数据科学基础第三次作业

王晨曦

2018 年 5 月 3 日

1 问题重述

对 Boston 数据集进行降维，并进行探索。

注：Boston 数据集包含 506 组数据，每条数据包含房屋以及房屋周围的详细信息。其中包含城镇犯罪率、一氧化氮浓度、住宅平均房间数、到中心区域的加权距离等 13 个属性，以及自住房平均房价。

2 主成分分析

主成分分析（Principal Component Analysis，简称 PCA）是最常用的一种降维方法。它是一种设法将原来变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法。

在 PCA 中，数据中心化后从原来的坐标系转换到新的坐标系。转换坐标系时，以方差最大的方向作为坐标轴方向，因为数据的最大方差给出了数据的最重要的信息。第一个新坐标轴选择的是原始数据中方差最大的方法，第二个新坐标轴选择的是与第一个新坐标轴正交且方差次大的方向。重复该过程，重复次数为原始数据的特征维数。

PCA 的主要过程如下：

- 各属性零中心化；
- 求协方差矩阵；
- 求协方差矩阵的特征值及对应特征向量；
- 对特征值从大到小排序；
- 将特征向量按对应特征值大小从上到下按行排列成矩阵；
- 将数据转换到 k 个特征向量构建的新空间中（k 为指定的维度数）。

通过 PCA 对数据进行降维，我们可以将数据中最主要的独立分量抽取出来，去除掉数据的一些冗余信息和噪声，使数据变得更加简单高效，提高其他机器学习任务的计算效率。

本次作业采用 PCA 进行数据降维，并以此为基础对数据集进行探索。

3 问题解答

本部分首先介绍数据集降维的 PCA 代码实现，根据特征值排序并统计分量累计占比和原始属性权重系数，接着以线性回归为基础探索降维后的维度数量对回归结果的影响，最后展示了 1-3 维的结果可视化。

3.1 降维实现

根据PCA的原理，对其进行代码实现，以便接下来的分析，函数定义如下：

```python
def pca_reduction(data, num_dims=1):
    '''
    PCA降维的实现
    INPUT:
        data: 原始数据, float (numpy array of N x D)
        num_dims: 降维后的维度数, int
    OUTPUT:
        pca_nd: 降维后的数据, float (numpy array of N x num_dims)
        w: 特征值
        v: 特征向量
    '''
    # 数据中心化
    col_mean = np.mean(data, axis=0)
    centered_data = data - col_mean
    # 求特征值和特征向量
    cov_data = np.cov(centered_data.T)
    w,v = np.linalg.eig(cov_data)
    # 求降维后的数据
    pca_nd = np.dot(centered_data, v[:,:num_dims])

    return pca_nd, w, v
```

基于以上实现，可以求出原始数据的特征值，各成分特征值如表1所示，其中Ci表示降序排列后的第i个成分。

表1: 矩阵特征值

C1	C2	C3	C4	C5	C6	C7
30889.518	6250.274	818.365	266.683	50.135	27.897	16.101
C8	C9	C10	C11	C12	C13	
9.534	3.273	1.181	0.255	0.060	0.003	

3.2 主成分累计占比

令λ_i为第i个分量的特征值，则前k个分量的累计占比为

$$\eta_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \times 100\%$$

分量的累计占比可以反映数据的主成分分量个数，从而帮助降维。图1展示了1-13个分量的累计占比，可以看出，前三个分量所占的比重较大（合计99%），后面的分量占比越来越小，最后几个分量可以理解为数据中的噪声，实际降维时应该舍去。

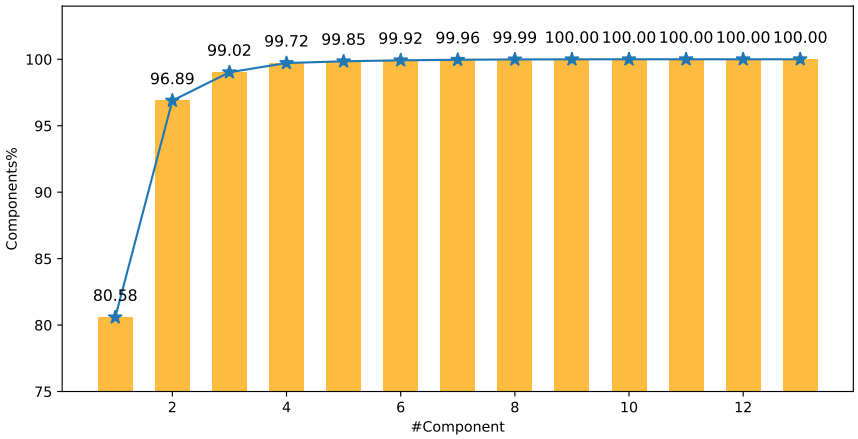


图 1: 主成分累计占比

3.3 各属性所占权重

除了分析各分量所占比重，还可以研究一下原始数据集中的各属性在降维过程中的重要程度，这里我们把重要程度量化为属性的总权值。假设将数据降至 K 维，第 i 个维度对应的特征值和特征向量分别为 λ_i 和 w_i，则第 j 个属性的总权值为

$$W_j = \sum_{i=1}^{K} \lambda_i |w_i(j)|$$

根据上述计算方式，可以计算出各属性的权重系数，图2展示了 $K = 5$ 时各属性的权重系数（没有归一化），为了方便展示，坐标为非均匀刻度。

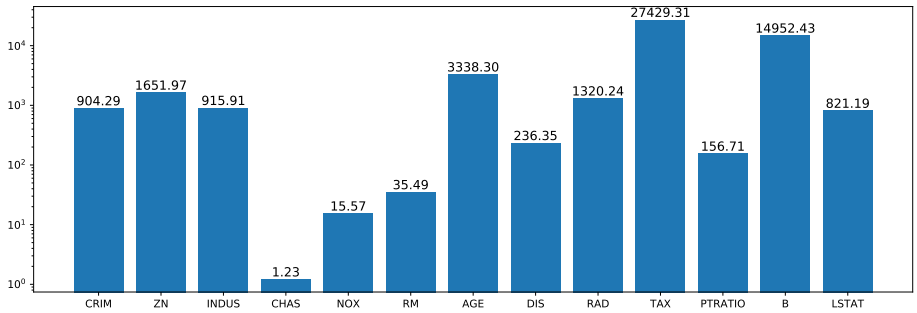


图 2: 各属性权重系数

根据图2，可以找到权重最大的前五个属性，如表2所示。

3.4 降维程度对预测误差的影响

我们从另一个角度来看看降维对数据的影响，Boston 数据集的主要任务是根据各类条件回归出一个合理的房价，因此降维后效果如何值得探讨。

从数据集中随机抽取 300 个样本作为训练集样本，其余放入测试集，使用简单的线性回归对数据进行训练，并在测试机上进行预测，观察预测值和真实值之间的误差，这里误差

表 2: 权重较大的属性

属性名称	属性说明
TAX	全价物业税税率（/10000 美元）
B	$1000(B_k - 0.63)^2$，B_k 是城镇黑人人口比例
AGE	1940 年前建成单位占有比例
ZN	地段住宅用地比例超过 25000 平方英尺
RAD	径向公路可达性指标

度量定为均方根误差（MSE）。图3展示了降至不同维度后的回归误差，可以看到，误差随着维度的增加，整体趋势表现为逐渐下降，且 5 个维度与 6 个维度的误差有明显下降（图中红色虚线），其他位置则平缓变化，这说明 6 个维度已经能将数据集的特征很好的表征出来，结合图1，前六个分量占到了超过 99.9% 的比例，也能验证这个推断。

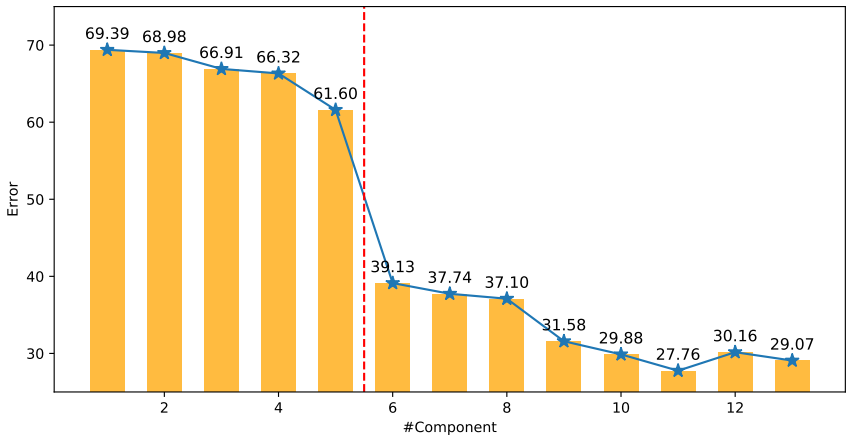


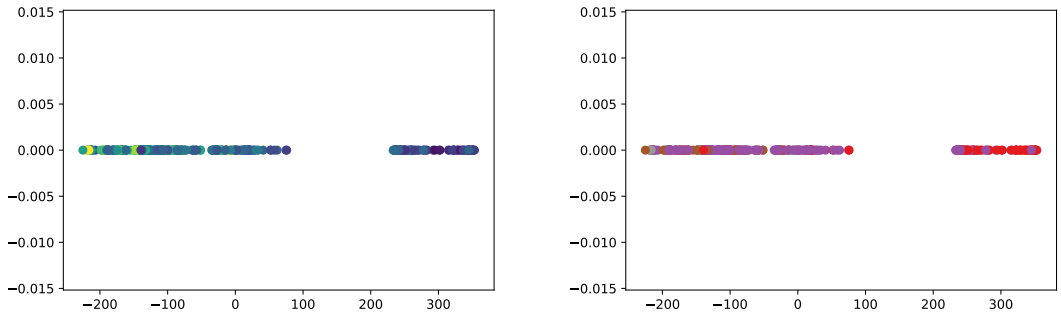
图 3: 测试集误差

在图3中，误差是逐渐下降的，这并不是说降维对回归起了冲突的效果，之所以维度数越多越好，是因为 Boston 数据集相对来说是个体量比较小的数据集，数据的维度也只有 13 维，噪声没有成为影响数据分布的关键因素，如果换成大数据集，降维的优越性就能更好的体现出来了。

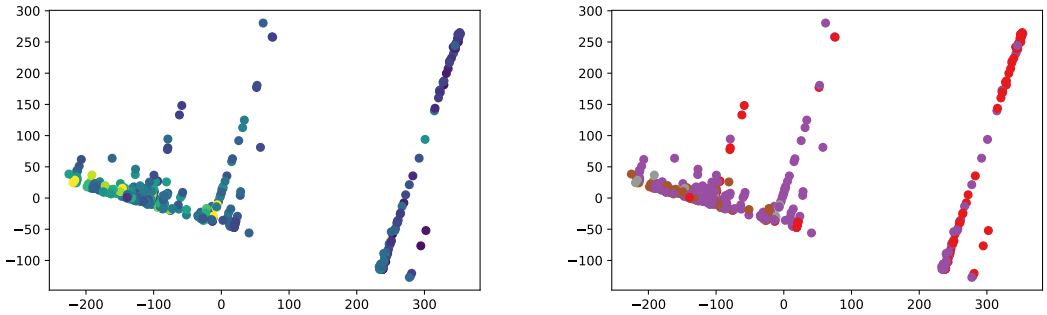
3.5 可视化

图4左侧三张图分别展示了 Boston 数据集降至 1、2、3 维时的分布情况，颜色明暗表示价格高低，为了更方便观看，我们以 15 为间隔对房价进行类别划分，从而得到四个离散的类别，并进行着色，结果见右侧的三张图。可以发现，1 维时已经能将数据划分成两类，拓展至 2 维和 3 维时则更清晰的展示了不同价位的数据在特征上的分布。

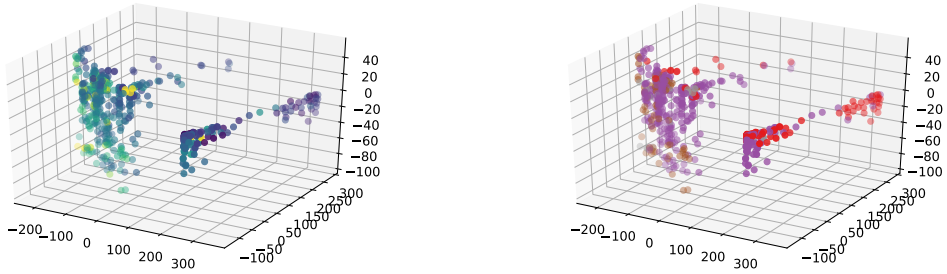
我们还可以观察到，正如定义的一样，降至 3 维的数据可通过投影依次继续降为 2 维和 1 维。



#Dims = 1



#Dims = 2



#Dims = 3

图 4: 降维结果可视化