PCA主成分分析

赵海臣

原理

- 主成分分析(Principal Component Analysis，PCA)，在正交空间中寻找信息量最大的方向。
- 通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。
  - 人们希望变量个数较少而得到的信息较多。
  - 变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，说明这两个变量的信息有重叠。
  - 主成分分析是对于原先提出的所有变量，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

基本思想

- 基本思想
  - 通过协方差矩阵映射到向量(线性组合)上的方差来衡量在该向量方向上的信息量大小。
    - 最经典的做法就是用F1（选取的第一个线性组合，即第一个综合指标）的方差来表达，即Var(F1)越大，表示F1包含的信息越多。
    - 因此在所有的线性组合中选取的F1应该是方差最大的，故称F1为第一主成分。
    - 如果第一主成分不足以代表原来P个指标的信息，再考虑选取F2即选第二个线性组合，为了有效地反映原来信息，F1已有的信息就不需要再出现在F2中，F1与F2两个向量垂直，用数学语言表达就是要求协方差Cov(F1, F2)=0，则称F2为第二主成分，依此类推可以构造出第三、第四，……，第P个主成分。

PCA计算例子

- 假定数据是二维的：
  - x=[2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2, 1, 1.5, 1.1]T
  - y=[2.4, 0.7, 2.9, 2.2, 3.0, 2.7, 1.6, 1.1, 1.6, 0.9]T
- 基本统计概念：
  - 均值：
    $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
  - 方差：
    $$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$
  - 标准差：
    $$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$$

协方差

标准差和方差一般是用来描述一维数据的，协方差，就是一种用来度量两个随机变量关系的统计量：

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

- 不难发现：
  - 相同变量协方差就是变量的方差：

$$\mathrm{cov}(X,X)=\mathrm{var}(X)$$

  - 两个变量的协方差满足交换律：

$$\mathrm{cov}(X,Y)=\mathrm{cov}(Y,X)$$

协方差矩阵

- 协方差也只能处理二维问题，那维数多了自然就需要计算多个协方差，那自然而然的我们会想到使用矩阵来组织这些数据。
  - 协方差矩阵的意义是n个维度之间相互的协方差，利用矩阵来组织起来：

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

  - 因此，协方差是对称的矩阵，对角线上是变量自身的协方差，即方差。

协方差的特征向量和特征值

- 矩阵乘法 $A\cdot\vec{b}$ 对应了一个变换，是把任意一个向量 $\vec{b}$ 变成另一个方向或长度都大多不同的新向量。

$$\vec{b}' = A\cdot\vec{b}$$

- 在这个变换的过程中，原向量 $\vec{b}$ 主要发生旋转、伸缩的变化。如果矩阵 A 对某一个向量或某些向量只发生伸缩变换，不对这些向量产生旋转的效果，那么这些向量就称为这个矩阵的特征向量，伸缩的比例 λ 就是特征值。

$$\lambda\cdot\vec{b}' = A\cdot\vec{b}$$

- 在PCA中，特征值的大小意味着该特征向量方向上的方差大小，特征值越大，该特征向量上的方差越大，信息越多。

选择特征向量

- 求出协方差矩阵的特征值及特征向量之后，按照特征值由大到小进行排列，这将给出成分的重要性级别。
  - 可以忽略那些重要性很小的成分，当然这会丢失一些信息，但是如果对应的特征值很小，不会丢失很多信息。
  - 如果忽略掉一些低重要性维度，最后的数据集将有更少的维数：如果原始数据是n维的，选择了前p个主要成分，那么现在的数据将仅有p维。

- 最后将选择出的特征向量形成模式矢量：
  - 几个具有较大特征值的特征向量组成的矩阵，它由选择出的特征向量构成，每一个特征向量是这个矩阵的一列。

降维处理

- 将原来的n维数据与模式矢量 $M_{n\times p}$ 做乘法，得到降维后的p维数据。

$$AdjustedData_{i\times p} = Data_{i\times n} \times M_{n\times p}$$

- i为数据长度，n为原数据维度，p为降维后维度。

THE END

THANK YOU！

聚美优品
JUMEI.COM