# 第四章 降维

## 1.PCA

PCA目标：最大化投影方差

X投影后的方差就是协方差矩阵的特征值，要找到最大的方差也就是协方差矩阵最大的特征正，最佳投影方向就是最大特征值所对应的特征向量。

PCA的求解方法：

（1）对样本数据进行中心化操作

（2）求样本协方差 矩阵

（3）对协方差矩阵进行特征值分解，将特征值从大到小排列

（4）取特征值前d大对应的特征向量w1,w2,...,wd，通过以下映射将n维样本映射到d维

x_i' = [w1^T*x_i, w2^T*x_i, ..., wd^T*x_i]

## 2.线性判别分析

LDA是一种有监督学习算法。在PCA中，算法没有考虑数据的标签（类别），只是把原数据映射到一些方差比较大的方向上而已。

LDA目的是找到一个投影方向w，使得投影后的样本尽可能按照原始类别分开。中心思想：最大化类间距离和最小化类内距离。

目标函数定义为类间距离和类内距离的比值，参考书p.84

最大化目标对应了一个矩阵的特征值，LDA降维变成了一个求矩阵特征向量的问题，投影方向就是(S_w)^(-1)*S_B最大特征值对应的特征向量。

对于一般的二分类问题，只需样本的均值和类内方差，就可以马上得到最佳的投影方向w。

S_w为类内散度矩阵，S_B为类间散度矩阵，参考书p.84

## 3.线性判别分析与主成分分析

当LDA应用到多类别的时候，解法会发生变化，主要是类间离散度求解方式改变，参考书p.86~88

PCA和LDA的不同之处：

（1）PCA选择的是投影后数据方差最大的方向，是无监督的，因此PCA假设方差越大，信息量越多，用主成分来表示原始数据可以去除冗余的维度，达到降维。

（2）LDA选择的是投影后类内方差小，类间方差大的方向，利用了类别标签信息，为了找到数据中具有判别性维度维度，使得原始数据在这些方向上投影后，不同类别尽可能区分开。

（3）从应用角度选择的基本原则——无监督任务使用PCA进行降维，有监督的使用LDA。