PCA

2019年3月31日 23:01

PCA

PCA原理 https://blog.csdn.net/hustqb/article/details/78394058

代码实现

https://github.com/wepe/MachineLearning/blob/master/PCA/pca.py

伪码（对行为观测，列为变量的数据，shape=（n,m））：

1.对数据按列均值为0

2.对数据求协方差矩阵，shape=（m，m）

3.计算协方差矩阵的特征向量和特征值

4.按特征值升序排列特征向量

5.输出前n个特征向量

np.cov 计算协方差矩阵

https://blog.csdn.net/jeffery0207/article/details/83032325

协方差：用于衡量两个变量“协同变异”的情况。而方差是协方差的一种特殊情况，即当两个变量是相同的情况，即单个变量“自身变异的情况”。

$$cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

协方差矩阵

$$C=\begin{pmatrix} cov(1,1) & cov(1,2) & cov(1,3) & \cdots & cov(1,n) \\ cov(2,1) & cov(2,2) & cov(2,3) & \cdots & cov(2,n) \\ cov(3,1) & cov(3,2) & cov(3,3) & \cdots & cov(3,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(n,1) & cov(n,2) & cov(n,3) & \cdots & cov(n,n) \end{pmatrix}$$

方差：统计学采用平均离均差平方和来描述变量的变异程度。

$$\sigma^2=\frac{\sum(X-\mu)^2}{N}$$

相关系数

$$\rho_{\xi\eta} = \frac{cov(\xi, \eta)}{\sqrt{var(\xi)}\sqrt{var(\eta)}}$$

MJ：方差是变量内的变异程度，而协方差是变量间的相关性。协方差越大，即你大的地方我也大，越相关。

MJ：PCA计算的是不同维度之间的协方差，而不是不同样本间的协方差。所以通常情况的数据，每行是一个观测，每列是一个变量/维度/属性，需要np.cov(data, rowvar=False)

特征值和特征向量

https://blog.csdn.net/weixin_37721518/article/details/79016226

不是太能理解，得有空再看看。

特征向量之间相互正交，Ax=λx。

特征值越大，说明特征向量越重要。

PCA算法的思想：

主成分分析是利用降维的思想，将多个变量转化为少数几个综合变量（即主成分），其中每个主成分都是原始变量的线性组合，各主成分之间互不相关，从而这些主成分能够反映始变量的绝大部分信息，且所含的信息互不重叠。它是一个线性变换，这个变换把数据变换到一个新的坐标系统中，使得任何数据投影的第一大方差在第一个坐标(称为第一主成分)上，第二大方差在第二个坐标(第二主成分)上，依次类推。主成分分析经常用减少数据集的维数，同时保持数据集的对方差贡献最大的特征。