主成分分析

机器学习笔记 create by siwanghu v1.0

主成分分析，将高维的特征用低维的特征表示。本质上来讲，主成分分析法是一种空间映射的方法，将在常规正交坐标系的变量通过矩阵变换操作映射到另一个正交坐标系中

假设样本集合$\mathrm{X}=\{x^1,x^2,x^i,\ldots,x^{\mathrm{m}}\}$，每个样本$x^i$为 n 维的列向量，代表每个样本有 n 个特征，则样本集合 X 是一个 n*m 的矩阵。

主成分分析的本质是将 X 映射到$\mathrm{Y}=\{y^1,y^2,y^i,\ldots,y^m\}$，样本 y^i为l维的列向量，代表样本有l个特征，其中$l<\mathrm{n}$。意味着样本集合 Y 中的样本特征数目少于样本集合 X 中的样本特征数目，起到了特征降维的过程。为了达到这个目的，则我们需要找到一个中间转换矩阵$\mathrm{D}_{\mathrm{n},\ l}$：

$$\mathrm{X}_{\mathrm{n},\ \mathrm{m}}=\mathrm{D}_{\mathrm{n},\ l}*\mathrm{Y}_{l,\ \mathrm{m}}$$

PCA 算法流程

输入：$\mathrm{X}_{\mathrm{n},\ \mathrm{m}}=\{x^1,x^2,x^i,\ldots,x^{\mathrm{m}}\}$

对数据$\mathrm{X}_{\mathrm{n},\ \mathrm{m}}$做中心化处理（变量减去它的均值）

对输入协方差矩阵X^TX进行特征值求解，将求得的协方差矩阵X^TX特征值从大到小排序

$$\lambda_1\ge\lambda_2\ge\lambda_3\ge\cdots\ge\lambda_k$$

取前l个特征值对应的特征向量，组成矩阵$\mathrm{D}_{\mathrm{n},\ l}$，其中$l$代表要压缩到的特征数目，$l$=1 代表将特征数目压缩到 1。

假设有样本集合$\mathrm{X}=\begin{pmatrix}1, & 3, & -1\\ 2, & 3, & -2\end{pmatrix}$，中心化处理$\mathrm{X}=\begin{pmatrix}0, & 2, & -2\\ 1, & 2, & -3\end{pmatrix}$

假设我们有一个二维数据，我们要通过 PCA 的方法来将这个二维数据降到一维

$$\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}$$

因为数据已经中心化，所以我们就省去了中心化的步奏，求 x 协方差

$$C=\frac{1}{5}\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}\begin{pmatrix}-1 & -2\\ -1 & 0\\ 0 & 0\\ 2 & 1\\ 0 & 1\end{pmatrix}=\begin{pmatrix}\frac{6}{5} & \frac{4}{5}\\ \frac{4}{5} & \frac{6}{5}\end{pmatrix}$$

求得其特征值和特征向量：

$$\lambda_1=2,\lambda_2=2/5$$

$$c_1\begin{pmatrix}1\\ 1\end{pmatrix},c_2\begin{pmatrix}-1\\ 1\end{pmatrix}$$

对特征向量进行标准化：

$$\begin{pmatrix}1/\sqrt{2}\\ 1/\sqrt{2}\end{pmatrix},\begin{pmatrix}-1/\sqrt{2}\\ 1/\sqrt{2}\end{pmatrix}$$

因为将数据降到一维，所以取向量c_1

$$\begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\end{pmatrix}\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}=\begin{pmatrix}-3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & -1/\sqrt{2}\end{pmatrix}$$

http://blog.csdn.net/weixin_37824397