# 2.1KPCA

KPCA，中文名称”核主成分分析“，是对PCA算法的非线性扩展，言外之意，PCA是线性的，其对于非线性数据往往显得无能为力。**KPCA能够挖掘到数据集中蕴含的非线性信息。**

其实很好理解，就是原始的数据线性不可分，那么就升维变成线性可分。这个过程就需要使用核函数。升维后线性可分那么就采用PCA。

从 XX^T 操作变成对 $\phi(x)\phi(x^T)$

## 1.理论部分

1. 为了更好处理非线性数据，**引入非线性映射函数 $\phi(x)$，将原空间中的数据映射到高维空间。**

2. 引入了一个定理：**空间中的任一向量（哪怕是基向量），都可以由该空间中的所有样本线性表示。**

假设中心化后的样本集合X（d*N，N个样本，维数d维，样本”按列排列“），现将X映射到高维空间，得到 $\phi(x)$，**假设在这个高维空间中，本来在原空间中线性不可分的样本现在线性可分了，然后呢？想啥呢！果断上PCA啊！**

于是乎！假设D（D >> d）维向量 **Wi 为高维空间中的特征向量**，**为对应的特征值** $\lambda_i$，高维空间中的PCA如下：

$$\Phi(X)\Phi(X)^T w_i = \lambda_i w_i \quad (1)$$

这个时候，在利用刚才的定理，将**特征向量Wi 利用样本集合 $\phi(x)$ 线性表示**，如下：

$$w_i = \sum_{k=1}^{N} \alpha_i \Phi(x_i) = \Phi(X)\alpha \quad (2)$$

然后，在把 $w_i\ (i=1,...,d)$ 代入上上公式，得到如下的形式：

$$\Phi(X)\Phi(X)^T \Phi(X)\alpha = \lambda_i \Phi(X)\alpha \quad (3)$$

进一步，等式两边同时左乘 $\Phi(X)^T$，得到如下公式：

$$\Phi(X)^T \Phi(X)\Phi(X)^T \Phi(X)\alpha = \lambda_i \Phi(X)^T \Phi(X)\alpha \quad (4)$$

这样做的目的是，构造两个$\Phi(X)^T\Phi(X)$出来，进一步用核矩阵K（为对称矩阵）替代，其中：

$$\Phi(X)^T\Phi(X)\Phi(X)^T\Phi(X)\alpha=\lambda_i\Phi(X)^T\Phi(X)\alpha \tag{5}$$

于是，公式进一步变为如下形式：

$$K^2\alpha=\lambda_i K\alpha \tag{6}$$

两边同时去除K，得到了PCA相似度极高的求解公式：

$$K\alpha=\lambda_i\alpha \tag{7}$$

求解公式的含义就是**求K最大的几个特征值所对应的特征向量**，由于K为对称矩阵，所得的解向量彼此之间肯定是正交的。

但是，请注意，这里的$\alpha$只是K的特征向量，但是其不是高维空间中的特征向量，回看公式（2），**高维空间中的特征向量w应该是由$\alpha$进一步求出。**

## 2.核函数

(1) 线性核函数(可视为特例)

$$K(x,x_i)=x\cdot x_i;$$

(2) $p$ 阶多项式核函数

$$K(x,x_i)=[(x\cdot x_i)+1]^p;$$

(3) 高斯径向基函数(RBF)核函数

$$K(x,x_i)=\exp\left(-\frac{\|x-x_i\|}{\sigma^2}\right);$$

(4) 多层感知器(MLP)核函数

$$K(x,x_i)=\tanh[v(x\cdot x_i)+c];$$