一个很好的工具，可以用来降低特征空间的维数。PCA的显著优点是它能产生不相关的特征，并能提高模型的性能。

它可以帮助你深入了解数据的分类能力。在本文中，我将带你了解如何使用PCA，同时提供Python代码，完整的项目可以在GitHub链接：https://github.com/conorosully/medium-articles。

什么是PCA

我们先来复习一下这个理论，但是如果你想确切了解PCA是如何工作的，我们不会详细介绍，网上有大量学习资源。

PCA用于减少用于训练模型的特征维度数量，它通过从多个特征构造所谓的主成分（PC）来实现这一点。

PC的构造方式使得PC1方向在最大变化上尽可能地解释了你的特征，然后PC2在最大变化上尽可能地解释剩余特征，PC1和PC2通常可以解释总体特征变化中的绝大部分信息。

另一种思考方法是，前两个PC可以很好地概括大部分特征。这很重要，因为正如我们将看到的，它允许我们在二维平面上可视化数据的分类能力。

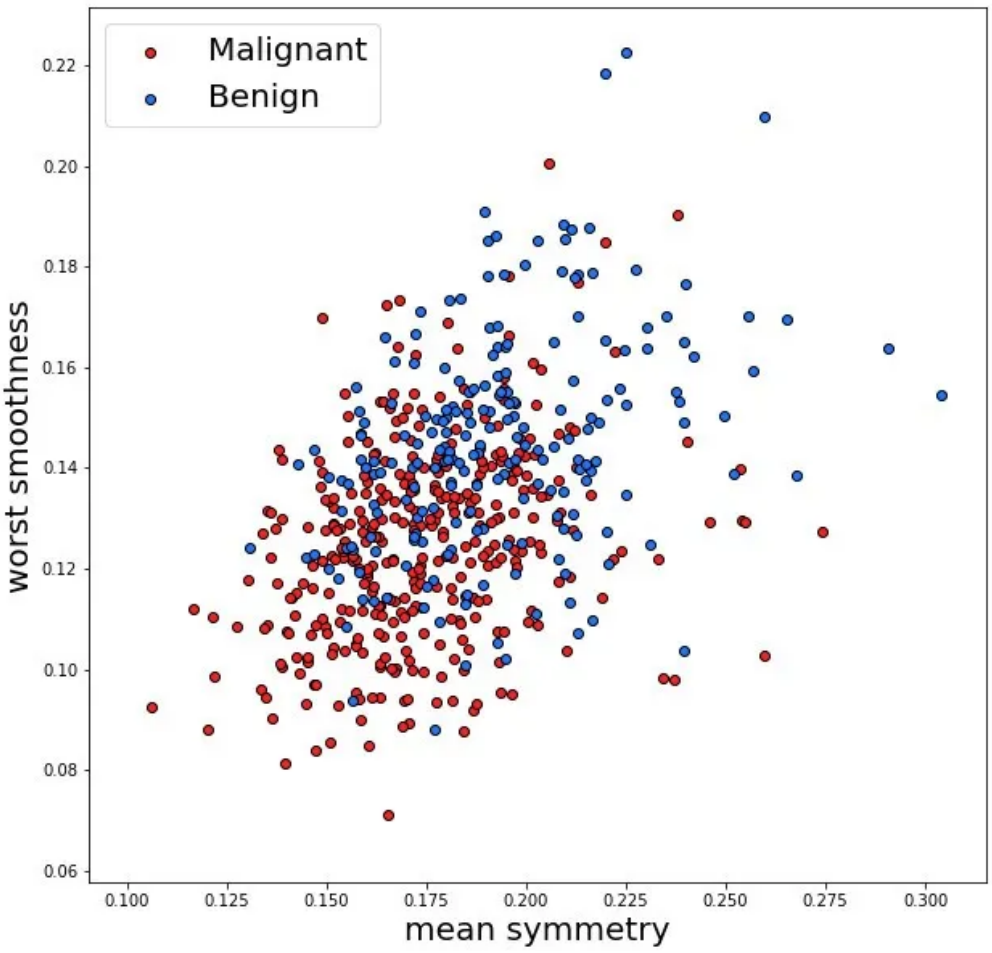
数据集

让我们来看看一个实际的例子，我们将使用PCA来探索乳腺癌数据集(http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))，我们使用下面的代码导入该数据集。

```
import numpy as np
import pandas as pd
from sklearn.datasets import load_breast_cancer
cancer = load_breast_cancer()

data = pd.DataFrame(cancer['data'],columns=cancer['feature_names'])
data['y'] = cancer['target']
```

目标变量是乳腺癌检测的结果，恶性或良性。每次测试，都要取多个癌细胞；然后从每个癌细胞中采取10种不同的测量，这些测量包括细胞半径和细胞对称性；最后，为了得到特征值，我们计算了每个度量值的平均值、标准误差和最大值(不太好的)，这样我们总共得到30个特征值。

在图中，我们仔细观察了其中两个特征——细胞的平均对称性(Benign)和最差平滑度(worst smoothness)。



在图中，我们看到这两个特征可以帮助区分这两个类，即良性肿瘤往往更为对称和光滑，但是，仍然有很多重叠的，所以仅仅使用这些特征的模型不会做得很好。

我们可以创建这样的图来了解每个特征的预测能力，但是有30个特征，这意味着有相当多的图要分析，他们也没有告诉我们如何作为一个整体来预测数据集。这就需要我们可以引入PCA。

PCA-整个数据集

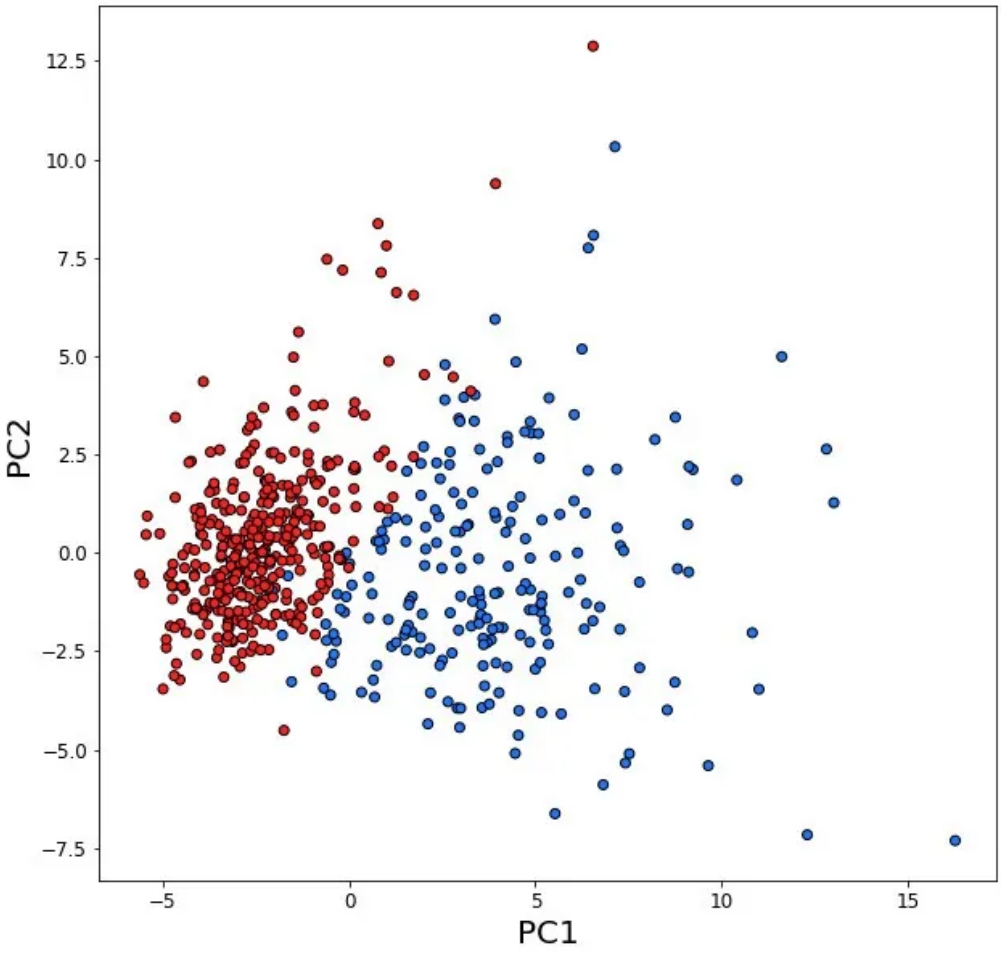
首先，我们对整个数据集进行主成分分析。我们使用下面的代码。

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

#标准化
scaler = StandardScaler()
scaler.fit(data)
scaled = scaler.transform(data)

#PCA
pca = PCA().fit(scaled)

pc = pca.transform(scaled)
pc1 = pc[:,0]
pc2 = pc[:,1]

#画出主成分
plt.figure(figsize=(10,10))

colour = ['#ff2121' if y == 1 else '#2176ff' for y in data['y']]
plt.scatter(pc1,pc2 ,c=colour,edgecolors='#000000')
plt.ylabel("Glucose",size=20)
plt.xlabel('Age',size=20)
plt.yticks(size=12)
plt.xticks(size=12)
plt.xlabel('PC1')
```

我们首先标准化特征，使它们的平均值为0，方差为1，这过程是很重要的，因为主成分分析通过最大化主成分分析所解释的方差来工作。

一些特征由于其没有经过标准化自然会有更高的方差，例如，以厘米为单位测量的距离将比以公里为单位测量的相同距离具有更高的方差。在不缩放特征的情况下，主成分分析将被那些高方差特征“吸引”。

缩放完成后，我们会拟合PCA模型，并将我们的特征转换为PC。由于我们有30个特征，我们最多可以有30个PC，对于可视化，我们只对前两个感兴趣，最后使用PC1和PC2创建如图所示的散点图。



在图2中，我们可以看到两个不同的簇。虽然仍然有一些重叠，但是比我们在之前的图中要清晰得多。这告诉我们，作为一个整体，这个数据集在区分恶性肿瘤和良性肿瘤方面会做得更好。

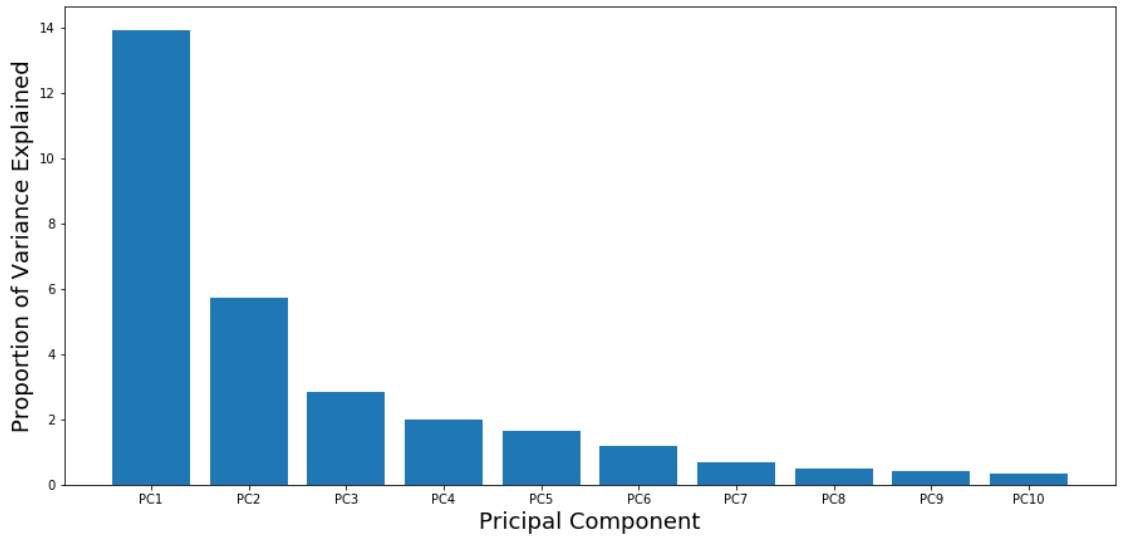
我们还应该考虑的是，我们只关注了前两个PC，因此并不是所有特征的变化都被捕获，这意味着使用所有特征训练的模型仍然可以正确预测异常值（即聚类中不清楚的点）。

在这一点上，我们应该注意这种方法的一个缺陷。我们提到PC1和PC2可以解释特征中很大一部分的差异，然而，这并不总是真的。在某种情况下，这些PC可以被认为是对特征的错误总结，这意味着，即使你的数据能够很好地分离，你也可能无法获得如上图所示的清晰簇。

我们可以通过查看PCA-scree图来确定。我们使用下面的代码为这个分析创建scree图，

```
var = pca.explained_variance_[0:10] #percentage of variance explained
labels = ['PC1','PC2','PC3','PC4','PC5','PC6','PC7','PC8','PC9','PC10']

plt.figure(figsize=(15,7))
plt.bar(labels,var,)
plt.xlabel('Pricipal Component')
plt.ylabel('Proportion of Variance Explained')
```

它本质上是一个柱状图，其中每个柱状图的高度是相关PC解释的方差百分比。我们看到，总共只有大约20%的特征方差是由PC1和PC2解释的，即使只解释了20%，我们仍然得到两个不同的集群，这强调了数据的预测能力。

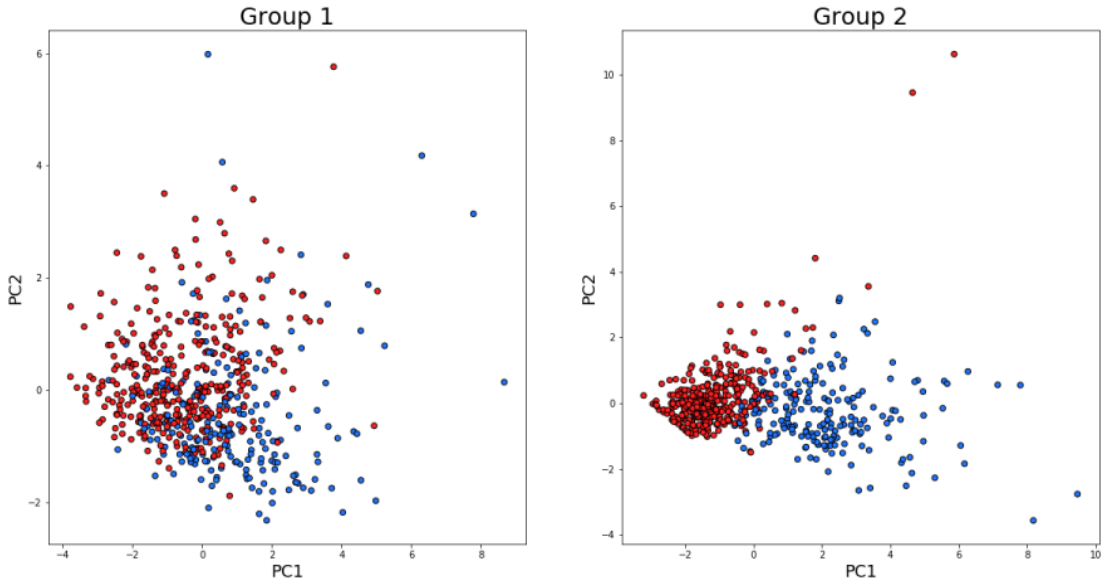


PCA-特征组

到目前为止，我们已经使用主成分分析来了解整个特征集对数据的分类效果，我们还可以使用这个过程来比较不同的特征组，例如，假设我们想知道细胞的对称性和光滑性是否比细胞的周长和凹陷性更好。

```
group_1 = ['mean symmetry', 'symmetry error','worst symmetry',
'mean smoothness','smoothness error','worst smoothness']

group_2 = ['mean perimeter','perimeter error','worst perimeter',
'mean concavity','concavity error','worst concavity']
```

我们首先创建两组特征，第一组包含所有与对称性和光滑性有关的特征，第二组包含所有与周长和凹陷性有关的特征；然后，除了使用这两组特征外，我们以与之前相同的方式进行主成分分析。这个过程的结果如下图所示。



我们可以看到，对于第一组，有一些分离，但仍然有很多重叠；相比之下，第2组有两个不同的簇。因此，从这些图中，我们可以预期第2组的特征（即细胞周长和凹陷）将是预测肿瘤是恶性还是良性的更好指标。

这也将意味着使用组2中特征的模型比使用组1中特征的模型具有更高的精度。现在，让我们来验证这个假设。

我们使用下面的代码来训练一个使用两组特征的logistic回归模型。在每种情况下，我们使用70%的数据来训练模型，剩下的30%用来测试模型。

```
from sklearn.model_selection import train_test_split
import sklearn.metrics as metric
import statsmodels.api as sm

for i,g in enumerate(group):

    x = data[g]
    x = sm.add_constant(x)
    y = data['y']
    x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.3,
```

```
                                                    random_state =
101)

    model = sm.Logit(y_train,x_train).fit() #fit logistic regression model

    predictions = np.around(model.predict(x_test))
    accuracy = metric.accuracy_score(y_test,predictions)

    print("Accuracy of Group {}: {}".format(i+1,accuracy))
```

第一组测试集的准确率为74%，相比之下，第二组的准确率为97%。因此，组2的特征明显是更好的预测因子，这正是我们从主成分分析结果中所看到的。

本文，我们了解了如何在开始建模之前使用PCA来加深对数据的理解，了解哪些特征是可预测的，将在特征选择方面给你带来优势，此外，查看特征的总体分类能力将使你了解预期的分类精度。

如前所述，这种方法并不能完全证明，因此应与其他数据勘探图和汇总统计一起使用。一般来说，在开始建模之前，最好从尽可能多的不同角度查看数据。

参考链接：https://towardsdatascience.com/visualising-the-classification-power-of-data-54f5273f640

☆ **END** ☆

如果看到这里，说明你喜欢这篇文章，请转发、点赞。微信搜索「uncle_pn」，欢迎添加小编微信「mthler」，每日朋友圈更新一篇高质量博文。

↓扫描二维码添加小编↓