PCA原理与实践

在机器学习的领域中，对原始数据进行特征提取，经常会得到高维度的特征向量。这些高维的特征不仅增加计算量，而且可能会包含一些冗余和噪声。所以我们希望通过降维的方式来寻找数据内部的特性，提升特征表达能力，降低模型的训练成本。PCA是一种降维的经典算法，属于线性、无监督、全局的降维方法。PCA作为一种降维方法，常用于数据预处理阶段。

|  |
| --- |
| 降维有两个主要算法：线性判别分析（Linear Discriminant Analysis，简称LDA）和主成分分析。这两者之间最根本的区别在于，LDA利用类的信息去找新特征，以最大化类的可分离性，而PCA利用每个特征的方差最大化不同特征的可分离性。因此，LDA是有监督算法，而PCA是无监督算法。 |

# PCA原理

PCA全名principal component analysis，即主成分分析。简单地说，主成分分析是一组变量通过正交变换转变成另一组变量的分析方法，来实现数据降维的目的。转换后得到的这一组变量，即是主成分。假设最终得到r个主成分，这r个主成分会依据方差的大小进行排序，称作主成分（PC）1、主成分2、……主成分r。而每个主成分的方差在这一组变量中的总方差中所占的比例，即是主成分的贡献度。

PCA的原理是线性映射，简单的说就是将高维空间数据投影到低维空间上，然后将数据包含信息量大的主成分保留下来，忽略掉对数据描述不重要的信息。PCA的目标是既可以减少特征空间，又尽可能多的保持原来数据的信息（因此，降维之后必定会损失一些信息）。

从宏观上理解PCA降维。就是说找到一些方差大的特征代表全部特征，这是因为再我们分析问题的时候，如果一个特征所有的值差异都很小，那么这个特征对我们分析问题的帮助就不会很大，而那些特征值变化较大的特征才会对我们分析问题带来更多的信息。因此，可以说方差越大的特征，包含的信息就越多。而PCA找主成分的时候其实在寻找K个尽可能的把样本区分开方向，即方差尽可能大的方向作为主成分，这样就可以做到在保留尽可能多的数据信息的情况下把数据的维度降到了K维（原来肯定是比K维大）

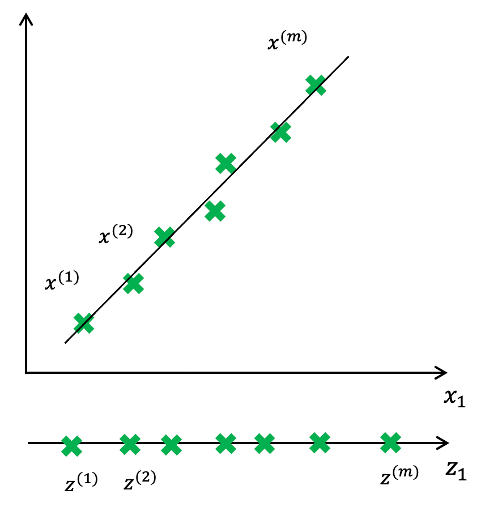
PCA降维是把数据看成空间中的点，然后尝试去寻找几个方向，把样本点进行投影，投影之后让这些点离得尽可能的远。这样这几个方向就是主成分，空间中的样本点就可以通过这几个新的方向进行描述了。但是找的这几个方向需要是互不干扰，没有线性关系，就像x轴和y轴那样，这样才能更好的去描述这些数据。主成分之间没有冗余。所以，主成分的标准两个条件，一是互不相关，二是用来描述数据的时候，方差尽可能大。

而对于正交属性空间中的样本，如何用一个超平面对所有样本进行恰当合适的表达呢？若存在这样的超平面，应该具有两种性质：

* 所有样本点到超平面的距离最近
* 样本点在这个超平面的投影尽可能分开

上面两种性质便是主成分分析的两种等价的推导，即PCA最小平方误差理论和PCA最大方差理论。

PCA的降维操作是选取数据离散程度最大的方向(方差最大的方向)作为第一主成分，第二主成分选择方差次大的方向，并且与第一个主成分正交。重复这个过程直到找到k个主成分。



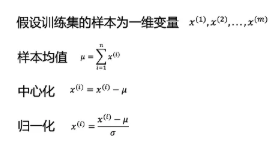
数据点分布在主成分方向上的离散程度最大，且主成分向量彼此之间正交。

# PCA步骤



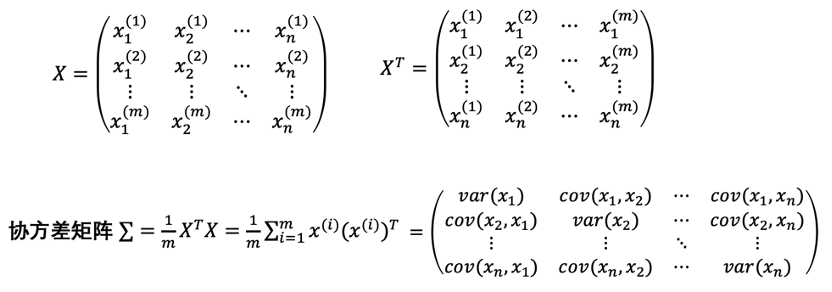
1、对所有数据特征进行中心化和归一化

对样本进行平移使其重心在原点，并且消除不同特征数值大小的影响，转换为统一量纲：

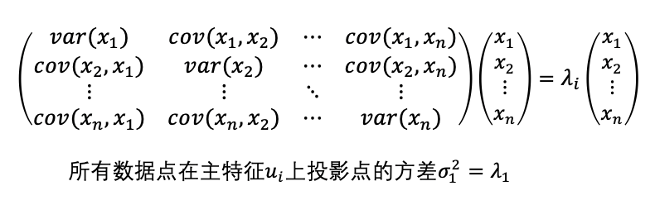


2、计算样本的协方差矩阵

协方差是对两个随机变量联合分布线性相关程度的一种度量；



3、对协方差矩阵求解特征值和特征向量



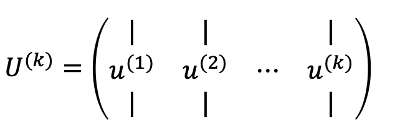
注意点：

1. 对称矩阵的特征向量相互正交，其点乘为0

② 数据点在特征向量上投影的方差，为对应的特征值，选择特征值大的特征向量，就是选择点投影方差大的方向，即是具有高信息量的主成分；次佳投影方向位于最佳投影方向的正交空间，是第二大特征值对应的特征向量，以此类推。

特征向量代表一组主成分空间的新坐标轴，特征值携带每个特征向量的方差数量的信息。因此，为了降低数据集的维度，我们准备选择那些具有较大方差的特征向量，丢弃那些具有较小方差的特征向量。

4、选取k个最大大特征值对应的特征向量，即是k个主成分



U是协方差矩阵所有的特征向量构成的矩阵，对应的特征值满足：λ1>λ2>⋯>λn，同时使其满足在主成分向量上投影的方差和占总方差的99%或者95%以上，即确定了k的选取。

|  |
| --- |
| PCA 不是从数据集选择某些特征而丢弃其他特征。PCA是在特征的组合基础上构造了一组新特征。从数学上讲，PCA 执行了一个线性转换，从原始特征集合转到由主成分组成的新空间。 |

# PCA应用

数据可视化

特征提取

异常值检测和聚类

参考

<https://mp.weixin.qq.com/s/flJRWMPgu5GUCGTlP1fMWA>

<https://mp.weixin.qq.com/s/7Azblz_KgBsM03oE9_RWPw>

<https://mp.weixin.qq.com/s/R5_agyfByJRVvs5GsgO2Qg>

<https://mp.weixin.qq.com/s/9-nNNhhDWSYWy46u0hTazQ>

<https://mp.weixin.qq.com/s/uAlBtGTmtBSjcnp9bWQr5Q>