廖浩 2025 年 4 月 25 日 日报

1. **工作内容：**

了解高PCA原理和方法并且整理。

1. **笔记整理：**

主成分分析法（Principal Component Analysis, PCA）是一种常见的数据分析方式，常用于高维数据的降维，可用于提取数据的主要特征分量。

PCA多用于对数据特征集进行降维，也方便对数据集进行可视化操作，说白了最后进行结果展示那么多特征向量要一起表示的话肯定很难展示，超过三维的数据就很难展示了。而PCA可对特征集进行简化，通俗的来讲也就是合并好理解。

1. **PCA简介**

主成分分析法（PCA）是设法将原来变量重新组合成一组新的相互无关的几个综合变量，同时根据实际需要从中可以取出几个较少的总和变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上处理降维的一种方法。主成分分析是设法将原来众多具有一定相关性（比如P个指标），重新组合成一组新的互相无关的综合指标来代替原来的指标。通常数学上的处理就是将原来P个指标作线性组合，作为新的综合指标。

在数学上更简单的理解我们可以把它想象为对很多个坐标点通过映射方法到一根函数直线上，如所示。通过这种方法我们肯定会得到新的特征用来表示这个事件，新的特征剔除了原有特征的冗余信息，因此更有区分度。新的特征基于原有特征，它能够重建原有特征。主成分分析要保留最有可能重建原有特征的新特征，从而达到数据降维的作用。

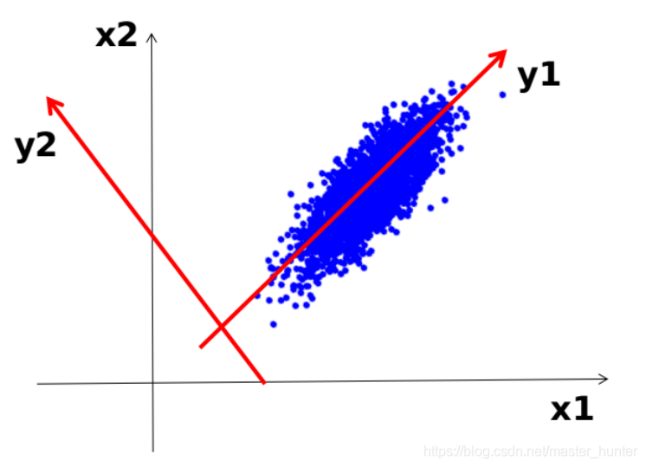
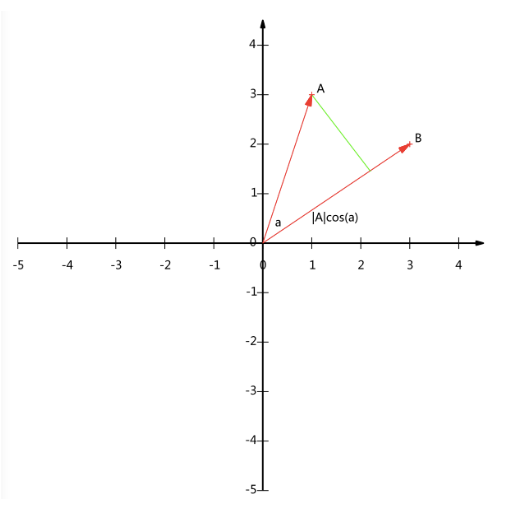


图 1 坐标映射

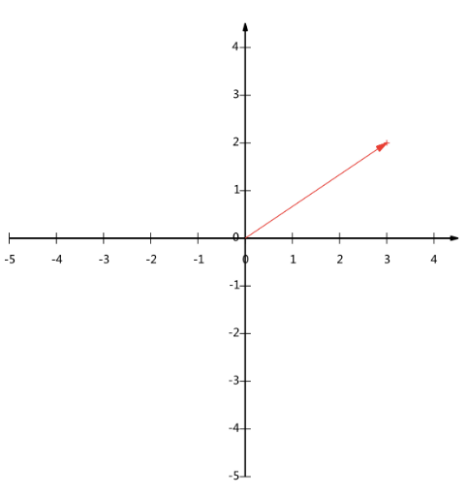
1. **PCA的数学原理**

首先在笛卡尔直角坐标系中上面的点都可以用一个二维向量表示。当然如果是n维向量可以等价表示为n维空间中的一条从原点发射的有向线段。假设我们就在二维笛卡尔直角坐标系中，有,表示。表示。那么在二维平面上A和B可以用两条发自原点的有向线段表示，如下图：

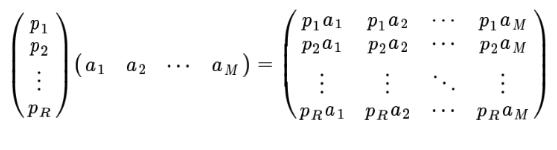


我们知道垂线与B的交点叫做A在B上的投影，再假设A与B的夹角为a，则投影的矢量长度为（这里假设向量B的模为1，,就是A点的模，也就是A线段的标量长度。而且我们知道：如果B向量的模为1的话，我们就可以发现得到。

一个二维向量可以对应二维笛卡尔直角坐标系中从原点出发的一条有向线段，对于来说，那么他们对应的坐标轴就是：

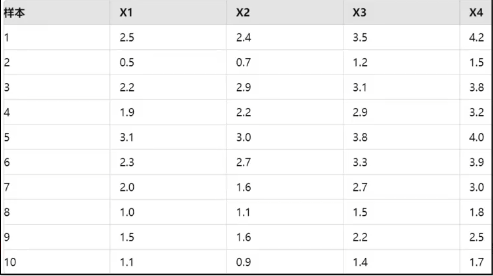


一般地，如果我们有M个N维向量，想将其变换为由R个N维向量表示的新空间中，那么首先将R个基按照行组成矩阵A,，然后将向量按照列组成矩阵B，那么两个矩阵的乘积AB就是变换结果，其中AB的第m列为A中的第M列变换后的结果。



特别要注意的是，这里R可以小于N，而R决定了变换后数据的维数，也就是说，我们可以将一个N维数据变换到更低维度的空间中去，变换后的维度取决于基的数量，因此这种矩阵相乘的表示也可以表示为降维变换。

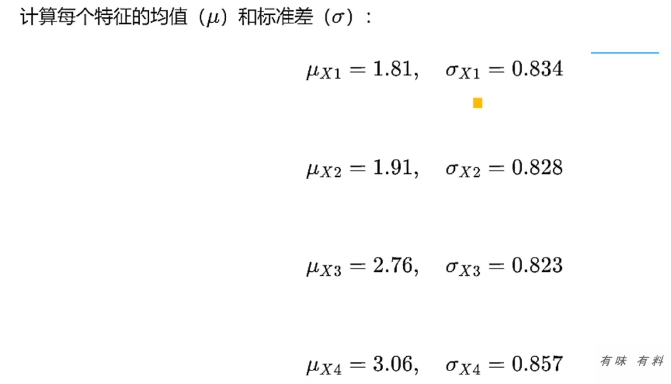
1. **PCA的应用**
2. **原始数据集矩阵X:**



1. **标准化:**

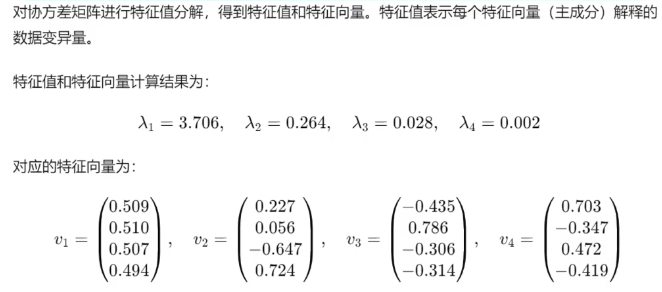
**标准化公式如下：**

**其中为均值，为标准差。**

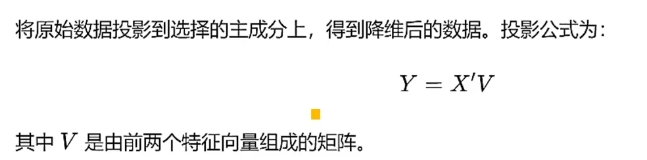




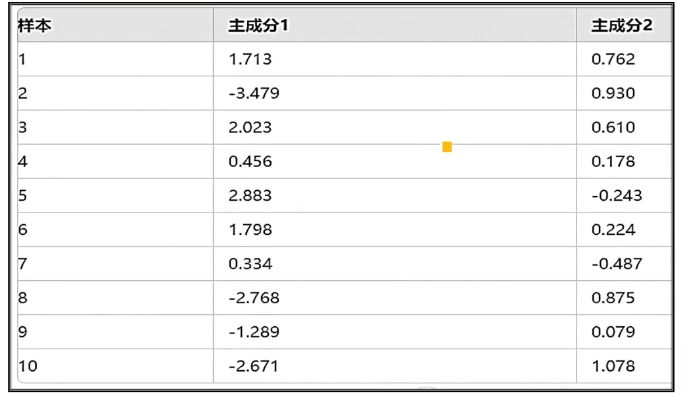
1. **再求协方差矩阵：**
2. **计算对应特征值和特征向量：**

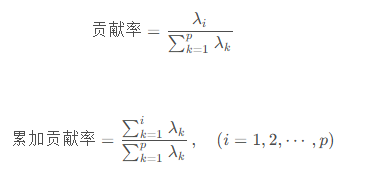


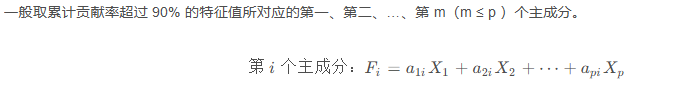
1. **选择较大特征值对应的特征向量，执行PCA变换：**



**得到的Y就是PCA降维后的值：**







**一些计算公式补充：**

**1.相关系数**



**2.协方差**

