PCA

主成分分析(Principal Components Analysis，简称PCA)，用于数据降维

PCA与最大可分性

对于$X=\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$，希望将$X$从$n$维降到$n'$维，同时希望信息损失最少

基变换

基变换(线性变换)：$Y=PX$，其中P是基向量，X是原始样本，Y是新表达。

$$\begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_r \end{bmatrix}_{r\times n} \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}_{n\times m} = \begin{bmatrix} p_1x_1 & p_1x_2 & \cdots & p_1x_m \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ p_rx_1 & p_rx_2 & \cdots & p_rx_m \end{bmatrix}_{r\times m}$$

其中p_i表示行向量，表示第i个基；x_j是列向量，表示第j个原始数据记录。当基的维度r<数据维度n时，可以达到降维的目的

方差

我们希望投影后的数据尽量分散开，用方差来表达 分散程度 ，方差越大，数据越分散。$Var(a)=\frac{1}{m}\sum\limits_{i=1}^{m}(a_i-\mu)^2$，为了方便处理，一般将每个字段内所有值都减去字段的平均值

协方差与协方差矩阵

1. $Cov(a,b)=\frac{1}{m}\sum\limits_{i=1}^{m}a_ib_i$

2. 对于n维随机变量，$x_i=\begin{bmatrix}x_1\\x_2\\\cdots\\x_n\end{bmatrix}$ $C=\begin{bmatrix}Var(x_1) & Cov(x_1,x_2) & \cdots & Cov(x_1,x_n)\\Cov(x_2,x_1) & Var(x_2) & \cdots & Cov(x_1,x_n)\\\vdots & \vdots & \ddots & \vdots\\Cov(x_n,x_1) & \cdots & \cdots & Var(x_n)\end{bmatrix}$即可得到协方差矩阵。

   > 假设有m个n维数组记录，将其按列排成$n\times m$的矩阵$\mathbf{X}$，则$C=\frac{1}{m}\mathbf{X}\mathbf{X}^T$就是协方差矩阵。

3. 协方差矩阵对角化

   假设原始数据矩阵为$X_{n\times m}$，想要找到基$P_{r\times n}$使得$Y_{r\times m}=P_{r\times n}X_{n\times m}$实现降维的目的。假设$X$的协方差矩阵为$C$，$Y$的协方差矩阵为$D$，且有$Y=PX$。目标是让协方差矩阵$D$的各个方向方差最大。

   则经过化简有：$D=\frac{1}{m}YY^T=PCP^T$。存在单位向量$E=[\,e_1\quad e_2\quad\cdots\quad e_n\,]$，$E^TCE=\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{bmatrix}$，因此$P=E^T$

PCA算法过程

设有m条n维数据

1. 将原始数据按列组成n行m列矩阵X
2. 将X的每一行(代表一个特征)进行零均值化(减去这一行的均值)
3. 求出协方差矩阵$C=\frac{1}{m}XX^T$
4. 求出协方差矩阵C的特征值以及对应的特征向量
5. 将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P
6. $Y=PX$即为降维到k维后的矩阵