PCA 和白化

2019 年 7 月 21 日

PCA 和白化是一种常用的数据预处理方法。PCA 是一种去除变量之间的相关性，使之线性无关最常用的方法。PCA 只能去除线性相关性不能去除非线性相关性。线性相关性是指变量之间满足线性函数关系，非线性相关性是指变量之间满足函数关系，但不是线性函数。

1 PCA 介绍

1.1 从去除线性相关性推导 PCA

假设样本只有两个属性 x 和 y, 我们采集一系列样本 (x_i, y_i), 表示为数据矩阵 D，每一行为一个样本数据，每一列为一个属性取值，设两列向量为 X 与 Y。若两个属性之间存在线性关系，则可表示为 $y = kx + b(k \neq 0)$。可以利用最小二乘法进行拟合得到 k:

$$k = 1/m \sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y}) / \left[m \sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i \right)^2 \right]$$

令 $\operatorname{cov}(x, y) = 1/m \sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y})$ ，则当 $\operatorname{cov}(x, y) = 0$ 时，x 和 y 线性无关；当 $\operatorname{cov}(x, y) > 0$ 时,x 和 y 线性正相关；当 $\operatorname{cov}(x, y) < 0$ 时,x 和 y 线性负相关。因为 $\operatorname{var}(x) = \operatorname{cov}(x, x)$ 时变量 x 的方差，所以称 $\operatorname{cov}(x, y)$ 为协方差。

现在已经解决了两个变量线性无关的度量（协方差为 0），下考虑存在一线性变换 C, 对属性进行变换，使得新变量的协方差为 0。

现假设已经对属性进行中心化，即每个属性的均值为 0，则协方差为 $\operatorname{cov}(x, y) = 1/m \sum_{i=1}^{m} x_i y_i$ 恰好是内积：$\operatorname{cov}(x, y) = 1/m X^{\mathrm{T}} Y$。线性变换 C 对数据矩阵 D 进行变换，变换后的新数据矩阵为 $D' = DC = [DC_1, DC_2]$，

新的列向量为 $X' = DC_1$,$Y' = DC_2$, 其中 $C_!$ 和 C_2 是 C 的列向量，则新向量的方差为：$X'^T X' = C_1^T D^T DC_1$,$Y'^T Y' = C_2^T D^T DC_2$, 协方差为 $X'^{\mathrm{T}}\boldsymbol{Y}' = C_1^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{DC}_2$。

矩阵 $D^T D$ 是维度为 2*2 的方阵，维数等于属性数量。该矩阵的对角线元素为原始属性的方差，非对角线元素为协方差，故该矩阵为协方差矩阵，且 $(D^T D)^T = D^T D$ 是对称矩阵。

令 $\boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{DC}_1 = \lambda_1$, $\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{DC}_2 = \lambda_2$，分别是两个新属性的方差（大于等于 0）。考虑新属性需要线性无关，则此时协方差 $X'^{\mathrm{T}}Y' = \boldsymbol{C}_1^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{DC}_2 = 0$，可以发现在此时 C_1 和 C_2 正交，如果将向量 C_1、C_2 与对应的 λ_1、λ_2 看作协方差矩阵的特征向量和特征值，即：$\boldsymbol{D}^{\mathrm{T}}\boldsymbol{DC}_1 = \lambda_1\boldsymbol{C}_1$, $\boldsymbol{D}^{\mathrm{T}}\boldsymbol{DC}_2 = \lambda_2\boldsymbol{C}_2$ 代入上面三个灯饰，得到 $C_1^T C_1 = 1$, $C_2^T C_2 = 1$, $C_1^T C_2 = 0$, 即线性变换为标准正交变换，故将特征向量正交化即可。

若由 d 个属性，推理过程完全一致。

去除样本属性之间的线性相关性：假设样本有 d 个属性，采集 n 个样本，数据矩阵为 D, 其中每行是一个样本，则 D 的尺寸为 $n \times d$。

第一步：计算协方差矩阵 $COV = D^T D$, 对角线元素为属性的方差，非对角线元素为协方差。

第二步：对协方差矩阵 COV 进行特征值分解，得到特征值和对应特征向量 (λ_i, C_i), 其中特征向量需要规范化（向量长度为 1 且两两正交），即：$C_i^T C_i = 1$, $C_i^T C_j = 0$。

第三步：利用向量 C_i 得到新属性 DC_i，新属性的方差为 $\lambda_i(\geq 0)$，协方差为 0，此时数据矩阵 $D' = DC$, 由于矩阵 C 是正交矩阵，所以变换是可逆的，可得 $D = D'C^T$。

1.2 PCA 代码实现

实践中，一般使用奇异值分解代替特征值分解，以获得更稳定的数值解和更快的速度。例：代码实现

```
import numpy as np
D = 784
N = 128
X = np.random.randn(N, D)
mean = np.mean(X, axis=0)
```

```
6         X -= mean
7         COV = np.dot(X.T, X) / X.shape(0)  # 协方差矩
              阵
8         C, S, V = np.linalg.svg(COV)  # 奇异值分解
9         X_decor = np.dot(X, C)  # 得到去除线性相关性后
              的数据矩阵（对角阵）
```

X_dector 与 X 尺寸相同，每行为一个样本，每列为所有样本的新属性值。X_decor 是中心化的（每个属性均值为 0）。第八行代码利用 numpy 实现奇异值分解得到正交矩阵 $C = (C_1, C_2, \cdots, C_d)$ 和特征值向量 $S = (\lambda_1, \lambda_2, \cdots, \lambda_d)$，其中 S 中的特征值是降序排列的，所以 X_dector 中第一个属性的方差最大，后面依次减小。

1.3 PCA 降维

PCA 降维原理：如果协方差矩阵的某个特征值为 0，说明该属性的方差为 0，属性值是恒定值，即各个样本在该属性上没有任何差别，都那么该属性对分类没有任何价值，可以去掉该属性。

实际中，方差很小的新属性都可以去掉，实现数据降维。

数据降维也实现了数据压缩。由于变换是可逆的，降维后的数据可以恢复到原始数据。

例：代码实现

```
    X_dector_reduce = np.dot(X, C[;, :127])  # 降维
    zero_matric = np.zeros((X.shape[0], X.shape[1] -
        127))
    X_decor_reduce_zero = np.hstack((X_decor_reduce,
        zero_matic))  # 除去近似0的属性数值
    X_denoise = np.dot(X_decor_reduce_zero, C.T)  # 逆
        变换
    X_denoise += mean   # 加上原均值，得到降维后重构的
        数据矩阵
```

在实践中，只有样本属性特别多，或者属性之间存在明细那的线性相关性时，才采用 PCA 降维。

具体降维时一般采用经验法则：

观察法：画出方差 S 的曲线，当曲线十分接近 x 轴时，取此点之前的属性。

比例法：降维后数据信息量占原始信息量的比例要大于阈值（如 95%），衡量信息量可以利用方差，属性方差越大表示该属性信息量越大。具体做法：计算前 d' 个方差之和占总方差之和的比例，取最小的 d' 使比例大于 95%，则保留前 d' 个属性即可。

注：去除方差小的属性，是一种数据去噪技术。

1.4 PCA 的几何意义

PCA 的几何意义：先堆数据点云进行平移（减去均值），使得重心和原点重合，然后旋转点云（乘以正交矩阵），使 X_decor" 内嵌" 到尽可能低的维度空间内。

点云内嵌到低维空间可以用更少的坐标表示点，达到 PCA 的降维效果。

2 白化

若需要对数据属性进行规范化（去除单位量纲的影响），则点云需要改变形状，这时可以利用属性标准差进行规范化，即属性值除以对应标准差。例：代码实现

```
X_white = X_decor / np.sqrt(S + 10**(-5))
```

这种规范化称为白化。几何意义是使每个属性的方差相同，数据点云完全处于高维球内。

X_white 的协方差矩阵为单位矩阵，X_white 变成近似独立同分布（i.i.d）的数据，每个属性都是均值为 0、方差为 1 的分布（近似同分布）。规范化后的数据是近似同分布但没有去除线性相关性（无旋转操作）。

分母通常加小常数，如：10^{-5}，这样是为了防止被除数为 0。原因在于当协方差矩阵不满秩时，存在为 0 的方差。白化最大的缺点是会扩大噪音，因为它把所有属性的标准差都拉伸为 1。小常数可以达到去噪的目的，但数据也会被过度平滑，因此最优小常数需要交叉验证。