主成分分析[1]

简介

在因子分析中我们介绍了一种找到高维数据中低维隐含因子的方法，这是一种基于概率模型的方法，而且需要我们利用EM算法去估计参数。在本notes中，我们将学习一种新的降维方法——主成分分析（Principal Components Analysis, PCA），它比因子分析来得更为直接，而且最终只涉及到关于特征向量的计算。

1 问题引入

真实的数据集总存在各种各样的问题。

- 假定关于汽车的数据中既包含以『千米/每小时』为单位的最大速度特征，也包含以『英里/小时』为单位的最大速度特征，显然这两个特征有一个多余。
- 假定关于学生数学成绩的数据中包含两列，第一列是学生对数学的兴趣程度，第二列是学生的考试分数，这两列是强相关关系，能否合并成一列呢？
- 在信号传输过程中，由于信道不理想，另一端接收到的信号会存在噪音干扰，那么我们该如何滤去噪音呢？

像这种数据特征中存在冗余和噪音的时候，我们应该想办法降低特征维度以去掉它们。不过与之前在learning theory中介绍的特征选择方法不同，特征选择是利用互信息剔除掉与类标签无关的特征，而在主成分分析中，我们是将n维特征映射到k维上，构造了一个全新的k维正交特征。

2 PCA算法和理论

2.1 数据预处理

给定数据集$\left\{x^{(i)}; i=1,\cdots,m\right\}$，其中$x^{(i)} \in \mathbb{R}^n$。在介绍PCA算法之前，我们需要对原始数据做一下预处理

1. Let $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
2. Replace each $x^{(i)}$ with $x^{(i)} - \mu$
3. Let $\sigma_j^2 = \frac{1}{m}\sum_i \left(x_j^{(i)}\right)^2$
4. Replace each $x_j^{(i)}$ with $x_j^{(i)}/\sigma_j$

第1、2步是为了令数据的平均值归零，如果已知数据平局值为零，这两步可省略。第3、4步是为了缩放每个坐标轴以得到单位方差，这样我们就可以在相同的『尺度』上处理每一维特征，如果已知所有特征『尺度』相同，这两步可省略。

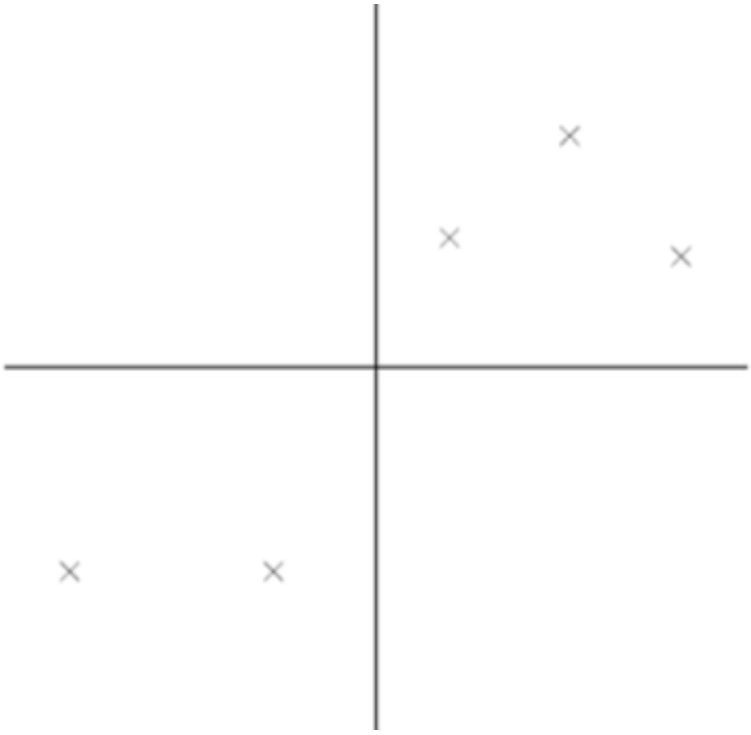
2.2 最大方差理论

在完成数据预处理后，我们希望找到k个单位向量$\mu \in \mathbb{R}^n$，并将每一条数据$x^{(i)}$投影在这k个向量上得到k个投影值$y_j^{(i)}; j=1,\cdots,k$，也就是说$y^{(i)} \in \mathbb{R}^k$。如果我们用新数据$y^{(i)}$代替原始数据$x^{(i)}$就达到了我们想要的降维的目的，那么该如何选取μ才是最优的呢？

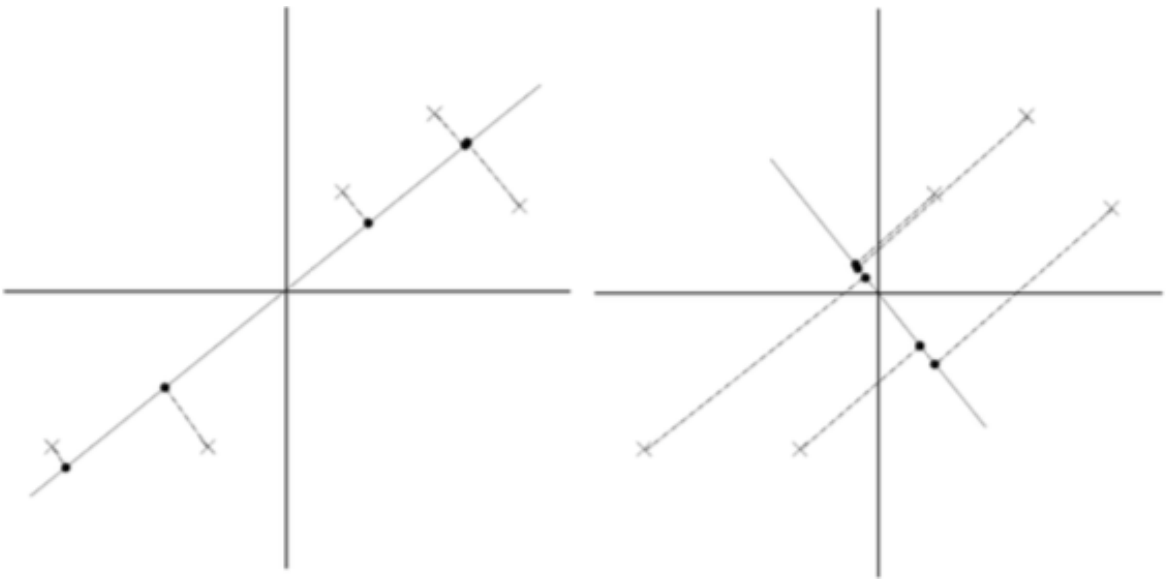
信息论认为信号具有较大方差，噪音具有较小方差，信噪比就是信号与噪音的方差比，这个值越

大越好。所以我们需要寻找的μ应该使数据的投影的方差最大化。这就是最大方差理论（maximum variance theory）。

为了更清楚地说明这一过程，假设我们拥有包含5个二维样本的已预处理数据集[2]，如下图所示



接下来让我们选择两个不同的方向μ，这里用一条通过原点的直线表示，并作出原始数据在μ上的投影



可以看出，左图的投影十分分散，方差较大；右图的投影十分集中，方差较小。根据最大方差理论，我们应该选择左图中的μ。这个过程我们该如何定量的表示呢？对于单位向量μ以及输入x，$x^T\mu = < x, \mu > = |x| \cdot |\mu| \cos\theta = |x| \cos\theta$表示投影长度，其中$\theta$表示$x, \mu$之间的夹角。对于上图的输入$x^{(i)}$，投影长度即投影点到原点的『距离』，它实际表达的是投影点到中心的差值，因此

我们的最大化目标函数为

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)^T}\mu\right)^2 &= \frac{1}{m}\sum_{i=1}^{m}\mu^T x^{(i)} x^{(i)^T}\mu \\ &= \mu^T\left(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)^T}\right)\mu\end{aligned}$$

又$\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)^T} = E[xx^T] = E[x]E[x^T] + \Sigma = \Sigma$，其中$\Sigma$表示协方差矩阵。令

$$\begin{aligned}\lambda &= \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)^T}\mu\right)^2 = \mu^T\Sigma\mu \\ &\Rightarrow \mu\lambda = \lambda\mu = \mu\mu^T\Sigma\mu \\ &\Rightarrow \lambda\mu = \Sigma\mu \quad (\mu^T\mu = 1)\end{aligned}$$

可以发现，λ是Σ的特征值，而μ是对应的特征向量。接着按照最大化的目标，我们找到top k特征值λ所对应的特征向量μ即为所求，它们被称为数据的主成分（principal components）且是正交的。得到k个向量后，我们作如下投影

$$y^{(i)} = \begin{bmatrix}\mu_1^T x^{(i)} \\ \mu_2^T x^{(i)} \\ \vdots \\ \mu_k^T x^{(i)}\end{bmatrix} \in \mathbb{R}^k$$

就得到了降维后的新数据。

---

1. Wriiten by Jimmy on 2016/03/16.↩
2. 预处理的目的包含两个：第一，使得原点移动到所有数据的中心；第二，缩放坐标轴使得数据在每个坐标上的分量是归一化的，具有可比较性。↩