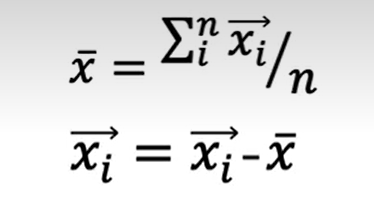
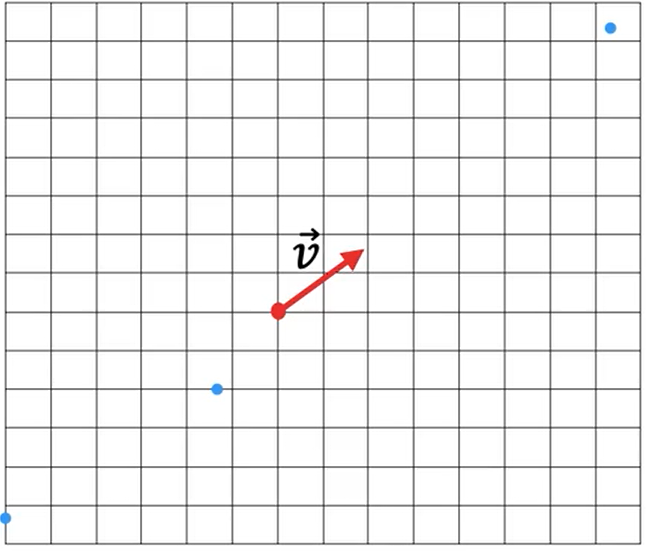
**PCA-主成成分分析**

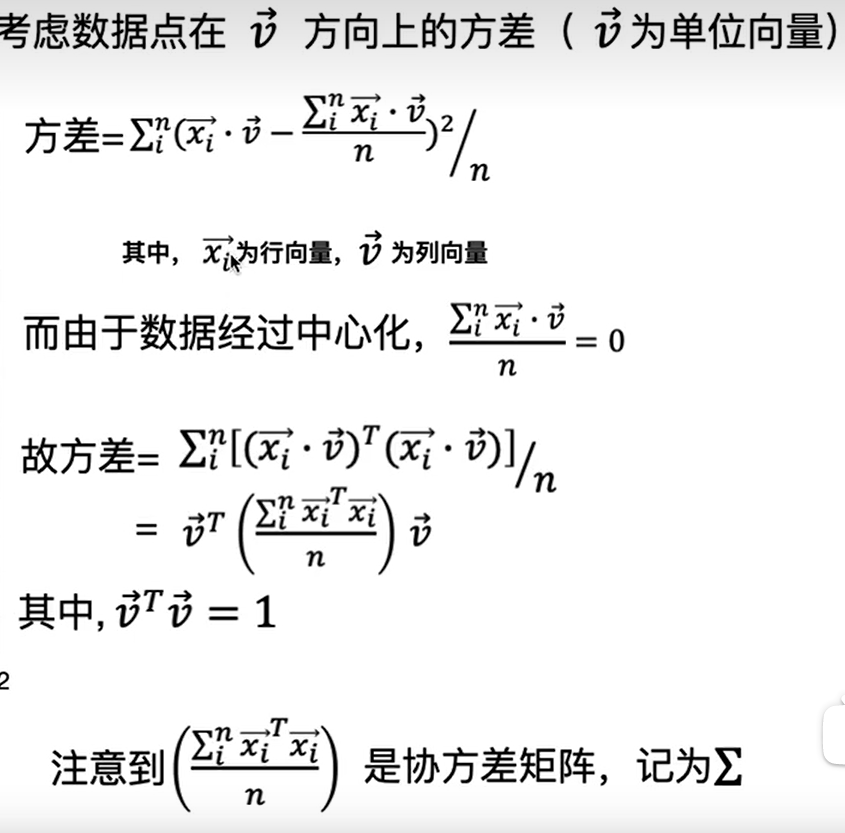
PCA降维的目的，就是为了在尽量保证“信息量不丢失”的情况下，对原始特征进行降维，也就是尽可能将原始特征往具有最大投影信息量的维度上进行投影。将原特征投影到这些维度上，使降维后信息量损失最小。

1. **数据中心化**

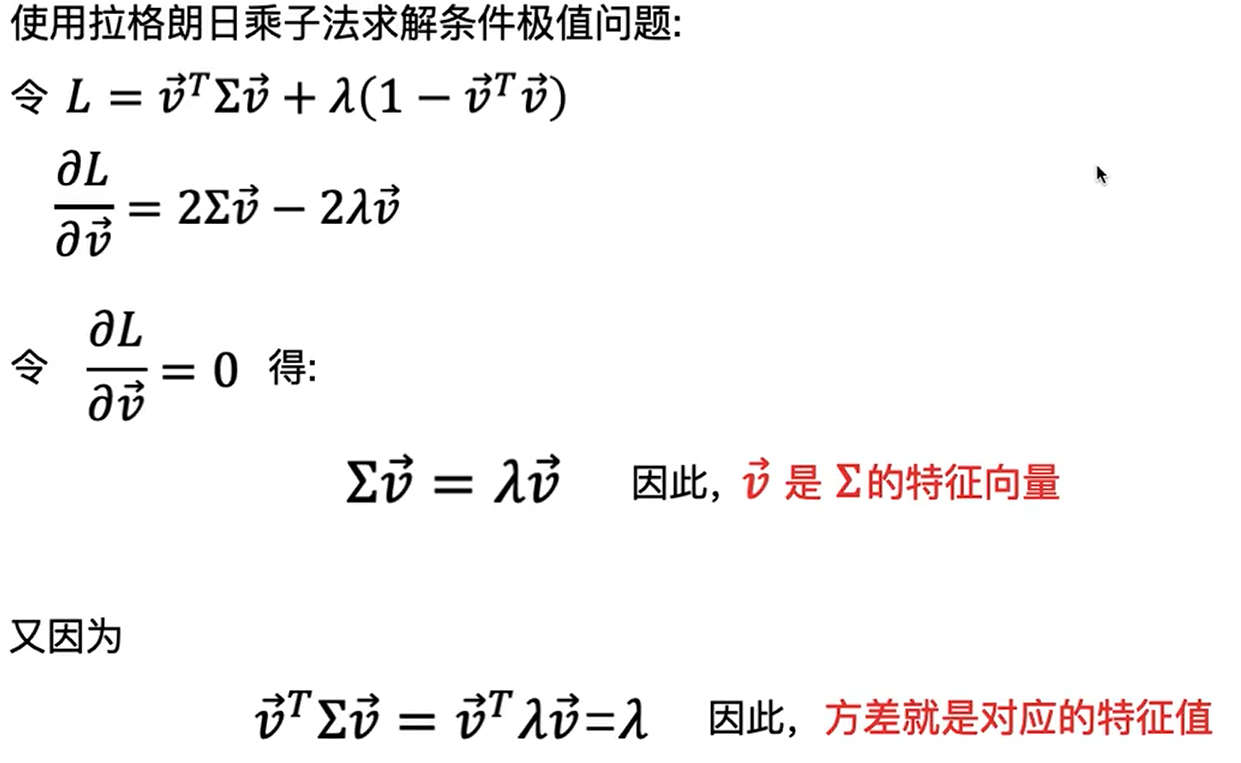


算出均值, 数据用原始数据减去均值来表示





求使得方差最大的V的方向,



标量对向量求导, 即对向量的每一个分量求导, 最后拼接成向量的形式

所以关键还是协方差矩阵的特征值

具体步骤:

* 去除平均值
* 计算协方差矩阵
* 计算协方差矩阵的特征值和特征向量
* 将特征值排序，保留前k个最大的特征值对应的特征向量
* 通过前k个特征向量构建映射矩阵W。
* 通过映射矩阵W将d维的输入数据集X转换到新的k维特征子空间中。

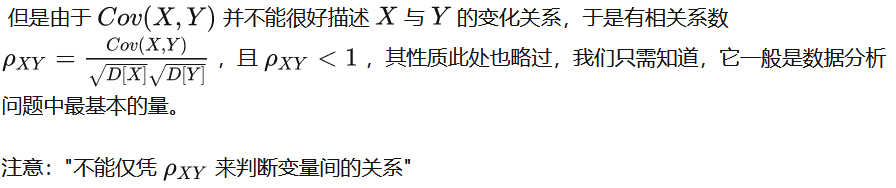
**信息保留率**: 采用N个特征值的和除以所有特征值的和。

前置知识

**协方差:**

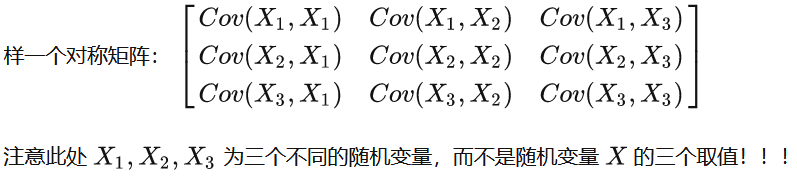
在统计学上, 协方差用来刻画两个随机变量之间的**相关性**。

基础知识： Cov(X,Y) = E[X-E[X]]\*E[Y-E[Y]], 且需要知道Cov(X,Y) = Cov(Y,X), Cov(X,X) = D(X)。



**协方差矩阵：**

Cov(X,Y)只表示两个随机变量X,Y之间的**相关性问题**, 那如果有n个随机变量呢, 这时候要用到协方差矩阵了。



上面这个矩阵， 就被称为X1, X2, X3的**协方差矩阵**/**协方差阵。**

**PCA降维的作用**

1. 降低算法时间与空间复杂度。
2. 去除数据集中夹杂的冗余信息与噪声信息。
3. 使模型在小数据集上有更强的鲁棒性。

在项目中以保留99%信息目标, 从1280维降到96维, 使样本维度与数量相差不大, 便于使用rbf高斯核函数。