PCA算法

xhj

November 8, 2018

1 最近重构性

最近重构性要求样本点到降维后的超平面的距离之和最短，以此来确定该超平面。

2 最大可分性

2.1 基于特征值和特征向量求主成分

1、中心化（均值化）。目的是为了方便之后的求解。因为需要求解协方差矩阵，协方差的公式为$cov(x,y)=E\{(x-Ex)(y-Ey)\}$。利用中心化之后的矩阵就可以直接求协方差。

2、求协方差矩阵。协方差矩阵的维数由数据的特征的数目决定。

$$cov=\begin{pmatrix} cov(x_0,x_0) & cov(x_0,x_1) & ... & cov(x_0,x_n) \\ cov(x_1,x_0) & cov(x_1,x_1) & ... & cov(x_1,x_n) \\ ... & ... & ... & ... \\ cov(x_n,x_0) & cov(x_n,x_1) & ... & cov(x_n,x_n) \end{pmatrix} \tag{1}$$

通过求解协方差矩阵，可以观察样本中各个特征之间的相关度。通过协方差也可以求解相关系数ρ，$\rho=\frac{cov(x,y)}{\sqrt{D(x)D(y)}}$。

3、求解协方差矩阵的特征值和特征向量。在协方差矩阵的特征值中选取最大的k个，并且提取它们对应的特征向量。此步骤的目的就是构建一个正交的新的坐标系，因为特征向量都是正交的单位向量，所以可以作为坐标系。其中协方差矩阵的特征值越大，说明特征之间的差别越大，越能将样本数据区分开。若选出前k个特征值，则对应的特征向量的维数为：(n,k)。

4、将中心化之后的样本与与选出的特征向量点乘（向量的点乘就是做投影运算）。该运算之后得到的向量，就是得到的主成分。

2.2 基于方差最大